Multiscale data assimilation approaches and error characterisation applied to the inverse modelling ofatmospheric constituent emission fields

Mohammad Reza Koohkan

UNIVERSITÉ PARIS-EST

Thèse présentée pour obtenir le grade de

Docteur de l'Université Paris-Est

Spécialité: Sciences et Techniques de l'Environnement

par

Mohammad Reza Koohkan

École Doctorale : SCIENCES, INGÉNIERIE ET ENVIRONNEMENT

Multiscale data assimilation approaches and error characterisation applied to the inverse modelling of atmospheric constituent emission fields.

Thèse soutenue le 20 décembre 2012 devant le jury composé de:

Dr Olivier Talagrand	CNRS/LMD	Président
Dr Slimane Bekki	CNRS/LATMOS	Rapporteur
Dr Frédéric Chevallier	CEA/LSCE	Rapporteur
Dr Gilles Forêt	UPEC/LISA	Examinateur
Dr Sébastien Massart	ECMWF	Examinateur
Dr Marc Bocquet	École des Ponts ParisTech/CEREA	Directeur de thèse

"with all the affection, devotion and love that
a son can ever express to his dear mum".

À ma petite maman

Acknowledgment

This Ph.D. thesis was made possible through an École des Ponts ParisTech scholarship with the help of the Agence Nationale de la Recherche (MSDAG project), the INSU/LEFE council (ADOMOCA-2 project) and finally the French Ministry of Ecology and ADEME (CARBOSOR project, Primequal research program). Many thanks to all these institutions.

Though only my name appears on the cover of this dissertation, a great many people have contributed to its production. First and foremost, I would like to express my sincere gratitude to my supervisor Marc Bocquet for his patience, motivation, enthusiasm and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better mentor.

Besides my supervisor, I am truly indebted to Christian Signieur and Yelva Roustan for the long discussions that helped me sorting out the technical details of my work.

I am also deeply grateful to my thesis committee, Dr Slimane Bekki, Dr Frédéric Chevallier, Dr Sébastien Massart, Dr Gilles Forêt and Dr Olivier Talagrand, for their encouragement.

I would like to thank sincerely my fellow lab-mates of École des Pont ParisTech, Lin Wu, Victor Winiarek, Kim Younsoub and my office-mate, Nora Duhanyan for her moral support throughout these three years. Many thanks also to Monika Krysta and Stéphane Sauvage and everyone who helped me writing my dissertation.

Finally, I would like to dedicate this entire work to the one and unique person who, wherever I could be, guides me through life: my Mother.

Abstract

Data assimilation in geophysical sciences aims at optimally estimating the state of the system or some parameters of the system's physical model. To do so, data assimilation needs three types of information: observations and background information, a physical/numerical model, and some statistical description that prescribes uncertainties to each componenent of the system.

In my dissertation, new methodologies of data assimilation are used in atmospheric chemistry and physics: the joint use of a 4D-Var with a subgrid statistical model to consistently account for representativeness errors, accounting for multiple scale in the BLUE estimation principle, and a better estimation of prior errors using objective estimation of hyperparameters. These three approaches will be specifically applied to inverse modelling problems focussing on the emission fields of tracers or pollutants.

First, in order to estimate the emission inventories of carbon monoxide over France, in-situ stations which are impacted by the representativeness errors are used. A subgrid model is introduced and coupled with a 4D-Var to reduce the *representativeness error*. Indeed, the results of inverse modelling showed that the 4D-Var routine was not fit to handle the representativeness issues. The coupled data assimilation system led to a much better representation of the CO concentration variability, with a significant improvement of statistical indicators, and more consistent estimation of the CO emission inventory.

Second, the evaluation of the potential of the IMS (International Monitoring System) radionuclide network is performed for the inversion of an accidental source. In order to assess the performance of the global network, a multiscale adaptive grid is optimised using a criterion based on *degrees of freedom for the signal* (DFS). The results show that several specific regions remain poorly observed by the IMS network.

Finally, the inversion of the surface fluxes of Volatile Organic Compounds (VOC) are carried out over Western Europe using EMEP stations. The uncertainties of the background values of the emissions, as well as the covariance matrix of the observation errors, are estimated according to the *maximum likelihood principle*. The prior probability density function of the control parameters is chosen to be Gaussian or truncated Gaussian distributed. Grid-size emission inventories are inverted under these two statistical assumptions. The two kinds of approaches are compared. With the Gaussian assumption, the departure between the posterior and the prior emission inventories is higher than when using the truncated Gaussian assumption, but that method does not provide better scores than the truncated Gaussian in a forecast experiment.

Keywords: Data assimilation, inverse modelling, 4D-Var, multiscale, representativeness errors, maximum likelihood principle.

Résumé

Dans les études géophysiques, l'assimilation de données a pour but d'estimer l'état d'un système ou les paramètres d'un modèle physique de façon optimale. Pour ce faire, l'assimilation de données a besoin de trois types d'informations : des observations, un modèle physique/-numérique et une description statistique de l'incertitude associée aux paramètres du système.

Dans ma thèse, de nouvelles méthodes d'assimilation de données sont utilisées pour l'étude de la physico-chimie de l'atmosphère : (i) On y utilise de manière conjointe la méthode 4D-Var avec un modèle sous-maille statistique pour tenir compte des erreurs de représentativité. (ii) Des échelles multiples sont prises en compte dans la méthode d'estimation BLUE. (iii) Enfin, la méthode du maximum de vraisemblance est appliquée pour estimer des hyper-paramètres qui paramètrisent les erreurs à priori. Ces trois approches sont appliquées de manière spécifique à des problèmes de modélisation inverse des sources de polluant atmosphérique.

Dans une première partie, la modélisation inverse est utilisée afin d'estimer les émissions de monoxyde de carbone sur un domaine représentant la France. Les stations du réseau d'observation considérées sont impactées par les erreurs de représentativité. Un modèle statistique sous-maille est introduit. Il est couplé au système 4D-Var afin de réduire les erreurs de représentativité. En particulier, les résultats de la modélisation inverse montrent que la méthode 4D-Var seule n'est pas adaptée pour gérer le problème de représentativité. Le système d'assimilation des données couplé conduit à une meilleure représentation de la variabilité de la concentration de CO avec une amélioration très significatives des indicateurs statistiques.

Dans une deuxième partie, on évalue le potentiel du réseau IMS (International Monitoring System) du CTBTO (preparatory commission for the Comprehensive nuclear-Test-Ban Treaty Organization) pour l'inversion d'une source accidentelle de radionucléides. Pour évaluer la performance du réseau, une grille multi-échelle adaptative de l'espace de contrôle est optimisée selon un critère basé sur les degrés de liberté du signal (DFS). Les résultats montrent que plusieurs régions restent sous-observées par le réseau IMS.

Dans la troisième et dernière partie, sont estimés les émissions de Composés Organiques Volatils (COVs) sur l'Europe de l'ouest. Cette étude d'inversion est faite sur la base des observations de 14 COVs extraites du réseau EMEP. L'évaluation des incertitudes des valeurs des inventaires d'émission et des erreurs d'observation sont faites selon le principe du maximum de vraisemblance. La distribution des inventaires d'émission a été supposée tantôt gaussienne et tantôt gaussienne tronquée. Ces deux hypothèses sont appliquées pour inverser le champs des inventaires d'émission. Les résultats de ces deux approches sont comparés. Bien que la correction apportée sur les inventaires est plus forte avec l'hypothèse gaussienne que gaussienne tronquée, les indicateurs statistiques montrent que l'hypothèse de la distribution gaussienne tronquée donne de meilleurs résultats de concentrations que celle gaussienne.

Mots-clés : assimilation de données, modélisation inverse, 4D-Var, multi-échelle, erreurs de représentativité, principe du maximum de vraisemblance.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Contents

In the present study, data assimilation methods are used to estimate the sources of pollutants which are part of the input data in a model of atmospheric physics. Such methods are very useful because they provide reliable information on some physical parameters which would be difficult or impossible to access otherwise (for instance experimentally) and help to calibrate the numerical models to predict the future. These methods are not only in need of a physical model, but also of observations and of some knowledge of the statistics on the errors related to the model and to the observations.

During the two last decades, many studies have been published on the assimilation of in-situ and satellite observations of pollutant concentrations. In particular, the present study focuses on the assimilation of the in-situ observations of concentrations. The objective pursued is to estimate the emission fluxes of carbon monoxide (CO) and the ability of the IMS (International Monitoring System) of Comprehensive Nuclear-Test-Ban Treaty (CTBTO) to reconstruct radionuclide sources. A final purpose is to correct the emission inventories of Volatile Organic Compounds (VOCs) with the help of the hyper-parameters (the errors related to the parameters

and related to the observations) computed via the maximum likelihood method.

This chapter aims at introducing the necessary elements of data assimilation. Section 1.1 presents the different physical models of the atmosphere, as well as, the sources of uncertainty related to them. In Section 1.2, the observations and their importance are discussed. The different kinds of errors between the observations and the simulations (modelling errors) will also be explained. Section 1.3 gives a detailed description of inverse modelling, of the estimation of the hyper-parameters of the objective function for data assimilation and finally, of the method of multiscale data assimilation. The outline of the study will be presented in Section 1.4.

1.1 Chemistry-transport models

The assessment of the pollutant concentrations in the atmospheric boundary layer is essential to improve air quality and prevent harmful impacts on human health. For many years, numerical codes have been developed to compute the spatio-temporal concentrations of atmospheric species. Most of them use Eulerian, Lagrangian and Gaussian numerical models. As will be explained below, each have their strengths and weaknesses and are used in different, though complementary, simulation contexts.

- *The Eulerian models*- are deterministic models. The pollutant motion is studied at specific locations in space (control volumes), through which air flows as time passes. In these models each physical quantity such as velocity and acceleration of the fluid is expressed as a function depending on space and time. As the Eulerian models are not focused directly on each particle, the computed concentrations are continuous quantities.

- *The Lagrangian models*- are stochastic models which follow the particle (the pollutants) trajectories in time. Each particle is labelled with a number. The concentration of a pollutant is computed with the help of the number and mass of the particle in a specific area. The computed concentrations that are discrete quantities are accurate near the sources. The computational load increases linearly with the number of sources. Thus, in an air quality context, Lagrangian models are suitable for accidental case studies.

- *The Gaussian models*- are based on analytical approximate solutions of the advection-diffusion equation. The latter is not solved and the physical and chemical processes are taken into account through parametrisations. They are suitable for operational studies as the computing time required is short. However, they cannot easily handle complex physical and chemical processes.

In this study the Eulerian code POLAIR3D of the POLYPHEMUS platform [Boutahar et al., 2004; Quélo et al., 2007] and the Lagrangian code FLEXPART [Stohl et al., 2005] will be used.

1.1.1 Eulerian models

An Eulerian model is based on the spatial and temporal resolution of the equations describing a physical system. In atmospheric studies, the equations are solved under the assumption of incompressible flow and for diluted species (neglecting the pollutant actions on the fluid flow). The concentration of the studied species is computed taking the flux, the production and the loss of the species in the cell into account according to the following equation:

$$\begin{aligned}\frac{\partial c(\mathbf{x},t)}{\partial t} &= -\mathrm{div}\left(\mathbf{u}(\mathbf{x},t)c(\mathbf{x},t)\right) + \nabla\cdot\left(\rho\mathbf{K}(\mathbf{x},t)\nabla\frac{c(\mathbf{x},t)}{\rho}\right)\\ &\quad -\Lambda(\mathbf{x},t)c(\mathbf{x},t) + \chi(c(\mathbf{x},t),\mathbf{x},t) + \sigma(\mathbf{x},t)\,. \end{aligned} \tag{1.1}$$

In this equation, $c(\mathbf{x},t)$ is the average concentration of the species at coordinate $\mathbf{x}$ and time t and $\mathbf{u}$ their average velocity. $\mathbf{K}$ is the turbulent diffusion matrix. ρ is the density of the fluid. Λ is the scavenging coefficient. Finally, χ and σ are the chemical reaction and the emission terms, respectively. The different terms in Eq. (1.1) are:

- the transport term following two principles:
  - the advection ($\mathrm{div}\left(\mathbf{u}(\mathbf{x},t)c(\mathbf{x},t)\right)$) which accounts for the transport of the species with the average fluid motion.
  - the turbulent diffusion ($\nabla\cdot\left(\rho\mathbf{K}(\mathbf{x},t)\nabla\frac{c(\mathbf{x},t)}{\rho}\right)$) which accounts for the transport of the species with the fluctuating fluid motion assuming a first order closure model. The molecular diffusion of the species is neglected compared to their turbulent diffusion.
- the wet scavenging term ($\Lambda(\mathbf{x},t)c(\mathbf{x},t)$) models the loss of species by absorption in hydrometeors. The pollutants incorporated, for instance, in raindrops are transferred from the atmosphere to the ground. The scavenging term is a sink term in the mass transport equation.
- the chemistry term ($\chi(c(\mathbf{x},t),\mathbf{x},t)$) accounts for the chemical reactions the species undergo. It is a source term when the species are produced or a sink term when they are consumed, however, it is zero for the inert pollutants.
- the volume emission term ($\sigma(\mathbf{x},t)$); it is a source of species in the mentioned equation. It includes the emissions of pollutants due to human activities (anthropogenic sources), e.g., traffic and industries, and due to natural (biogenic) sources, e.g., vegetation emission and uplake biomass burning and volcanic eruptions.

One can show the existence and the uniqueness of the solution for the above evolutionary equation (Eq. (1.1)) under the following conditions:

- The initial conditions are the concentrations at time t=0,

$$c(\mathbf{x},0) = c_0(\mathbf{x})\,, \tag{1.2}$$

  and show the state of the atmosphere at the beginning of the modelling process.
- The boundary conditions are the concentrations at the borders of the numerical domain:
  - Boundary conditions at the ground level ($z=0$):

$$\mathbf{K}(\mathbf{x}_{z=0},t)\nabla c(\mathbf{x}_{z=0},t)\cdot\mathbf{n} = v_d c(\mathbf{x}_{z=0},t) - E(\mathbf{x}_{z=0},t)\,, \tag{1.3}$$

    where, $\mathbf{n}$ is the unity vector normal to the surface $z=0$ and directed towards the outside of the domain. v_d is the dry deposition velocity. The left side of Eq. (1.3) displays the variation of the concentration at the ground with respect to time. $E(\mathbf{x}_{z=0},t)$ is the surface emission.

– The concentrations at the borders of the numerical domain, where the wind is incoming, $\partial\mathbf{\Omega}_{in}$, are depicted by:

$$c(\mathbf{x},t) = c_{\partial\mathbf{\Omega}_{in}}(\mathbf{x},t) \quad (\mathbf{x},t) \in \partial\mathbf{\Omega}_{in}\,. \tag{1.4}$$

$\partial\mathbf{\Omega}_{in} = \bigcup_t \partial\mathcal{D}_{in}^t$ and $\mathcal{D}_{in}^t$ is the border of the spatial domain when the wind is incoming at time t.

In POLAIR3D of the POLYPHEMUS, three numerical schemes, advection-diffusion-chemistry, are integrated using the splitting principle. The suitable order for the set of operator is advection, diffusion and then chemistry. A Third-order Direct Space and Time (DST-3) scheme is used to compute the advection term with flux limiter. The Rosenbrock method (second order) is used to integrate the turbulent diffusion and the chemistry schemes. The positivity of solution is guaranteed by the clipping condition.

1.1.2 Lagrangian models

A Lagrangian model is based on the computation in time of the labelled particle positions and trajectory. The motion of particles is described by the following equation,

$$\mathbf{x}(t+\Delta t) = \mathbf{x}(t) + \mathbf{v}\left(\mathbf{x}(t)\right)\Delta t\,, \tag{1.5}$$

where, $\mathbf{x}(t)$ is the position of the particle at time t and $\mathbf{v}$ is its velocity. The latter is the addition of the wind vector averaged over the entire grid cell, $\bar{\mathbf{v}}$, and of the turbulent wind fluctuation, $\mathbf{v}_t$. The variation of the turbulent wind fluctuations in the i^{th} direction is given by the Langevin equation,

$$\Delta\mathbf{v}_{ti}(t) = a_i(\mathbf{x},\mathbf{v}_t,t)\Delta t + b_{ij}(\mathbf{x},\mathbf{v}_t,t)\Delta\mathbf{w}_j\,. \tag{1.6}$$

In the above equation, $\mathbf{w}_j$, is the incremental component of the Wiener's stochastic process. The mean value of $\mathbf{w}_j$ is zero and its variance is Δt. a_i and b_{ij} are the drift and diffusion terms, respectively (see Stohl et al. [2005]).

The radioactive decay, the wet and dry depositions are taken into account while describing the mass of particles. Therefore, the mass of the particle k at time t, $m_k(t)$, can be written as follows,

$$m_k(t+\Delta t) = m_k(t)\exp(-r\Delta t)\,. \tag{1.7}$$

The coefficient r is defined as:

- for radioactive decay
$$r = \frac{\ln(2)}{T_{1/2}}\,, \tag{1.8}$$
where, $T_{1/2}$ is the half life of the particles.

- For wet scavenging
$$r = \Lambda\,. \tag{1.9}$$

- For dry deposition
$$r = \frac{v_d(h_{\text{ref}})}{2h_{\text{ref}}}\,, \tag{1.10}$$
where, v_d is the dry deposition velocity and h_{ref} is a height of reference (usually 15 metres, see Stohl et al. [2005]).

The sources of emission are also taken into account with the mass of particles in a grid cell at each instant.

Finally, the average concentration of pollutants at time t and position $\mathbf{x}$ is given by the following Lagrangian formula,

$$c(\mathbf{x},t) = \frac{1}{|\Delta V(\mathbf{x})|}\sum_{k=1}^{N}\delta_k(\mathbf{x})m_k(t)\,. \tag{1.11}$$

Eq. (1.11) $|\Delta V(\mathbf{x})|$, is the volume of the grid cell to which the vector $\mathbf{x}$ is pointing. $\delta_k(\mathbf{x})$ is the Dirac distribution that is unity if the particle k is located inside the grid cell to which x is pointing and zero otherwise. N is the total number of particles. m_k is the mass of the species which are produced inside the mentioned grid cell. This is the source/sink term for the particle k. Eq. (1.11) shows that the concentration is a discrete quantity in the Lagrangian model.

Instead of Eqs. (1.5) and (1.6), one can also use the following Fokker-Planck equation in which the turbulent diffusion matrix, $\mathbf{K}$ (described in Sec. 1.1.1) appears:

$$\mathbf{x}_i(t+\Delta t) = \mathbf{x}_i(t) + \left(\mathbf{v}_i(t) + \frac{\partial \mathbf{K}_i}{\partial \mathbf{x}_i}\right)\Delta t + \sqrt{2\mathbf{K}_i}\Delta\mathbf{w}_i\,, \tag{1.12}$$

where, i is the vector component index.

1.1.3 Uncertainty of the parameters of the model

The accuracy of any of the previously presented models depends strongly on the uncertainty of the parameters [Mallet and Sportisse, 2006] and the sensitivity of the models to them. As it is difficult to correct each source of uncertainty, it is important to identify those of them which have the strongest effect on the results.

Here are listed the sources of uncertainty that can be met in the numerical simulations of air quality:

- *Input data*: boundary and initial conditions, dry deposition velocity, emission inventories, meteorological fields (temperature, pressure, wind, etc). These parameters are closely dependent on the spatial and temporal domain which is chosen for the computation.
- *Sub-grid parameterisations* : mesoscale meteorological parameterisations, mesoscale emission fields, etc.
- *The accuracy in the description of the physical phenomena* : the lack of understanding of some phenomena can lead to a non comprehensive and incorrect theoretical representation of the reality.
- *The numerical errors*: errors due to the methods of discretisation and to the solvers.

It is important to know the sensitivity of the model to each of the above mentioned points in order to get numerical results as close to reality as possible.

1.2 Observations

1.2.1 Air pollution observations

Measurements are necessary to quantify the state and quality of the atmosphere. They are useful to estimate the needed parameters to run the numerical simulations and also to validate

their results. The measurements can be performed by ground based stations, marine monitoring buoy networks, airplanes, radiosondes, LIDARs and satellites. The in-situ stations provide measurements at one single place and a network of several of them is needed to have wider spatial sight of the air quality. The other instruments (airplane, Lidar, radiosonde and satellites) can provide a spatial (vertical and horizontal) picture of the state of the atmosphere [Lahoz et al., 2010]. The fixed measurement stations usually provide information with a greater accuracy. Furthermore, at a given place, they can describe the evolution in time of the collected information. They remain necessary in boundary layer studies.

1.2.2 Measurement stations: strengths and weaknesses

Stations provide in-situ measurements (observations) of a given pollutant in air at a given location. The pollutant concentrations can be monitored in real time or samples of air can be collected and analysed in laboratories. The relevance of the measurements depends on the instruments which are used and also on the spatial representativity of the stations for the monitored pollutant. The stations are organised into networks which enables regular, if not continuous, temporal information for a whole surface area. The spatial relevance of the measurements is increased with the density of the stations in the network.

To perform a numerical simulation and check its results with the help of the in-situ measurements, the following points should be investigated:

- Assessment of the instrumental error.
- Spatial and temporal representativeness of the observations for the selected numerical spatio-temporal domain.
- Observability or the ability of the measurement network, to provide as much information as possible useful for the numerical model.

1.2.2.1 Instrumental error

The errors of measurement can arise from the data-recording facilities, the methods or processes carried out and even from the interference of inexpert operators.

If $\boldsymbol{\mu}_{\mathrm{true}}$ is a vector of physical parameters (for instance, concentrations) dependent on some continuous fields x_{true} (for instance in the frame work of air quality studies, emission inventories and meteorological data), the following relation can be written,

$$\boldsymbol{\mu}_{\mathrm{true}} = \mathcal{H}(x_{\mathrm{true}})\,, \tag{1.13}$$

where $\mathcal{H}$ is a continuous operator linking $\boldsymbol{\mu}_{\mathrm{true}}$ to the continuous true state x_{true}. The instrumental error ($\boldsymbol{\epsilon}_{\mathrm{meas}}$) is the departure of the measured value of $\boldsymbol{\mu}$ from its true value.

$$\boldsymbol{\mu} = \boldsymbol{\mu}_{\mathrm{true}} + \boldsymbol{\epsilon}_{\mathrm{meas}}\,. \tag{1.14}$$

1.2.2.2 Representativeness error

Errors of representativeness arises of shifting from a continuous space to a discrete one. For numerical modelling purposes, the continuous operator $\mathcal{H}$ is replaced by the discrete operator $\mathbf{H}$ and the continuous true state x_{true} by the discrete one $\mathbf{x}^t$. In that purpose, the restriction operator $\boldsymbol{\Gamma}_s$ enables to shift from the continuous to the discrete spaces with a resolution s,

$$x_{\mathrm{true}} \xrightarrow{\boldsymbol{\Gamma}_s} \mathbf{x}^t\,. \tag{1.15}$$

The discrete operator $\mathbf{H}$ can be formalised as follows:

$$\mathbf{H} = \mathcal{H}\mathbf{\Gamma}_s^\star \,, \tag{1.16}$$

where $\mathbf{\Gamma}_s^\star$ is the prolongation operator (see section 1.3.4).

While writing the vector of physical parameters, $\boldsymbol{\mu}_{\rm true}$, this time in the discrete space, an additional error is brought into the theoretical relation, where $\boldsymbol{\epsilon}_{\rm rep}$ is called the representativeness error:

$$\boldsymbol{\mu}_{\rm true} = \mathbf{H}\mathbf{x}^t + \boldsymbol{\epsilon}_{\rm rep} \,. \tag{1.17}$$

The summation of $\boldsymbol{\epsilon}_{\rm meas}$ and $\boldsymbol{\epsilon}_{\rm rep}$ gives:

$$\boldsymbol{\mu} = \mathbf{H}\mathbf{x}^t + \boldsymbol{\epsilon}_{\rm meas} + \boldsymbol{\epsilon}_{\rm rep} \,. \tag{1.18}$$

Note that in data assimilation textbooks, the above equation is written as:

$$\boldsymbol{\mu} = \mathbf{H}\mathbf{x}^t + \boldsymbol{\epsilon}_t \,, \tag{1.19}$$

where, the modelling error, $\boldsymbol{\epsilon}_t$, does not only include the measurement and representativeness errors but also the error, $\boldsymbol{\epsilon}_{\rm model}$, due to the physical model $\mathcal{H}(x_{\rm true})$.

Let us specify that the loss of information arises from the discretisation of $x_{\rm true}$ only, and not from the operator $\mathbf{H}$. In other words, the error of representativeness is generated by the restriction operation.

As long as the results of the discrete model (at the available measurement points) remain unaffected by the grid resolution ($\mathbf{\Gamma}_s$), the representativeness error ($\boldsymbol{\epsilon}_{\rm rep}$) can be assumed to be small. Therefore, in the frame of air quality modelling, two categories of measurement stations can be identified :

- *The background stations*- are located far from the pollution sources (for instance, regional and rural stations). These stations help to measure the average quality of ambient air and are not affected by the immediate impact of pollution sources. These are good stations to fulfil the statement just above.
- *The proximity stations*- are located close to the pollution sources (for instance, urban, traffic and industrial stations). The discrete model results at these stations are strongly dependent on the grid resolution.

The proximity stations cannot be used for regional and global scale modelling as they can not provide relevant information at low grid resolution. To use the proximity stations, the resolution needs to be higher, which at global scale modelling this would challenge the performances of the computer. Furthermore, at large scale, there won't be enough available meteorological and emission related data.

1.2.2.3 Observability

The observability of a network of stations for a given study is its ability to provide relevant measurements for that particular study.

The observability of a network depends on the climatology of the area covered by the network [Mason and Bohlin, 1995]. It also depends on the location of the sources of pollution and their distance from the measurement stations. For instance, for pollutants with a small life time, the network should be dense enough to be able to detect the pollution plume.

The observability of the networks is an essential issue in inverse modelling to reconstruct the sources of pollution from the observations. Furthermore, inverse modelling can help to check the observability of the network and improve the network design. For instance, combination of optimisation studies can be performed in order to design the spread of a network [Abida and Bocquet, 2009]. Geostatistical methods, such as Kriging, are among the simplest methods used for network design. Data assimilation can also be used at a higher level of complexity.

1.3 Inverse modelling

Inverse modelling is a way to use the available information (e.g. measurements) in order to determine some specific parameters of the model (e.g. emissions fields, initial conditions, vertical diffusion, etc). The space of these parameters is called control space. Inverse modelling is not only applied to estimate the parameters of the model, but also it is used to increase the ability of the model of predicting a physical phenomena. For instance, in atmospheric chemistry, inversion studies are focused on the estimation of the parameters that impact the species concentration fields. The parameters of the model which should be estimated are the initial state for short time simulations, or the emission fields, boundary conditions or diffusion fields in long simulations.

In geophysical literature, data assimilation is the word commonly used for the methods that help to find the true state of the parameters describing a phenomenon. Data assimilation is a technique of inverse modelling for very large scale systems which are ill-posed. Although, inverse modelling focuses on the parameters, data assimilation focuses on the outputs of the model. The main problem in inverse modelling is the lack of data or the lack of observability of the computed parameters. Therefore, having an initial idea about the background information is essential. The latter are used to regulate (or adjust) the model parameters.

1.3.1 General description

In order to estimate the vector $\mathbf{x} \in \mathbb{R}^n$, n being the number of model parameters, from a set of m observations represented by a vector $\boldsymbol{\mu} \in \mathbb{R}^m$, the linear map $\mathbf{H} \in \mathbb{R}^{m \times n}$ (which depicts a linear physical model) is used to link the model parameters to the observation vector,

$$\boldsymbol{\mu} = \mathbf{H}\mathbf{x} + \boldsymbol{\epsilon}\,. \tag{1.20}$$

In Eq.(1.20), $\boldsymbol{\epsilon} \in \mathbb{R}^m$ is the *modelling error*. The latter represents the mismatch between the observations and the model results. The information arising from the observations is the key to get reliable results from the model. However, the extraction of information using inverse modelling in atmospheric studies is difficult for several reasons.

As it is often question in geophysical data assimilations, the system under study is often largely underdetermined: i.e. $m \ll n$. One of the techniques commonly used is to aggregate the control parameters $\mathbf{x}$ into coarser variables to reduce the number of effective parameters. One can assume a map $g : \mathbb{R}^n \rightarrow \mathbb{R}^l$, where: $\mathbf{x} = g(\boldsymbol{\alpha})$. The dimension of $\boldsymbol{\alpha} \in \mathbb{R}^l$ is a nip-and-tuck of the dimension of the observations vector, $\boldsymbol{\mu}$. Although this methodology helps to reduce the control parameters, it may lead to a loss in the resolution of the results.

Moreover, the results of the model ($\mathbf{H}$) are not always sensitive enough with respect to all the model parameters. In that case, the extraction of significant information in order to adjust the parameters from the observations is difficult. For instance, in atmospheric studies, while assimilating the source emission parameters, the model does not always retrieve the information far from the observation. In atmospheric transport models, this is a common problem due to the effective diffusion term generated by the turbulent mixing.

Furthermore, the errors coming from the estimations of the model make it still more complicated to find a reliable solution for $\mathbf{x}$.

For all these reasons, the inverse problem is often *ill-posed* in atmospheric studies. Therefore, it is mandatory that the a priori information (background or first guess) should be accounted for, in order to resolve the inverse problem. The Bayesian approach (based on the Bayes formalism) allows to consistantly include the statistics of the errors and the model parameters in the inversion system [Bennett, 1992; Kaipio and Somersalo, 2010; Rodgers, 2000].

1.3.2 Bayesian approach

1.3.2.1 Bayesian inference and maximum likelihood

The Bayesian inference is based on two antecedents, the probability of the a priori model parameters, denoted $p_b(\mathbf{x})$, and the modelling error probability, denoted $p_e(\boldsymbol{\epsilon})$. The approach leads to compute the posterior probability following the Bayes' rule. The difficulties encountered by this approach are the estimations of the distribution, the uncertainties involved in the background of the parameters and the statistical parameters of the modelling error. The probability density function (*pdf*) of the modelling error (or observation mismatch) is called the likelihood function. The latter can be interpreted as the probability of the observations, given a vector of parameters evidence, $p_e(\boldsymbol{\mu}|\mathbf{x}) = p_e(\boldsymbol{\mu} - \mathbf{H}\mathbf{x})$. According to the Bayes' rule

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\mu}) &= \frac{p_e(\boldsymbol{\mu}|\mathbf{x})p_b(\mathbf{x})}{p(\boldsymbol{\mu})} \\ &= \frac{p_e(\boldsymbol{\mu} - \mathbf{H}\mathbf{x})p_b(\mathbf{x})}{p(\boldsymbol{\mu})} \,. \end{aligned} \tag{1.21}$$

The denominator term in Eq. (1.21), $p(\boldsymbol{\mu})$, is disconnected from the models and the control parameters. This is the so-called marginal-likelihood or model evidence.

Different methods can be used to compute a reliable value for the vector $\mathbf{x}$. One of them, stands on maximising the probability of the variable $\mathbf{x}$. The maximum a posteriori estimator (MAP) specifies the solution of the parameters of the model according to

$$\mathbf{x}_a = \mathrm{argmin}_{\mathbf{x}}\, p(\mathbf{x}|\boldsymbol{\mu})\,. \tag{1.22}$$

Often, the log-likelihood method is used for that purpose. Therefore, one can define the objective cost function as follows:

$$J(\mathbf{x}) = -\ln\left(p_e(\boldsymbol{\mu} - \mathbf{H}\mathbf{x})\right) - \ln\left(p_b(\mathbf{x})\right)\,. \tag{1.23}$$

The marginal log-likelihood term appears as a constant term in the cost function which can be eliminated. Eq. (1.23) is a simple form of the cost function, normally used in the three dimensional and four dimensional variational data assimilation (3D-Var and 4D-Var, respectively) approaches [Sasaki, 1958; Lorenc, 1986; Le Dimet and Talagrand, 1986].

1.3.2.2 Gaussian statistics

The argument of the minimum of Eq. (1.23), is closely dependent on the statistical assumptions for the likelihood function pdf, p_e, and the pdf of the prior model parameters, p_b. The assumption is commonly used for the modelling error in curve fitting method, such as the least

squares. Let us assume that $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and the model error covariance matrix is $\mathbf{R} = \mathbb{E}[\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t^{\mathrm{T}}]$. As a result, the pdf p_e is

$$p_e(\boldsymbol{\mu}|\mathbf{x}) = \frac{e^{-\frac{1}{2}\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{\epsilon}}}{\sqrt{(2\pi)^m|\mathbf{R}|}}\,. \tag{1.24}$$

The pdf of the a priori control variables can be chosen following different distributions. Specifically, the selection of the distribution depends on the nature of the parameters. In most data assimilation textbooks, this pdf is chosen to be Gaussian. Assume that $\mathbf{x}_b$ is the vector of a first guess of the control parameters such that $\mathbb{E}\left[\mathbf{x}_b - \mathbf{x}^t\right] = \mathbf{0}$. Let's also assume that $\mathbf{B} = \mathbb{E}\left[(\mathbf{x}_b - \mathbf{x}^t)(\mathbf{x}_b - \mathbf{x}^t)^{\mathrm{T}}\right]$ is the background error covariance matrix which represents the information about the uncertainty of the first guess. Therefore

$$p_b(\mathbf{x}) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b)}}{\sqrt{(2\pi)^n|\mathbf{B}|}}\,. \tag{1.25}$$

Using these two pdf, the cost function, Eq.(1.23), can be rewritten as:

$$J(\mathbf{x}) = \frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{H}\mathbf{x}) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_b) + \frac{1}{2}\ln\left((2\pi)^{m+n}|\mathbf{R}||\mathbf{B}|\right)\,. \tag{1.26}$$

The very last term in Eq. (1.26) is independent from $\mathbf{x}$, and the cost function can be reformulated as

$$J(\mathbf{x}) = \frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{\mu} - \mathbf{H}\mathbf{x}) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_b)\,. \tag{1.27}$$

The argument of the minimum of Eq. (1.27) is equivalent to the solution of the maximum likelihood method. The Gaussian assumption on the pdf of the a priori control variables leads to the creation of a regulation term of the *Tikhonov* kind [Tikhonov and Arsenin, 1977]. That regulation term guarantees the existence of a unique solution for the problem, even though the inverse problem is ill-defined. This solution can be written as

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{B}\mathbf{H}^{\mathrm{T}}\left(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right)^{-1}(\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b)\,, \tag{1.28}$$

where,

$$\mathbf{K} = \mathbf{B}\mathbf{H}^{\mathrm{T}}\left(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right)^{-1}\,. \tag{1.29}$$

$\mathbf{K}$ is the so-called *gain* matrix. This solution (Eq. (1.28)) is known as the *Best Linear Unbiased Estimator* (BLUE).

1.3.2.3 The posterior distribution

If $\mathbf{x}$ and $\boldsymbol{\epsilon}$ are two independent normal random vectors, then the posterior distribution $p(\mathbf{x}|\boldsymbol{\mu})$ follows the same distribution. According to Eq. (1.28),

$$\mathbf{x}_a - \mathbf{x}^t = (\mathbf{I} - \mathbf{K}\mathbf{H})(\mathbf{x}_b - \mathbf{x}^t) + \mathbf{K}\boldsymbol{\epsilon}_t\,. \tag{1.30}$$

The above equation is the key to estimate the statistical parameters of the posterior distribution $p(\mathbf{x}|\boldsymbol{\mu})$. Using Eq. (1.29), one can easily deduce that

$$\mathbb{E}\left[\mathbf{x}_a - \mathbf{x}^t\right] = 0\,, \tag{1.31}$$

and,

$$\begin{aligned}\mathbf{P}_a &= \mathbb{E}\left[(\mathbf{x}_a-\mathbf{x}_t)(\mathbf{x}_a-\mathbf{x}_t)^{\mathrm{T}}\right]\\ &= (\mathbf{I}-\mathbf{KH})\mathbf{B}\\ &= \mathbf{B}-\mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}\mathbf{H}\mathbf{B}\,.\end{aligned} \tag{1.32}$$

The a posteriori pdf, which constitutes the probabilistic Bayes' interference result reads

$$p(\mathbf{x}|\boldsymbol{\mu}) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_a)^{\mathrm{T}}\mathbf{P}_a^{-1}(\mathbf{x}-\mathbf{x}_a)}}{\sqrt{(2\pi)^n|\mathbf{P}_a|}}\,. \tag{1.33}$$

1.3.2.4 The marginal likelihood

The likelihood of the observation set, $p(\boldsymbol{\mu})$, is a key to estimate of the uncertainty matrix, $\mathbf{B}$ and $\mathbf{R}$ [Desroziers and Ivanov, 2001]. That's why its computation is important. According to the law of total probability:

$$p(\boldsymbol{\mu}) = \int_{\mathbb{R}^n\times\mathbb{R}^m}\delta(\boldsymbol{\mu}-\mathbf{H}\mathbf{x})p_e(\boldsymbol{\epsilon})p_b(\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\epsilon} = \int_{\mathbb{R}^n} p_e(\boldsymbol{\mu}|\mathbf{x})p_b(\mathbf{x})\mathrm{d}\mathbf{x}\,, \tag{1.34}$$

where δ is Dirac delta function. Replacing Eq. (1.24) and Eq. (1.25) in the above formula gives

$$p(\boldsymbol{\mu}) = \frac{1}{\sqrt{(2\pi)^{m+n}|\mathbf{R}||\mathbf{B}|}}\int_{\mathbb{R}^n} e^{-\frac{1}{2}\left((\boldsymbol{\mu}-\mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\boldsymbol{\mu}-\mathbf{H}\mathbf{x})+(\mathbf{x}-\mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b)\right)}\mathrm{d}\mathbf{x}\,. \tag{1.35}$$

The above equation (Eq. 1.24) can be reformulated as below (see Appendix A)

$$p(\boldsymbol{\mu}) = \frac{1}{\sqrt{(2\pi)^{m+n}|\mathbf{R}||\mathbf{B}|}}\int_{\mathbb{R}^n} e^{-\frac{1}{2}\left((\boldsymbol{\mu}-\mathbf{H}\mathbf{x}_b)^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}(\boldsymbol{\mu}-\mathbf{H}\mathbf{x}_b)+(\mathbf{x}-\mathbf{x}_a)^{\mathrm{T}}\mathbf{P}_a^{-1}(\mathbf{x}-\mathbf{x}_a)\right)}\mathrm{d}\mathbf{x}\,. \tag{1.36}$$

Finally:

$$p(\boldsymbol{\mu}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{H}\mathbf{x}_b)^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}(\boldsymbol{\mu}-\mathbf{H}\mathbf{x}_b)}}{\sqrt{(2\pi)^m|\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}|}}\int_{\mathbb{R}^n}\frac{e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_a)^{\mathrm{T}}\mathbf{P}_a^{-1}(\mathbf{x}-\mathbf{x}_a)}}{\sqrt{(2\pi)^n|\mathbf{P}_a|}}\mathrm{d}\mathbf{x}\,. \tag{1.37}$$

The term presented in the integral denotes the a posteriori pdf, $p(\mathbf{x}|\boldsymbol{\mu})$. The integral in the above formulation (Eq. (1.37)) is equal to unity. Then the pdf of the observation reads

$$p(\boldsymbol{\mu}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{H}\mathbf{x}_b)^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}(\boldsymbol{\mu}-\mathbf{H}\mathbf{x}_b)}}{\sqrt{(2\pi)^m|\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}|}}\,. \tag{1.38}$$

1.3.2.5 Degrees of freedom for the signal

The goal of inverse modelling is to estimate the model parameters as reliably as possible. That means the posterior uncertainty of the model parameters, compacted in the term $\mathbf{P}_a$, is smaller for a given prior uncertainty $\mathbf{B}$. According to Eq. (1.32), if the term $\mathbf{KH}$ is close to the identity matrix, the posterior value of the model parameters becomes more certain. The symmetric matrix, $\mathbf{A} = \mathbf{KH}$, is the so-called *averaging kernel*. The *degrees of freedom for the signal* (DFS) is a quantity closed to the a posteriori uncertainty of the model parameters. This value extracts the fraction of the observations used in the data assimilation system to retrieve the solution. It reads,

$$\mathrm{DFS} = \mathbb{E}\left[(\mathbf{x}_a-\mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}_a-\mathbf{x}_b)\right] = \mathrm{Tr}(\mathbf{A}) = \mathrm{Tr}(\mathbf{HK}) = \mathrm{Tr}(\mathbf{KH})\,. \tag{1.39}$$

The singular vector decomposition of the averaging kernel gives:

$$\mathrm{DFS} = \sum_{l=1}^{m} w_l \,. \tag{1.40}$$

where w_l is the l^{th} eigenvalue of $\mathbf{A}$.

When the observations are representative enough, the value of the DFS shows the quality of the analysis. The higher is the DFS, the higher is the information recovered from the observations. The computation of the DFS is inconsistent under the non-Gaussian assumption for the a priori pdf of the control parameters. However, under the positivity enforcing, in the presence of the positive background of the control parameters and when the uncertainty of the control parameters is not very high, that can be also applied. Note that, in a non-Gaussian context, the computation of relative entropy can be used in order to understand the quality of assimilation [Bocquet, 2008].

1.3.3 Prior error estimation

The results of the inverse system depend not only on the model and on the observations but also on the prior error estimations. One of the difficulties for inverse modelling studies is the assessment of the error covariance matrix, $\mathbf{R}$ and $\mathbf{B}$. This section introduces the maximum likelihood method in order to estimate the prior errors.

1.3.3.1 Gaussian assumption

Let us assume that the two error covariance matrix, $\mathbf{B}$ and $\mathbf{R}$, are as follows

$$\mathbf{R} = r^2\mathbf{R}_0 \quad , \quad \mathbf{B} = \beta^2\mathbf{B}_0 \,. \tag{1.41}$$

where $\mathbf{R}_0$ and $\mathbf{B}_0$ are the first estimation of the information about the prior parameters. r and β, called hyper-parameters, are the parameters used to fix these two covariance matrices. In order to obtain the more likely values for r and β, one has to maximise the pdf of the observations, Eq. (1.38), with respect to r and β. The log-likelihood can be written as

$$\ln p(\boldsymbol{\mu}|r,\beta) = -\frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b)^{\mathrm{T}}(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}(\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b) - \ln|\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}| + C \,, \tag{1.42}$$

where C is a constant parameter. The optimisation of the above log-likelihood function, Eq. (1.42), with respect to the hyper-parameters, gives

$$r^2 = \frac{(\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_a)^{\mathrm{T}}\mathbf{R}_0^{-1}(\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_a)}{\mathrm{Tr}(\mathbf{I}_m - \mathbf{H}\mathbf{K})} \,, \tag{1.43}$$

and,

$$\beta^2 = \frac{(\mathbf{x}_a - \mathbf{x}_b)^{\mathrm{T}}\mathbf{B}_0^{-1}(\mathbf{x}_a - \mathbf{x}_b)}{\mathrm{Tr}(\mathbf{K}\mathbf{H})} \,. \tag{1.44}$$

The above two equations (1.43 and 1.43) can be used in an iterative system which converges to a fixed point. At each iteration, $\mathbf{x}_a$ and $\mathbf{K}$ are obtained from equations (1.28 and 1.29). This method was first presented by Desroziers and Ivanov [2001]. They also show that the method is equivalent to the maximum likelihood.

The χ^2 method (see Ménard et al. [2000]; Tarantola [2005]) can be derived from Deroziers method (see Chapnik et al. [2006]) and when one of the hyper-parameters is assumed to be fixed. The method is useful in the variational data assimilation method [Koohkan and Bocquet, 2012; Michalak et al., 2005].

1.3.3.2 Semi-normal assumption

When the model parameters that should be retrieved are all positive, the Gaussian pdf is not appropriate. In that case, we assume the following Gaussian truncated pdf for the a priori model parameters, is assumed:

$$p_b(\mathbf{x}) = \frac{e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b)}}{\sqrt{(2\pi)^n|\mathbf{B}|}\,(1-\Phi(\mathbf{B},\mathbf{x}_b,\mathbf{0}))}\mathbb{I}_{\mathbf{x}\geq\mathbf{0}}\,. \tag{1.45}$$

In Eq. (1.45), $\Phi(\mathbf{B},\mathbf{x}_b,\mathbf{0})$ is the cumulative distribution function (*cdf*), $\mathcal{N}(\mathbf{x}_b,\mathbf{B})$, computed over the integral from minus infinity to zero. $\mathbb{I}_{\mathbf{x}\geq\mathbf{0}}$ is a function with the value zero if $\mathbf{x}_i < 0$ for each $i = 0,..,n$ and with the value unity elsewise. According to the semi-normal assumption, the marginal probability function can be written as

$$p(\boldsymbol{\mu}|r,\beta) = \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{H}\mathbf{x}_b)^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}(\boldsymbol{\mu}-\mathbf{H}\mathbf{x}_b)}}{(1-\Phi(\mathbf{B},\mathbf{x}_b,\mathbf{0}))\sqrt{(2\pi)^m|\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}|}}\int\frac{e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_a)^{\mathrm{T}}\mathbf{P}_a^{-1}(\mathbf{x}-\mathbf{x}_a)}}{\sqrt{(2\pi)^n|\mathbf{P}_a|}}\mathbb{I}_{\mathbf{x}\geq\mathbf{0}}\mathrm{d}\mathbf{x}\,. \tag{1.46}$$

The analytical computation of the hyper-parameters is difficult from Eq. (1.46). However, the latter can be used by choosing the input values from within an allowed set of hyper-parameters and computing the value of the function (Winiarek et al. [2011]; also in chapter 5). This method is expensive for very large systems, but it can be used for a few thousand variables.

1.3.4 Multiscale data assimilation

1.3.4.1 Scaling operators

For a given domain, a regular grid, $\boldsymbol{\Omega}$, can be defined by a discretisation in space and time. For instance, in a surface space and time discretisation (2D+T), N_x denotes the number of grid cells along the longitude, N_y denotes the number of grid cells along the latitudes axis, and N_t is the number of time step. Let us assume that $N_{\mathrm{fg}} = N_x \times N_y \times N_t$ is the finest grid resolution of the spatio-temporal domain. $\mathbf{x} \in \mathbb{R}^{N_{\mathrm{fg}}}$ is the vector which gives the control parameters for the finest resolution. Now, let us assume that $\mathcal{R}(\boldsymbol{\Omega})$ is the dictionary of all of the adaptative grids representing the domain. A *representation* ω is a member of $\mathcal{R}(\boldsymbol{\Omega})$, which gives a spatio-temporal discretisation of the domain, such that, $\mathbf{x}^\omega \in \mathbb{R}^N$, $N \leq N_{\mathrm{fg}}$ (see Bocquet et al. [2011]; Bocquet [2009]). A restriction operator, $\boldsymbol{\Gamma}_\omega : \mathbb{R}^N \rightarrow \mathbb{R}^{N_{\mathrm{fg}}}$ denotes how the vector of control parameters $\mathbf{x}$ is coarse-grained into the vector $\mathbf{x}^\omega$. Vice-versa, the prolongation operator, $\boldsymbol{\Gamma}_\omega^\star : \mathbb{R}^{N_{\mathrm{fg}}} \rightarrow \mathbb{R}^N$, refines the vector $\mathbf{x}^\omega$ into $\widetilde{\mathbf{x}} \in \mathbb{R}^{N_{\mathrm{fg}}}$, where

$$\mathbf{x}^\omega = \boldsymbol{\Gamma}_\omega\mathbf{x} \quad , \quad \widetilde{\mathbf{x}} = \boldsymbol{\Gamma}_\omega^\star\mathbf{x}^\omega\,. \tag{1.47}$$

Since $N \leq N_{\mathrm{fg}}$, the loss of information occurs during the restriction operation. The composition of the prolongation and the restriction operator is the identity $\boldsymbol{\Gamma}_\omega\boldsymbol{\Gamma}_\omega^\star = \mathbf{I}_N$. The operator $\boldsymbol{\Gamma}_\omega^\star$ is ambiguous since additional information is needed to reconstruct the vector $\mathbf{x}$ from $\widetilde{\mathbf{x}}$. A simplest choice for that operator is to set $\boldsymbol{\Gamma}_\omega^\star = \boldsymbol{\Gamma}_\omega^{\mathrm{T}}$. The best choice to determine this operator is the use of the Bayes' rule. The method is based on maximising the probability of $\mathbf{x}$, for a given representation $\mathbf{x}_\omega$. As mentioned before (Eq. (1.25)), the random variable $\mathbf{x}$ can be assumed to be Gaussian: $\mathbf{x} \sim \mathcal{N}(\mathbf{x}_b,\mathbf{B})$. From Bayes' rule, one can write

$$p(\mathbf{x}|\mathbf{x}^\omega) = \frac{p(\mathbf{x})\delta(\mathbf{x}^\omega - \boldsymbol{\Gamma}_\omega\mathbf{x})}{p_\omega(\mathbf{x}^\omega)}\,, \tag{1.48}$$

where δ is the Dirac distribution. For a linear operator $\mathbf{\Gamma}_\omega$, the pdf of $\mathbf{x}$ in a representation ω, $p_\omega(\mathbf{x}^\omega)$ remains Gaussian: $\mathbf{x}^w \sim \mathcal{N}(\mathbf{x}_b^\omega, \mathbf{B}_\omega)$,

$$\mathbf{x}_b^\omega = \mathbf{\Gamma}_\omega \mathbf{x}_b \quad , \quad \mathbf{B}_\omega = \mathbf{\Gamma}_\omega \mathbf{B} \mathbf{\Gamma}_\omega^{\mathrm{T}} \,. \tag{1.49}$$

The optimum solution of Eq. (1.48) can be computed with the help of the BLUE analysis

$$\mathbf{x}^\star = \mathbf{x}_b + \mathbf{\Lambda}_\omega^\star (\mathbf{x}_\omega - \mathbf{\Gamma}_\omega \mathbf{x}_b) \,, \tag{1.50}$$

where

$$\mathbf{\Lambda}_\omega^\star = \mathbf{B} \mathbf{\Gamma}_\omega^{\mathrm{T}} (\mathbf{\Gamma}_\omega \mathbf{B} \mathbf{\Gamma}_\omega^{\mathrm{T}})^{-1} \,. \tag{1.51}$$

Moreover, the projection operator is defined as,

$$\mathbf{\Pi}_\omega = \mathbf{\Lambda}_\omega^\star \mathbf{\Gamma}_\omega \,, \tag{1.52}$$

so that, we can choose the prolongation operator to be:

$$\mathbf{\Gamma}_\omega^\star : \mathbf{x}^\omega \rightarrow (\mathbf{I}_{N_{\mathrm{fg}}} - \mathbf{\Pi}_\omega) \mathbf{x}_b + \mathbf{\Lambda}_\omega^\star \mathbf{x}^\omega \,. \tag{1.53}$$

The composition of the restriction and the prolongation operator gives

$$\mathbf{\Gamma}_\omega^\star \mathbf{\Gamma}_\omega : \mathbf{x} \rightarrow (\mathbf{I}_{N_{\mathrm{fg}}} - \mathbf{\Pi}_\omega) \mathbf{x}_b + \mathbf{\Pi}_\omega \mathbf{x} \,. \tag{1.54}$$

When $\mathbf{x}_b = \mathbf{0}$, the projection operator $\mathbf{\Pi}_\omega$ is equal to $\mathbf{\Gamma}_\omega^\star \mathbf{\Gamma}_\omega$. This operator satisfies the following equations:

$$\mathbf{\Pi}_\omega^2 = \mathbf{\Pi}_\omega \quad , \quad \mathbf{\Pi}_\omega \mathbf{B} = \mathbf{B} \mathbf{\Pi}_\omega^{\mathrm{T}} \tag{1.55}$$

If the representation ω is coarse, $\mathrm{Tr}(\mathbf{\Pi}_\omega) \ll N_{\mathrm{fg}}$. For a representation ω close to the finest grid $N_{\mathrm{cg}} \ll \mathrm{Tr}(\mathbf{\Pi}_\omega)$ (N_{cg} is the number of cells in the coarsest grid resolution). The higher $\mathrm{Tr}(\mathbf{\Pi}_\omega)$, the better the recovered information. However, this operator cannot be the identity because the coarse-graining implies a loss of information.

1.3.4.2 Multiscale source receptor model

The source receptor model, Eq. (1.20), can be written in any representation ω. The Jacobian matrix $\mathbf{H}$ in the representation ω changes to $\mathbf{H}_\omega = \mathbf{H} \mathbf{\Gamma}_\omega^\star$. The scale-dependent source-receptor model is defined as:

$$\begin{aligned} \boldsymbol{\mu} &= \mathbf{H}\mathbf{x} + \boldsymbol{\epsilon} \\ &= \mathbf{H} \mathbf{\Gamma}_\omega^\star \mathbf{\Gamma}_\omega \mathbf{x} + \mathbf{H}(\mathbf{I}_{N_{\mathrm{fg}}} - \mathbf{\Gamma}_\omega^\star \mathbf{\Gamma}_\omega)\mathbf{x} + \boldsymbol{\epsilon} \\ &= \mathbf{H}_\omega \mathbf{x}^\omega + \boldsymbol{\epsilon}_\omega \,. \end{aligned} \tag{1.56}$$

Where, the $\boldsymbol{\epsilon}_\omega$ is a scale-made dependent error:

$$\boldsymbol{\epsilon}_\omega = \mathbf{H}(\mathbf{I}_{N_{\mathrm{fg}}} - \mathbf{\Gamma}_\omega^\star \mathbf{\Gamma}_\omega)\mathbf{x} + \boldsymbol{\epsilon} = \mathbf{H}(\mathbf{I}_{N_{\mathrm{fg}}} - \mathbf{\Pi}_\omega)(\mathbf{x} - \mathbf{x}_b) + \boldsymbol{\epsilon} \tag{1.57}$$

Using Eq. (1.54), the source receptor model can be reformulated as

$$\boldsymbol{\mu} = \mathbf{H}\mathbf{x}_b + \mathbf{H}\mathbf{\Pi}_\omega(\mathbf{x} - \mathbf{x}_b) + \boldsymbol{\epsilon}_\omega \,. \tag{1.58}$$

The observation covariance matrix in a representation ω is different from that one in the finest grid

$$\mathbf{R}_\omega = \mathbf{R} + \mathbf{H}(\mathbf{I}_{N_{\mathrm{fg}}} - \mathbf{\Pi}_\omega)\mathbf{B}\mathbf{H}^{\mathrm{T}} \,. \tag{1.59}$$

The term $\mathbf{H}(\mathbf{I}_{N_{\mathrm{fg}}} - \mathbf{\Pi}_\omega)\mathbf{B}\mathbf{H}^{\mathrm{T}}$, in the above equation, Eq. (1.59), leads to an increase of the observation covariance matrix term. The term $\mathbf{H}(\mathbf{I}_{N_{\mathrm{fg}}} - \mathbf{\Pi}_\omega)(\mathbf{x} - \mathbf{x}_b)$ is identified as the aggregation error.

1.3.4.3 Design criteria

DFS criterion

The degrees of freedom for the signal quantify the quality of the analysis. As presented in Section 1.3.2.5, the DFS value is computed by the trace of the averaging kernel ($\mathrm{Tr}(\mathbf{KH})$). In a multi-scale context, one hopes to find a representation ω, which maximises the DFS. This criterion is expressed as:

$$\mathcal{J}_\omega = \mathrm{Tr}(\mathbf{I}_N - \mathbf{B}_\omega^{-1}\mathbf{P}_a^\omega) = \mathrm{Tr}\left(\mathbf{\Pi}_\omega \mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}\mathbf{H}\right) \tag{1.60}$$

Fisher criterion

This criterion measures the reduction of uncertainty granted by the observations. This criterion can be computed in the finest grid, $\mathbf{\Omega}$, according to

$$\mathcal{J} = \mathrm{Tr}\left(\mathbf{B}\mathbf{P}_a^{-1}\right) = \mathrm{Tr}\left(\mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right) . \tag{1.61}$$

For a given representation ω, the criterion reads

$$\mathcal{J}_\omega = \mathrm{Tr}\left(\mathbf{B}_\omega \mathbf{H}_\omega^{\mathrm{T}}\mathbf{R}_\omega^{-1}\mathbf{H}_\omega\right) = \mathrm{Tr}\left(\mathbf{\Pi}_\omega \mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right) . \tag{1.62}$$

The aggregation error is related to the representation. The two equations above (1.61-1.62) give an assessment of that error. The following equation (1.63), presents the normalised aggregation error for a given representation ω [Koohkan et al., 2012; Wu et al., 2011]:

$$\mathrm{Tr}\left(\mathbf{R}^{-1}(\mathbf{R}_\omega - \mathbf{R})\right) = \mathrm{Tr}\left(\mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right) - \mathrm{Tr}(\mathbf{\Pi}_\omega \mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}) . \tag{1.63}$$

Note that the Fisher criterion is the limiting case of the DFS criterion when $\mathbf{R}$ is inflating or $\mathbf{B}$ is vanishing.

1.3.4.4 Adaptive tiling

To apply the multiscale extension in 2D+T, the dictionary of representations, $\mathcal{R}(\mathbf{\Omega})$ should be handled mathematically. Let us assume that N_x, N_y and N_t are multiples of 2^{n_x}, 2^{n_y} and 2^{n_t}, respectively. For each scale $\mathbf{l} = (l_x, l_y, l_t)$, such that $0 \leq l_x \leq n_x$, $0 \leq l_y \leq n_y$ and $0 \leq l_t \leq n_t$, the domain is presented with $N_{\mathrm{fg}} \times 2^{-(l_x+l_y+l_t)}$ coarser cells. The latter are called the *tiles*. In all directions, a coarser grid is made when two adjacent grids in each direction are gathered into one. The physical quantities of the coarser cell are the average of those from the finest grid cells.

A physical quantity in each cell of the finest grid, indexed by k is attached to a base vector $\mathbf{u}_{i,j,h} \in \mathbb{R}^{N_{\mathrm{fg}}}$ with $1 \leq i \leq N_x$, $1 \leq j \leq N_y$ and $1 \leq h \leq N_t$. At a coarser scale $\mathbf{l}$, this quantity in the finest grid, which is enumerated by k, attaches to the vector: $\mathbf{v}_{\mathbf{l},k} = \sum_{\delta_i=0}^{2^{l_x}-1} \sum_{\delta_j=0}^{2^{l_y}-1} \sum_{\delta_h=0}^{2^{l_t}-1} \mathbf{u}_{i_k+\delta_i, j_k+\delta_j, h_k+\delta_h}$, where (i_k, j_k, h_k) denotes the index of cell k in the coarser representation. For a representation ω of $\mathbf{\Omega}$, the projection operator defines:

$$\mathbf{\Pi}_\omega = \sum_l \sum_{k=1}^{n_\mathbf{l}} \alpha_{\mathbf{l},k}^\omega \frac{\mathbf{v}_{\mathbf{l},k}\mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}}{\mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}\mathbf{v}_{\mathbf{l},k}} , \tag{1.64}$$

where $\alpha_{\mathbf{l},k}^\omega$ are the coefficients which define the representation ω. To obtain an *admissible* representation ω, each parameter $\alpha_{\mathbf{l},k}^\omega$ is set to 0 or 1. Each cell of the finest grid cell should be attached in the representation ω. Therefore,

$$\sum_l \sum_{k=1}^{n_\mathbf{l}} \alpha_{\mathbf{l},k}^\omega \mathbf{v}_{\mathbf{l},k} = (1, ..., 1)^{\mathrm{T}} . \tag{1.65}$$

The number of multiscale grid cells N should be imposed

$$N_{\mathrm{fg}} \times 2^{-(n_x+n_y+n_t)} \leq \sum_{k=1}^{n_{\mathbf{l}}} \alpha^{\omega}_{\mathbf{l},k} = N \leq N_{\mathrm{fg}}\,. \tag{1.66}$$

1.3.4.5 Optimisation

In order to optimise the cost function, $\mathcal{J}_\omega$, in a fixed number of tiles, the following Lagrangian function is defined:

$$\mathcal{L}(\omega) = \sum_{l}\sum_{k=1}^{n_l} \alpha^{\omega}_{\mathbf{l},k} \frac{\mathbf{v}_{\mathbf{l},k}\mathbf{Q}\mathbf{v}^{\mathrm{T}}_{\mathbf{l},k}}{\mathbf{v}^{\mathrm{T}}_{\mathbf{l},k}\mathbf{v}_{\mathbf{l},k}} + \sum_{k=1}^{N_{\mathrm{fg}}} \boldsymbol{\lambda}_k \left(\sum_{l} \alpha^{\omega}_{\mathbf{l},k} - 1 \right) + \xi \left(\sum_{l}\sum_{k=1}^{n_l} \alpha^{\omega}_{\mathbf{l},k} - N \right)\,. \tag{1.67}$$

The first term on the right hand side of this equation stands for $\mathcal{J}_\omega$. $\mathbf{Q}$ is the average kernel matrix for the DFS criterion and is equal to $\mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ for the Fisher criterion. Since the value of $\alpha^{\omega}_{\mathbf{l},k}$ should be 0 or 1, a second term is added to the right hand side of Eq. (1.67). The vector $\boldsymbol{\lambda}$ is the Lagrangian multiplier. The very last term of the above equation aims to satisfy the condition of Eq. (1.66). ξ is a scalar and is called the Lagrange multiplier. The Lagrangian objective function can be also written as follows:

$$\mathcal{L}(\omega) = \sum_{l}\sum_{k=1}^{n_l} \left(\frac{\mathbf{v}_{\mathbf{l},k}\mathbf{Q}\mathbf{v}^{\mathrm{T}}_{\mathbf{l},k}}{\mathbf{v}^{\mathrm{T}}_{\mathbf{l},k}\mathbf{v}_{\mathbf{l},k}} + \mathbf{v}^{\mathrm{T}}_{\mathbf{l},k}\boldsymbol{\lambda} + \xi \right) \alpha^{\omega}_{\mathbf{l},k} - \sum_{k=1}^{N_{\mathrm{fg}}} \boldsymbol{\lambda}_k - \xi N \tag{1.68}$$

The optimisation of the representation ω can be done by minimising the dual function of $\mathcal{L}$. Therefore, it comes:

$$\hat{\mathcal{L}}(\boldsymbol{\lambda},\xi) = \sum_{l}\sum_{k=1}^{n_l} \max\left(0, \frac{\mathbf{v}_{\mathbf{l},k}\mathbf{Q}\mathbf{v}^{\mathrm{T}}_{\mathbf{l},k}}{\mathbf{v}^{\mathrm{T}}_{\mathbf{l},k}\mathbf{v}_{\mathbf{l},k}} + \mathbf{v}^{\mathrm{T}}_{\mathbf{l},k}\boldsymbol{\lambda} + \xi \right) \alpha^{\omega}_{\mathbf{l},k} - \sum_{k=1}^{N_{\mathrm{fg}}} \boldsymbol{\lambda}_k - \xi N \tag{1.69}$$

The above introduced cost function, $\hat{\mathcal{L}}(\boldsymbol{\lambda},\xi)$, cannot be optimised directly using a gradient-based minimisation algorithm. Besides, the uniqueness of the solution is not guaranteed. To overcome that difficulty, a regularisation method for the cost function is used in Bocquet [2009].

1.4 Outline

New methodologies of data assimilation are presented in the following chapters:

- In *chapter* 2 are detailed the adjoint of the Eulerian Chemistry Transport Model and the 4D-Var method.
- In *chapter* 3, the 4D-Var model is used to invert the emission inventories of carbon monoxide provided by the EMEP (*European Monitoring and Evaluation Program*). The observations used for the inversion are impacted with the representativeness error. As the 4D-Var routine is not fit to handle the representativeness error, a subgrid model is developed and coupled to the 4D-Var algorithm.
- *chapter* 4- introduces the observations to be retrieved from the International Monitoring System (IMS) radionuclide network. The compatibility of the observation network with data assimilation, in other words, the ability of the network to observe the radionuclide

pollutants, is discussed. In order to build the Jacobian matrix of the atmospheric transport model, the Lagrangian FLEXPART model and also the adjoint of the Eulerian POLAIR3D model of POLYPHEMUSare used.

- In *chapter* 5 is shown the application which can be made of the maximum likelihood method in order to estimate the uncertainty of the model parameters, as well as, the covariance matrix of the model errors. A fast version of POLAIR3D CTM and its adjoint are developed. The emission inventories of the Volatile Organic Compounds (VOCs) are inverted. The observations extracted from the EMEP database are assimilated.
- *Chapter* 6 concludes on the achievements of this work and presents some interesting points to investigate.

The points dealt with in chapter 2 and 3 are presented in Koohkan and Bocquet [2012]. The contents of chapter 4 was published in Koohkan et al. [2012]. Chapter 5 was submitted to Atmospheric Chemistry and Physics.

Chapter 2

Four-Dimensional Variational Data Assimilation

Summary

The present chapter describes the 4D-Var method that we have used. First of all, an approximate adjoint model of the chemistry transport model is introduced and developed. The adjoint solution is validated via a duality test. Then, the way this adjoint is taken into account in the 4D-Var algorithm is described. The latter algorithm is also checked through two gradient tests. The duality test shows that the concentrations computed with the help of the adjoint solution are in good agreement with the concentrations computed using the CTM, directly. The Pearson correlation between the solutions of the two models for a tracer species is of 99.8%. The gradient of the cost function obtained with the adjoint model is compared with the one obtained with the finite difference method. The results show that the gradient of the cost function obtained via the adjoint solution is correct enough to be used in the optimisation algorithm.

Contents

2.1 Introduction

Five decades after the appearance of data assimilation [Cressman, 1959; Gandin, 1963], Four-Dimensional Variational method (4D-Var) has become one of the most important tools to estimate the parameters of a physical model. Data assimilation combines experimental data, information coming from models (chemistry and transport model in our case) and statistics of the errors in order to find the optimum values of the parameters which minimise the observations mismatch.

Variational data assimilation is a powerful method when it comes to constraining dynamical systems by numerous observations. In variational data assimilation, all types of information mentioned above are accounted for altogether in a two-term objective cost function $\mathcal{J} = \mathcal{J}_o + \mathcal{J}_b$. The first term $\mathcal{J}_o$ is a measure of the discrepancy between the observed and simulated concentrations. The second term $\mathcal{J}_b$ evaluates the departure of the control parameters from their first guess (background). By minimising the sum of these two terms, 4D-Var makes in our case an optimal compromise while enforcing the fact that the simulated concentrations are obtained from a given numerical transport model.

One way to minimise the cost function, $\mathcal{J}$, is to use an analytical method, granted that the function in question is continuous and derivable. That way supposes to solve the algebraic equations equal to zero of the gradient function. The solution obtained is a minimum point if the function is locally convex. This analytical method includes a system of n equations, where n is the number of control parameters. It leads to find an analytical solution for the optimal control parameters and actually performs well when the dimension of the problem is small. Another way to minimise the cost function consists in using an iterative descent algorithm, such as, the conjugate gradient or the quasi-Newton algorithm. In the descent algorithm, an initial value for the control parameters is chosen and the local gradient of the cost function is computed. The latter gradient shows the direction of decrease of the cost function. Therefore, the optimal point (control parameters) for the cost function in that direction can be found by the line search method. Using the updated parameters, the above described procedure is repeated until the cost function is minimised.

In order to find the gradient of the cost function with respect to the control parameters, an adjoint of the numerical model is needed. To compute that adjoint, besides hand calculation which is almost impossible in air quality context, two methods can be used: the automatic differentiation and the approximated adjoint [Krysta and Bocquet, 2007]. The former is used to evaluate the derivative of a function specified by a computer program. The latter is an analytical method which gives the adjoint solution of the chemistry transport model. The adjoint of a dynamical model was first used by Kontarev [1980] and Hall and Cacuci [1983] for sensitivity studies. Le Dimet and Talagrand [1986] proposed an algorithm for minimising the 4D-Var cost function with the help of the adjoint dynamical equation in the framework of meteorology.

The adjoint model is described in Section 2.2. In Section 2.3, a continuous expansion of the control optimal problem is introduced and control variables space are defined. Then, the numerical discretisation of the 4D-Var system is presented. Finally, the optimisation algorithm is described. The adjoint model, as well as, the descent algorithm are validated in Section 2.4. A short conclusion follows in section 2.5.

2.2 The adjoint model

In air quality modelling, the concentrations of pollutants at each time and location can be obtained by a chemistry transport model (CTM). For the species with a linear chemistry and physics, the Jacobian of the CTM can be computed either using the solution generated by the CTM, or the one generated by the adjoint CTM model. For instance, in the studies of accidental cases, when the source is pointwise, the Jacobian matrix of the model is easier to compute using the solution of the CTM. In case the number of sources is higher than the number of measurements, the Jacobian is easier to build via the adjoint model. Furtheremore, the adjoint model can be used in the 4D-Var algorithm to compute the gradient of the cost function. That is why, a particular attention has to be paid to the adjoint model when assimilating atmospheric observations.

We can assume that $c(\mathbf{x}, t)$ is continuously differentiable on the space-time domain $\Omega = \mathcal{D} \times [0, T]$ ($\mathcal{D}$ denotes the spatial domain) and the chemical reaction operator $\boldsymbol{\chi}$ is linear (which is not always the case). Then Eq. (1.1) is multiplied by a sufficiently smooth function $\phi(\mathbf{x}, t)$ and integrated. The "weak" form of the CTM is:

$$\int_{\Omega} \phi \left(\frac{\partial c}{\partial t} + \nabla \cdot (\mathbf{u}c) - \nabla \cdot (\mathbf{K} \nabla c) + \Lambda c - \boldsymbol{\chi}(c) - \sigma \right) d\mathbf{x}dt = 0 \tag{2.1}$$

Note that the density of air is assumed to be constant. The above equation can be transformed into (see Roustan and Bocquet [2006a]):

$$\begin{aligned} \int_{\Omega} c \left(-\frac{\partial \phi}{\partial t} - \nabla \cdot (\mathbf{u}\phi) - \nabla \cdot (\mathbf{K} \nabla \phi) + \Lambda \phi - \boldsymbol{\chi}^{\dagger}(\phi) \right) d\mathbf{x}dt = \\ - \int_{\mathcal{D}} \phi(T)c(T) - \phi(0)c(0) d\mathbf{x} - \int_{\partial \mathcal{D}_{bc} \times [0,T]} (\phi c \mathbf{u}) \cdot d\mathbf{S}dt \\ - \int_{\partial \mathcal{D}_b \times [0,T]} (c\mathbf{K}\nabla\phi - \phi \mathbf{K} \nabla c) \cdot d\mathbf{S}dt + \int_{\Omega} \phi \sigma d\mathbf{x}dt \end{aligned} \tag{2.2}$$

where $\boldsymbol{\chi}^{\dagger}$ denotes the adjoint operator of the chemical reactions term, $\boldsymbol{\chi}$. $\partial \mathcal{D}_b$ stands for the ground surface boundary of the domain. $\partial \mathcal{D}_{bc}$ are the boundaries of the domain with the exception of the ground surface. Note also that $d\mathbf{S} = dS\mathbf{n}$.

Before further developping Eq. (2.2), let us assume the following equation:

$$\begin{aligned} \frac{\partial \phi(\mathbf{x}, \tau)}{\partial \tau} &= \operatorname{div} (\mathbf{u}(\mathbf{x}, \tau)\phi(\mathbf{x}, \tau)) + \nabla \cdot \left(\rho \mathbf{K}(\mathbf{x}, \tau) \nabla \frac{\phi(\mathbf{x}, \tau)}{\rho} \right) \\ &\quad - \Lambda(\mathbf{x}, \tau)\phi(\mathbf{x}, \tau) + \boldsymbol{\chi}^{\dagger}(\phi(\mathbf{x}, \tau), \mathbf{x}, \tau) + \pi(\mathbf{x}, \tau) \,. \end{aligned} \tag{2.3}$$

$\pi_i(\mathbf{x}, \tau)$ is a continuous sampling function defined over the time-space domain and τ is the reverse time variable: $\tau = T - t$. Eq.(2.3) can be seen as a CTM, backward in time with a reversed wind direction. One can set the following conditions for the above equation (Eq. (2.3):

- final conditions:

$$\phi(\mathbf{x}, T) = 0 \quad \mathbf{x} \in \mathcal{D} \tag{2.4}$$

- border conditions (top, left, right, front and back sides):

$$\phi(\mathbf{x},\tau) = 0 \qquad \forall(\mathbf{x},\tau) \in \partial\Omega_{out} \tag{2.5}$$

$\partial\Omega_{out} = \bigcup_t \partial\mathcal{D}^t_{out}$ denotes the border when the wind is outgoing. $\mathcal{D}^t_{out}$ is the border of the spatial domain when the wind is outgoing at instance t.

- Boundary conditions at ground level ($z = 0$):

$$\mathbf{K}(\mathbf{x}_{z=0},\tau)\nabla\phi(\mathbf{x}_{z=0},\tau)\cdot\mathbf{n} = v_d\phi(\mathbf{x}_{z=0},\tau) \tag{2.6}$$

Using Eqs. (2.3 to 2.6), Eq. (2.2) can be rewritten as follows:

$$\begin{aligned}\int_\Omega c\pi d\mathbf{x}dt &= \int_{\mathcal{D}} \phi(0)c(0)d\mathbf{x} - \int_{\partial\Omega_{in}} (\phi c\mathbf{u})\cdot d\mathbf{S}dt \\ &\quad + \int_{\partial\mathcal{D}_b\times[0,T]} \phi\mathbf{E}\cdot d\mathbf{S}dt + \int_\Omega \phi\sigma d\mathbf{x}dt\,,\end{aligned} \tag{2.7}$$

where, $\mathbf{E} = -E\cdot\mathbf{n}$ is the surface emission vector. Note that in the above equation, the term $\mathbf{K}\nabla c$ and $\mathbf{K}\nabla\phi$ are assumed to be zero at the top and around the domain.

2.3 The 4D-Var model

Four dimensional variational data assimilation (4D-Var) is used to invert the surface fluxes of non-reactive species (tracers). The analysis of the emissions is achieved by computing the minimum value of the following Lagrangian cost function:

$$\begin{aligned}\mathcal{J} &= \frac{1}{2}\int_{\mathcal{D}\times\Omega} (\sigma(\mathbf{x},t)-\sigma^b(\mathbf{x},t))B_\sigma^{-1}(\mathbf{x},\hat{\mathbf{x}},t)(\sigma(\hat{\mathbf{x}},t)-\sigma^b(\hat{\mathbf{x}},t))d\mathbf{x}d\hat{\mathbf{x}}dt \\ &\quad + \frac{1}{2}\int_{\mathcal{D}_b\times\Omega_b} (E(\mathbf{x}_p,t)-E^b(\mathbf{x}_p,t))B_E^{-1}(\mathbf{x}_p,\hat{\mathbf{x}}_p,t)(E(\hat{\mathbf{x}}_p,t)-E^b(\hat{\mathbf{x}}_p,t))d\mathbf{x}_p d\hat{\mathbf{x}}_p dt \\ &\quad + \frac{1}{2}\int_0^T (\mathbf{y}(t)-\mathcal{H}_{t,\mathbf{x}}[c])\,\mathbf{R}^{-1}(t)\,(\mathbf{y}(t)-\mathcal{H}_{t,\mathbf{x}}[c])\,dt \\ &\quad + \int_\Omega \phi\left(\frac{\partial c}{\partial t} + \nabla\cdot(\mathbf{u}c) - \nabla\cdot(\mathbf{K}\nabla c) + \Lambda c - \sigma\right) d\mathbf{x}dt\,.\end{aligned} \tag{2.8}$$

The first and the second terms on the right hand-side of the equation above are representative of the cost of the background emissions. $\sigma(\mathbf{x},t)$ is the volume emission inventories at time t and coordinate $\mathbf{x}$. $\sigma^b(\mathbf{x},t)$ denotes its first guess. $B_\sigma(\mathbf{x},\hat{\mathbf{x}},t)$ is called the background covariance of the volume emissions. $E(\mathbf{x}_p,t)$ and $E^b(\mathbf{x}_p,t)$ are the surface emission function and the background surface emission function at time t and at the surface coordinate $\mathbf{x}_p$. $B_E(\mathbf{x}_p,\hat{\mathbf{x}}_p,t)$ is called the background covariance of the surface emissions. Ω_b denotes the surface-time domain $\mathcal{D}_b\times[0,T]$. The third term represents the cost of the observation mismatch. $\mathbf{y}(t)$ is the vector of observations at time t and $\mathcal{H}_{t,\mathbf{x}}$ denotes the observation operator. $\mathbf{R}(t)$ is the application related to the observations error covariance. The integral in the very last

term is the model constraint. ϕ is the Lagrange multiplicatier. One can assume that ϕ belongs to the following set:

$$\phi \in \left\{ w \in \mathcal{H}^2(\Omega) | \mathbf{K}\nabla w = v_d w, w(T) = 0 \right\} \tag{2.9}$$

In the above set, $\mathcal{H}^2(\Omega)$ is a Sobolev space [Nikodym, 1933]. According to Sec. 2.2, the equation (2.10) can be transformed into:

$$\begin{aligned}
\mathcal{J} &= \frac{1}{2}\int_{\mathcal{D}\times\Omega} (\sigma(\mathbf{x},t) - \sigma^b(\mathbf{x},t)) B_\sigma^{-1}(\mathbf{x},\hat{\mathbf{x}},t)(\sigma(\hat{\mathbf{x}},t) - \sigma^b(\hat{\mathbf{x}},t)) d\mathbf{x} d\hat{\mathbf{x}} dt \\
&+ \frac{1}{2}\int_{\mathcal{D}_b\times\Omega_b} (E(\mathbf{x}_p,t) - E^b(\mathbf{x}_p,t)) B_E^{-1}(\mathbf{x}_p,\hat{\mathbf{x}}_p,t)(E(\hat{\mathbf{x}}_p,t) - E^b(\hat{\mathbf{x}}_p,t)) d\mathbf{x}_p d\hat{\mathbf{x}}_p dt \\
&+ \frac{1}{2}\int_0^T (\mathbf{y}(t) - \mathcal{H}_{t,\mathbf{x}}[c])\, \mathbf{R}^{-1}(t)\, (\mathbf{y}(t) - \mathcal{H}_{t,\mathbf{x}}[c])\, dt \\
&+ \int_\Omega c \left(-\frac{\partial \phi}{\partial t} - \nabla\cdot(\mathbf{u}\phi) - \nabla\cdot(\mathbf{K}\nabla\phi) + \Lambda\phi + \chi^\dagger(\phi) \right) d\mathbf{x} dt \\
&- \int_\mathcal{D} \phi(0) c(0) d\mathbf{x} + \int_{\partial\mathcal{D}_{in}\times[0,T]} (\phi c \mathbf{u}) \cdot d\mathbf{S} dt - \int_{\partial\mathcal{D}_b\times[0,T]} \mathbf{E}\cdot d\mathbf{S} dt - \int_\Omega \phi\sigma d\mathbf{x} dt\,.
\end{aligned} \tag{2.10}$$

The optimisation of Eq. (2.10) with respect to the concentrations at time t and point $\mathbf{x}$, gives:

$$\begin{aligned}
\frac{\delta \mathcal{J}}{\delta c(\mathbf{x},t)} &= -\mathcal{H}_{\mathbf{x},t}^\dagger \mathbf{R}(t)^{-1} \left(\mathbf{y}(t) - \mathcal{H}_{t,\mathbf{x}}[c]\right) \\
&\quad -\tfrac{\partial\phi}{\partial t} - \nabla\cdot(\mathbf{u}\phi) - \nabla\cdot(\mathbf{K}\nabla\phi) + \Lambda\phi = 0\,.
\end{aligned} \tag{2.11}$$

The positivity of the operator $\mathcal{H}_{\mathbf{x},t}^\dagger \mathbf{R}(t)^{-1}\mathcal{H}_{t,\mathbf{x}}$ shows that the extrema function, $c(\mathbf{x},t)$, is a minimiser solution for the cost function. Setting

$$\pi(\mathbf{x},t) = \mathcal{H}_{t,\mathbf{x}}^\dagger \mathbf{R}(t)^{-1} \left(\mathbf{y}(t) - \mathcal{H}_{\mathbf{x},t}[c]\right)\,, \tag{2.12}$$

Eq. (2.11) can be seen as the adjoint of the CTM (Eq. (2.3)) and ϕ is the solution of the adjoint model. The gradient of the cost function, Eq. (2.10), with respect to the initial conditions, to the surface and to the volume emissions, can be obtained with the following set of equations:

$$\frac{\partial \mathcal{J}}{\partial c_0} = -\int_\Omega \phi_0 d\mathbf{x} \tag{2.13}$$

and,

$$\frac{\delta \mathcal{J}}{\delta \sigma(t,\mathbf{x})} = -\int_\Omega \phi d\mathbf{x} + \int_\Omega B_\sigma^{-1}(\mathbf{x},\hat{\mathbf{x}},t)(\sigma(\hat{\mathbf{x}},t) - \sigma^b(\hat{\mathbf{x}},t)) d\hat{\mathbf{x}} \tag{2.14}$$

$$\frac{\delta \mathcal{J}}{\delta E(t,\mathbf{x}_p)} = -\int_{\partial\Omega_b} \phi d\mathbf{x}_p + \int_{\Omega_b} B_E^{-1}(\mathbf{x}_p,\hat{\mathbf{x}}_p,t)(E(\hat{\mathbf{x}}_p,t) - E^b(\hat{\mathbf{x}}_p,t)) d\hat{\mathbf{x}}_p\,. \tag{2.15}$$

2.3.1 The control space

The dimension of the observation space is often smaller than that of the control space. The system of equations is then under-determined. The background term of the cost function guaranties the uniqueness of the optimum solution for control parameters. Now, even though that solution is found, the high uncertainty on the model parameters which should be optimised, impacts on its reliability. Therefore, an alternate way for estimating it is to aggregate some of

the parameters (of the same nature) together with the help of a secondary and coarser parameter called a *scaling factor*. This choice leads to parameters which are less uncertain and decrease the number of unknown parameters in the system. Besides, this methodology helps to estimate the parameters through an extrapolation, even though there is a lack of observability at a specific moment and location. When estimating the emission inventories, one can assume that the emission fields at a specific location depend on time and change periodically. Therefore, the following changes are brought to the expression of σ and E.

$$\sigma(t, \mathbf{x}) = \alpha(t, \mathbf{x}_p)\sigma_b(t, \mathbf{x})\,, \tag{2.16}$$

$$E(t, \mathbf{x}_p) = \alpha(t, \mathbf{x}_p)E_b(t, \mathbf{x}_p)\,. \tag{2.17}$$

In the equations above, $\mathbf{x}_p$ is the image of the point $\mathbf{x}$ on the ground surface. $\alpha(t, \mathbf{x}_p)$ is a time-periodic function (with the period T_α). The scale factor function, $\alpha(t, \mathbf{x}_p)$, is the new state vector which will be optimised instead of σ and E.

2.3.1.1 Numerical discretisation

The analysis and equations introduced in section 2.3 are based on a continuous model. To build a numerical model, a discretisation of the equations is needed. First of all, a 3D grid is built to discretise the continuous geometric space. Let's consider a continuous function $f(t, \mathbf{x})$. The vector $\mathbf{f}_k$ includes the discrete values of the function f at time t_k all over the grid cells. The observation operator at time t_k can then be written as follows :

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{c}_k + \boldsymbol{\epsilon}_k \tag{2.18}$$

$\mathbf{H}_k$ is the linear observation operator that maps the concentrations from the state space to the observation space. In Eq. (2.18), $\mathbf{y}_k \in \mathbb{R}^{m_k}$ is the vector of the observed concentrations (m_k observations at time t_k), $\boldsymbol{\epsilon}_k$ is the vector of the observation errors at time t_k, and $\mathbf{c}_k$ is the vector of the concentrations. The discrete form of the CTM equation, Eq. (1.1), can be written as:

$$\mathbf{c}_k = \mathbf{M}_k\mathbf{c}_{k-1} + \Delta t\mathbf{e}_k\,, \tag{2.19}$$

where $\mathbf{M}_k$ denotes the dynamical operator of the model from t_{k-1} to t_k and Δt is the model integration time step. When there is no observation at the intermediate time t_k, $m_k = 0$. Vector $\mathbf{e}_k$ is representative of both the volume sources $\boldsymbol{\sigma}_k$ and of the fluxes $\mathbf{E}_k$. Assuming $\mathbf{e}_k^b$ is the first guess of the emissions, one has:

$$\mathbf{e}_{k,l}^b = \boldsymbol{\sigma}_{k,l}^b + \delta^{l,1}\frac{\mathbf{E}_k^b}{\Delta}\,, \tag{2.20}$$

where, Δ is the height of the surface layer and l is the layer number. δ is the Kronecker's delta:

$$\delta^{i,j} = \begin{cases} 0, & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases} \tag{2.21}$$

According to Eq. (2.16) and Eq. (2.16):

$$\mathbf{e}_k = \sum_{i \in \partial\mathcal{D}_b} \boldsymbol{\alpha}_{(k-N_h[k/N_h]),i}\mathbf{e}_{k,i}^b \tag{2.22}$$

where N_h is the number of time steps included in the time discretisation of α.

The 4D-Var data assimilation is used to invert the non-dimensional control variable vector

$\boldsymbol{\alpha}$. Setting the same uncertainty for the surface and for the volume emissions, the first two terms of equation (2.10) can be included in a term dependent on $\boldsymbol{\alpha}$. The cost function to be minimised over the time-window $[t_0, t_N]$ becomes:

$$\begin{aligned} \mathcal{J}(\boldsymbol{\alpha}) &= \frac{1}{2}\sum_{h=0}^{N_h-1}(\boldsymbol{\alpha}_h-\mathbf{1})^{\mathrm{T}}\mathbf{B}_{\alpha_h}^{-1}(\boldsymbol{\alpha}_h-\mathbf{1}) \\ &+\frac{1}{2}\sum_{k=0}^{N}(\mathbf{y}_k-\mathbf{H}_k\mathbf{c}_k)^{\mathrm{T}}\mathbf{R}_k^{-1}(\mathbf{y}_k-\mathbf{H}_k\mathbf{c}_k) \\ &+\sum_{k=1}^{N}\boldsymbol{\phi}_k^{\mathrm{T}}(\mathbf{c}_k-\mathbf{M}_k\mathbf{c}_{k-1}-\Delta t\mathbf{e}_k)\,. \end{aligned} \tag{2.23}$$

$\mathbf{R}_k = \mathrm{E}\left[\boldsymbol{\epsilon}_k(\boldsymbol{\epsilon}_k)^{\mathrm{T}}\right]$ is the observation error covariance matrix, $\mathbf{B}_{\alpha_h} = \mathrm{E}\left[\boldsymbol{\epsilon}_h^b(\boldsymbol{\epsilon}_h^b)^{\mathrm{T}}\right]$ is the background error covariance matrix, and $\mathbf{1}$ is the vector with entries 1. The vector $\boldsymbol{\alpha}_h$ is a set of $[\boldsymbol{\alpha}]_{i,j,h}$ for which $0 \leq i \leq N_x-1$, $0 \leq j \leq N_y-1$, and $0 \leq h \leq N_h-1$. In addition, $\boldsymbol{\epsilon}_h^b = \boldsymbol{\alpha}_h^{\mathrm{t}} - \mathbf{1}$ is the background error, where $\boldsymbol{\alpha}_h^{\mathrm{t}}$ is the unknown true state of the scale factors at a given h. In order to minimise the cost function $\mathcal{J}$ with respect to $\boldsymbol{\alpha}$, with an iterative gradient-based minimiser, the gradient of the cost function can be computed as follows:

$$\begin{aligned} \nabla_{\boldsymbol{\alpha}}\mathcal{J} &= \frac{\partial\mathcal{J}}{\partial\boldsymbol{\alpha}}+\sum_{k=0}^{N-1}\left(\frac{\partial\mathbf{e}_k}{\partial\boldsymbol{\alpha}}\right)\frac{\partial\mathcal{J}}{\partial\mathbf{e}_k} \\ &= \mathbf{B}_{\boldsymbol{\alpha}}^{-1}(\boldsymbol{\alpha}-\mathbf{1})-\sum_{k=0}^{N-1}\Delta t\left(\frac{\partial\mathbf{e}_k}{\partial\boldsymbol{\alpha}}\right)\boldsymbol{\phi}_k\,. \end{aligned} \tag{2.24}$$

$\dfrac{\partial\mathbf{e}_k}{\partial\boldsymbol{\alpha}}$ is a matrix which describes the dependence of the source $\boldsymbol{\sigma}$ and emission $\mathbf{E}$ on the control variable vector $\boldsymbol{\alpha}$. Its entries can be read out from Eq. (2.16) and Eq. (2.17), and depend on $\mathbf{e}_k^b$.

The optimisation of Eq. (2.23) with respect to the concentrations field at time t_k gives:

$$\boldsymbol{\phi}_k = \mathbf{M}_{k+1}^{\mathrm{T}}\boldsymbol{\phi}_{k+1}+\boldsymbol{\Delta}_k\,, \tag{2.25}$$

where the normalised innovation $\boldsymbol{\Delta}_k$ is:

$$\boldsymbol{\Delta}_k = \mathbf{H}_k^{\mathrm{T}}\mathbf{R}_k^{-1}(\mathbf{y}_k-\mathbf{H}_k\mathbf{c}_k)\,. \tag{2.26}$$

Equation (2.25) is the adjoint model equation.

2.3.2 The optimisation algorithm

The limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) is an efficient quasi-Newton optimisation method which can used in 4D-Var. The optimisation procedure uses a limited memory variation update to approximate the inverse of the Hessian matrix. L-BFGS stores only a few vectors that represent the approximation of the Hessian matrix, implicitly, by keeping a history of the past m updates of the input variables and never explicitly forming it.

Let us assume that $\boldsymbol{\alpha}_k$ is the position at the k^{th} iteration and $g_k = \nabla\mathcal{J}(\boldsymbol{\alpha}_k)$. Updating $\Delta\boldsymbol{\alpha}_k = \boldsymbol{\alpha}_{k+1} - \boldsymbol{\alpha}_k$ and $\Delta g_k = g_{k+1} - g_k$, one can define $\rho_k = \frac{1}{(\Delta g_k)^{\mathrm{T}}\boldsymbol{\alpha}_k}$. The initial approximate of the inverse Hessian at iteration k is set to H_k^0. The algorithm starts with

- set $q = g_k$

- for $i = k-1, k-2, \ldots, k-m$ repeat:
  $r_i = \rho_i (\Delta \boldsymbol{\alpha}_i)^{\mathrm{T}} q$
  $q = q - r_i \Delta g_i$
- put $z = H_k^0 q$
- for $i = k-m, k-2, \ldots, k-1$ repeat:
  $\beta_i = \rho_i (\Delta g_i)^{\mathrm{T}} z$
  $z = z + \Delta \boldsymbol{\alpha}_k (r_i - \beta_i)$
- set $H_k g_k = z$
- obtain $\lambda^\star = \operatorname{argmin} \left(\mathcal{J}(\boldsymbol{\alpha}_k + \lambda H_k g_k) \right)$, using the line search minimisation.
- $\boldsymbol{\alpha}_{k+1} = \boldsymbol{\alpha}_k + \lambda^\star H_k g_k$.
- repeat the algorithm for the next iteration, $k+1$, until the convergence of the solution $\boldsymbol{\alpha}_k$.

2.4 The verification of the 4D-Var routine

2.4.1 Validation of the approximate adjoint model

Following Bocquet [2012], an approximate but fast POLAIR3D adjoint model of the platform POLYPHEMUS can be built. That adjoint model is the discretisation of the continuous adjoint. This allows to use the CTM model, but propagating the concentrations backward in time with reversed wind fields.

A simulated observation y_i can be computed with both the forward model and the adjoint model. When the numerical adjoint is correct, the results from the two method should coincide to the numerical round-off errors. Hence, the discrepancy between the two can be viewed as a measure of the quality of the adjoint. This is the so-called duality test [Davoine and Bocquet, 2007]. The duality test was generalised following Roustan and Bocquet [2006a]. The simulated observation can be computed according to the two following equations:

$$y_i = \sum_{j \in \mathcal{D}, k} c_{k,j} \pi^i_{k,j} \Delta v_j \Delta t \,, \tag{2.27}$$

$$\begin{aligned} y_i = & \sum_{j \in \mathcal{D}} \phi^i_{0,j} c_{0,j} \Delta v_j + \sum_{j \in \partial \mathcal{D}_{\mathrm{in}}, k} \phi^i_{k,j} c_{k,j} \mathbf{u}_{k,j} \cdot \Delta \mathbf{S}_j \Delta t \\ & + \sum_{j \in \partial \mathcal{D}_{\mathrm{b}}, k} \phi^i_{k,j} E_{k,j} \cdot \Delta \mathbf{S}_j \Delta t + \sum_{j \in \mathcal{D}, k} \phi^i_{k,j} \sigma_{k,j} \Delta v_j \Delta t \,. \end{aligned} \tag{2.28}$$

In Eq. (2.27), π^i is the sampling function of the measurement i. It describes the measurement process. Δv_j is the volume of the grid-cell j. $c_{k,j}$ is the concentration in cell j at time t_k. Hence, Eq. (2.27) connects the observation y_i to the simulated concentration field. The second equality Eq. (2.28) describes the same observation but using the adjoint model. $\phi^i_{k,j}$ is the value of the solution of the adjoint model forced with the sampling $\boldsymbol{\Delta}_k = \pi^i_k$ in cell j and time t_k.

In Eq. (2.28), the first term of the right-hand side describes the contribution of the initial conditions to y_i. $\mathcal{D}$ is the space domain. The second, the third and the fourth terms represent the contributions of the boundary conditions, of the surface emissions and of the volume emissions respectively, to the measurement y_i. $\partial \mathcal{D}_{\mathrm{in}}$ is the boundary where the wind field is incoming. $\partial \mathcal{D}_{\mathrm{b}}$ represents the ground.

The validation of the adjoint model is done, considering each term separately and setting the other terms to zero. Figure 2.1 compares the measurements obtained with the forward model and via the adjoint model for carbon monoxide. The latter model do not account for chemistry reactions. For the present test of validation, the domain extends between [41.75N, 5.25W] (the left bottom corner) and [52.75N;12.25E] (the right top corner). The model is run at a resolution of $0.5^\circ \times 0.5^\circ$. Nine vertical levels are considered from the surface up to an altitude of 2780 m. The intermediary levels are 30, 150, 350, 630, 975, 1360, 1800 and 2270m agl. The meteorological fields are provided by the European Centre for Medium Range Weather Forecasts (ECMWF). The Pearson correlation coefficient between the CTM-based concentrations and the adjoint-based concentrations is 99.8% in the surface emissions case (a), 99.8% in the volume emissions case (b), 99.7% in the initial conditions case (c), and 93% in the boundary conditions case (d).

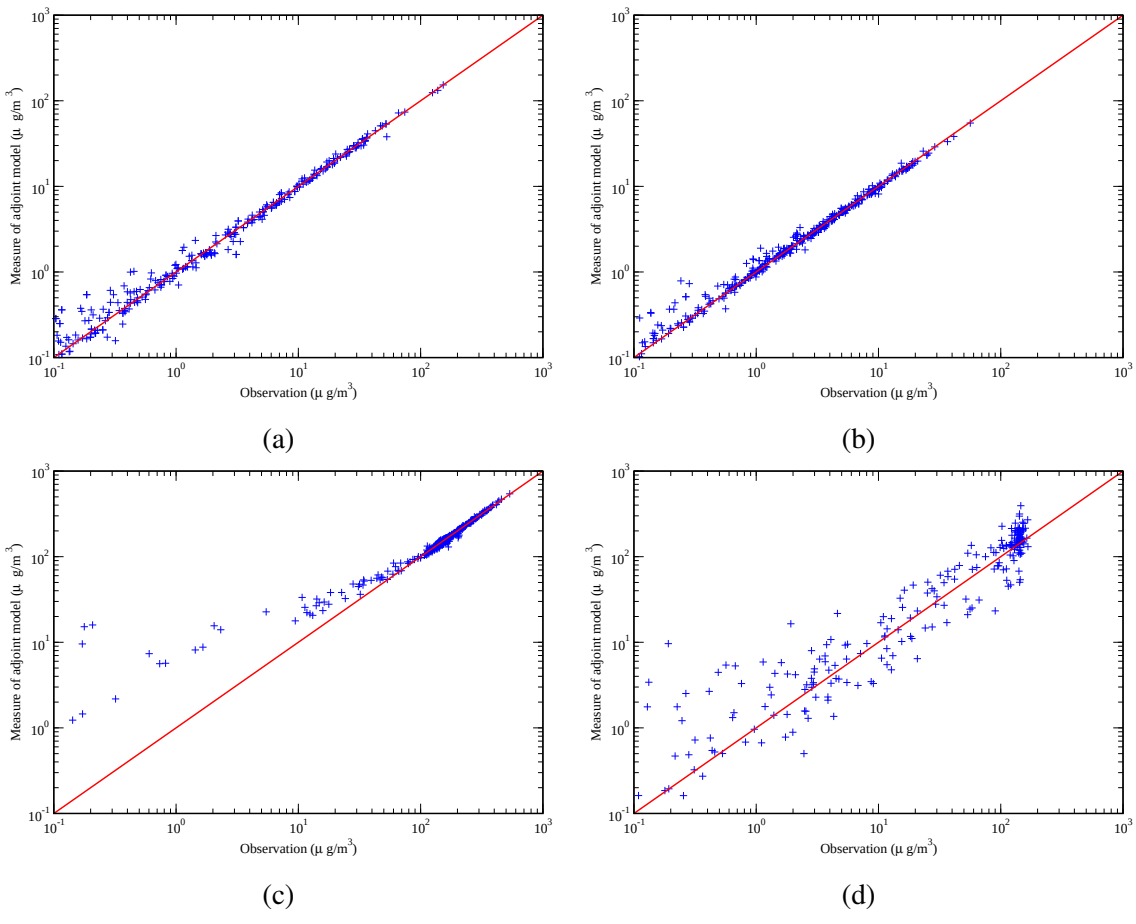


Figure 2.1: The computed concentrations via the adjoint model due to (a) the surface emissions, (b) the volume emissions, (c) the initial conditions and (d) the boundary conditions, versus the measured concentrations.

2.4.2 Verification of the gradient

In addition to the duality test, two kinds of gradient tests were carried out. In the first one, based on random perturbations, the following ratio ρ was computed:

$$\rho = \frac{\mathcal{J}\left(\mathbf{e} + \beta\mathbf{h_e}, \mathbf{c}_0 + \beta\mathbf{h}_0\right) - \mathcal{J}\left(\mathbf{e}, \mathbf{c}_0\right)}{\beta\left(\left(\nabla_{\mathbf{e}}\mathcal{J}\right)^{\mathrm{T}}\mathbf{h_e} + \left(\nabla_{\mathbf{c}_0}\mathcal{J}\right)^{\mathrm{T}}\mathbf{h}_0\right)}\,. \tag{2.29}$$

In Eq. (2.29), the cost function is seen as a function of both $\mathbf{e}$ and $\mathbf{c}_0$. $\mathbf{h_e}$ and $\mathbf{h}_0$ are the perturbation vectors of the emissions and initial conditions. β is the perturbation coefficient. When β tends towards zero, the ratio ρ must tend towards 1. In Fig. 2.2, ρ is plot as a function of β. The instability which is observed for very low values of β is due to the round-off errors [Zou et al., 1997].

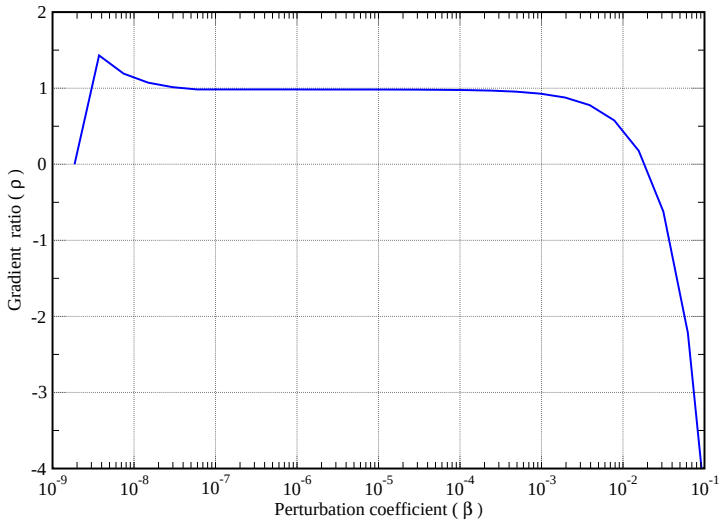


Figure 2.2: The perturbation test: Variation of the gradient ratio ρ with respect to the perturbation coefficient β.

A second test focuses on the cost function $\mathcal{J}(\boldsymbol{\alpha})$. In particular, the derivative of $\mathcal{J}(\beta\mathbf{1})$ with respect to β is computed with a β varying between 0 and 6. It is obtained either using the gradient via the adjoint model or using the finite difference method. The results presented in Fig. 2.3, show that the gradient computed via the adjoint model is well approximated [Chao and Chang, 1992].

2.5 Conclusion

The adjoint solution of the CTM is essential to build the Jacobian matrix in the case that the sources of pollutant are wide-spread (unlike the pointwise sources, for which the Jacobian matrix is easy to build from the CTM). An approximate adjoint model of POLAIR3D (of POLYPHEMUS) is developed and validated for a tracer species. The concentrations computed with the help of the adjoint solution (see Eq. 2.28) are compared to the CTM concentrations. The statistical indicators show that the two set of simulated concentrations are consistent enough. The adjoint model in question is used in the 4D-Var algorithm to compute the gradient of the cost function. The 4D-Var algorithm is checked with two grandient tests. When it is question of inverse modelling with high frequency observations and there are no fast enough

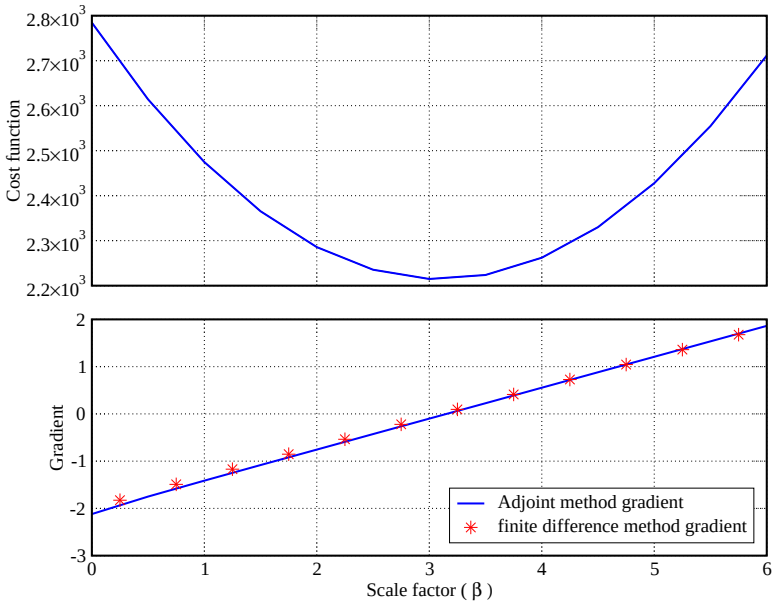


Figure 2.3: Scale factor test: Variation of the gradient and cost function with respect to the scale factor β.

models available, the 4D-Var appears to be a convenient tool. These newly developed tools will be used in the following chapters.

.

Chapter 3

Inversion of regional carbon monoxide fluxes: Coupling 4D-Var with a simple subgrid statistical model

Summary

A four-dimensional variational data assimilation system (4D-Var) is employed to retrieve carbon monoxide (CO) fluxes at regional scale, using an air quality network. The air quality stations that monitor CO are proximity stations located close to industrial, urban or traffic sources. The mismatch between the coarsely discretized Eulerian transport model and the observations, inferred to be mainly due to representativeness errors in this context, lead to a bias (average simulated concentrations minus observed concentrations) of the same order of magnitude as the concentrations. 4D-Var leads to a mild improvement in the bias because it does not adequately handle the representativeness issue. For this reason, a simple statistical subgrid model is introduced and is coupled to 4D-Var. In addition to CO fluxes, the optimisation seeks to jointly retrieve *influence coefficients*, which quantify each station's representativeness. The method leads to a much better representation of the CO concentration variability, with a significant improvement of statistical indicators. The resulting increase in the total inventory estimate is close to the one obtained from remote sensing data assimilation. This methodology and experiments suggest that information useful at coarse scales can be better extracted from atmospheric constituent observations strongly impacted by representativeness errors.

Contents

3.1 Introduction

In tracer transport studies, observations are infrequent in time and, for ground-measurements, sparse in space. Furthermore, they do not intrinsically carry any information about the future. That is why, complementarity, numerical models are used to assess the meteorological and chemical state of the atmosphere. In air quality modelling, input data, such as initial and boundary conditions, emission fluxes, and vertical diffusion coefficients are necessary to run proper simulations. The uncertainties of these input data and perhaps the lack of understanding of the underlying physical processes induce model errors in the simulations. To minimise them, data assimilation (DA) methods can be used. They combine observational data, and information coming from chemistry and transport models and their related error statistics in order to find the optimal values of the parameters which minimise the errors.

Introducing optimal control theory ideas in geophysics, Le Dimet and Talagrand [1986] used 4D-Var to assimilate meteorological observations. Fisher and Leny [1995] used 4D-Var for the analysis of some chemically active tracer species. Lately, variational data assimilation studies have focused on the inverse modelling of pollutant emission fields (e.g. Elbern et al. [2007] and other references within Zhang et al. [2012]).

Focusing on carbon monoxide (CO), several modelling studies pointed out to the discrepancy between the observations and the simulated concentrations. Using the Emission Database for Global Atmospheric Research 3 (EDGAR3) inventory, before any correction, the model global run of Fortems-Cheiney et al. [2011] underestimates the CO concentrations of about 5 to 10% with respect to the satellite observations for January, February and March 2005. Emmons et al. [2010] compared the satellite observations to simulations of the the Model for OZone And Related chemical Tracers, version 4 (MOZART-4), using the EDGAR3 inventory. Displaying a similar trend, their results exhibit an underestimation of the CO concentrations over Europe of about 10 to 20% for the same period.

That is why inverse modelling experiments have been carried out to update the CO flux inventories. For instance, Mulholland and Seinfeld [1995]; Saide et al. [2011] focused on urban scale. Yumimotoa and Uno [2006]; Kopacz et al. [2009] used 4D-Var or analytical methods to invert the emissions at regional scale. Other studies have also been performed on global scale: e.g. Pétron et al. [2002]; Stavrakou and Müller [2006]; Arellano and Hess [2006]; Fortems-Cheiney et al. [2009]; Kopacz et al. [2010]. These studies make use of ground-based instruments that measure concentrations, or they make use of satellite instruments to infer satellite-derived retrieval of CO. The former instruments are mostly used in conjunction with regional scale models whereas the latter instruments are mostly used with global scale models.

In the case of an assimilation of observations over a short period (i.e. a few hours to a few days), the parameters to be optimised are usually the initial conditions. With larger data assimilation windows (i.e. a few days to a few months), the model is more sensitive to other parameters, such as the emissions inventory, the meteorological fields and the boundary conditions.

In most top-down (i.e. inverse modelling) studies related to the global scale, the CO emissions fluxes were found to be underestimated in the Northern hemisphere whereas they are quite consistent with the measurements in the Southern hemisphere (e.g. Müller and Stavrakou [2005]), or slightly overestimated (e.g. Arellano and Hess [2006]). This underestimation in the Northern hemisphere is also found in the modelling studies (e.g. Emmons et al. [2010]).

Satellite and in-situ measurements require specific care when compared to transport models. The discrepancy between the observations and the model forecast of these observations are known to be due to instrumental errors, deficiencies of the model and of the forcing fields (model error), and the *representativeness error*. The assessment of this representativeness error becomes a key issue when assimilating in-situ observations, which are the focus of this study. Indeed, the model is operative at coarser scale and by construction cannot simulate subgrid events. The in-situ observations do capture both the coarser scale pollutant plumes, but also subgrid plumes that are not accounted for by the model. Therefore, there is a residual mismatch due to unresolved scales known as the representativeness error. In data assimilation, it is often considered part of model error, but formally ascribed to the observation error.

Due to the complexity of its estimation, an experience-based value is usually assumed for that error. This value is often chosen to be the same for all measurements. Yet that is certainly not true, because the nature of the measurements can be different (urban, rural, etc.). The maximum possible representativeness error is often chosen for all observations. Alternatively a χ^2 criterion (used by Ménard et al. [2000] in tracer studies) can be implemented to estimate the proper magnitude of the observational errors.

In this chapter, our goal is to estimate carbon monoxide surface emissions with inverse modelling, using in-situ measurements from an air quality network. This network operates in France and we wish to retrieve the emissions over France. Hence, as opposed to most of the studies mentioned earlier, the focus is on mesoscale and lower troposphere modelling. These measurements are abundant, but strongly impacted by representativeness errors since many of them are influenced by nearby industrial, traffic or urban sources. Most of them aim at measuring (some of) those influential sources. To perform emission inverse modelling in this context, this lack of representativeness must be accounted for. One needs to demonstrate that observations obtained at fine scale, and strongly impacted by representativeness errors, can be assimilated with the aim of correcting a pollutant inventory defined at larger scale.

In Section 4.2, the atmospheric transport model (ATM) is introduced, as well as, a detailed description of the observational data. The specifications of the control space are presented. An investigation of the modelling of errors and of the uncertainties of the control parameters is also reported. In Section 3.4, 4D-Var is used to optimise the spatio-temporal parameters of the inventories with unsatisfactory results. Since there is a dramatic lack of representativeness of the measurements, a simple subgrid statistical model is built in order to improve the 4D-Var numerical results. The statistical model aims at taking into account the impact of close-by sources on monitoring stations. Section 3.5 introduced and justifies this statistical model and its tight coupling to 4D-Var. In Section 4.3, the inverse modelling experiment is performed with the combination of 4D-Var and the subgrid statistical model, which will be called 4D-Var-ξ. The analysis produced by the retrieval is studied. Validations with independent observations are performed, notably using cross-validation and a long-term forecast of the CO concentrations. In Section 5.5, the findings of this study are summarised. The potential and limitation of the approach are discussed.

3.2 Inverse modelling setup

First, details are given about the necessary ingredients of the inverse modelling study: the transport model setup, the observations, the control variables (which are the scale factors of the emission inventories) and the first guess provided by the initial inventory. How to incorporate them in a 4D-Var-ξ system is then described, as well as the necessary statistical assumptions on the errors present in the system.

3.3 Experiment setup

All runs of the model will be performed over France. The domain extends between [41.75N, 5.25W] (the left bottom corner) and [52.75N;12.25E] (the right top corner). The grid has the resolution of $0.25^\circ \times 0.25^\circ$. Nine vertical levels are considered from the surface up to an altitude of 2780 m. The intermediary levels are 30, 150, 350, 630, 975, 1360, 1800 and 2270 m. The meteorological fields are provided by the European Centre for Medium Range Weather Forecasts (ECMWF). These fields have a resolution of $0.36^\circ \times 0.36^\circ$ and 60 vertical levels. The time step is 3 hours. Concentrations from the global chemistry-transport model MOZART, version 2 [Horowitz et al., 2003] are used to provide boundary conditions, and the initial condition. A calibration factor of 1.2 is used to correct a global underestimation of incoming carbon monoxide, following the global estimations of Emmons et al. [2010].

It has initially been examined that within our regional, lower troposphere setup, and for our timescale, carbon monoxide is barely reactive. To do so, we have compared the photochemical version of POLAIR3D to the tracer version (validated in Quélo et al. [2007]). A small bias of $5.8\,\mu$g m^{-3} is observed between the CO concentrations with or without reactions, i.e. about 2% of the average measurements. As a consequence, neglecting the reactions, we chose to use the faster tracer version of the model.

3.3.1 Observations

The BDQA (Base de Données de la Qualité de l'Air[1]) is a database listing the concentrations of several air quality pollutants over France. The (mostly hourly) collected observations are provided by 600 monitoring stations distributed all over France. For carbon monoxide, 89 stations provide hourly measurements at ground level (with an average of 75 observations per hour for the year 2005). These stations belong to one of the four different categories: industrial, traffic, urban and suburban. This gives an indication of their environment but not necessarily of their representativeness in an ATM. Larssen et al. [1999] define an area of representativeness for a station as being an area in which the concentrations do not differ from the ones measured at the station by more than a specified amount. This amount can be set to the total uncertainty of the measurement or to a value not to be exceeded in order to fulfil data quality objectives. Nappo et al. [1982] further precise that more than 90% of the concentrations measured in that area should satisfy that definition. When these conditions cannot be satisfied for a station, the latter is not deemed representative of its area.

In the case of carbon monoxide, the stations belonging to the BDQA network are far from representative as it is very difficult to determine an area of representativeness for most of them. These receptors are likely to be influenced by nearby surface fluxes [Henne et al., 2010]. Background stations, far from pollution sources, are missing.

For the experiments performed in this study, 8 weeks of BDQA observations will be assimilated from January the 1^{st} 2005 to February the 26^{th} 2005, for a total of $107,914$ observations,

[1] details available at http://www.atmonet.org

while up to more than 10 months of observations $(548, 964)$, corresponding to the rest of the year, will be used for validation. In another experiment, about 55% of the $107, 914$ observations will be assimilated and the rest of the $107, 914$ observations will be used for validation.

The locations of the BDQA network CO monitoring stations, are shown in Fig. 3.1.

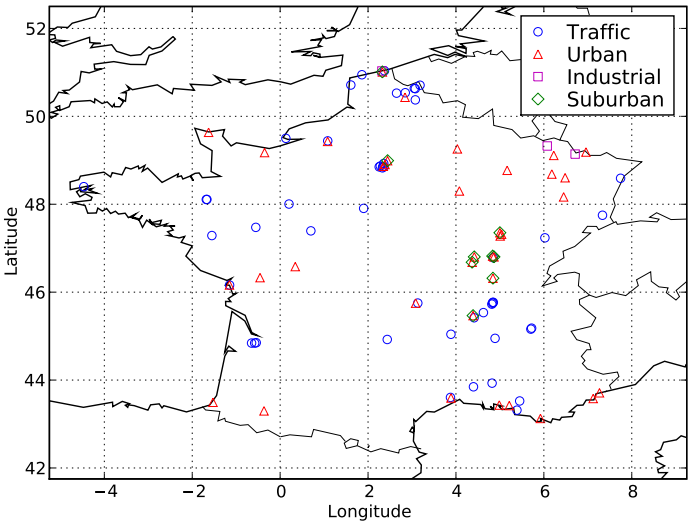


Figure 3.1: The carbon monoxide monitoring stations of the BDQA network, sorted out by their official type.

3.3.2 Inventory and control variables

The first guess (background information) on the fluxes needed to perform the model runs and the inversions is provided by the anthropogenic emission from the European Monitoring and Evaluation Programme (EMEP, details can be found at http://www.ceip.at) inventory, and the biogenic emissions of the Model of Emissions of Gases and Aerosols from Nature (MEGAN) model [Guenther et al., 2006]. The EMEP inventory is modulated using hourly, weekly and monthly distribution coefficients. These coefficients are provided by the GENEMIS project [GENEMIS, 1994]. The EMEP inventory has a resolution of $0.50°$ and the MEGAN inventory has a resolution of $0.04°$. We have checked that the vegetation fire emissions over the domain defined earlier and time window of this study can be neglected.

The aim of the present study is to determine the hourly grid-size optimal sources of carbon monoxide, for both the volume source σ in Eq. (1.1), and the emission fluxes E of Eq. (1.3). An estimation of the number of independent control variables over a data assimilation window of 8 weeks, a domain of 58×43 grid-cells ($0.25° \times 0.25°$ resolution), and six levels for the volume source, yield about 2×10^7 independent variables to retrieve. That is why we have chosen to constrain the number of degrees of freedom of control space in the following way.

The year is divided into weeks, indexed by $w = 0, \ldots, N_w - 1$ where $N_w = 52$. Each week is divided into $N_h = 56$ 3-hour periods, indexed by $h = 0, \ldots, N_h - 1$. Each 3-hour period is divided into $N_s = 3$ hours, indexed by $s = 0, \ldots, N_s - 1$. A grid-cell has space coordinates i, j, l (indices related to longitude, latitude and altitude respectively) and time coordinates h, w, s (or using the global time index $k = s + N_s(h + N_h w)$). In order to reduce the number of control variables to deal with, the discrete hourly grid-size volume

sources $\boldsymbol{\sigma}$ and emissions $\mathbf{E}$ are parameterised according to

$$[\boldsymbol{\sigma}]_{i,j,l,h,w,s} = [\boldsymbol{\alpha}]_{i,j,h}\,[\boldsymbol{\sigma}_b]_{i,j,l,h,w,s}\ , \tag{3.1}$$

$$[\mathbf{E}]_{i,j,h,w,s} = [\boldsymbol{\alpha}]_{i,j,h}[\mathbf{E}_b]_{i,j,h,w,s}\ , \tag{3.2}$$

where $[\boldsymbol{\alpha}]_{i,j,h}$ are the non-dimensional effective control variables corresponding to the residual degrees of freedom. They represent $58 \times 43 \times 56 = 139,664$ scalars. The first guesses $\boldsymbol{\sigma}_b$ and $\mathbf{E}_b$ are the background sources stemming from the inventory. Let us make a remark on the temporal cycles of the inventory, that are for instance due to vehicles traffic, urban heating, industry, etc. Because the control variables $[\boldsymbol{\alpha}]_{i,j,h}$ are indexed by h, the intra-week temporal cycles will be solved for in the inverse modelling experiments. However the longer cycles will not be solved for, but are determined by the built-in cycles of the inventory: $[\boldsymbol{\sigma}_b]_{i,j,l,h,w,s}$ depends on the indexes w and s. For instance, seasonal cycles of urban heating are prescribed by $[\boldsymbol{\sigma}_b]_{i,j,l,h,w,s}$.

The surface $\mathbf{E}$ and volume emission $\boldsymbol{\sigma}$ variables have a similar local signature and would have a similar impact on a distant observation site, so that they would appear as ill-determined variables in an inverse problem. That is the reason why they were parameterised in Eq. (3.1) and Eq. (3.2) in terms of the same control vector $\boldsymbol{\alpha}$. It is convenient to introduce a composite emission vector $\mathbf{e}$, defined in the surface layer by

$$\mathbf{e}_{l=0} = \boldsymbol{\sigma}_{l=0} + \frac{\mathbf{E}}{\Delta}\,, \tag{3.3}$$

where Δ is the height of the surface layer. Note that this equality assumes a well-mixed surface layer. In the upper layers $l \geq 1$, it is defined by

$$\mathbf{e}_l = \boldsymbol{\sigma}_l\,. \tag{3.4}$$

In the following, the first guess about $\mathbf{e}$ (background) will be denoted $\mathbf{e}_b$. Correspondingly, one has:

$$[\mathbf{e}_b]_{i,j,l=0,h,w,s} = [\boldsymbol{\sigma}_b]_{i,j,l=0,h,w,s} + \frac{[\mathbf{E}_b]_{i,j,h,w,s}}{\Delta} \quad \text{and} \quad [\mathbf{e}_b]_{i,j,l\neq 0,h,w,s} = [\boldsymbol{\sigma}_b]_{i,j,l\neq 0,h,w,s}\,. \tag{3.5}$$

As a result, Eq. (3.1) and Eq. (3.2) can be synthesised into

$$[\mathbf{e}]_{i,j,l,h,w,s} = [\boldsymbol{\alpha}]_{i,j,h}[\mathbf{e}_b]_{i,j,l,h,w,s}\,. \tag{3.6}$$

3.3.3 4D-Var

In spite of the quasi-linear physics of carbon monoxide (at these space and time scales), the computation of the Jacobian matrix is difficult to afford because of the very large set of data and control variables we intend to use. 4D-Var is meant to handle such a computational problem [Chevallier et al., 2005]. The details of the 4D-Var are reported in chapter 2.

3.3.4 Error modelling

In this section, we describe how the background and observation errors are statistically modeled. The background errors on the independent variables $\boldsymbol{\alpha}$ are first related to the traditional background errors on $\mathbf{e}$ (hence $\boldsymbol{\sigma}$ and $\mathbf{E}$). While the background error variances will be chosen a priori, the observation errors will be determined through a χ^2 diagnosis.

3.3.4.1 Background error covariance matrix

The background error covariance matrix $\mathbf{B}_{\boldsymbol{\alpha}}$ defines the variances-covariances between the different components of the departure of the scale factors $\boldsymbol{\alpha}$ from $\boldsymbol{\alpha}_b = \mathbf{1}$. In the inventory, anthropogenic emissions significantly dominates the biogenic emissions (1.8% of the total inventory over France). Assuming the anthropogenic sources (such as the individual industrial sources, or urban heating sources) have errors that are barely spatially correlated, the error correlation between grid-cells are taken as negligible, so that the covariance terms of that matrix are set to zero. Note that other sources of anthropogenic sources, such as traffic might have extended correlated errors. We also neglect temporal correlations, which is a weaker assumption even though the emission are mostly anthropogenic. As a consequence of our assumptions, the prior errors are essentially represented by the variances of the prior emissions (diagonal assumption for $\mathbf{B}_{\boldsymbol{\alpha}}$).

Assuming that the emission errors are not time dependent, the variance of control variable $[\boldsymbol{\alpha}]_{i,j,h}$ is:

$$[\mathbf{B}_{\boldsymbol{\alpha}}]^{i,j,h} = \frac{\sum_{w=0}^{N_w-1}\sum_{s=0}^{N_s-1}\sum_{l=0}^{N_l-1}[\mathbf{B}_{\mathbf{e}}]^{i,j,l,h,w,s}}{\left(\sum_{w=0}^{N_w-1}\sum_{s=0}^{N_s-1}\sum_{l=0}^{N_l-1}[\mathbf{e}^b]_{i,j,l,h,w,s}\right)^2}\,, \tag{3.7}$$

where

$$[\mathbf{B}_{\mathbf{e}}]^{i,j,l,h,w,s} = \mathrm{E}\left[\left([\mathbf{e}]_{i,j,l,h,w,s} - [\mathbf{e}^b]_{i,j,l,h,w,s}\right)^2\right] \tag{3.8}$$

is the background error variance of the emission fluxes in the grid-cell of coordinates i, j, l at time h, w, s. Since the data assimilation window of the experiments ahead is 8-week long, N_w is now set to 8.

3.3.4.2 Observation error covariance matrix

In Eq. (2.18), $\boldsymbol{\epsilon}_k$ includes the instrumental error and representativeness error of the observations. It is assumed that they are independent from site to site, and from observation time to observation time. At this stage the variances are assumed to be the same for all observations, which is crude since the representativeness error is expected to significantly vary between stations. Accordingly, $\mathbf{R}_k$ is modeled as a diagonal matrix:

$$\mathbf{R}_k = r^2\mathbf{I}_{m_k}\,, \tag{3.9}$$

where $\mathbf{I}_{m_k}$ is the identity matrix in observation space at time t_k, and

$$r^2 = \varepsilon_{\mathrm{repr}}^2 + \varepsilon_{\mathrm{meas}}^2\,. \tag{3.10}$$

$\varepsilon_{\mathrm{meas}}$ is the standard deviation of instrumental error, and $\varepsilon_{\mathrm{repr}}$ is the standard deviation of the representativeness errors, which depends on the species, the station type, and the grid size [Elbern et al., 2007].

To estimate the standard deviation parameter r, we resort to a χ^2 diagnosis ([Ménard et al., 2000; Elbern et al., 2007] for instance in the context of atmospheric chemistry). When the statistics of the errors are consistent with the innovations, then one should expect that the average value of the cost function is equal to half of the number of assimilated observations. Accordingly r should be chosen such that:

$$\left\{\min_{\boldsymbol{\alpha}} \mathcal{J}(\boldsymbol{\alpha})\right\}(r) \simeq \frac{m}{2} \tag{3.11}$$

where $m = \sum_{k=0}^{k=N} m_k$ is the number of observations. Based on this diagnosis, an iterative process can be used to estimate r. The algorithm begins by assuming an initial value, r_0, for r. At each iteration, r_{i+1} is computed by

$$r_{i+1}^2 = \frac{d_n^i}{m - d_s^i} r_i^2 \tag{3.12}$$

where d_s^i and d_n^i are twice the background part $\mathcal{J}_b$ of the cost function, and twice the observation departure part $\mathcal{J}_o$ of the cost function respectively at the i^{th} step. They respectively converge to d_s the number of degrees of freedom for the signal (hence the s), and to d_n the number of degrees of freedom for the noise (hence the n). The value of r is thus obtained when the sequence of r_i has converged. The method needs iterating because the minimum of the cost function does not linearly depend on r.

We note that this iterative scheme is equivalent to that of Desroziers and Ivanov [2001]: Eq. (3.12) coincides with Eq. (4) of Desroziers and Ivanov [2001] when the background term is fixed. Since the method of Desroziers and Ivanov [2001] converges to one maximum of a parameter likelihood, we conclude that so does our χ^2 approach.

3.4 Application of 4D-Var

Following these assumptions, we perform the 4D-Var inversion of the $\boldsymbol{\alpha}$ parameters. The assimilation window of the experiment is in the winter period, from January 1 2005 to February 26 2005. For comparison, a free simulation is first performed using the inventories and boundary conditions described earlier. Then, the $\boldsymbol{\alpha}$ variables of Section 3.3.2 are inverted using 4D-Var.

At each grid-cell, the standard deviation of the prior error in the emission is set to 50% of the prior emission. This value is consistent with Pétron et al. [2002] and Kopacz et al. [2010]. In Yumimotoa and Uno [2006], Pétron et al. [2004] and Fortems-Cheiney et al. [2009], the standard deviations are set to 100% of the prior emissions in each grid-cell, but using the EDGAR3 inventory and not over the Western Europe where the inventories are more ascertained.

An iterative test (χ^2 criterion) for the same period is applied to estimate the observational error variance. We found a standard deviation of $r \simeq 652.5\,\mu\text{g m}^{-3}$ for the observational error using the χ^2 method. It is very significant since it is of the order as the average observation ($662\,\mu\text{g m}^{-3}$).

A comparison of the observations with the results of the model free run, as well as a comparison to the results of the data assimilation experiment (optimisation of $\boldsymbol{\alpha}$) are presented in Tab. 3.1. The scores of this DA run show that the consistency between the analysed concentrations and the observations is low, in spite of a Pearson correlation coefficient increasing from 0.16 to 0.36. Furthermore, the reduction of the bias $\overline{O} - \overline{C}$ is unsatisfyingly small.

The total emission of the background inventory between January 1 to February 26 is 1.06 Tg. From the computation of the analysed fluxes using inverse modelling, we obtain 1.44 Tg, 36% higher than the total a priori emission. However, Fortems-Cheiney et al. [2011], estimated that value to be 17% for Western Europe, during 2005, with the reference being the EDGAR3 inventory, using biomass and anthropogenic emissions, and a spatial resolution of $2.5^\circ \times 3.5^\circ$. Kopacz et al. [2010] estimated it between $16 - 24\%$ from May 2004 to April 2005. This indicates a possible over-estimation of the emission by the 4D-Var analysis. In Fig. 3.8 on page 68 are plotted 300 hours of the simulation and 4D-Var runs in the DA window, for four stations. The four corresponding profiles are too smooth to represent the peaks of the observation profile. This supports our assumption on the impact of representativeness error.

Table 3.1: Comparison of the observations and the simulated or analysed concentrations. $\overline{C}$ is the mean concentration, $\overline{O}$ is the mean observation, and NB= $2(\overline{C}-\overline{O})/(\overline{C}+\overline{O})$ is the normalised bias. RMSE stands for root-mean square error. R is the Pearson correlation. FA_x is the fraction of the simulated concentrations that are within a factor x of the corresponding observations. $\overline{C}$, $\overline{O}$, and the RMSE are given in $\mu\mathrm{g\,m}^{-3}$.

	$\overline{C}$	$\overline{O}$	NB	RMSE	R	FA_2	FA_5
Simulation (01/01–02/26 2005)	303	662	-0.74	701	0.16	0.52	0.90
Optimisation of $\boldsymbol{\alpha}$ (4D-Var)	396	662	-0.50	633	0.36	0.59	0.92
Optimisation of $\boldsymbol{\xi}$	615	662	-0.07	503	0.57	0.73	0.96
Coupled optimisation of $\boldsymbol{\alpha}$, $\boldsymbol{\xi}$ (4D-Var-ξ)	671	662	0.01	418	0.73	0.79	0.97

The BDQA CO network is mostly composed of proximity stations, whose observations are likely to be influenced by local sources. Therefore, the lack of consistency between the model and the observations could be explained by the direct impact of nearby pollution sources on observations. The 4D-Var analysis cannot account for the local peaks of CO concentrations since it uses a model that cannot resolve those subgrid-scale processes. However, we believe that there is some useful signal to extract from these observations. To do so, one needs to account for the subgrid processes. At least two state-of-the-art options are possible. The deterministic route consists in using explicit representations of partial information that one may have about the subgrid processes, emissions, etc. These representations are incorporated into the coarser model. This is what typically does a plume-in-grid model that uses some additional information about short-range dispersion (e.g. Karamchandani et al. [2009] for an application to CO subgrid traffic emission). A second route is of statistical nature. The aim is to make a statistical regression between the observations and the coarse resolution model output, which results in a fitted linear correspondence between the model to the observations. In geosciences, downscaling techniques have taken this path (e.g. Guillas et al. [2008] for an application to ozone concentrations). In this study, a statistical approach is chosen to represent the subgrid effects. A deterministic modelling approach of the subgrid processes would theoretically be desirable, but it requires additional subgrid information that we do not have here, and it would be computationally more expensive.

3.5 Coupling 4D-Var with a subgrid statistical model

3.5.1 A simple subgrid statistical model

Assume that s is a continuous source field: it describes the emission at any spatial scale. Recall that $\mathbf{e}$ is the discrete coarse-grained source that we use to drive the model. Ideally, s and $\mathbf{e}$ should be related through a restriction, coarse-graining operator $\boldsymbol{\Gamma}$, which acts as a low-pass filter, filtering out the fine details of the source:

$$\mathbf{e} = \boldsymbol{\Gamma} s\,. \tag{3.13}$$

Following Bocquet et al. [2011], we can consider a prolongation operator $\boldsymbol{\Gamma}^\star$, which refines a coarse emission field $\mathbf{e}$ to a continuous field $s^\star$:

$$s^\star = \boldsymbol{\Gamma}^\star \mathbf{e}\,. \tag{3.14}$$

There is freedom in choosing $\boldsymbol{\Gamma}^\star$. It could be a basic subgrid spatial interpolation operator, or it could rely on additional subgrid information, or it could be obtained from a Bayesian inference

[Bocquet et al., 2011]. For the purpose of this derivation, we do not have to specify a precise form for $\mathbf{\Gamma}^\star$. However, it is reasonable to assume $\mathbf{\Gamma}\mathbf{\Gamma}^\star = \mathbf{I}$. Besides, $\mathbf{\Gamma}^\star\mathbf{\Gamma}$ is a projection operator, not the identity, because of some details of the real fine scale emission field are lost in the restriction process $\mathbf{\Gamma}$.

If $\mathcal{H}$ is the Jacobian of a continuous multiscale hypothetical carbon monoxide model that relates s to the measurements $\mathbf{y}$, the vector collecting all measurements, then

$$\begin{aligned}\mathbf{y} &= \mathcal{H}s + \boldsymbol{\epsilon} \\ &= \mathcal{H}\mathbf{\Gamma}^\star\mathbf{\Gamma}s + \mathcal{H}\left(\mathbf{I} - \mathbf{\Gamma}^\star\mathbf{\Gamma}\right)s + \boldsymbol{\epsilon} \\ &= \left(\mathcal{H}\mathbf{\Gamma}^\star\right)\mathbf{e} + \mathcal{H}\left(\mathbf{I} - \mathbf{\Gamma}^\star\mathbf{\Gamma}\right)s + \boldsymbol{\epsilon}\,. \end{aligned} \tag{3.15}$$

Assume $\mathbf{\Gamma}$ operates the coarse-graining at the finest scale accessible by the model. Therefore $\mathcal{H}\mathbf{\Gamma}^\star$ could be identified with the Jacobian of our Eulerian ATM. Since $\mathbf{I} - \mathbf{\Gamma}^\star\mathbf{\Gamma}$ is a high-pass projector (it retains the short-scale fluctuations of the real emission field), $\mathcal{H}(\mathbf{I} - \mathbf{\Gamma}^\star\mathbf{\Gamma})s$ theoretically stands for the representativeness error [Wu et al., 2011].

Unfortunately, we do not have access to s or a multiscale model $\mathcal{H}$, and one needs a simple subgrid scale model to approximate $\mathcal{H}(\mathbf{I} - \mathbf{\Gamma}^\star\mathbf{\Gamma})s$ and close the equation. We assume this representativeness error is mostly due to subgrid/nearby sources that have a strong impact on the measurements which are not representative of the background carbon monoxide concentration level. Another possibly significant source of error is the weakness of current vertical turbulent diffusion parameterisations. Notice that part of it may be categorised as representativeness errors when for instance the boundary layer height varies significantly within grid-cells.

Guided by the structure of $\mathcal{H}(\mathbf{I} - \mathbf{\Gamma}^\star\mathbf{\Gamma})s$, we choose to model this nearby source influence by the term

$$\xi_i \mathbf{\Pi}_{i,k}\mathbf{e} \tag{3.16}$$

where ξ_i is a positive scalar attached to a station indexed by i. Similarly to $\mathcal{H}(\mathbf{I} - \mathbf{\Gamma}^\star\mathbf{\Gamma})s$, $\xi_i\mathbf{\Pi}_{i,k}\mathbf{e}$ has a linear explicit dependence on the emission $\mathbf{e}$. The *influence coefficient* ξ_i quantifies the influence of local nearby sources onto the station. It can be interpreted as the time (given in hours in the following) required to reach a CO concentration level equivalent to the subgrid part of the measurement $[\mathbf{y} - \mathbf{H}\mathbf{c}]_{i,k}$, by emitting $\mathbf{\Pi}_{i,k}\mathbf{e}$ which is based on the coarse-grained inventory. This influence factor is assumed constant in time and it is a priori unknown. $\mathbf{\Pi}_{i,k}$ is an operator that linearly interpolates $\mathbf{e}$ at the station location and at time t_k. If ξ_i is vanishing, then the representativeness of the station is deemed good. Otherwise, a significant ξ_i (a few hours and beyond) indicates a possible significant impact of nearby sources. Figure 3.2 illustrates this rationale.

This term is enforced in the observation model Eq. (2.18), which becomes, at any given time:

$$\mathbf{y} = \mathbf{H}\mathbf{c} + \boldsymbol{\xi} \cdot \mathbf{\Pi}\mathbf{e} + \widehat{\boldsymbol{\epsilon}}\,, \tag{3.17}$$

where $\boldsymbol{\xi} \cdot \mathbf{\Pi}\mathbf{e}$ is the vector of entries $[\boldsymbol{\xi} \cdot \mathbf{\Pi}\mathbf{e}]_{i,k} = \xi_i\mathbf{\Pi}_{i,k}\mathbf{e}$. The residual error $\widehat{\boldsymbol{\epsilon}}$ should statistically be smaller than $\boldsymbol{\epsilon}$ of Eq. (2.18) since part of the representativeness error should now be accounted for by the subgrid term. We denote its covariance matrix with $\widehat{\mathbf{R}} = \mathrm{E}\left[\widehat{\boldsymbol{\epsilon}}\widehat{\boldsymbol{\epsilon}}^{\mathrm{T}}\right]$. Under independence assumptions, the two are connected by

$$\mathbf{R} = \mathrm{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\right] = \boldsymbol{\xi} \cdot \mathbf{\Pi}\mathrm{E}\left[\mathbf{e}\mathbf{e}^{\mathrm{T}}\right]\mathbf{\Pi}^{\mathrm{T}} \cdot \boldsymbol{\xi}^{\mathrm{T}} + \widehat{\mathbf{R}}\,. \tag{3.18}$$

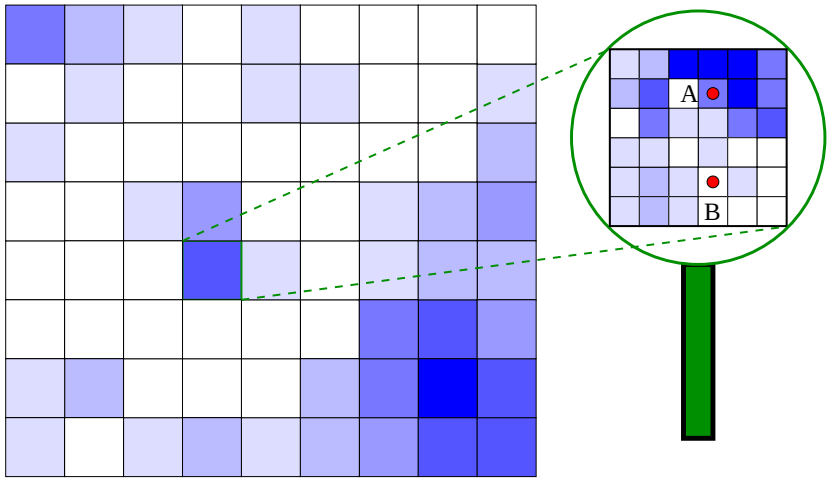


Figure 3.2: Possible physical interpretation of the subgrid model. This mesh represents the CO inventory of a spatial domain. The darker the blue shade, the bigger the emission in the grid-cell. Notice the high emission zone in the south-east corner. A zoom is performed on one of the central grid-cell (see in the magnifier). Inside this grid-cell is represented a finer scale inventory inaccessible to the modeller that may represent the true multiscale inventory. Two CO monitoring stations are considered. Station A is under the direct influence of a nearby active emission zone that represents a significant contribution to the grid-cell flux. The model, operating at coarser scales cannot scale the influence of this active zone onto station A, even though it has an estimation of its total contribution through the grid-cell total emission. Differently, station B which is located in the same grid-cell, does not feel the active zone as much as station A. Our subgrid statistical model assumes that the influence of the active subgrid zone onto A or B has a magnitude quantified by the influence factors ξ_A and ξ_B. Obviously, in this case, one has $\xi_A \gg \xi_B$. Notice that both station A and station B are under the influence of the south-east corner of the whole domain. But this influence is meant to be represented through the Eulerian coarser ATM.

3.5.2 Coupling to the 4D-Var system

Taking into account the statistical subgrid model, the 4D-Var cost function becomes:

$$\begin{aligned}\mathcal{J}(\boldsymbol{\alpha},\boldsymbol{\xi}) &= \frac{1}{2}\sum_{h=0}^{N_h-1}(\boldsymbol{\alpha}_h-\mathbf{1})^{\mathrm{T}}\mathbf{B}_{\alpha_h}^{-1}(\boldsymbol{\alpha}_h-\mathbf{1}) \\ &+\frac{1}{2}\sum_{k=0}^{N}(\mathbf{y}_k-\mathbf{H}_k\mathbf{c}_k-\boldsymbol{\xi}\cdot\boldsymbol{\Pi}\mathbf{e}_k)^{\mathrm{T}}\widehat{\mathbf{R}}_k^{-1}(\mathbf{y}_k-\mathbf{H}_k\mathbf{c}_k-\boldsymbol{\xi}\cdot\boldsymbol{\Pi}\mathbf{e}_k) \\ &+\sum_{k=1}^{N}\boldsymbol{\phi}_k^{\mathrm{T}}(\mathbf{c}_k-\mathbf{M}_k\mathbf{c}_{k-1}-\Delta t\mathbf{e}_k)\ . \end{aligned} \tag{3.19}$$

As mentioned in the previous section, if the subgrid model does account for a significant part of the representativeness error, the error covariance matrix $\widehat{\mathbf{R}}_k$ should differ from $\mathbf{R}_k$ since it accounts for the residual errors. Its magnitude will be determined by the χ^2 method.

A joint iterative optimisation of the scale factors $\boldsymbol{\alpha}$ and the influence factor vector $\boldsymbol{\xi}$ is used to minimise the cost function. Within each iteration, $\boldsymbol{\xi}$ is obtained by a minimisation of the cost function under the constraint of positivity of the ξ_i. To perform the minimisation, one needs the gradient with respect to $\boldsymbol{\xi}$

$$\nabla_{\boldsymbol{\xi}}\mathcal{J}(\boldsymbol{\alpha},\boldsymbol{\xi})=\sum_{k=0}^{N}\mathbf{e}_k^{\mathrm{T}}\boldsymbol{\Pi}^{\mathrm{T}}\widehat{\mathbf{R}}_k^{-1}(\mathbf{y}_k-\mathbf{H}_k\mathbf{c}_k-\boldsymbol{\xi}\cdot\boldsymbol{\Pi}\mathbf{e}_k)\ , \tag{3.20}$$

and the innovation vector of Eq. (2.26) becomes

$$\boldsymbol{\Delta}_k=\mathbf{H}_k^{\mathrm{T}}\widehat{\mathbf{R}}_k^{-1}(\mathbf{y}_k-\mathbf{H}_k\mathbf{c}_k-\boldsymbol{\xi}\cdot\boldsymbol{\Pi}\mathbf{e}_k)\ . \tag{3.21}$$

After the ξ_i are optimised, the χ^2 method is used to rescale the new observational error covariance matrices $\widehat{\mathbf{R}}_k=\widehat{r}\mathbf{I}_{m_k}$. It is used iteratively until convergence of $\widehat{r}$. For each cycle within this loop, the $\boldsymbol{\alpha}$ are first optimised using 4D-Var for the current value of $\boldsymbol{\xi}$ and of the $\widehat{\mathbf{R}}_k$. Then the $\widehat{\mathbf{R}}_k$ are updated. Figure 3.3 summarises the minimisation procedure for the coupled DA system (in short 4D-Var-ξ). Note that the first step of the minimisation can begin by optimising either the influence factors $\boldsymbol{\xi}$ or the scale factor vector $\boldsymbol{\alpha}$. Our tests show that the final results of both minimisations are consistent. However, the former approach shows a faster convergence.

3.6 Application of 4D-Var-ξ

In this section, the 4D-Var-ξ system is first applied to the same setup as the 4D-Var analysis of Section 3.4. The resulting analysis is discussed both in terms of retrieved emission and in terms of analysed CO concentrations. Then, the system is validated with a comparison, a cross-validation and a forecast experiments.

3.6.1 Analysis

3.6.1.1 Minimisation of the cost function

Figure 3.4 shows the minimisation of the cost function $\mathcal{J}$ in the two following cases: the optimisation of the scale factor vector $\boldsymbol{\alpha}$ (4D-Var alone), and the optimisation of $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$ with 4D-Var-ξ. In the latter case, several cycles of 9 iterations each are run. In each cycle, the influence factors are first optimised and 8 other iterations are used to optimise the scale factors.

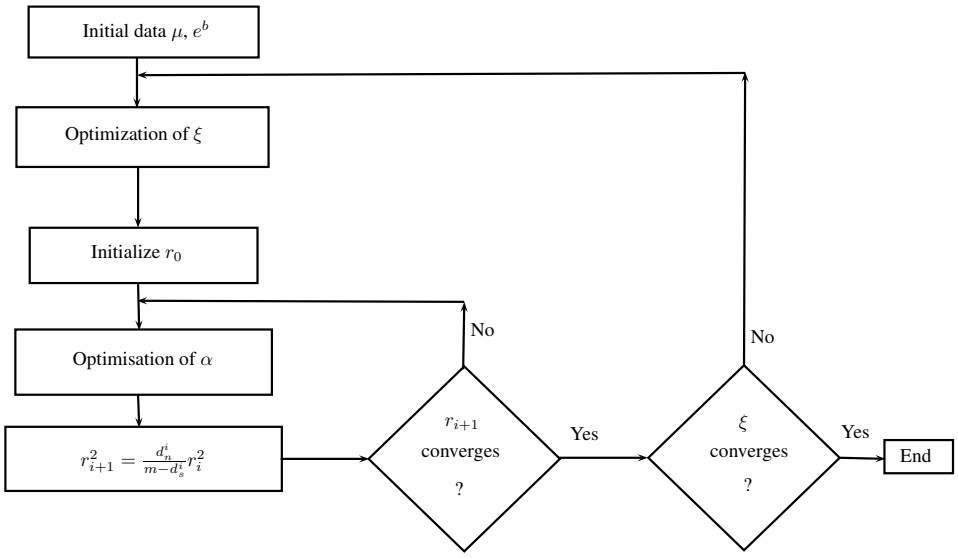


Figure 3.3: Schematic of the minimisation algorithm for the 4D-Var-ξ system.

This cycle is repeated 9 times, beyond which convergence is reached. For the first iteration of a cycle, the diagonal elements ($\widehat{r}$) of the observational covariance matrix are diagnosed with χ^2. This may lead to a temporary increase of the cost function value as seen in Fig. 3.4. In both cases the cost function $\mathcal{J}$ consistently converges to half of the observation numbers (that is $m/2 = 53,957$). The values of the observation and background terms of the cost function, $\mathcal{J}_o$ and $\mathcal{J}_b$ respectively, have also been plotted (cf. Fig. 3.4).

The $\mathcal{J}_o$ of 4D-Var-ξ convergences to a higher value than the $\mathcal{J}_o$ of 4D-Var because the coupled scheme is able to identify a higher fraction of the degrees of freedom as noise (representativeness errors). The $\mathcal{J}_b$ of 4D-Var-ξ convergences to a smaller value than the $\mathcal{J}_b$ of 4D-Var because the coupled scheme recognises that the degrees of freedom for the signal present in the observations are significantly less important than what 4D-Var would assume. Specifically the number of degrees of freedom for the signal is $d_s = 6,316$ with 4D-Var, whereas it is $d_s = 2,367$ with 4D-Var-ξ. They stand for about 2% of the information load of the in-situ observations. This shows that ignoring the representativeness issue leads to a severe overestimation of the information content of the dataset. The standard deviation of the residual diagnosed observation error that was $r \simeq 652.5\,\mu$g m^{-3} without the implementation of the subgrid scheme is now $\widehat{r} \simeq 422\,\mu$g m^{-3}.

3.6.1.2 Results: Scores

Statistical indicators are computed for the output of an 8-week experiment using the 4D-Var-ξ scheme. They are reported in Tab. 3.1 (joint optimisation of $\boldsymbol{\xi}$ and $\boldsymbol{\alpha}$). A significantly better agreement is obtained between the analysis and the observations. The large underestimation of the CO concentrations (see the means in Tab. 3.1), is significantly reduced: the normalised bias is as small as 1.4%. The total emission is diagnosed to be 1.16 Tg. This is an inventory increase of about 9%, which is rather consistent with studies performed over Western Europe using remote sensing. In addition to the bias reduction, it also leads to an increase of the Pearson correlation coefficient up to 0.73. The optimisation of the influence coefficients, using the a priori fluxes, leads to decrease the root mean square error (RMSE) from 701 μg m^{-3} to 503 μg m^{-3}. The emission optimisation decreases this number down to 418 μg m^{-3}. The impact of the subgrid model on the RMSE is consistent with the predominance of the local

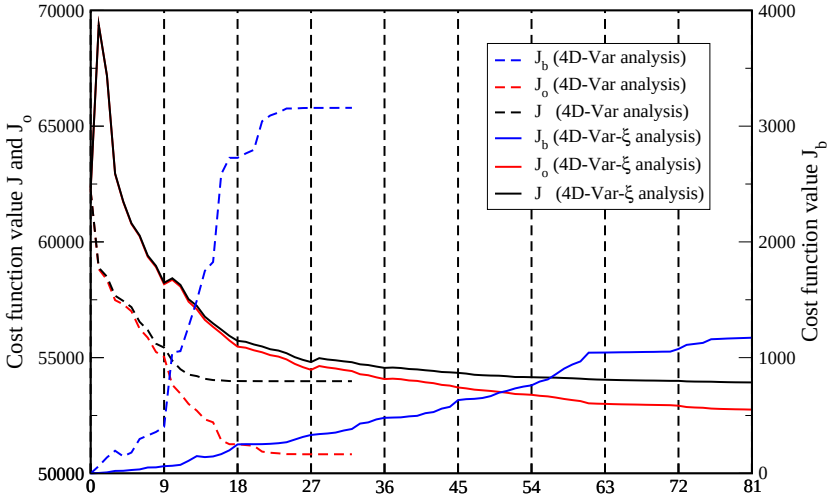


Figure 3.4: Iterative decrease of the full cost function (black lines), of the background term of the cost function $\mathcal{J}_b$ (blue lines), and of the observation departure term of the cost function $\mathcal{J}_o$ (red lines). For the sake of clarity, the $\mathcal{J}_b$ values are to be read on the right y-axis. Two optimisations are considered: with 4D-Var (dashed lines), and joint 4D-Var and ξ optimisation (full lines), within the assimilation window of the first 8 weeks of 2005.

sources on the observations.

3.6.1.3 Results: Total stations scores

The scores for the simulations at each station are presented in Appendix B.1. The value of the bias between the observations and the simulations lies between $10\mu g/m^3$ and $1922\mu g/m^3$. The RMSE spreads from $195.5\mu g/m^3$ to $2500\mu g/m^3$. The Pearson coefficient changes from -0.12 to 0.51. The FA_2 coefficient varies between 0.05 and 0.86.

The same statistical indicators are displayed in Appendix B.2, as regards the results of the 4D-Var simulation. In this case, the bias between the simulated results and the corresponding observations is between $0.8\mu g/m^3$ and $1852\mu g/m^3$. The RMSE changes from $183.7\mu g/m^3$ to $2422\mu g/m^3$. The correlation between the simulated results and the measurements varies between -0.07 and 0.73. FA_2 is changed from 0.07 to 0.93.

The third set of indicators coresponds to the 4D-Var-ξ results (see Appendix B.3). The bias is decreased. It is now between $1\mu g/m^3$ and $301.3\mu g/m^3$. The RMSE ranges between $181.2\mu g/m^3$ and $1176\mu g/m^3$. The correlation is also increased and ranges between -0.01 and 0.78. In this case, FA_2 varies between 0.31 and 0.98.

3.6.1.4 Results: Spatial distribution of the retrieval

The values of the scale factors $\boldsymbol{\alpha}$ of the 4D-Var-ξ system range between 0.01 and 19.5, with an average value of 1, showing that some important correction can be made to the inventory. Figure 3.5 displays the carbon monoxide EMEP+MEGAN inventory (the first guess) integrated over the first 8 weeks of 2005, for each grid-cell. Figure 3.6 displays the ratio of time-integrated retrievals to the time-integrated EMEP+MEGAN inventory, for each grid-cell. Figure 3.6a displays the retrieval obtained using 4D-Var, whereas Fig. 3.6b displays the retrieval obtained using 4D-Var-ξ. 4D-Var-ξ shows a much less pronounced correction than the 4D-Var retrieval, which is consistent with the findings from the statistics discussed in the previous section. The joint inverse modelling retrieval suggests an increase of the emissions in the South of Paris

area, Lyons, La Rochelle, Lille and in the Mediterranean coast of France, pointing to an underestimation of the inventory. It suggests a decrease of the emissions in the area of Dunkerque, in the area of Metz, and in the North of Paris area, pointing to an overestimation of the inventory. The values of the scale factors ($\boldsymbol{\alpha}$) of the 4D-Var-ξ system range

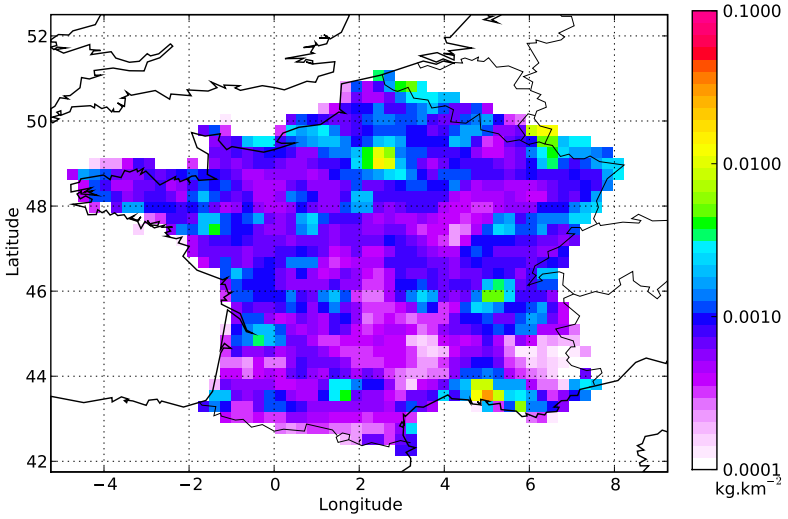


Figure 3.5: Time-integrated spatial distribution of the carbon monoxide EMEP+MEGAN inventory over the first 8 weeks of 2005.

3.6.1.5 Results: scatterplots

In Fig. 3.7a, a scatterplot compares the observations to the concentrations simulated by the model using the a priori emissions. It is clearly impacted by the representativeness errors, since the variability of the observations is much stronger than that of the simulated concentrations. In Fig. 3.7b, a second scatterplot compares the observations to the ATM concentrations using the a posteriori emissions from 4D-Var. Even though 4D-Var corrects the shape of the scatterplot, it is still highly impacted by representativeness errors. Figure 3.7c is a scatterplot of the observations versus the concentrations diagnosed by the 4D-Var-ξ system. The representativeness errors have been significantly reduced. However, there is still a residual impact for the smallest observations. This may be due to situations where carbon monoxide emitted locally is not advected nearby monitoring station i, whereas ξ_i may be significant because of the impact of the local source when the winds are blowing in the direction of the instrument. Indeed, our simple statistical model cannot account for the changes in the local micro-meteorology, only for its indirect impact.

3.6.1.6 Results: On-site profiles

Here, the focus is on the analysis at individual stations. The values of the station-dependent influence factors ξ_i range between 0 and 97.5 h, with a median value of 5.9 h, and a mean value of 11.3 h (Table 3.2 presents the value of the influence factor for each of the station).

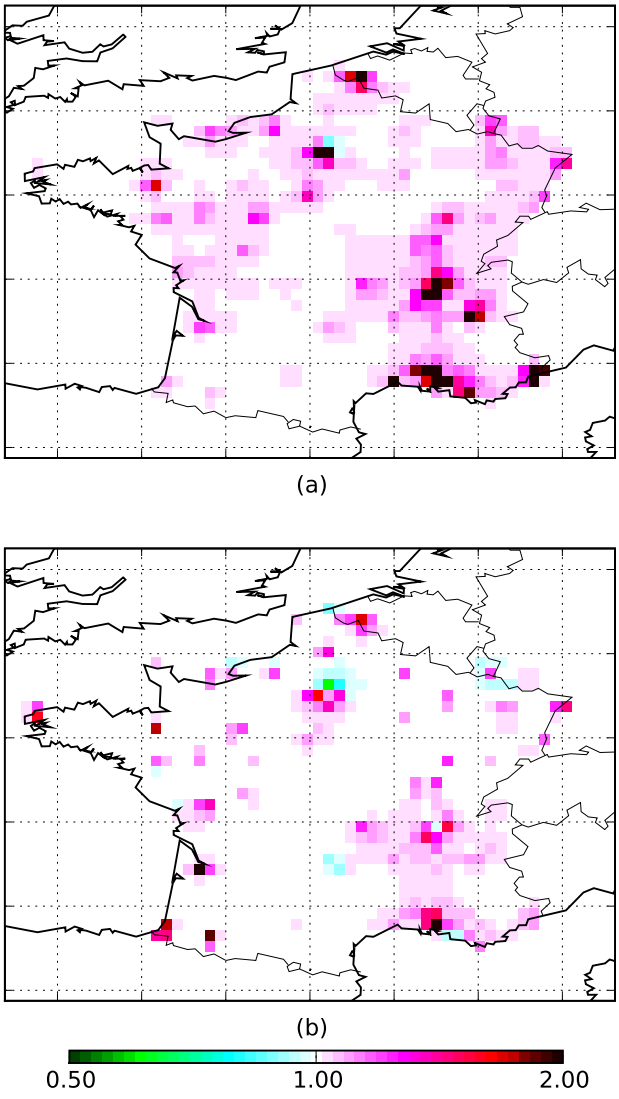


Figure 3.6: Ratio of the time-integrated CO flux retrieval to the EMEP+MEGAN time-integrated CO flux for each grid-cell, in the 4D-Var case (a) and in the joint 4D-Var and subgrid model case (b).

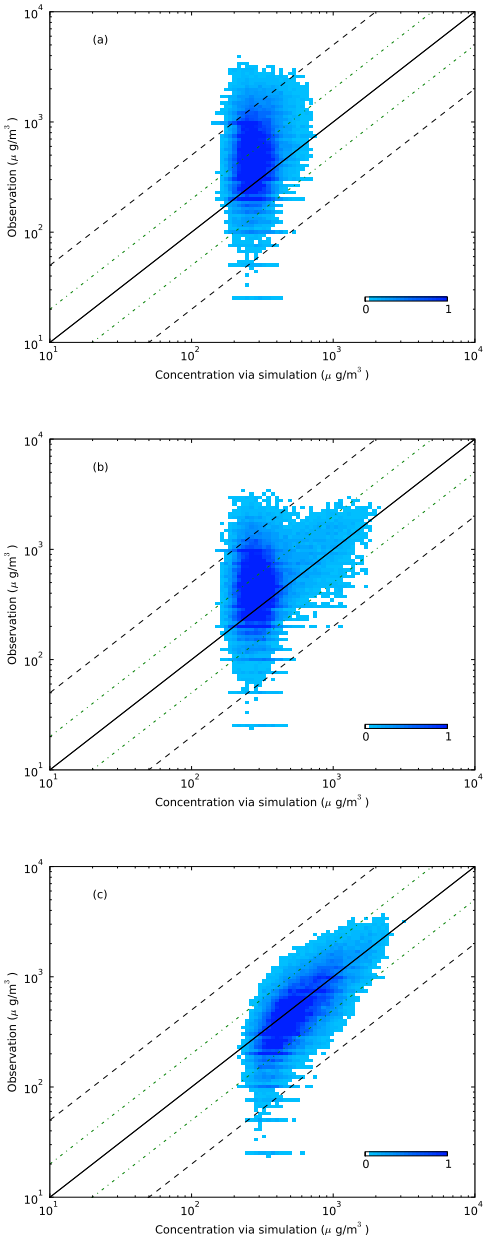


Figure 3.7: Scatterplot during 8-week: (a) comparison between the concentrations via the model and the observations, (b) comparison between the concentrations via the model using the a posteriori emissions retrieved from 4D-Var and the observations, (c) comparison between the concentrations diagnosed by the 4D-Var-ξ system and the observations. The colour bars show the correspondence between the blue shade and the density of points of the scatterplot. This density has been normalised so that its maximum is 1. Dashed lines are the FA_5 dividing lines, and dashed-dotted lines are the FA_2 dividing lines.

Table 3.2: The values of the influence factors ξ_i for the stations.

Station	ξ	Station	ξ	Station	ξ
HAYANGE	3.06	Liane Boulogne Sud	11.91	DAIX	0.56
Marignane Ville	8.00	LIBERTE	1.34	Station MARSANNAY	4.54
Port de Bouc EDF	0.99	PASTEUR	0.59	BETHUNE PROX AUTO	5.66
PLOMBIERES	45.62	LA BASSEE/CENTRE	1.13	COUBERTIN	3.04
AIX CENTRE	5.21	Roubaix/Serres	1.01	ST ETIENNE ROND PT	15.84
TOULON FOCH	97.52	Hotel de ville	7.15	RIVE DE GIER	3.70
AVIGNON ROCADE	2.34	Rue de la Tour	7.62	Hotel Districal	11.56
Place Victor Basch	2.98	Grenoble Foch	24.75	Epinal	11.96
AUBERVILLIERS	0.00	Le Rondeau	16.29	Bar-le-Duc	20.27
Avenue des Champs Elysees	1.19	Strasbourg Clemenceau	8.07	Luneville	16.75
Boulevard peripherique Auteuil	2.67	Muhl.ASPA	3.33	BORDEAUX-BASTIDE	1.26
PARIS 1er Les Halles	0.00	CTRE VILLE MEGEVAND	11.19	MERIGNAC	4.56
Autoroute A1 - Saint-Denis	1.67	Amiens Saint Leu	2.89	SAMONZET	20.47
Rue Bonaparte	0.89	LAENNEC	2.84	ANGLET	9.26
Quai des Celestins	1.99	Halles centralles	5.86	Chalon centre ville	19.30
LeHavre Republique	23.11	PUITS GAILLOT	5.96	Champforgeuil	1.34
ESQUERCHIN A DOUAI	3.33	BERTHELOT	8.54	Hilaire Chardonnet	5.17
Jardin Lecoq CF	4.07	GARIBALDI	13.65	Montceau-les-Mines	9.87
Aurillac Centre	61.46	LA MULATIERE	6.84	Macon Paul Bert	10.31
Le Puy Fayolle	58.74	VAUCELLES	10.35	Le Creusot Molette	9.35
Rousillon	14.51	Cherbourg Paul Doume	8.03	Gambeta	11.92
Saint Denis	10.64	Batiment ELF-ATO	0.00	Mirabeau	14.96
Pres Arenes	5.20	Forbach(12)	1.51	Valence Trafic	10.71
Planas	9.52	GENERAL DE GAULLE	25.57	GONESSE	0.16
rue de la GRILLE	73.93	LA ROE	20.55	VICTOR HUGO	3.70
Place du Marche	13.30	Nice Pellos	45.84	METZ-BORNY	0.00
Mairie MALO	0.40	ANTIBES GUYNEMER	26.80	Brest 3 CDM	34.12
FORT-MARDYCK	0.04	ALEXIS CARREL ROUENG	1.16	Ecole Jules Ferry	5.23
Petite Synthe	0.09	Rouen Le Conquerant	3.88	place de VERDUN	2.59
Calais Centre	0.44	Pasteur	8.21		

In Fig. 3.8, four different time series of concentrations are displayed for four different stations: the observations, the concentrations simulated with the a priori emissions, the concentrations obtained from 4D-Var, and 4D-Var-ξ concentrations. The traffic station of Lille Pasteur, can be cited as an example of small influence factor value with $\xi_i = 0.6$ h. In that station, the simulation concentrations are in quite good agreement with the observations. The correlation between the observations and the simulated concentrations reaches 0.49. It is 0.74 for the 4D-Var-ξ results. At the station Paris, boulevard périphérique Auteuil (suburban), for which ξ_i is of 2.7 h, the correlation increases from 0.29 up to 0.77. Orléans Gambetta (traffic zone) station can be cited as an example with a moderate influence factor value of $\xi_i = 11.9$ h. At this station, the Pearson correlation coefficient increases from 0.11 to 0.67 when using the 4D-Var-ξ system. The dependence of the observations and the local emissions is clearly shown in Fig. 3.8c. The model simulation gives a smooth curve, whereas the observations are highly fluctuating. The 4D-Var system is able to anticipate the trend of the concentrations, but cannot predict the peaks. Furthermore, it over-estimates the inventory by trying to adjust to the peaks.

Figure 3.8d shows the concentrations in Nice Pellos (urban station) with a high influence factor value of $\xi = 45.8$ h. The results of 4D-Var-ξ are in good agreement with the observations whereas neither the simulation, nor 4D-Var are able to match the observations. The correlation value is significantly increased from 0.32 to 0.68. It is also clear that although 4D-Var-ξ is able to account for a substantial part of the peaks, it underestimates their maxima and overestimates the minima, which may be due to residual representativeness error.

3.6.1.7 Results: sensitivity to the background standard deviation

The whole inverse modelling study using a background standard deviation is performed for the carbon monoxide of 100%, instead of 50%. The results are qualitatively unchanged. They are barely quantitatively changed. For instance, one retrieves a total of 1.18 Tg instead of 1.16 Tg over the 8-week winter period. This relative insensitivity is mostly due to the use of the χ^2 criterion.

3.6.2 Validation

A direct and reliable validation of a spatial emission inventory is currently out of reach for most pollutants (see the in-depth discussion of Vestreng et al. [2007] about SO_2). It is only possible to compare with another independent estimation (top-down or bottom-up), which, as a relative comparison approach, may not be as satisfying as a straight comparison to observations. Local flux measurements are possible (e.g. for CO_2) in some media but these are sparse and cannot fully validate a spatial inventory. Therefore, a CO emission inventory can only be indirectly validated. For instance one can compare the CO concentrations simulated with the inventory to real measurements.

We shall first compare the total emitted carbon monoxide to an independent bottom-up inventory over France. We will then compare simulated concentrations obtained with an inventory retrieved from a training network, on a distinct validation network. Finally, after an assimilation period of 8 weeks, we shall make a 10-month CO concentration forecast. The forecasted concentrations will be compared to independent observations (that have not been assimilated).

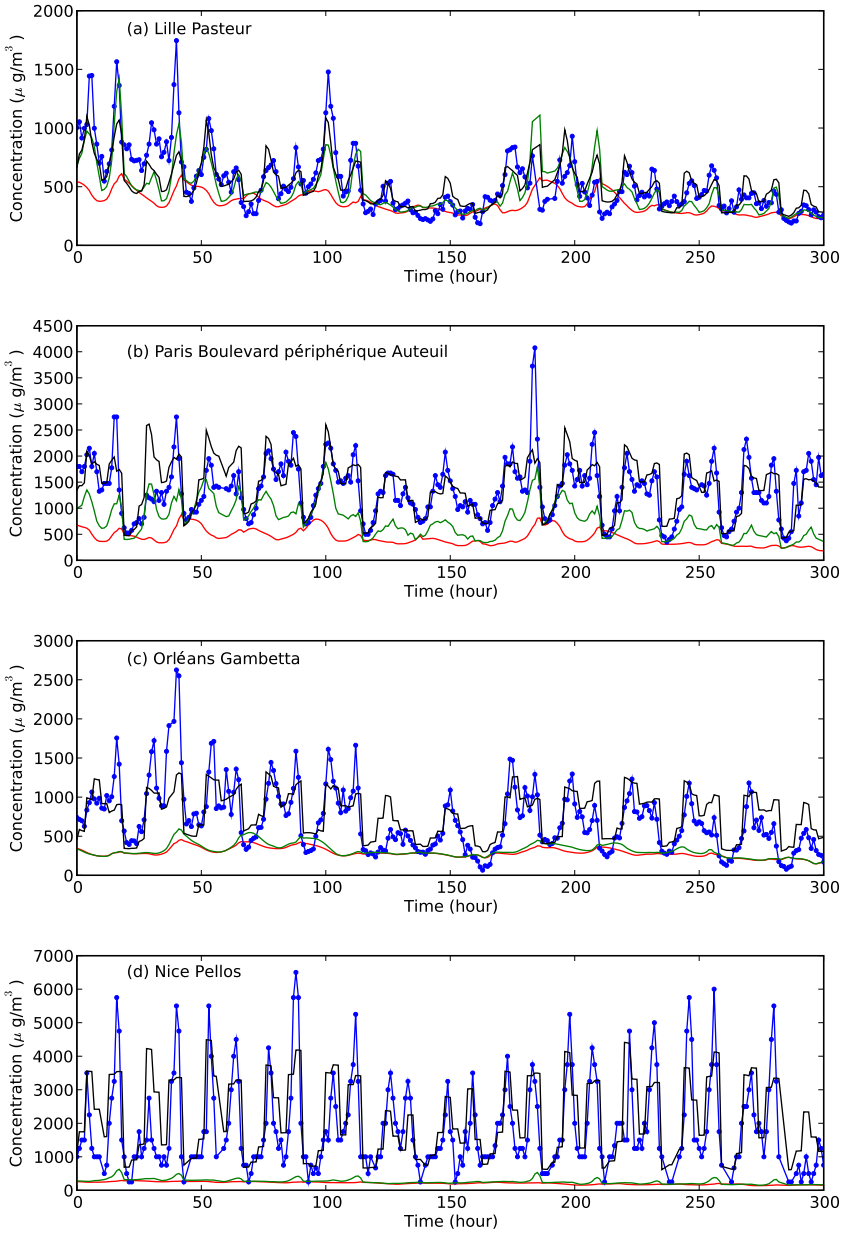


Figure 3.8: Time series of CO concentrations for the first 300 hours of 2005, at four stations: observations (blue), simulation using the prior emissions (red), simulation using the posterior emissions of data assimilation (green) and simulation using the posterior emissions of 4D-Var-ξ (black) with adjusted observations using the statistical subgrid model.

3.6.2.1 Global comparison with the CITEPA inventory

The total retrieved CO emitted mass from 4D-Var-ξ is compared to the inventory of the Centre Interprofessionnel Technique d'Études de la Pollution Atmosphérique (CITEPA[2]). According to CITEPA, the total French inventory for 2005 is 5.3 Tg. We have inferred the total emitted mass for the first 8 weeks of 2005 using the weekly and the monthly coefficients of GENEMIS for each of the 11 sectors of the SNAP nomenclature of emitting activities. The contribution of each SNAP sector to the total emission is estimated following EMEP distribution for this year. Following this rationale, the total CO emitted mass of the CITEPA inventory is found to be 1.15 Tg between January 1st, and February 26. This value is very close to 1.16 Tg obtained with 4D-Var-ξ.

3.6.2.2 Cross-validation experiment

49 BDQA stations have been randomly selected as a training network. Inverse modelling will be performed using the CO observations of this subnetwork for the first 8 weeks of 2005. The rest of the stations of the BDQA network forms a 40-station validation network. The observations of these stations will be compared to the simulated CO concentrations obtained using the retrieved emission field inferred from the training set. The partition between the BDQA stations is displayed in Fig. 3.9.

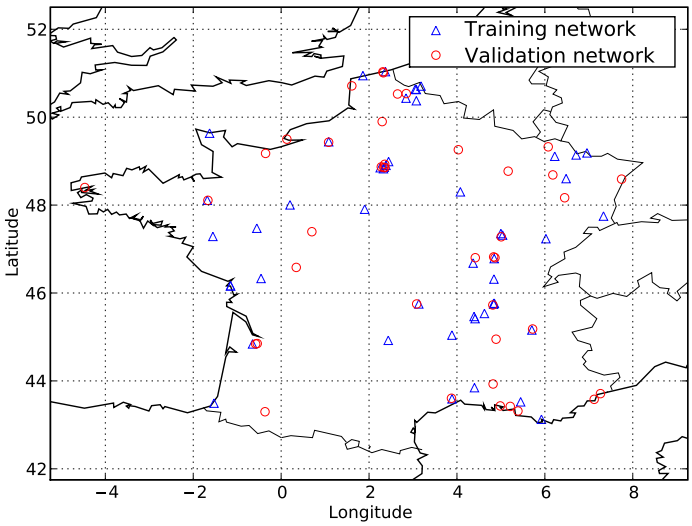


Figure 3.9: The training (triangle) and validation (circle) subnetworks that partition the BDQA stations measuring carbon monoxide. This partition is randomly generated for the cross-validation experiment.

Three simulations for validation are performed: a simulation using the EMEP+MEGAN background inventory; a simulation using the emissions retrieved with 4D-Var; and a simulation using the emissions retrieved with 4D-Var-ξ. In addition to these three simulations, we shall use the influence coefficients ξ_i attached to the stations of the validation network to correct the concentrations, using the background emissions, the 4D-Var retrieved emissions, and the 4D-Var-ξ retrieved emissions. Even though these 40 factors have been inferred (in the previous

[2] `http://www.citepa.org/emissions/nationale/Aep/aep_co.htm`

section) using observations of the full network, we believe they are intrinsic to the stations. Inferring them from a different (sufficiently large) observation set would yield close values. We have checked this by comparing the ξ_i of the training network obtained from a 89-station (full network) optimisation, with the ξ_i of the training network obtained from a 49-station (training network) optimisation. The results, that are reported in a scatterplot Fig. 3.10, confirm that the values are close, and support that they are intrinsic to each station.

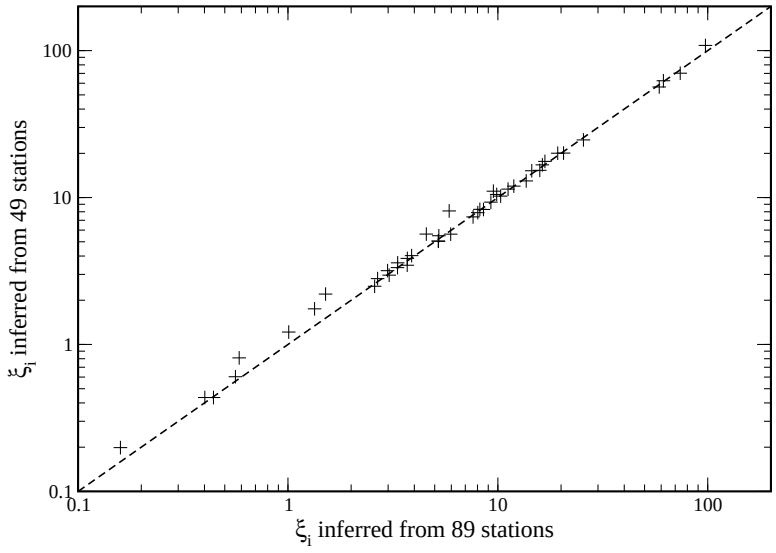


Figure 3.10: Scatterplot of the 49 ξ_i of the training network inferred from either the training network or the full network (89 stations). Four $\xi_i = 0$ crosses are missing. In the four cases, they were concordantly diagnosed to be 0 by the two inferences.

The statistical scores, as well as the total emitted mass, for these six validation experiments are reported in table 3.3.

Table 3.3: Comparison of the observations and the forecasted concentrations on the validation network for the first 8 weeks of 2005. The statistical indicators are described in Tab.3.1. Additionally, the total retrieved emitted mass is given (in Tg). The corresponding value for the retrieved mass using the full network is recalled in parenthesis.

Used inventory	$\overline{C}$	$\overline{O}$	NB	RMSE	R	FA_2	FA_5	Total mass
Background	296	697	-0.81	771	0.16	0.51	0.88	1.06 (1.06)
4D-Var	357	697	-0.65	726	0.28	0.57	0.89	1.25 (1.44)
4D-Var-ξ	310	697	-0.77	758	0.22	0.52	0.89	1.14 (1.16)
Background + climatological $\boldsymbol{\xi}$	644	697	-0.08	538	0.60	0.73	0.96	1.06 (1.06)
4D-Var + climatological $\boldsymbol{\xi}$	968	697	0.33	1216	0.40	0.67	0.94	1.25 (1.44)
4D-Var-ξ + climatological $\boldsymbol{\xi}$	674	697	-0.03	514	0.64	0.75	0.96	1.14 (1.16)

Firstly 4D-Var-ξ without correction at the validation stations performs poorly, with scores of the same order as 4D-Var. This is to be expected since 4D-Var-ξ is meant to be used in conjunction with the $\boldsymbol{\xi}$ coefficients, which is not the case for this experiment. Secondly, 4D-Var yields sensibly better scores than 4D-Var-ξ. This is due to the excessive correction of 4D-Var that wrongly takes the CO peaks as a systematic bias. As should be, this bias correction equally applies to the validation set, leading to slightly better scores than 4D-Var-ξ, but for the wrong

reasons.

Applying the ξ_i coefficients of the validation stations to the concentrations obtained with the first guess emissions considerably reduces the bias and improves all the other statistical indicators as compared to the reference simulation. Applying the ξ_i coefficients of the validation stations to the concentrations obtained with the 4D-Var retrieved emissions leads to a very large positive bias. Even though the approach is by construction inconsistent, it yields significantly better scores as compared to using the 4D-Var retrieval without corrections on the validation stations. Lastly, the ξ_i coefficients of the validation stations are used in conjunction with the 4D-Var-ξ retrieved emission field. This leads to much higher scores than the other experiments. These indicators are consistent with the scores obtained using the full network data (in Tab. 3.1).

It is remarkable that the total retrieved mass of this last experiment, 1.14 Tg, is consistent with that obtained by 4D-Var-ξ using all stations, that is 1.16 Tg. A convincing validation of such a retrieval methodology would require such a consistency. The same is not true for 4D-Var with 1.25 Tg obtained using the training subnetwork and 1.44 Tg using the full network, pointing to the inconsistency of the method that does not properly account for the representativeness errors.

3.6.2.3 Forecast experiments

A validation forecast is performed over the year 2005. This second indirect validation is demanding since no new observation are assimilated over a ten-month period. That is why in atmospheric chemistry/air quality a forecast is often considered a more stringent validation test [Zhang et al., 2012]. However, our validation by a forecast has a limitation due to the statistical subgrid model. It is meant to efficiently apply to the observational network employed in the initial assimilation time-window. Notice that this limitation is inherent to any forecasting system making use of some form of statistical adaptation.

Four runs are considered. They all use the ECMWF meteorological fields and the MOZART, version 2, output for the initial and boundary conditions. The first run is a direct simulation over 2005 that is driven by the EMEP+MEGAN inventory. The second one is a direct run from February 26 to December 31, but using the optimal $\boldsymbol{\alpha}$ obtained from the 4D-Var analysis from January 1 to February 25, and Eq. (3.6) to generate the inventory. The third one is a direct run from February 26 to December 31, using the EMEP+MEGAN inventory but using the optimal $\boldsymbol{\xi}$ obtained from an optimisation over $\boldsymbol{\xi}$ of the total cost function from January 1 to February 25. The fourth one is a direct run from February 26 to December 31, but using the optimal $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$ parameters obtained from the 4D-Var-ξ analysis from January 1 to February 25, and Eq. (3.6) to generate the inventory. None of the observations from February 26 to December 31 are assimilated. They are exclusively used for validation.

Such forecast requires a forecast of the emissions. The parameterisation of the emission by the $\boldsymbol{\alpha}$ allow us to do so. In particular some of the temporal (but not spatial) seasonal variability is implicitly accounted for thanks to the GENEMIS temporal modulation present in the first guess $\mathbf{e}_b$.

Firstly, we have focused on the first month forecast, from February 26 to March 26, where one can assume that the winter emission trend endures. The results are in very good agreement with the observations. For the forecast period, the correlation coefficient between the observations and 4D-Var-ξ increases from 0.13 to 0.68. The RMSE is improved by about 40% during the analysis period. Almost 68% of that improvement is due to the optimisation of the influence factors ξ_i.

Secondly, we have extended the forecast period, from February 26 to December 31 across seasons. The monthly results for the RMSE and the correlation coefficients, over the year 2005

are presented in Fig. 3.11. Using 4D-Var-ξ, the RMSE decreases decreases by 282 μg m^{-3} within the analysis period, January 1 to February 26 (left side of the vertical dashed line). It decreases by 172 μg m^{-3} during the forecast period, from February 26 to December 31 (right side of the vertical dashed line). The improvement is remarkably persistent during the whole 10-month forecast period. It shows that choosing $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$ as control vectors has a good prognostic value. In spring and summer, the RMSE decreases for all four experiments. This can be due to the decrease of urban heating during that period which is accounted for in the cycles of the inventory but which reduces a source of uncertainty. It can also be seen that the RMSE gain in the spring and summer is essentially due to the subgrid model identification, and not the emission estimation, since 4D-Var-ξ and the optimal-ξ forecast yield the same RMSE. Unsurprisingly, this means that the emission retrieval carried out over two winter months are not optimal for the spring and summer months. Another possible explanation is the emergence of new source of errors in the spring-summer time, such as the higher OH concentration that leads to a higher reactivity of CO, or a stronger turbulent mixing in the boundary layer. However, this should be balanced by a persistent gain in the spring-summer period of the correlation due to the emission retrieval.

3.7 Conclusion

In this article, a 4D-Var data assimilation system was developed to estimate carbon monoxide fluxes at regional scale. An approximate adjoint of the POLAIR3D model has been built and validated for this 4D-Var system. A study over France, at a resolution of $0.25^\circ \times 0.25^\circ$ is conducted. We used the in-situ observations of the BDQA database that includes the observations from industrial, traffic, urban and suburban stations. They are strongly impacted by local sources that the stations are meant to monitor. Hence, although the number of observations is very significant, their information load is impacted by large representativeness errors. The Pearson correlation coefficient between the simulated concentrations and the observations is computed to be 0.16. A first 4D-Var inversion of the CO fluxes leads to a mild improvement of the skill. The Pearson correlation climbs to 0.36. However looking at stations profile, it is clear that the representativeness errors are not accounted for, since the analysis from 4D-Var cannot reproduce the intense CO peaks. Besides, it leads to an artificially large increase of the retrieved emissions.

Therefore, a simple model is developed to statistically represent the subgrid effects of nearby sources. A coefficient attached to each station is used to estimate this influence. The 4D-Var system is coupled to this subgrid model and the fluxes are determined altogether with the influence coefficients. The correlation coefficient reaches 0.73, while the bias between the observations and the analysed concentrations is considerably reduced. The net increase of the CO inventory is estimated to be 9%, consistent with other top-down approaches using satellite data. Cross-validation experiments using a training subnetwork and a validation subnetwork demonstrates the consistency of the inventory estimation, whereas, in this context, the traditional 4D-Var does not deliver consistent estimations with different training subnetworks. Forecast experiments with the analysed coefficients and fluxes over 10 months, after an assimilation window of 8 weeks, show remarkably persistent scores throughout the year. This emphasises the relevance of the choice of $\boldsymbol{\xi}$ and $\boldsymbol{\alpha}$ as joint control parameter vectors of the 4D-Var-ξ analysis.

We believe that this methodology and experiment show that, in this context, it is possible to extract relevant information from observations strongly impacted by representativeness errors. One limitation which is inherent to the statistical adaptation component of the system is that it is meant to be used on a given monitoring network. A validation forecast can safely be made

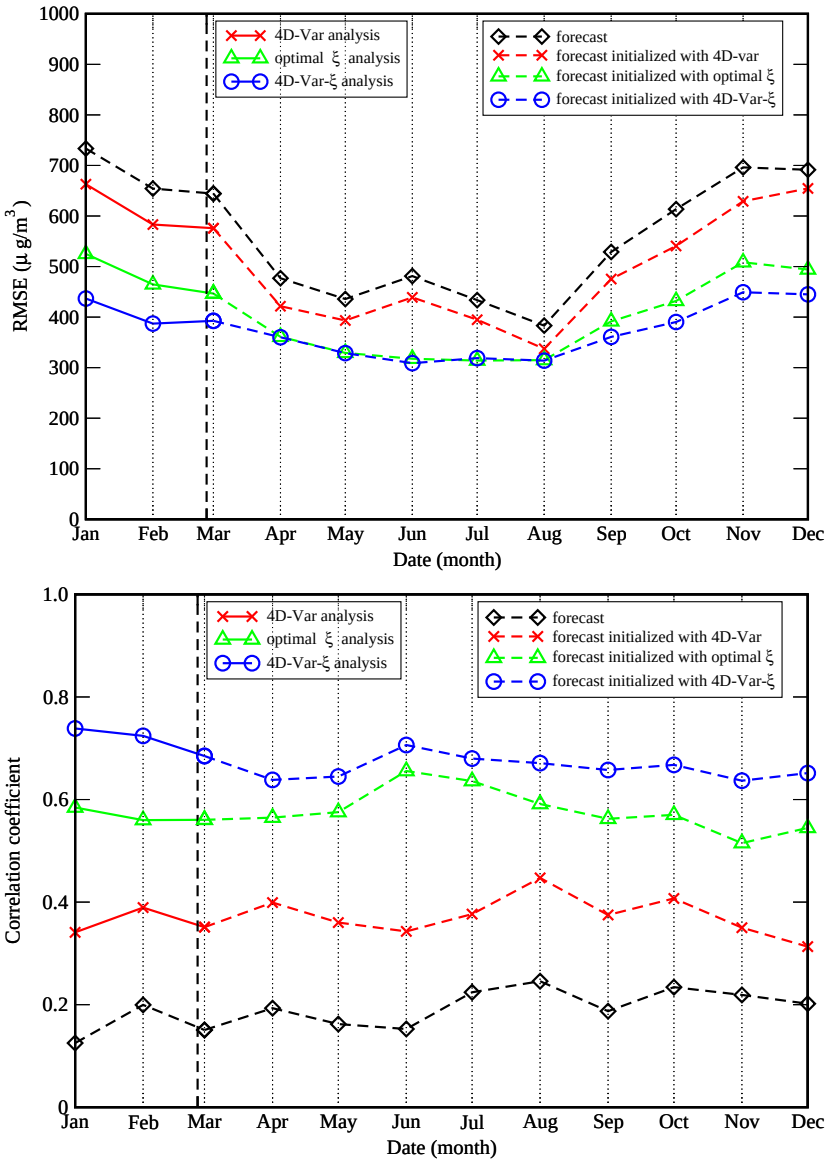


Figure 3.11: Monthly RMSE (left panel) and Pearson correlation (right panel) of four runs: a pure forecast, a ten-month forecast initialised by an 8-week 4D-Var assimilation, a ten-month forecast initialised by an 8-week window where the ξ are optimised and a ten-month forecast initialised with an 8-week joint 4D-Var and ξ optimisation. The vertical dashed line indicates the end of the assimilation window and the start of the forecasts.

to additional stations, but statistical adaptation cannot be performed to these stations, if the related influence factor ξ_i were not previously estimated.

To improve the present statistical subgrid model, which uses the influence factors to estimate the immediate impact of the emissions on the observations, a more comprehensive statistical subgrid model could be used. For instance, that model could include the effects of the wind direction, deposition parameters, etc, that are used or diagnosed in the coarse resolution model. Computationally, it would not be as cheap as the subgrid model used here.

Beyond the carbon monoxide context of this study, it is believed that the integration of the simple statistical subgrid scale into a 4D-Var can be generalised to pollutants whose observations could highly be impacted by representativeness errors.

Chapter 4

Potential of the International Monitoring System radionuclide network for inverse modelling

Summary

The International Monitoring System (IMS) radionuclide network enforces the Comprehensive Nuclear-Test-Ban Treaty which bans nuclear explosions. We have evaluated the potential of the IMS radionuclide network for inverse modelling of the source, whereas it is usually assessed by its detection capability. To do so, we have chosen the *degrees of freedom for the signal* (DFS), a well established criterion in remote sensing, in order to assess the performance of an inverse modelling system. Using a recent multiscale data assimilation technique, we have computed optimal adaptive grids of the source parameter space by maximising the DFS. This optimisation takes into account the monitoring network, the meteorology over one year (2009) and the relationship between the source parameters and the observations derived from the FLEXPART Lagrangian transport model. Areas of the domain where the grid-cells of the optimal adaptive grid are large emphasise zones where the retrieval is more uncertain, whereas areas where the grid-cells are smaller and denser stress regions where more source variables can be resolved.

The observability of the globe through inverse modelling is studied in strong, realistic and small model error cases. The strong error and realistic error cases yield heterogeneous adaptive grids, indicating that information does not propagate far from the monitoring stations, whereas in the small error case, the grid is much more homogeneous. In all cases, several specific continental regions remain poorly observed such as Africa as well as the tropics, because of the trade winds. The northern hemisphere is better observed through inverse modelling (more than 60% of the total DFS) mostly because it contains more IMS stations. This unbalance leads to a better performance of inverse modelling in the northern hemisphere winter. The methodology is also applied to the subnetwork composed of the stations of the IMS network which measure noble gases.

Contents

4.1 Introduction

4.1.1 The IMS network and the CTBT enforcement

The Comprehensive Nuclear-Test-Ban Treaty (CTBT) signed by 182 states bans nuclear explosions [United Nations, 1996]. The monitoring of the treaty is implemented by the United Nations CTBT Organisation (CTBTO), based in Vienna, Austria. It operates an International Monitoring System (IMS) and collects seismic, infrasound, hydroacoustic data as well as radionuclide (particulate matter and noble gases) activity concentrations. This article focuses on the latter. Upon completion of the installation, the radionuclide IMS network will have 80 stations. As of June 2011, 60 stations are certified and operational. The instruments are radionuclide gamma detectors coupled to particle filters. They allow to deliver 24 hour-averaged activity concentrations for several particulate/aerosol species: caesium-137, caesium-134, iodine-131 (aerosol form), etc. In the long term, 40 of those stations will also be able to measure noble gases (xenon-131m, xenon-133, xenon-133m, xenon-135), among which 24 are operating as of June 2011.

The locations of 79 (among 80) stations are detailed in the treaty, even though the actual locations could slightly differ (see `http://www.ctbto.org/map`). The design of the network has been validated using dispersion modelling. For instance, using a global atmospheric transport model (ATM), one can compute the ability of the monitoring network to detect a release stemming from any location on Earth [Ringbom and Miley, 2009]. Recently, the radionuclide IMS network has measured the Fukushima Dai-ichi plume throughout the world, although only part of the observations has been disclosed.

The observations of the IMS network can be used to detect a nuclear tests, and to discriminate nuclear test among underground explosions. They could also help to characterise a test (location, signature and intensity) using inverse modelling techniques. The objective of this article is to determine the potential of the IMS radionuclide network for inverse modelling of the source term, using rigorous mathematical tools in conjunction with global or regional ATMs.

4.1.2 Inverse modelling of tracers

The application of inverse modelling techniques to the reconstruction of the source term is recent in atmospheric dispersion. The European Tracer Experiment (ETEX, Nodop et al. [1998]), initially triggered by the Chernobyl accident, served as a playground to test inverse methodologies [Robertson and Langner, 1998; Pudykiewicz, 1998; Seibert and Stohl, 2000; Issartel and Baverel, 2003; Bocquet, 2005a, b]. Full reconstructions using real data with results close to the known characteristics of the source have been obtained [Bocquet, 2007; Krysta et al., 2008]. These authors used methodologies inspired by geophysical data assimilation techniques: the field to retrieve is discretised into a spatially organised large set of source variables/parameters. Alternatively, the so-called *parametric* methods rely on the optimisation of a restricted set of variables that parametrise the source term. In the specific case of accidental dispersion, the lat-lon coordinates and the emission rate parametrise the source [Delle Monache et al., 2008; Yee et al., 2008].

As far as real radionuclide dispersion events are concerned, these methodologies have been tested on the Algeciras dispersion event [Krysta and Bocquet, 2007; Delle Monache et al., 2008], with about hundred caesium-137 integrated activity concentration measurements. The results are satisfying but mostly because of the very simple shape of the source (a single peak). The inverse modelling approach was also applied to the atmospheric source term of Chernobyl (caesium-137, caesium-134, and iodine-131) by Davoine and Bocquet [2007]; Bocquet [2012] with an estimation of the source terms consistent with the official UNSCEAR source term [United Nations, 2000]. Reconstruction of the source term was also performed for a North Korea nuclear test measured by the IMS radionuclide network [Becker et al., 2010], although the reconstruction was not strictly based on inverse modelling.

In this context, the inverse modelling approach remains a difficult one, because:

- the observations are ground-based and local. Activity concentration measurements are sparse, infrequent or integrated, as opposed to gamma dose measurements. Moreover, point-wise observations may lead to representativeness errors, depending on whether the dispersion model is Eulerian or Lagrangian.

- The dispersion models remain imprecise. They are driven by meteorological fields of increasing precision and reliability at a given resolution, but the planetary boundary layer remains difficult to model, and the vertical turbulent diffusion is still uncertain. With only a few documented field experiments, the microphysical properties of the radionuclides in the atmosphere, are still difficult to grasp. Therefore the physical parametrisations implemented in the ATMs (dry deposition, wet scavenging, aerosol modelling, granulometry of particles) remain gross.

Beyond its own interest, inverse modelling of the source term is also the *sine qua non* condition for a proper forecasting of the resulting plume, as was illustrated by Politis and Robertson [2004]; Bocquet [2007]; Abida and Bocquet [2009].

4.1.3 Objectives and outline

Detectability has been used to assess the performance of the IMS radionuclide network [Hourdin and Issartel, 2000; Wotawa et al., 2003; Ringbom and Miley, 2009]. A more complex criterion is a measure of the ability to interpolate activity concentrations in between the stations of the network, using geostatistical techniques (Wu and Bocquet [2011] and references therein). It has been used to assess and even design a radionuclide monitoring network [Abida et al., 2008]. One step further in complexity, our goal is to evaluate the potential of the IMS radionuclide network for inverse modelling, using an objective quantitative criterion: the degrees of freedom for the signal.

In Section 4.2, we define the typical inverse modelling experiment that could serve the CTBT enforcement. The average quality of an inversion is rigorously defined by the the degrees of freedom for the signal. We do not focus on the particular results of specific inverse modelling experiments. This was done for instance by Winiarek et al. [2011] in the same context. Instead, we focus on the average ability of inverse modelling to extract information from the measurements. A multiscale formalism is used to rigorously diagnose how the information contained in the observations should optimally be spread in regions of the world. In Section 4.3, the formalism is applied to the IMS radionuclide network using all influence functions of year 2009 computed by the CTBTO. Adaptive grids that maximise the degrees of freedom for the signal are computed. By construction, they are optimal for the assimilation of observations. For a given number of grid-cells, they are numerically more efficient, and bear less aggregation errors than regular grids with the same number of grid-cells. They rigorously determine the ability of the monitoring network to resolve source variables through data assimilation. Consequently, they allow to pinpoint well observed (from inverse modelling) as well as poorly observed regions of the world. They have indirect implications on the way to optimise the design of the network. The technique is also applied to the subnetwork of the stations that monitor noble gases. The difference between an Eulerian and a Lagrangian model in the design of those adaptive grids is examined, using a specific region of the globe. Conclusions are given in Section 5.5.

4.2 Methodology of data assimilation

4.2.1 Inverse modelling with Gaussian statistical assumptions

The source parameters are the unknown variables. Each one of them is attached to a grid-cell in a domain Ω, and to a time interval. At first, Ω will be the globe. At the end of Section 4.3, Ω will be a limited area of the globe. We assume that the domain Ω is discretised. We shall use unprojected (lat-lon) coordinates in the following, with N_x meridians and $N_y - 1$ parallels. The source vector $\boldsymbol{\sigma}$ is defined on this grid. It has an extension in time of N_t time-steps, so that $\boldsymbol{\sigma}$ is a vector of dimension $N_x N_y N_t$. The radionuclide plume is observed by the monitoring network. The observations yield a measurement vector $\boldsymbol{\mu}$ in $\mathbb{R}^d$.

The physics of dispersion is assumed linear. This is the case for most transport and physical processes: advection, diffusion, radioactive decay, dry deposition, and wet scavenging. This assumption is true for noble gases or particulate matter, but could be breached for aged parcels of radionuclides which can lead to the formation of aerosols, whose modelling implies complex nonlinear equations.

With this assumption of linearity, and in the absence of boundary conditions, or using clean air boundary conditions which are suitable for accidental release, a source-receptor relationship between the observation vector $\boldsymbol{\mu}$ and the source $\boldsymbol{\sigma}$ is established. It is formalised by the

Jacobian matrix $\mathbf{H}$

$$\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\sigma} + \boldsymbol{\epsilon}\,, \tag{4.1}$$

where the vector $\boldsymbol{\epsilon}$ represents errors of all kinds: instrumental error, representativeness error, and model errors.

The simplest approach for non-parametric inverse modelling is to minimise the discrepancy

$$\mathcal{L}(\boldsymbol{\sigma}) = \frac{1}{2}\left(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}\right)^{\mathrm{T}} \mathbf{R}^{-1} \left(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}\right)\,, \tag{4.2}$$

where $\mathbf{R} = \mathrm{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\right]$ is the observation error covariance matrix, which, in this ground observation context, is almost always assumed diagonal, even though transport model errors could induce some cross-correlations. Following the Bayesian paradigm of geophysical data assimilation, a background term (also called regularisation term in this inverse modelling context) should be added to the cost function Eq. (4.2). This term is obviously unavoidable when the number of variables to retrieve is greater than the number of observations. However, even with a larger set of observations, a regularisation may be needed because of the errors that impoverish the information content of the observations, and because of the lack of observability of some regions of the source space. It is often said that these inverse problems are ill-posed.

It was shown in Winiarek et al. [2011], that even when the location of the source is well known (anticipated in the case of Fukushima Dai-ichi, or with delay in the case of the Chernobyl) so that only a temporal rate profile should be retrieved, and even when the observations are abundant, a background term is still necessary in a significant fraction of the cases. In the case of Chernobyl, where the location is supposed to be known in re-analysis, Bocquet [2012] has demonstrated that a problem without a properly defined background but with much more observations than source parameters can lead to aberrant total retrieved activity for the source term.

Therefore, it is often safer to use the objective function with a regularisation term:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\sigma}) =& \frac{1}{2}\left(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}\right)^{\mathrm{T}} \mathbf{R}^{-1} \left(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}\right) \\ &+ \frac{1}{2}\left(\boldsymbol{\sigma} - \boldsymbol{\sigma}_b\right)^{\mathrm{T}} \mathbf{B}^{-1} \left(\boldsymbol{\sigma} - \boldsymbol{\sigma}_b\right)\,,\end{aligned} \tag{4.3}$$

where $\boldsymbol{\sigma}_b$ is the first guess (or background), an estimation of the source before the observations are assimilated, and $\mathbf{B}$ is the background error covariance matrix. In the context of an accidental release, it is reasonable to assume $\boldsymbol{\sigma}_b = \mathbf{0}$ for the accidental source, since so little is known about it. In the case of noble gases, there could be significant diffuse natural (radon) or anthropogenic (xenon) emissions, that would have to be taken into account through an offset term in Eq. (4.1), or incorporated into the inverse modelling scheme. In that latter case, a non-zero diffuse background $\boldsymbol{\sigma}_b$ would be defined from their emission inventories.

Matrix $\mathbf{B}$ is a rather well studied object in meteorological and oceanographical data assimilation, even though its modelling is complex. In our context, $\mathbf{B}$ is very poorly known, since it is meant to measure our ignorance on the source term before the accident or the nuclear test, which is difficult to quantify. The $\mathbf{B}$ matrix related to noble gas with an estimated background which measures the errors in the inventory, may be better known. In the following, we are not considering such non-trivial background, and we will focus on the accidental release part. However the formalism used in this article can cope with more complex situations.

A posteriori parameter estimation techniques, such as L-curve, maximum-likelihood, generalised cross-validation [Vogel, 2002; Hansen, 2010], can efficiently help to assess the background term in an accidental context [Davoine and Bocquet, 2007; Krysta et al., 2008; Saide et al., 2011], where a single realisation of the set of observations is available (as opposed to

routine pollution). However it should be clear that the errors represented by $\mathbf{R}$ and $\mathbf{B}$ are very difficult to assess in this context.

In the absence of any constraint such as the positivity of $\boldsymbol{\sigma}$, the best linear unbiased estimator of the source is given by the argument of the minimum of Eq. (4.3)

$$\boldsymbol{\sigma}_a = \boldsymbol{\sigma}_b + \mathbf{B}\mathbf{H}^{\mathrm{T}} \left(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right)^{-1} \left(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_b\right) . \tag{4.4}$$

The uncertainty of this estimator is given by the analysis error covariance matrix

$$\mathbf{P}_a = \left(\mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right)^{-1} . \tag{4.5}$$

which is obtained as the inverse matrix of the Hessian of Eq. (4.3), which represents the precision matrix of the estimator. It is often equivalently rewritten as

$$\mathbf{P}_a = \mathbf{B} - \mathbf{B}\mathbf{H}^{\mathrm{T}} \left(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right)^{-1} \mathbf{H}\mathbf{B} , \tag{4.6}$$

which is to be used later.

More advanced methodologies that are able to handle the non-Gaussianity of errors, can lead to more sophisticated estimators of the a posteriori errors (see Bocquet et al. [2010] and references therein). However, second-order moments of the error distribution still provide an approximation of the posterior error statistics. In this case, $\mathbf{P}_a$ is approximately obtained as the inverse of the Hessian of the cost function,

4.2.2 Information content and DFS

After the analysis, a scalar residual posterior uncertainty is given by $\mathrm{Tr}(\mathbf{P}_a)$. The reduction of uncertainty in the data assimilation process can be measured by a related quantity: $\mathrm{Tr}(\mathbf{I}_N - \mathbf{P}_a\mathbf{B}^{-1})$, which identifies with the *degrees of freedom for the signal*, abbreviated DFS in the following [Rodgers, 2000]. The DFS are often used in the inversion of satelite-based instrument radiances. In our context, it measures the fractional number of observations that are effectively used in the inversion to retrieve the source. Explicitly, one has

$$\begin{aligned} \mathcal{J}_{\mathrm{DFS}} &= \mathrm{Tr}\left(\mathbf{I}_N - \mathbf{P}_a\mathbf{B}^{-1}\right) \\ &= \mathrm{Tr}\left(\mathbf{B}\mathbf{H}^{\mathrm{T}} \left(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right)^{-1} \mathbf{H}\right) . \end{aligned} \tag{4.7}$$

It is always lower or equal to the total number of observations d: $0 \leq \mathcal{J}_{\mathrm{DFS}} \leq d$.

As explained earlier, it is difficult to specify $\mathbf{R}$ and $\mathbf{B}$, especially in the retrieval of sources in atmospheric dispersion. Besides these matrices are context-dependent. In the absence of significant correlations in-between observation errors, and in-between background errors, they can be both chosen proportional to the identity matrix: $\mathbf{R} = \chi^2\mathbf{I}_d$ and $\mathbf{B} = m^2\mathbf{I}_N$. Yet, χ and especially m still need to be estimated. However, in this article, we are not interested in the precise values of χ and m. We are more interested in the degrees of freedom for the signal available for the inversion. They depend on the ratio χ/m as can be checked on Eq. (4.7). To some extent, reasoning in terms of DFS circumvents the necessity to reason on $\mathbf{R}$ and $\mathbf{B}$.

Using the results of inverse modelling of actual dispersion problems: ETEX, Chernobyl, Algeciras, or from the results of carbon dioxide inverse modelling [Krysta and Bocquet, 2007; Krysta et al., 2008; Wu et al., 2011], we have found that the DFS usually represents 5% to 15% for the total number of observation, for this kind of dispersion problem. In the following of this study, rather than specifying χ, m, or the ratio χ/m, we shall assume that when dealing with real observations, one should expect to reach a DFS of about 10% of the total number of observations. In the future, with the reduction of model errors, this fraction of the DFS may increase. However a strong reduction of the model or representativeness errors may not necessarily lead to a strong increase of the ratio $\rho = \mathrm{DFS}/d$, because of the ill-posed nature of dispersion.

4.2.3 Multiscale data assimilation

One usually considers a regular mesh, with grid-cells of constant size in one system of coordinates, to discretise the source $\boldsymbol{\sigma}$. However, adaptive grids can also be considered to model the transport of pollutant [Constantinescu et al., 2008], or to perform source inversion [Bocquet, 2009; Bocquet et al., 2011; Bocquet and Wu, 2011]. Such grids are relevant to atmospheric chemistry modelling because of the high heterogeneity of the emission fields. They are especially relevant in data assimilation for atmospheric dispersion when the observations are sparse, because the (adjoint) model can carry information from the observations in a very heterogeneous manner. We shall adopt such an adaptive grid formalism following the methodology developed in [Bocquet, 2009; Bocquet et al., 2011]. Details can be found in these references, and we shall focus here on what is necessary to interpret the results.

The activity concentrations of the numerical transport model are defined on, or interpolated to, a regular grid, which is the finest available grid in the rest of this article. In the case of the CTBT problematic, the finest grid will be lat-lon, with $N_x = 512$ and $N_y = 256$. In particular the Jacobian $\mathbf{H}$ computed with the numerical model, or possibly its adjoint, is defined in this grid. The background error covariance matrix $\mathbf{B}$ is defined in this grid too.

One can define a restriction operator that coarse-grains a source $\boldsymbol{\sigma}$ defined in the finest grid into a coarser $\boldsymbol{\sigma}_\omega$ defined in an adaptive grid ω with grid-cells of various sizes but all assembled from grid-cells of the finest regular grid. A prolongation operator refines a coarse $\boldsymbol{\sigma}_\omega$ defined in the adaptive grid ω into a source $\boldsymbol{\sigma}$ defined in the finest regular grid. Coarse-graining a vector $\boldsymbol{\sigma}$ defined in the finest grid, then refining the result to project back to the finest grid does not give $\boldsymbol{\sigma}$ back, because information is lost in the coarse-graining. Rather, it gives

$$\boldsymbol{\sigma} \longrightarrow (\mathbf{I}_{N_{\rm fg}} - \boldsymbol{\Pi}_\omega)\boldsymbol{\sigma}_b + \boldsymbol{\Pi}_\omega \boldsymbol{\sigma} \tag{4.8}$$

where $\boldsymbol{\Pi}_\omega$ is a projection operator that can be defined from the action of the restriction and the prolongation operators. $N_{\rm fg}$ is the number of grid-cells in the finest grid, so that $\mathbf{I}_{N_{\rm fg}}$ is the identify operator defined in the corresponding vector space. A Bayesian construction of the prolongation operator leads to a $\boldsymbol{\Pi}_\omega$ which is $\mathbf{B}-$symmetric: $\boldsymbol{\Pi}_\omega \mathbf{B} = \mathbf{B}\boldsymbol{\Pi}_\omega^{\rm T}$. In the accidental context, the assumption $\boldsymbol{\sigma}_b = \mathbf{0}$ sets the constant term in Eq. (4.8) to zero. A schematic representation of the action of $\boldsymbol{\Pi}_\omega$ is drawn in Fig. 4.1. The errors caused only by the aggregation of grid-cells can be formally computed [Bocquet et al., 2011]

$$\boldsymbol{\epsilon}_\omega = \mathbf{H}\left(\mathbf{I}_{N_{\rm fg}} - \boldsymbol{\Pi}_\omega\right)(\boldsymbol{\sigma} - \boldsymbol{\sigma}_b) \ . \tag{4.9}$$

Performing inverse modelling in the finest regular grid yields the DFS given by Eq. (4.7). Bocquet et al. [2011] have shown that performing inverse modelling in the adaptive grid ω yields the DFS

$$\mathcal{J}_{\rm DFS}^\omega = {\rm Tr}\left(\boldsymbol{\Pi}_\omega \mathbf{B}\mathbf{H}^{\rm T}\left(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\rm T}\right)^{-1}\mathbf{H}\right) \ . \tag{4.10}$$

This result assumes that representativeness errors, such as Eq. (4.9), are taken into account when changing resolution. The DFS can be used as a criterion to find the optimal adaptive grid given a fixed total number of grid-cells N. The algorithm to perform this optimisation is described in Bocquet [2009]; Bocquet et al. [2011]. In the context of atmospheric dispersion with ground-based point-wise observations, an optimal adaptive grid can deliver much more DFS than a regular grid with about the same number of grid-cells. The grid is usually refined close to the observation sites. It also depends on the dispersion itself and the meteorology. The size of a grid-cell offers a rigorous measure of the *resolution* defined by Rodgers [2000], that is to say the capacity to resolve a variable from the observations. As opposed to using the inverse of the diagonal entries of $\mathbf{P}_a$, this measure does not rely on any approximation. In practice,

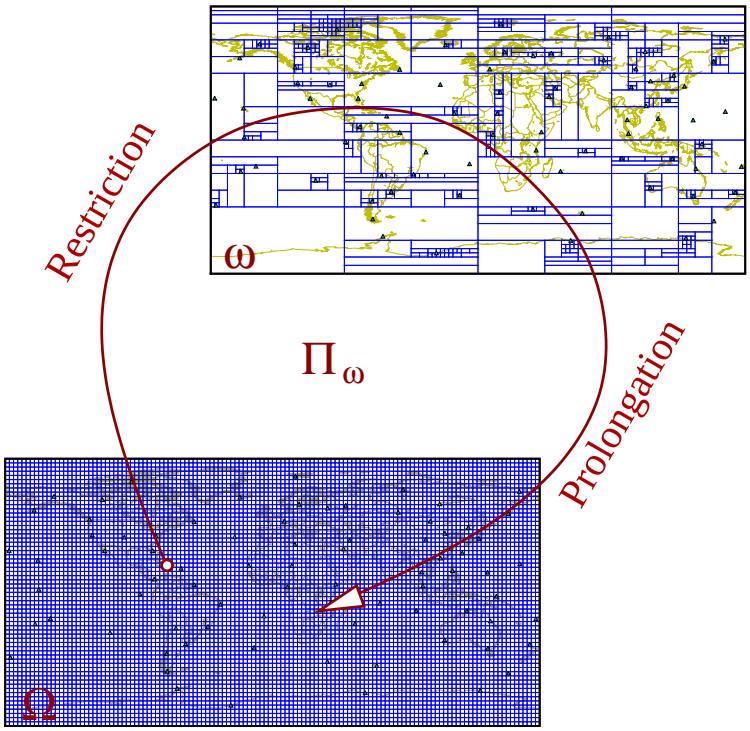


Figure 4.1: Schematic for the projector $\mathbf{\Pi}_\omega$ which operates in the finest regular grid cell.

if a location is encompassed in a large (respectively small) grid-cell, little (respectively much) information will be obtained at this point from inverse modelling.

In the regime where χ/m is high (large error limit), it is clear that the objective function Eq. (4.10) can be approximated by the simpler

$$\mathcal{J}^\omega_{\text{fisher}} = \text{Tr}\left(\mathbf{\Pi}_\omega \mathbf{B}\mathbf{H}^{\text{T}}\mathbf{R}^{-1}\mathbf{H}\right) , \tag{4.11}$$

called the Fisher criterion in Bocquet et al. [2011]. Although the value of Eq. (4.11) can be different from that of Eq. (4.10), it was observed that the optimal grids obtained with the two criteria are very similar in the large χ/m limit.

Practically, for real applications, the ratio $\text{DFS}/d \simeq 10\%$ corresponds to a large χ/m, so that the Fisher criterion can be used in place of the DFS criterion. Moreover, in that limit it can be shown that the optimal grid is the grid that minimises the aggregation errors. Indeed, from Eq. (4.9), and using the $\mathbf{B}-$symmetry of $\mathbf{\Pi}_\omega$, the aggregation error covariance matrix is

$$\mathbf{R}_\omega = \mathbf{H}\left(\mathbf{I}_{N_{\text{fg}}} - \mathbf{\Pi}_\omega\right)\mathbf{B}\mathbf{H}^{\text{T}} . \tag{4.12}$$

As a consequence, the normalised aggregation error can be assessed by

$$\begin{aligned}\text{Tr}\left(\mathbf{R}^{-1}\mathbf{R}_\omega\right) &= \text{Tr}\left(\mathbf{R}^{-1}\mathbf{H}\mathbf{B}\mathbf{H}^{\text{T}}\right) - \text{Tr}\left(\mathbf{R}^{-1}\mathbf{H}\mathbf{\Pi}_\omega\mathbf{B}\mathbf{H}^{\text{T}}\right)\\ &= \text{Tr}\left(\mathbf{B}\mathbf{H}^{\text{T}}\mathbf{R}^{-1}\mathbf{H}\right) - \text{Tr}\left(\mathbf{\Pi}_\omega\mathbf{B}\mathbf{H}^{\text{T}}\mathbf{R}^{-1}\mathbf{H}\right)\\ &= \mathcal{J}_{\text{fisher}} - \mathcal{J}^\omega_{\text{fisher}} .\end{aligned} \tag{4.13}$$

That is why the maximisation of $\mathcal{J}^\omega_{\text{fisher}}$ entails a minimisation of the aggregation errors.

The optimal adaptive grids that result from the optimisation of these criteria were shown to be numerically efficient to perform inverse modelling [Bocquet, 2009]. They have better data assimilation performance as compared to regular grids with the same number of grid-cells. Moreover, we have just shown that they entail little aggregation errors by construction. Therefore, building such grids could be used to perform inverse modelling of IMS data. However, in the following, we shall rather focus on the fact that these grids rigorously pinpoint well observed and poorly observed regions of the world for the purpose of inverse modelling.

4.3 Application to the IMS radionuclide network

4.3.1 Setup

The potential of the global IMS radionuclide network for data assimilation with an ATM is studied in this section, using the formalism recalled in Section 4.2. A drastically simplified version of the setup, with unrealistic physics and annual observations, was experimented in Bocquet and Wu [2011] as a proof of concept.

Among the 80 targeted stations, 79 have assigned locations, and we shall consider these 79 stations. The year 2009 is the focus of the study. As mentioned earlier, activity concentrations measurements are integrated over 24 hours. Therefore, $79 \times 365 = 28,835$ observations are considered. The Comprehensive Nuclear-Test-Ban Treaty Organisation has provided us with one year of influence functions (also known as adjoint solutions, or footprints, or retroplumes), attached to each one of these observations. They correspond to the rows of the Jacobian matrix $\mathbf{H}$ built over one year. Those influence functions have been generated using the Lagrangian ATM FLEXPART [Stohl et al., 2005], version 5 (with minor modifications by the CTBTO scientists) driven by ECMWF meteorological fields at a resolution of $1^\circ \times 1^\circ$. The tracer is completely inert: only transport is considered. Hence, these influence functions represent an upper bound of how far the influence of any radionuclide can reach. The temporal extend of each influence function is 14 days, with a time step of $\Delta t = 3$ hours.

Our goal is, given a fixed number of grid-cells N, to build the corresponding optimal adaptive grid for the IMS network. As explained in Bocquet [2009], the optimal grid can be chosen among a dictionary of adaptive grids. In this study, we shall choose the so-called dictionary of *tilings*: the grid-cells (*tiles*) are rectangles, and their zonal, meridional and even time lengths can be chosen independently. The adaptive grid in Fig. 4.1 is an example of a tiling.

With the multiscale formalism recalled earlier, it is possible to built a grid which is adaptive in space, but also in time (see the ETEX-I example of Bocquet [2009]). In this study we rather focus on a static grid, that would be optimal on average over the whole 2009 year. However, a simple average of $\mathbf{H}$ over the 365 days of the year is too naive an approach. Instead, it is necessary to average over the optimality criterion, which has a non-linear dependence in $\mathbf{H}$. In the following, we use $\mathcal{J}^\omega_{\mathrm{DFS}}$, or its limiting case $\mathcal{J}^\omega_{\mathrm{fisher}}$, where the non-linear dependence in $\mathbf{H}$ is obvious.

4.3.2 Daily-averaged criteria

More specifically, we look for optimal adaptive grids which are invariant by time translations of 24 hours, and whose grid-cell length in time is 24 hours. Therefore $\mathbf{\Pi}_\omega$ is invariant by

translations of 24 hours. The averaged criteria are

$$\langle \mathcal{J}^{\omega}_{\mathrm{DFS}} \rangle = \mathrm{Tr}\left(\mathbf{\Pi}_{\omega} \langle \mathbf{B}\mathbf{H}^{\mathrm{T}} \left(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} \right)^{-1} \mathbf{H} \rangle \right) , \tag{4.14}$$

$$\langle \mathcal{J}^{\omega}_{\mathrm{fisher}} \rangle = \mathrm{Tr}\left(\mathbf{\Pi}_{\omega} \langle \mathbf{B}\mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H} \rangle \right) . \tag{4.15}$$

The brackets $\langle \cdot \rangle$ represent the average over the 365 days of year 2009. For each one of the 365 contributions to the mean, one should identify the influence functions present in $\mathbf{H}$ that contribute to $\mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}\mathbf{H}$, or $\mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$. For each day t of the year, a source variable defined in a grid-cell can be causally connected through data assimilation to any of the 79 stations. However, it is causally connected to only 14 observations, at day $t + \tau$, with $\tau = 0, \ldots, 13$, per station, through 14, 14-day long, influence functions.

Hence, for a given day, the number of observations that are used in the computation of $\mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}\mathbf{H}$, or $\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ is $d = 14\times79 = 1106$. Then, the resulting matrices are averaged over the 365 days to obtain the criterion value: $\langle \mathcal{J}^{\omega}_{\mathrm{DFS}} \rangle$, and $\langle \mathcal{J}^{\omega}_{\mathrm{fisher}} \rangle$. The numerical parallelised computation of the this average matrix demands a 3-day run on a 12-core Intel Xeon machine. For the DFS criterion, $d = 1106$ represents the maximum possible DFS, since the DFS criterion is now averaged over 365 days. In the rest of the article, these averaged criteria $\langle \mathcal{J}^{\omega}_{\mathrm{DFS}} \rangle$, or $\langle \mathcal{J}^{\omega}_{\mathrm{fisher}} \rangle$ are used to determine time-invariant optimal adaptive grids.

4.3.3 Dependence of the DFS in the number of grid-cells

Each optimisation is performed at a given number of grid-cells N. Figure 4.2 shows the performance of the adaptive grids as compared to the regular grids at different resolution. The DFS criterion and the Fisher criterion are plotted as a function of the number N of tiles in the grid.

As mentioned in Section 4.2, we assume $\mathbf{R} = \chi^2\mathbf{I}_d$ and $\mathbf{B} = m^2\mathbf{I}_N$. Choosing a priori particular values for χ and m is difficult, and maybe even methodologically wrong since it was shown in Davoine and Bocquet [2007] that m should be determined a posteriori in such an accidental context. Instead, we choose the values of χ/m so as to match a given $\rho = \mathrm{DFS}/d$ ratio, which is more universal than the precise value of χ/m. In Fig. 4.2(b), we consider the cases where $\rho \simeq 10\%$, which is a good indication of the capability of current inverse modelling system in the accidental context with ground point-wise observations. In Fig. 4.2(c), we consider the case $\rho \simeq 90\%$, as an indication for distant future systems with very low errors (typically an error standard deviation 100 times smaller than in the current systems). Finally, in Fig. 4.2(a), we consider the limiting Fisher criterion case which corresponds to small ρ. This small ρ and conservative limit may be preferable, if one believes $\rho = 10\%$ is still too optimistic an assumption. As shown by Fig. 4.2, the gap between the optimal grid and a regular grid having the same number of grid-cells is increasing with the errors (instrumental, representativeness and model errors). This gives away an increase of heterogeneity of retrievals with the errors: information cannot propagate far from the network and help resolve source variables.

The fact that the curves in Fig. 4.2 are monotonically increasing functions of N has been proven in Bocquet et al. [2011]. However future complex inverse modelling experiments will deal with scale-dependent model error, or with models operating at different scales (Lagrangian at mesoscale and Eulerian at global scale, e.g. Rigby et al. [2011]). In that case, it is expected [Peylin et al., 2001; Bocquet et al., 2011] that a maximum DFS be reached which does not correspond to the finest regular grid.

In the following, the study is performed at finite N, i.e. a computationally affordable number of grid-cells $N \ll N_{\mathrm{fg}}$, where N_{fg} is the number of grid-cells in the finest grid. Besides, the qualitative results (interpretation of the adaptive grids) will essentially be insensitive to the

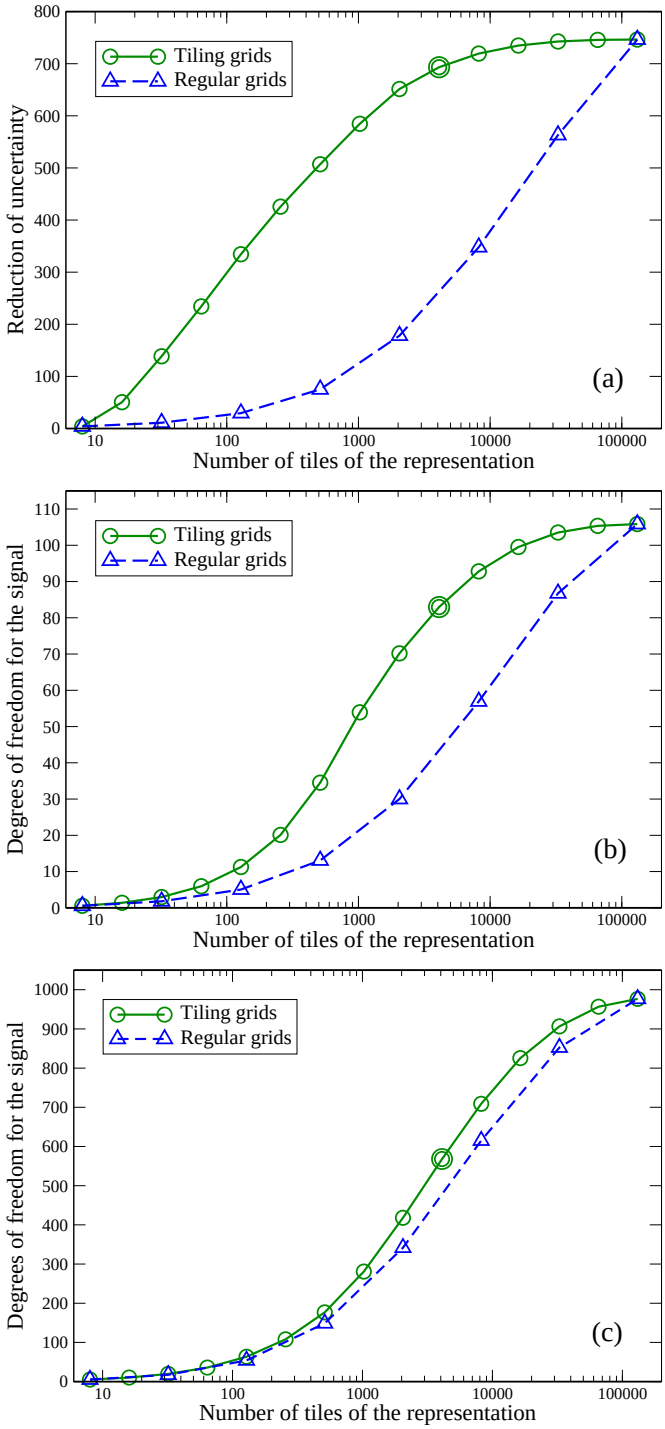


Figure 4.2: Fisher criterion (a), and degrees of freedom for the signal (b,c) of optimal tilings and regular grids against the number of grid-cells in the representation. Upper panel (a): χ/m is arbitrary (just a multiplicative factor). Middle panel (b): with $\chi/m = 100$. Lower panel (c): with $\chi/m = 1$. The illustrations of Fig. 4.3 correspond to the points indicated by double circles.

choice of N provided $N_{\rm cg} \ll N \ll N_{\rm fg}$, where $N_{\rm cg}$ is the number of grid-cells in the coarsest

regular grid ($N_{\mathrm{cg}} = 8$ in this study).

4.3.4 Interpretation of optimal grids

Typical optimal grids are displayed for $N = 4096$ in Fig. 4.3. Firstly, let us consider the

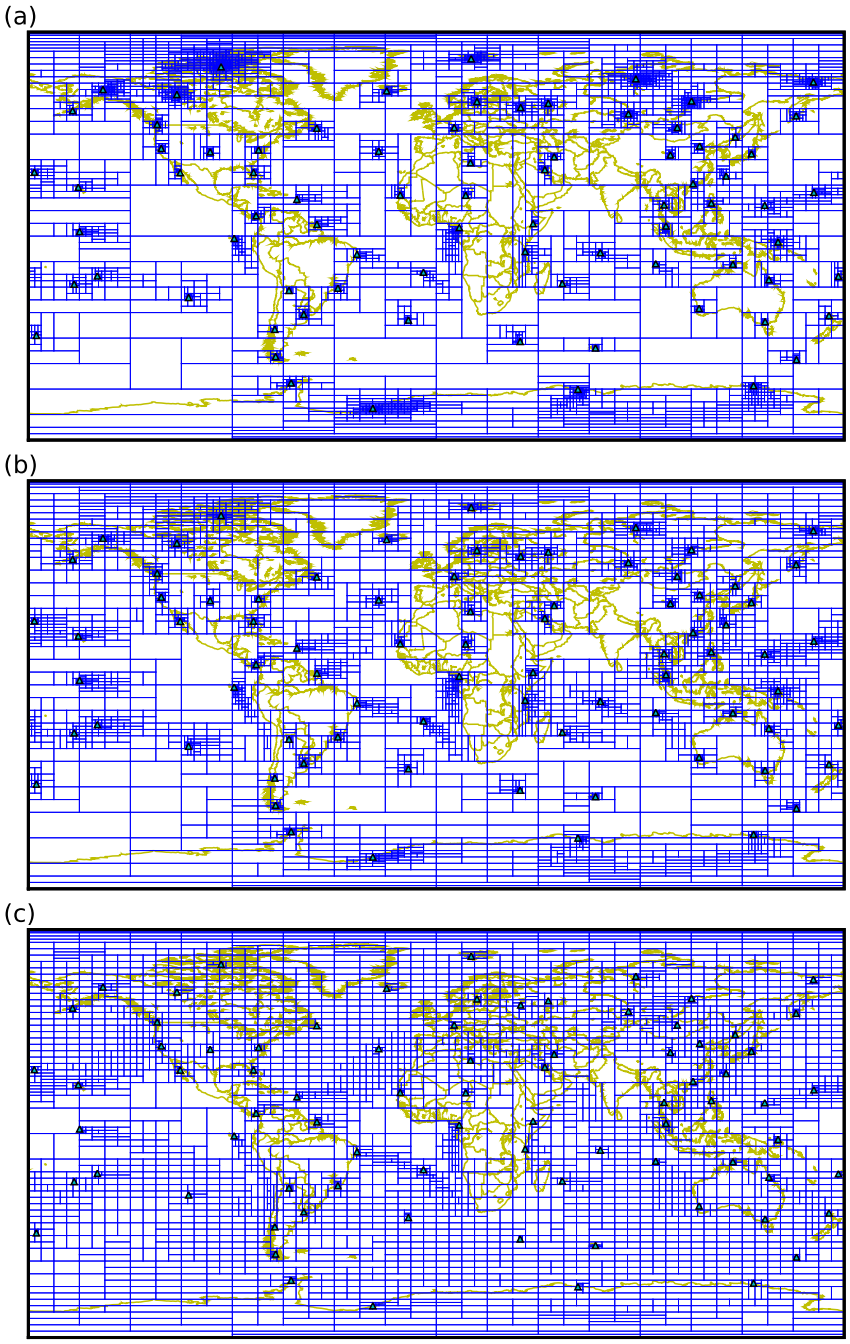


Figure 4.3: Optimal adaptive grids for $N = 4096$ grid-cells. Upper panel (a): from the Fisher criterion optimisation, (b): from the DFS optimisation in the realistic case $\chi/m = 100$. Lower panel (c): from the DFS optimisation with little error $\chi/m = 1$. The stations of the IMS radionuclide network are indicated by triangles.

Table 4.1: Distribution of the DFS over hemispheres and seasons.

DFS	Whole year 2009	Mar-Apr-May	Jun-Jul-Aug	Sep-Oct-Nov	Dec-Jan-Feb
NH	66.79	62.4	64.50	67.15	73.10
SH	39.06	41.49	33.75	35.88	45.23
Total	105.9	104.03	98.25	103.02	118.33

optimal grid based on the Fisher criterion. As explained earlier, it is impacted by the monitoring network distribution: the mesh is refined close to the stations indicating the ability of the data assimilation system to better resolve the source variables in these areas. It is also impacted by the meteorological climatology. For instance, in the polar regions, the information remains confined within the polar cells. As a result, stations in Antarctica do not significantly help in resolving variables over the Antarctic Ocean. In this high-error limit, the mesh is especially dense close to the observations: the information cannot propagate far from the stations.

Next, consider the *realistic* case where $\mathrm{DFS}/d \simeq 10\%$. Again, the grid is refined close to the observations site, but to a lesser extent. The impact of the trade winds is clear. Information is back-propagated from the tropical stations south-easterly in the South hemisphere, and north-easterly in the North hemisphere. Besides, since those winds are very directive, information cannot substantially reach the inter-tropical zone.

Finally, consider the case where model and representativeness errors are very small. The information can back-propagate much farther than in the previous cases. In particular, in the mid-latitude regions westerlies winds (maybe jets) efficiently back-propagate the information, so that the mesh is relatively even in these regions. In the tropics, the impact of the trade winds is even more obvious. Tropical regions that are not under direct observation are poorly resolved by inverse modelling.

In moderately large wind conditions, sea and land breezes may have an influence on the local climatological winds, near the shores. A clear impact on the optimal grids is the poor observability of Africa, even though 6 stations are installed on the continent. Besides, the Harmattan, which is a trade wind, leaves station RN13, Edea, Cameroon, with a poor visibility on the continent. In general, for stations well inside the continent the impact of an observation station is more isotropic, but also more short-ranged.

4.3.5 Distribution of the DFS over hemispheres and seasons

The degrees of freedom for the signal can be computed locally in the source space. To compute the DFS attached to a subset of source variables, it suffices to compute the corresponding subtrace in the DFS formula Eq. (4.7), that is to say a partial sum of the diagonal entries. We have computed the DFS for the northern and the southern hemispheres, as well as for the seasons, as defined by the four trimesters March-April-May, June-July-August, September-October-November, December-January-February. Because the length of these periods can slightly vary, their DFS are given as a mean and are therefore comparable. The results are reported in Tab. 4.1. On average, the DFS of the northern hemisphere captures about 63% of the total DFS, which is consistent with the fact that 48 stations out of 79 are in the northern hemisphere. Because of this unbalance, some seasonal effects become evident. Indeed the northern hemisphere winter shows a stronger DFS than in the summer. This might be explained by the stronger westerlies winds in the winter, that propagate tracers (and related information) farther away, as opposed to a more diffusive/stationary summer climatology.

4.3.6 Implication on the design of the network

Obviously this analysis has implications on the design of the network. The IMS radionuclide network has been evaluated and perhaps designed using detectability criteria. In the same context, other criteria could be based on the ability to map activity concentrations using the data available from the network and geostatistical techniques [Abida et al., 2008]. Our criteria are based on the ability of data assimilation to retrieve source parameters. As we pointed out, it is dependent on the instrumental error, on the representativeness errors, and especially on the modelling errors. However in all circumstances, some constant features have emerged and could help in the re-allocation of stations.

4.3.7 Noble gas network

We shall perform the same study but with the noble gas network which is a subnetwork of the IMS network, as stated in the treaty. Among the future noble gas 40 stations, 39 stations have a designated location, while the 40^{th} will be the currently unknown 35^{th} station of the 80-station radionuclide network. That is why we have chosen to perform the adaptive grid optimisation on this subset of 39 stations. The list of the stations can be found on the CTBTO website (`http://www.ctbto.org`) and are displayed in an interactive map (`http://www.ctbto.org/map`). Moreover, it is assumed that the measurement length are 24-hour long, while 12-hour long measurements are also performed for noble gases.

We assume the same χ/m ratio as for the realistic case of the full IMS network. The total number of observations is now $39 \times 365 = 14,235$. It leads to a similar ratio $\rho = \mathrm{DFS}/d \simeq 10\%$.

The behaviour of the DFS as a function of the number of grid-cells is plotted in Fig. 4.4. Even though it should represent a similar case to Fig. 4.2(b), the shape of the DFS curve stands

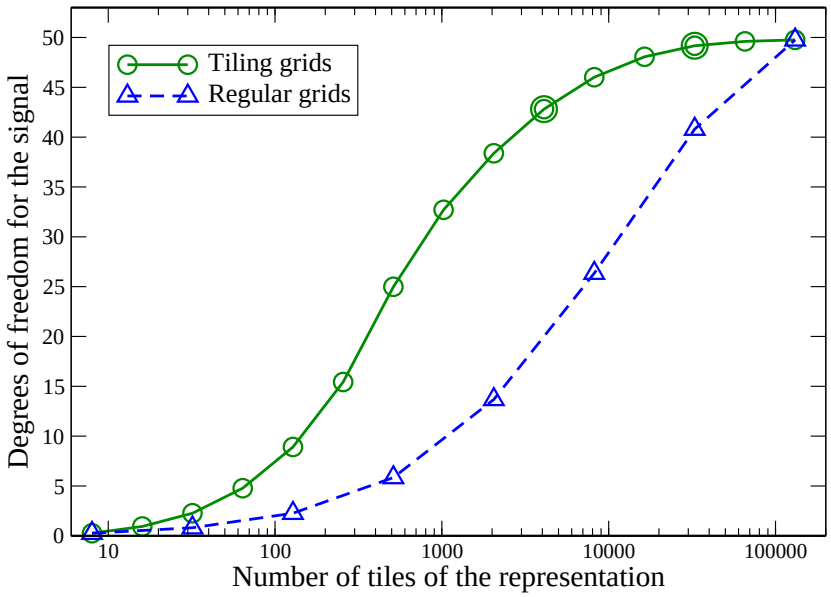


Figure 4.4: Degrees of freedom for the signal of optimal tilings and regular grids against the number of grid-cells in the representation, in the case of the noble gas subnetwork ($\chi/m = 100$). The illustrations of Fig. 4.5 correspond to the points indicated by double circles.

in between the Fig. 4.2(a) and Fig. 4.2(b) cases. The fact that only 39 stations are exploited

makes the globe significantly less observable by inverse modelling means. The information from the observations cannot reach remote areas, and this translates into a more rounded DFS curve, which makes optimal designed grids much more efficient than regular grid with the same number of grid-cells.

The optimal tilings for $N = 4096$ and $N = 32768$ are drawn in Fig. 4.5. The former allows a comparison with the full network case. The latter grid has a graphical interest since it underlies the poorly observed regions of the globe (clear/dark regions). In particular it is clear that the Pacific and the Intertropical Convergence Zone are much less observed than with the full network. The large cell over Antarctica of the first map should not be interpreted as a too significantly unobservable zone. Indeed, the genuine area of this cell is smaller than displayed in lat-lon coordinates. Specifically, the large cell over Antarctica has an area about three times smaller than the area of one of the two large grid-cells over the tropical Pacific.

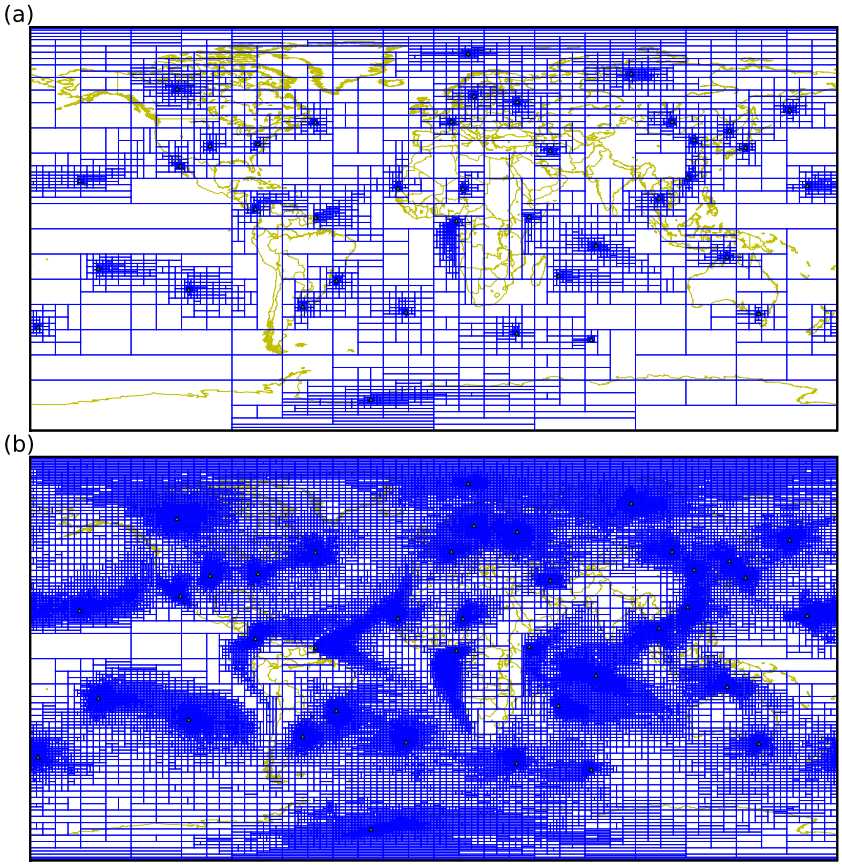


Figure 4.5: Optimal adaptive grids for the 39-station noble gas network, for (a): $N = 4096$ and (b): 32768 grid-cells, using $\chi/m = 100$. The 39 stations of the noble gas network are indicated by triangles.

4.3.8 Eulerian and Lagrangian models

As a final experiment, the inverse modelling potential of the IMS network was investigated in a limited-area domain (spanning [59°W-109°E]×[29°S-69°N]), using a Lagrangian ATM (FLEXPART as used by the CTBTO), and a regional Eulerian ATM (POLYPHEMUS/POLAIR3D, Quélo et al. [2007]). The limited area domain allowed us to use the regional model POLAIR3D,

but it also allowed to significantly reduce the computational cost, since only 18 stations of the IMS radionuclide network are considered in this domain. The influence functions obtained from both models over the year 2009 simulate an atmospheric inert tracer (such as xenon-133 but without decay). The magnitude of the errors chosen for the experiment corresponds to the realistic case, where $\mathrm{DFS}/d \simeq 10\%$. The maximal DFS (finest grid) for the Lagrangian model is about 21, while the maximal DFS for the Eulerian model is about 25. In the present context, the comparison of these two numbers should not be interpreted as a measure of the respective merit of two inverse modelling systems. Indeed, each model should in principle be endowed with its own magnitude of model errors in $\mathbf{R}$. However, a qualitative comparison can be made with the assumption that they both carry the same errors. The DFS curves of the two systems are plotted in Fig. (4.6). A difference between the two types of simulation, is that since the Lagrangian influence function are computed from the global footprints, they have re-entries of tracer within the domain. To minimise the differential impact of re-entries, only 18 stations in the domain, those within an angular distance of 10° away from the boundaries, have been kept. We have checked that, on average, these re-entries do not impact the following quantitative results.

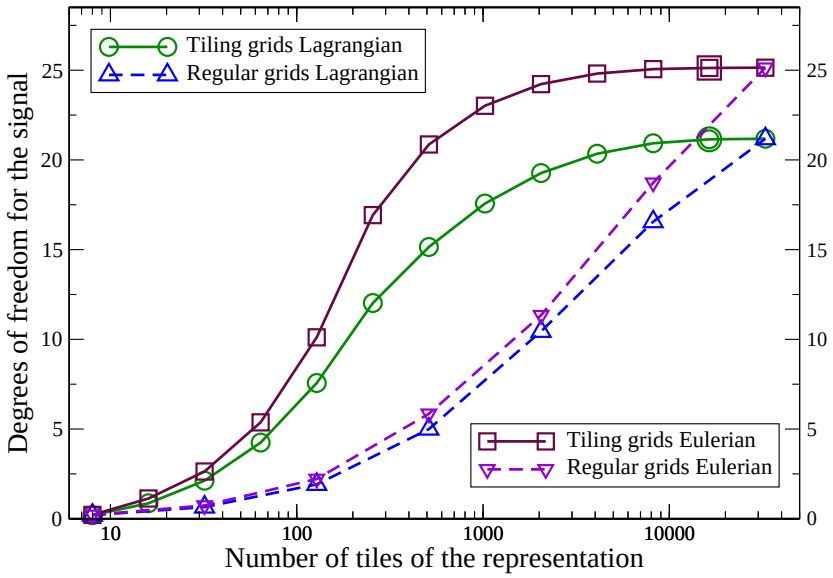


Figure 4.6: Degrees of freedom for the signal of optimal tilings and regular grids against the number of grid-cells in the representation, in the case of the limited area models ($\chi/m = 100$). The illustrations of Fig. 4.7 correspond to the points emphasised by a double circle and a double square.

There are differences between the two sets of curves. As the number of grid-cells N increases, the DFS of the Eulerian inverse modelling system increases more than those of the Lagrangian system. With the same $\mathbf{R}$ matrix, it does not imply that one system is better than the other, but that the physics of dispersion (of the tracer and the information) is somehow different. It shows that for a same N, the Eulerian grid is more heterogeneous than the Lagrangian grid. This means that, in the Eulerian case, the tracer extends less, leading to denser influence functions around the stations. This is confirmed by Fig. 4.7 where optimal grids with $N = 16384$ were chosen because it emphasises by contrast the poorly observed areas. It clearly shows that the Lagrangian grid is less dense around the stations and extend farther. But it also gives away sampling issues for the Lagrangian grid. By comparison of the two optimal grids,

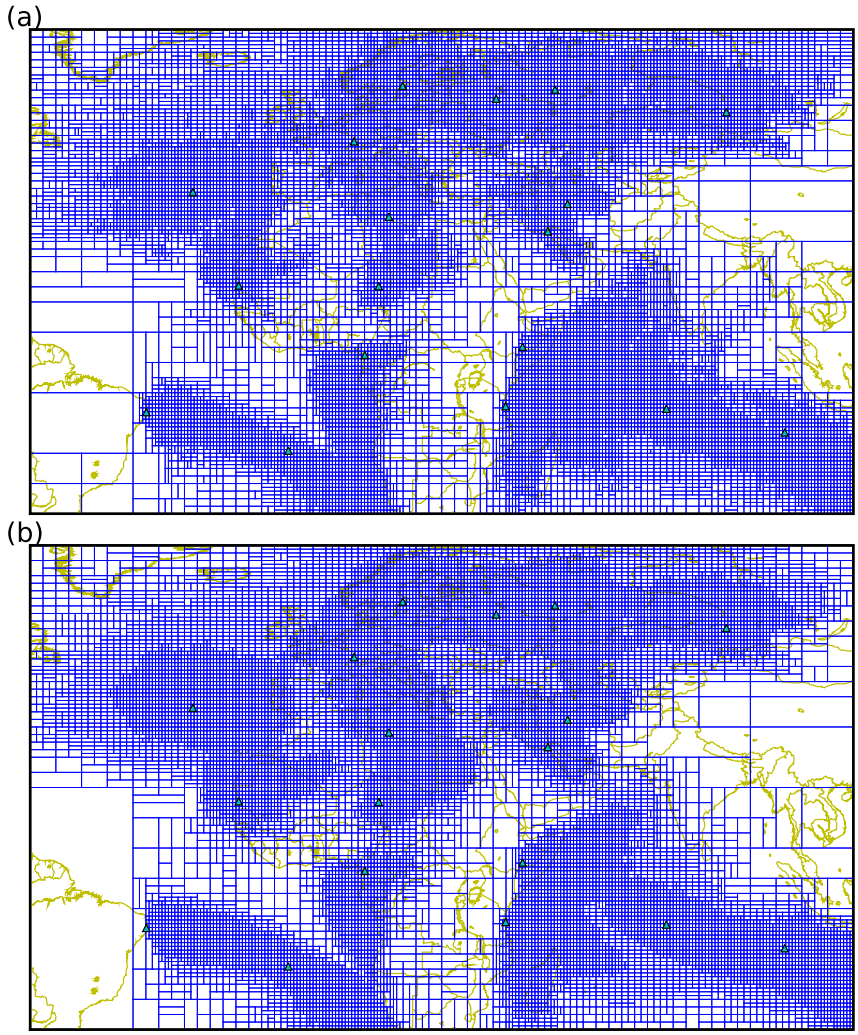


Figure 4.7: Optimal adaptive grids in the limited-area domain with $N = 16384$ grid-cells, computed from a Jacobian matrix $\mathbf{H}$ obtained from the influence function of (a): a Lagrangian model and (b): a Eulerian model, using $\chi/m = 100$. Within this domain only 18 stations away from the borders are considered. This helps to avoid re-entries of tracer in the Eulerian case.

it is clear that the Eulerian grid is more converged than the Lagrangian grid. We conjecture this is due to the fact that with 560×10^3 particles, a Lagrangian influence function may sample very well branches of the plume (better than the Eulerian model would do), but misses other possible branches of the plume. Hence, on the one hand, the Lagrangian inverse modelling system (with a limited number of particles) may capture information better than the Eulerian system does, but on the other hand it may miss information because of undersampling issues.

4.4 Conclusion

The potential of the International Monitoring System (IMS) radionuclide network has been evaluated for the inverse modelling of radionuclide releases: e.g. a nuclear explosion test or nuclear accidents like Chernobyl or Fukushima Dai-ichi, or the inverse modelling of the diffuse

sources of xenon radio-isotopes. We have proposed an evaluation methodology accounting for the performance of the inversion (or data assimilation) system. This methodology differs from the detection capability approach and the geostatistics approach in that: i) the degrees of freedom for the signal (DFS) is chosen to be the criterion that assesses the information gain from the observations to the whole domain through inversion; ii) optimal multiscale adaptive grids of sources are constructed by maximising the DFS criterion; and iii) the radionuclide network is evaluated by the spatial distribution of the grid-cells of the optimal adaptive grids.

For optimal grids, the inverse of the size of a grid cell measures the resolution defined as the capacity to locally resolve the source by the inversion system. Therefore the dense mesh indicates the regions where the source variables are well resolved. By contrast, for sparse mesh, the over-aggregations of the regular grid cells at finer scale result in high uncertainties of inverted sources.

We have constructed global optimal grids with the IMS radionuclide network for its evaluation. The influence functions, which relate the observations with the sources, were generated using the Lagrangian transport model FLEXPART driven by ECMWF meteorological fields at a resolution of $1^\circ \times 1^\circ$ over the 365 days of year 2009. The ratio ρ between the DFS and the total number of observations was used to control different error levels in the inversion, i.e. the error of a priori sources and the observational error that encapsulates the instrumental error, the representativity error and the transport model error.

Grid optimisations have been performed for three cases with $\rho \simeq 0$ (very large observational error), $\rho \simeq 10\%$ (realistic observational error), and $\rho \simeq 90\%$ (accurate transport) respectively. Some stable spatial patterns have emerged in the optimal grids with these different settings. In all cases, the trade winds carry information towards the Intertropical Convergence Zone along straight paths, leaving large unobserved areas in the tropics. For the case of large observational errors, the optimal grid is very heterogeneous. The mesh is dense close to most observation sites. The information propagation is not obvious except for the polar areas due to the impact of the polar vortices. For realistic observational errors, there are still many areas away from the information propagation path, e.g. the African continent.

When accurate atmospheric transport representation is assumed, the optimal grids become more uniformly distributed, especially for the mid-latitude regions where westerlies winds prevail. Such coverage is desirable. However, we do not believe current state-of-the-art ATM can reach that accuracy level. Moreover, even with high accuracy and a good coverage of the globe on average, the tropics will remain difficult to probe with inverse modelling.

The results obtained in this study can also serve as a basis for reallocation or installation of stations using the size of the adaptive grid-cells as an indication. It would be interesting to compare the results of this approach with the results based on the detectability criterion.

Chapter 5

Estimation of volatile organic compound emissions for Europe using data assimilation

Summary

The emission of volatile organic compounds (VOCs) over western Europe for the year 2005 are estimated via inverse modelling, by assimilation of in situ observations of concentration and compared to a standard emission inventory. The study focuses on fifteen VOC species: five aromatics, six alkanes, two alkenes, one alkyne and one biogenic diene. The inversion relies on a validated fast adjoint of the chemistry transport model used to simulate the fate and transport of these VOCs. The assimilated ground-based measurements over Europe are provided by the European Monitoring and Evaluation Programme (EMEP) network. The background emissions errors and the prior observational errors are estimated by maximum likelihood approaches. The positivity assumptions on the VOC emission fluxes is pivotal for a successful inversion and this maximum likelihood approach consistently accounts for the positivity of the fluxes. For most species, the retrieval leads to a significant reduction of the bias, which underlines the misfit between the standard inventories and the observed concentrations. The results are validated through a forecast test and a cross-validation test. It is shown that the statistically consistent non-Gaussian approach based on a reliable estimation of the errors offers the best performance. The efficiency in correcting the inventory depends on the lifetime of the VOCs. In particular, it is shown that the use of in-situ observations using a sparse monitoring network to estimate emissions of isoprene is inadequate because its short chemical lifetime significantly limits the spatial radius of influence of the monitoring data. For species with longer lifetime (a few days), successful, albeit partial, emission corrections can reach regions hundreds of kilometres away from the stations. Domainwide corrections of the emissions inventories of some VOCs are significant, with underestimations on order of a factor of two of propane, ethane, ethylene and acetylene.

Contents

5.1 Introduction

Volatile organic compounds (VOCs) are of particular environmental concern because they are precursors of secondary pollutants, such as ozone and fine particulate matter ($PM_{2.5}$), and some VOCs are pollutants in their own right due to their adverse carcinogenic and/or non-carcinogenic health effects. Therefore, it is essential to have accurate emission inventories of VOCs to conduct air quality modelling studies for the development of optimal emission control strategies and for air quality forecasting as well as to follow their temporal emission trends over the years as emission control strategies get implemented. The large number of emission sources, both anthropogenic and biogenic, and processes leading to those emissions (combustion, evaporation, vegetation metabolism) make the development of accurate VOC emission inventories difficult. Furthermore, VOC emissions cannot be derived from mass balances and they must be obtained from experimental measurements conducted at the source of the emissions. Also, emission testing is costly and some emission factors are developed for total VOCs rather than for individual VOCs. Therefore, a chemical speciation must be applied to total VOC emission factors using speciation data that are limited and uncertain. Although uncertainties in anthropogenic emissions have been reduced over the years as a result of better characterisation of major emission sources, uncertainties still remain. Furthermore, large uncertainties are associated with biogenic emissions due to the difficulty of estimating the meteorology-dependent emission rates for a large number of vegetation species as well as characterising the land-use/land-cover of the area of interest. Therefore, several approaches have been used to evaluate the accuracy of VOC emission inventories and, if appropriate, apply some correction.

Uncertainties in emission inventories have been estimated, for example, by comparing ambient air measurements in tunnels with vehicle exhaust emission estimates [e.g., Staehelin et al., 1998; Sawyer et al., 2000; Touaty and Bonsang, 2000; Stemmler et al., 2005; Ho et al., 2007]. However, such experiments characterise only one source category (on-road traffic) and focus on a single location and time period. Satellite measurements have been used to assess VOC emissions, but such techniques are limited by the number of VOCs that can be measured via satellite-borne instruments [Vijayaraghavan et al., 2008] and to areas that are specific to a major source category, e.g., the use of formaldehyde (an oxidation product of isoprene) to estimate isoprene emissions in remote areas where biogenic emissions dominate [e.g., Shim et al., 2005; Fu et al., 2007; Millet et al., 2008; Dufour et al., 2009]. Measurements of VOC concentrations aloft have also been used to estimate fluxes of VOCs originating from an area [e.g., Hopkins et al., 2009]; however, there are uncertainties associated with the mass balance method used to estimate the atmospheric transport flux and relate it to an emission flux. Furthermore, such an approach is limited to the estimation of an emission flux for a given region over a given period. Comparisons of the output concentrations of air quality model simulations with observations aloft [e.g., Xiao et al., 2008] or at ground level [e.g., Harley and Cass, 1995] provide some estimates of uncertainties in VOC emissions; however, such information is typically also limited to a specific region and period for which those measurements are available. Inverse modelling has been conducted using ground-level ambient concentrations to estimate emission inventories for some air pollutants, but such studies [e.g., Quélo et al., 2005; Elbern et al., 2007; Koohkan and Bocquet, 2012] have focused so far on regulated air pollutants with ambient concentration data available from routine monitoring networks and have not yet included VOCs.

It is, therefore, of great interest to investigate the current status of VOC emissions using an approach that provides information for several major VOCs with spatial distribution over a large domain and for a long period of time. To that end, we present here the first assessment of the emissions of several VOCs measured routinely at several remote sites, that covers all VOC sources over western and central Europe for an entire year.

In Section 5.2, the chemical transport model (CTM) used for modelling the VOCs is introduced and its reduced counterpart is described. The source-receptor relationship is built using an approximate adjoint model which is validated. The control variables (i.e., the emission fluxes), and the inversion modelling method are described. The method to estimate the so-called hyperparameters that parametrise the prior error statistics is introduced. In Section 5.3, the setup of the numerical experiments is described. Details about the observations set and the first guess inventory are provided. The optimal values of the hyperparameters to be used in the inversions are computed. In Section 5.4, the results of the inversions are presented and discussed. A forecast test and a cross-validation test are provided to validate the corrected emission fluxes using independent observations. Conclusions are presented in Section 5.5.

5.2 Methodology

5.2.1 Full chemical transport model and reduced VOC model

The chemical transport model (CTM) POLAIR3D [Sartelet et al., 2007] of the POLYPHEMUS air quality modelling system [Mallet et al., 2007] is chosen to model the atmospheric concentrations of chemical species. The numerical discretisation of the model for chemistry and transport, based on a first order time splitting algorithm, can be summarised as follows:

$$\boldsymbol{c}^{k+1} = \mathcal{X}_k \left(\boldsymbol{M}_k(\boldsymbol{c}^k) \right) + \boldsymbol{\Delta} t \, \mathbf{e}_{k+1} \tag{5.1}$$

where,

- $\boldsymbol{c}^k$ is the field of the concentrations of all simulated species at time step k.
- $\boldsymbol{M}_k$ is the linear advection-diffusion operator. It also includes the deposition processes.
- $\mathcal{X}_k$ is the chemical reaction operator
- $\mathbf{e}_k$ is the emission field at time step k.

Table 5.1 lists the fifteen VOC species for which experimental measurements are available from the European Monitoring and Evaluation Programme (EMEP) database. In order to simulate the concentrations of these species, the RACM 2 (Regional Atmospheric Chemistry Mechanism, version 2) chemical kinetic mechanism [Goliff and Stockwell, 2008] is used within the CTM [Kim et al., 2009]. The chemical reactions considered for these species and their typical lifetime are presented in Table 5.1. After undergoing oxidation reactions, these fifteen primary species result in secondary species. The latter are not presented in Table 5.1 because they are not relevant to our study: they are not measured in the EMEP network and, therefore, cannot be assimilated; they are, however, included in RACM 2 either explicitly or via surrogate species. Some of the primary VOC species (isoprene, acetylene, ethane, ethylene, benzene) are treated explicitly in RACM 2. The others are represented through a lumped molecule approach and, therefore, need a specific treatment to be followed separately. They are added as explicit species with their own oxidation reactions written in a way that does not affect RACM 2.

This chemical mechanism involves more than three hundred reactions that result in non-linear interactions among the chemical species. As a result, computational burden of inverse modelling studies is very large. To address this issue, we developed a reduced chemical mechanism, denoted $\boldsymbol{X}_k$, which uses the concentration fields of the oxidants, hydroxyl radicals (OH), ozone (O_3) and nitrate radicals (NO_3), provided as external data. The oxidant concentration fields are pre-computed with RACM 2 and used later in the reduced mechanism. This approximation makes sense if $\delta\mathcal{X}_k(c^k) = \mathcal{X}_k(c^k) - \boldsymbol{X}_k(c^k)$ is small with respect to $\boldsymbol{X}_k(c^k)$. The validation of this approximations is checked a posteriori in Section 5.4.1.1. When replacing $\mathcal{X}_k$ by $\boldsymbol{X}_k$, Eq. (5.1) becomes linear with respect to the emission fields and the computational cost of inverse modelling becomes manageable.

5.2.2 The source receptor model

The source-receptor model provides the relationship between the emissions and the observations. For species s, this can be written as follows:

$$\boldsymbol{\mu}^s = \mathcal{H}^s\mathbf{e}^s + \boldsymbol{\lambda}^s + \boldsymbol{\epsilon}^s \tag{5.2}$$

where $\boldsymbol{\mu}^s \in \mathbb{R}^{d_s}$ represents the vector of the observations (d_s is the number of observations for species s). $\mathcal{H}^s$ is the Jacobian operator with respect to $\mathbf{e}^s$ and $\mathbf{e}^s = (\mathbf{e}^s_0, \mathbf{e}^s_1, \ldots, \mathbf{e}^s_k, \ldots \mathbf{e}^s_{N_t}) \in \mathbb{R}^E$ defines the hourly and spatially discretised emission vector, where $E = N_t \times N_x \times N_y \times N_z$. N_t is the total number of time steps, and N_x, N_y, N_z are the total number of elements (grid cells) in the x, y and z directions. $\boldsymbol{\lambda}^s \in \mathbb{R}^{d_s}$ is the vector of the concentrations induced by the initial and the boundary conditions for species s. If $\boldsymbol{\mu}^{s,i_k}$is the observation of species s at time t_k at station i, $\boldsymbol{\lambda}^s_{i_k}$ is the concentration at the same time and location, computed with the full CTM, i.e., Eq. (5.1), with $\mathbf{e}^s = \mathbf{0}$. The vector $\boldsymbol{\epsilon}^s$ represents the errors: representativeness error, model error and instrumental error of the observations.

The Jacobian operator $\mathcal{H}^s$ can be built following two different methods. The first method consists in using the CTM. Let us assume that $\mathbf{e}^s = \delta_{l,h,k'}$ is the unity source at the surface coordinate $l \in [1; N_x \times N_y]$, altitude $h \in [1; N_z]$ and time $k' \in [1; N_t]$, and equals zero

Table 5.1: The volatile organic compounds, their (indicative) lifetime and reactions.

	species	symbol	lifetime (days)	reactions
	isoprene	ISO	0.07	ISO + OH →
				ISO+ O_3 →
				ISO + NO_3 →
RACM 2	acetylene	ACE	110	ACE + OH →
reactions	ethane	C_2H_6	60	C_2H_6 + OH →
	ethylene	C_2H_4	1.45	C_2H_4 + OH →
				C_2H_4 + O_3 →
				C_2H_4 + NO_3 →
	benzene	BEN	11	BEN + OH →
	propane	C_3H_8	14	C_3H_8 + OH →
	n-butane	NBUT	7	NBUT + OH →
	isobutane	IBUT	7.5	IBUT + OH →
	n-pentane	NPEN	5	NPEN + OH →
additional	isopentane	IPEN	4	IPEN + OH →
reactions	propene	C_3H_6	0.625	C_3H_6 + OH →
to RACM2				C_3H_6 + O_3 →
				C_3H_6 + NO_3 →
	toluene	TOLU	2.4	TOLU + OH →
	o-xylene	OXYL	1.1	OXYL + OH →
	m-xylene*	MXYL	0.625	MXYL + OH →
	p-xylene*	PXYL	1.05	PXYL + OH →

$\star$ measured jointly in the EMEP monitoring network and represented with the symbol MPXYL.

anywhere else. The CTM simulated concentration for clean air boundary and initial conditions at time t_k and station i with this source term is stored in $\mathcal{H}_{l,h}^{s,i_{k'}}$. In order to compute the $\mathcal{H}^s$ operator with this method, the CTM model needs to be run E times, which is computationally intensive. That is why this method is usually restricted to point-wise emission sources.

The second method consists in using the adjoint model of the CTM [e.g., Roustan and Bocquet, 2006b]. For a monitoring site i at time t_k, using the linearity of the CTM for the VOCs, the adjoint solution can be written as follows:

$$\boldsymbol{\phi}_i^k = \boldsymbol{M}_k^\dagger \left(\boldsymbol{X}_k^\dagger (\boldsymbol{\phi}_i^{k+1}) \right) + \boldsymbol{\Delta} t \pi_k^i \tag{5.3}$$

where, $\pi_k^i = \delta_{i,k}$ is the *sampling function* that represents the concentration measurement at station i and time t_k, $\boldsymbol{X}_k^\dagger$ is the adjoint of $\boldsymbol{X}_k$, and $\boldsymbol{M}_k^\dagger$ is the adjoint of $\boldsymbol{M}_k$. At the final time N_t, $\boldsymbol{\phi}_i^{N_t}$ is chosen to be $\mathbf{0}$. The adjoint model is also computed for clean air boundary conditions. Then, the Jacobian matrix is given by $\mathcal{H}_{l,h}^{s,i_{k'}} = [\boldsymbol{\phi}_{i,s}^{k'}]_{l,h}$. The adjoint model is run d times ($d \leq \sum_s d_s$). This method is of great interest to reduce the computational time when the problem is ill-posed ($d \ll E$). It is appropriate for estimating the emissions originating from spatially distributed sources.

The Jacobian matrix of the present study is computed row by row, that is using adjoint

solutions.

5.2.3 Control space

In order to reduce the dimension of the control space, that is the space of the fluxes to be estimated via inverse modelling, we introduce a relation between the effective control variables $\boldsymbol{\alpha}^s$ and the emission $\mathbf{e}^s$:

$$[\mathbf{e}_k^s]_{l,h} = [\boldsymbol{\alpha}^s]_l [\mathbf{e}_k^{s,b}]_{l,h} \,. \tag{5.4}$$

In this equation, $\mathbf{e}^{s,b}$ is the a priori (first guess or background) vector of emission for species s. Obviously the first guess value of the scaling factors $\boldsymbol{\alpha}^s$, is $\boldsymbol{\alpha}_b^s = \mathbf{1}$, where $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}}$. Indices k, l, h are respectively related to the time sequence, the horizontal space grid, and the vertical grid. The $\boldsymbol{\alpha}^s$ factors, rather than the full 3D time-dependent emission fields of Eq. (5.1), will be optimised. This choice of control parameters, which is only a function of l, implies that the correction of the emission fluxes is spatially distributed in the horizontal directions but that the vertical and the temporal distribution of the emission fluxes are not modified by the data assimilation analysis.

5.2.4 Inversion method

Combining Eq. (5.2) with Eq. (5.4), one obtains

$$\boldsymbol{\mu}^s = \mathbf{H}^s \boldsymbol{\alpha}^s + \boldsymbol{\lambda}^s + \boldsymbol{\epsilon}^s \tag{5.5}$$

where $\mathbf{H}^s$ is the Jacobian matrix that relates $\boldsymbol{\mu}^s$ to $\boldsymbol{\alpha}^s$:

$$\mathbf{H}_{i_k,l}^s = \sum_k [\boldsymbol{\phi}_{s,i}^k]_l [\mathbf{e}_k^{s,b}]_l \,. \tag{5.6}$$

In order to optimise the $\boldsymbol{\alpha}$ parameters, the following objective function with a regularisation term is used:

$$\begin{aligned} \mathcal{L}_s(\boldsymbol{\alpha}^s) \;=\; & \frac{1}{2} \left(\boldsymbol{\mu}^s - \mathbf{H}^s \boldsymbol{\alpha}^s - \boldsymbol{\lambda}^s\right)^{\mathrm{T}} \mathbf{R}_s^{-1} \left(\boldsymbol{\mu}_s - \mathbf{H}^s \boldsymbol{\alpha}^s - \boldsymbol{\lambda}^s\right) \\ & + \frac{1}{2} \left(\boldsymbol{\alpha}^s - \mathbf{1}\right)^{\mathrm{T}} \mathbf{B}_s^{-1} \left(\boldsymbol{\alpha}^s - \mathbf{1}\right) \,. \end{aligned} \tag{5.7}$$

The vector $\boldsymbol{\alpha}_b^s = \mathbf{1} = (1, \ldots, 1)^{\mathrm{T}}$ is the first guess of $\boldsymbol{\alpha}^s$. $\mathbf{R}_s$ is the observation error covariance matrix. $\mathbf{B}_s$ is the background error covariance matrix. For each species, these two matrices are both chosen to be diagonal with uniform variances, that is, $\mathbf{B}_s = m_s^2 \mathbf{I}_{N_x \times N_y}$, $\mathbf{R}_s = r_s^2 \mathbf{I}_{d_s}$. These statistical assumptions imply that we neglect any spatial and temporal correlations between grid cells in the errors. The anthropogenic emissions of VOCs are not expected to induce long-range correlation in the errors. However, potential important reasons for this hypothesis to fail are when the biogenic VOC emissions have correlated errors due for some VOCs common sources and similar emission model formulations, or when transport model errors induce temporal correlation in the error covariance matrix. That is why our diagonal assumption is an approximation.

In the following, two solutions of Eq. (5.7) are considered and compared. The first one assumes that the errors are Gaussian-distributed and the analysed parameters $\boldsymbol{\alpha}_a^s$ are given by the Best Linear Unbiased Estimator (BLUE):

$$\boldsymbol{\alpha}_a^s = \mathbf{1} + \mathbf{K}^s \left(\boldsymbol{\mu}^s - \mathbf{H}^s \mathbf{1} - \boldsymbol{\lambda}^s\right) \tag{5.8}$$

where $\mathbf{K}^s$ is the gain matrix:

$$\mathbf{K}^s = \mathbf{B}_s \mathbf{H}^{s\mathrm{T}} \left(\mathbf{R}_s + \mathbf{H}^s \mathbf{B}_s \mathbf{H}^{s\mathrm{T}} \right)^{-1} \tag{5.9}$$

The second solution of Eq. (5.7) is obtained assuming a truncated Gaussian distribution for the background error statistics, so that $\boldsymbol{\alpha}$ is optimised under a positivity constraint of each one of its entry $[\boldsymbol{\alpha}^s]_l$. As opposed to the Gaussian case, the retrieved scaling factors $[\boldsymbol{\alpha}^s]_l$ cannot be negative.

5.2.5 Estimation of hyperparameters

The parameters of the prior statistics, such as r_s and m_s, usually coined *hyperparameters*, often need to be estimated because their first guess is usually inaccurate, while the dependence of the retrieval on the hyperparameters can be dramatic [Davoine and Bocquet, 2007].

The estimation method for the hyperparameters depends on the statistical assumptions underlying Eq. (5.7). In the first case, the error $\boldsymbol{\epsilon}^s$ (in Eq. (5.5)) is assumed to be Gaussian-distributed, $\boldsymbol{\epsilon}^s \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_s)$. The same assumption applies to the control parameters: $\boldsymbol{\alpha}^s \sim \mathcal{N}(\mathbf{1}, \mathbf{B}_s)$. The probability density function (pdf) of the observation vector can be computed as follows:

$$p(\boldsymbol{\mu}^s | r_s, m_s) = \int p(\boldsymbol{\mu}^s | \boldsymbol{\alpha}^s, r_s, m_s) p(\boldsymbol{\alpha}^s | r_s, m_s) \mathrm{d}\boldsymbol{\alpha}^s = \int p(\boldsymbol{\epsilon}^s | r_s) p(\boldsymbol{\alpha}^s | m_s) \mathrm{d}\boldsymbol{\alpha}^s \tag{5.10}$$

or analytically,

$$p(\boldsymbol{\mu}^s | r_s, m_s) = \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}^s - \mathbf{H}^s\mathbf{1} - \lambda^s)^{\mathrm{T}} (\mathbf{R}_s + \mathbf{H}^s \mathbf{B}_s \mathbf{H}^{s\mathrm{T}})^{-1} (\boldsymbol{\mu}^s - \mathbf{H}^s\mathbf{1} - \lambda^s)}}{\sqrt{(2\pi)^{d_s} |\mathbf{R}_s + \mathbf{H}^s \mathbf{B}_s \mathbf{H}^{s\mathrm{T}}|}} \,. \tag{5.11}$$

This pdf is also proportional to the likelihood of r_s and m_s. In order to estimate the hyperparameters r_s and m_s, Desroziers and Ivanov [2001] suggested an iterative method to converge towards a fixed point. Chapnik et al. [2006] showed that this approach converges to one local maximum of the likelihood. The maximisation of $\log\left(p(\boldsymbol{\mu}^s | r_s, m_s)\right)$ with respect to the two hyperparameters gives the stationary conditions:

$$r_s^2 = \frac{\| \, \mu^s - \mathbf{H}^s \boldsymbol{\alpha}_a^s \, \|^2}{\mathrm{Tr}\left(\mathbf{I}_{d_s} - \mathbf{H}^s \mathbf{K}^s\right)} \,, \qquad m_s^2 = \frac{\| \, \boldsymbol{\alpha}_a^s - \mathbf{1} \, \|^2}{\mathrm{Tr}\left(\mathbf{H}^s \mathbf{K}^s\right)} \tag{5.12}$$

where $\| \cdot \|$ is the Euclidean norm.

However, the Desroziers method relies on Gaussian assumptions, and, for the sake of consistency, one needs another approach to compute the likelihood under the truncated Gaussian assumption [Winiarek et al., 2012]. In this case, the prior on the scaling factors is:

$$p(\boldsymbol{\alpha}^s) = \frac{e^{-\frac{1}{2}(\boldsymbol{\alpha}^s - \mathbf{1})^{\mathrm{T}} \mathbf{B}_s^{-1} (\boldsymbol{\alpha}^s - \mathbf{1})}}{\sqrt{(2\pi)^{N_x \times N_y} |\mathbf{B}_s|} \left(1 - \Phi_{\mathbf{1}, \mathbf{B}_s}(\mathbf{0})\right)} \mathbb{I}_{\boldsymbol{\alpha}^s \geq \mathbf{0}} \tag{5.13}$$

where $\Phi_{\mathbf{1}, \mathbf{B}_s}(\mathbf{0})$ is the Gaussian cumulative density function (cdf) of $\mathcal{N}(\mathbf{1}, \mathbf{B}_s)$. $\mathbb{I}_{\boldsymbol{\alpha}^s \geq \mathbf{0}}$ is a function equal to unity when $[\boldsymbol{\alpha}^s]_l \geq 0$ for each l, and equal to zero otherwise. This pdf is referred to as $\mathcal{N}(\mathbf{1}, \mathbf{B}_s, \mathbf{0})$ or the truncated Gaussian distribution. From Eq. (5.10) and Eq. (5.13), one can derive:

$$\begin{aligned} p(\boldsymbol{\mu}^s | r_s^+, m_s^+) \quad = \quad & \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}^s - \mathbf{H}^s\mathbf{1} - \lambda^s)^{\mathrm{T}} (\mathbf{R}_s + \mathbf{H}^s \mathbf{B}_s \mathbf{H}^{s\mathrm{T}})^{-1} (\boldsymbol{\mu}^s - \mathbf{H}^s\mathbf{1} - \lambda^s)}}{\sqrt{(2\pi)^{d_s} |\mathbf{R}_s + \mathbf{H}^s \mathbf{B}_s \mathbf{H}^{s\mathrm{T}}|}} \\ & \times \frac{\int_{\boldsymbol{\alpha}^s \geq \mathbf{0}} e^{-\frac{1}{2}(\boldsymbol{\alpha}^s - \mathbf{1})^{\mathrm{T}} (\mathbf{P}_s^a)^{-1} (\boldsymbol{\alpha}^s - \mathbf{1})}}{\sqrt{(2\pi)^{N_x \times N_y} |\mathbf{B}_s|} \left(1 - \Phi_{\mathbf{1}, \mathbf{B}_s}(\mathbf{0})\right)} \end{aligned} \tag{5.14}$$

where r_s^+ and m_s^+ refer respectively to the standard deviation of the error and of the emission noise (departure of the surface fluxes from their a priori values) according to the truncated Gaussian distribution. $\mathbf{P}_s^a$ is the a posteriori error covariance matrix of the control variables of the Gaussian case,

$$\mathbf{P}_s^a = \mathbf{B}_s \left(\mathbf{I} - \mathbf{K}^s \mathbf{H}^s\right) . \tag{5.15}$$

Even though it is of formal use in the truncated Gaussian case, $\mathbf{P}_s^a$ *is not* the error covariance matrix of the truncated Gaussian case. A mathematical hardship is that Eq. (5.15) requires the computation of a $1,768$-dimensional ($N_x \times N_y$) integral over the positive cone. To overcome this difficulty, we resort to the sampling technique used by Winiarek et al. [2012].

5.3 Setup of the numerical experiments

5.3.1 Observations

The in situ observations used in this study are extracted from the EMEP database [1]. The EMEP monitoring network covers most of Europe. Eleven stations of western Europe are used in this study, resulting in a rather sparse network. These stations measure the concentrations of fourteen different VOCs. Note that the m-xylene and p-xylene are combined in a lumped mp-xylene category. The observations are from January 11, 2005 to December 29, 2005.

Table 5.2 gives the number of observations used in the inversion per species and per station, for a total of 18,675 observations from 11 stations. A forecast test will also be performed using 19,746 observations of the year 2006. The VOC station Kollumerwaard, in the Netherlands (code NL0009R), does not provide any observation in 2005. However, this station provides 26,732 observations in 2006, and they will be used for cross-validation. The locations of the EMEP sampling stations for VOCs are shown in Fig. 5.1.

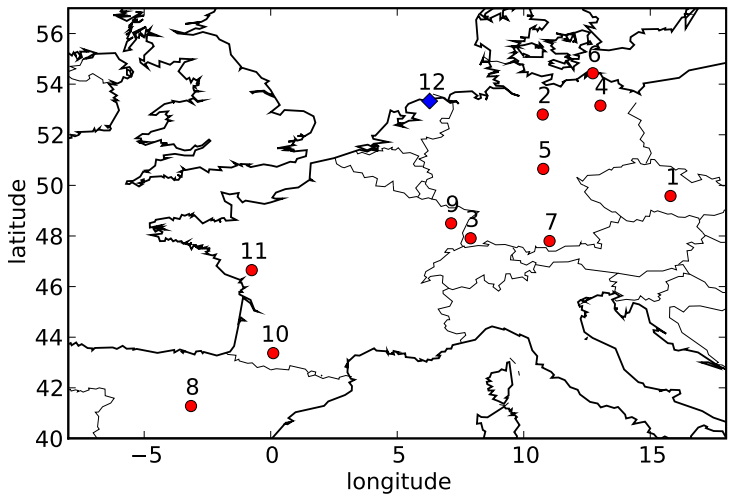


Figure 5.1: The 11 monitoring stations of the EMEP network for volatile organic compounds whose observations are assimilated are indicated with a circle. The Kollumerwaard station in the Netherlands used for validation only is indicated by a rhombus.

[1] details available at `http://www.nilu.no/projects/ccc/emepdata.html`

Table 5.2: Number of observations for each species ($N^{\text{obs}}_{\text{species}}$) and number of observation dates for each station ($N^{\text{obs}}_{\text{station}}$). The numbers 1-12 are given to help locate the stations on the map of Fig. 5.1.

species	$N^{\text{obs}}_{\text{species}}$			station code	$N^{\text{obs}}_{\text{station}}$	
	2005	2006			2005	2006
C_3H_8	1505	1415	1	CZ0003R	1397	1297
NBUT	1503	1414	2	DE0002R	948	1023
IBUT	1033	1453	3	DE0003R	837	990
NPEN	1501	1409	4	DE0007R	942	1029
IPEN	1610	1553	5	DE0008R	883	1054
C_3H_6	1497	1411	6	DE0009R	938	1020
TOLU	1441	1408	7	DE0043G	9206	9283
OXYL	868	1273	8	ES0009R	259	199
MPXYL	1318	1459	9	FR0008R	1215	1285
ISO	936	1309	10	FR0013R	769	1214
ACE	1034	1407	11	FR0015R	1281	1352
C_2H_6	1494	1409	12	NL0009R	0	26732
C_2H_4	1434	1416				
BEN	1503	1410				

5.3.2 Inversion and validation setup

Two simulation periods are considered. The first one is the the assimilation time window of the study, from January 11 to December 29, 2005. The second one is the subsequent validation of the inversion results and corresponds to the whole year 2006.

The simulation domain over western Europe extends between [40N, 8W] (the left bottom corner) and [57N;18E] (the right top corner). The grid resolution is $0.5^\circ \times 0.5^\circ$. Nine levels are considered above the surface: 50, 250, 600, 1000, 1500, 2100, 2800, 3600 and 4500 m agl. The control space discretisation follows the simulation grid, with $N_x = 52$ and $N_y = 34$. As a result, the number of control variable $[\boldsymbol{\alpha}^s]_l$ used in the inversion for each species s, is $1,768$.

The meteorological data are generated from the re-analysis fields of the European Centre for Medium Range Weather Forecast (ECMWF), delivered in 60 vertical levels and every 3 hours, with a horizontal resolution of $0.36^\circ \times 0.36^\circ$.

For anthropogenic emissions, the background emissions over the whole domain are provided by the EMEP inventory for the years 2005 and 2006 [Tarrasón et al., 2007; Fagerli et al., 2008]. The anthropogenic emissions of EMEP have a resolution of $0.5^\circ \times 0.5^\circ$. These emissions are modulated in time with the help of the hourly, weekly and monthly distribution coefficients, provided by the GENEMIS project [GENEMIS, 1994]. The biogenic emissions of isoprene are also taken into account using the model proposed by Simpson et al. [1999]. All these emissions are used as a first guess, $\mathbf{e}^b$, in the data assimilation experiments.

The initial and boundary conditions concentration fields are obtained from the global chemistry transport model MOZART 2 [Horowitz et al., 2003]. Since the species we are interested in are not all explicitly present in MOZART 2, the values for the VOCs not included in MOZART 2 were inferred from the concentration fields of some species present in MOZART 2. The fac-

Table 5.3: Factors applied to MOZART 2 explicit species concentrations to determine the initial and boundary conditions of the model species.

species	Factors and MOZART 2 species
C_3H_8	C_3H_8
NBUT	0.44 C_3H_8
IBUT	0.22 C_3H_8
NPEN	0.05 C_3H_8
IPEN	0.1 C_3H_8
C_3H_6	C_3H_6
TOLU	0.26 C_3H_8
OXYL	0.03 C_3H_8
MPXYL	0.03 C_3H_8
ISO	0.03 C_3H_8
ACE	0.35 C_3H_8
C_2H_6	C_2H_6
C_2H_4	C_2H_4
BEN	0.44 C_3H_8

tors applied for this inference are given in Table 5.3 [see Rudolph and Ehhalt, 1981; Rudolph and Johnen, 1990; Penkett et al., 1993].

5.3.3 Verification of the adjoint solutions

In order to generate the adjoint solutions at a low computational cost, we have used an approximate adjoint model, following the construction of Bocquet [2005a]. Moreover, we have assumed the lifetime of the species within the domain to be less than 10 days. After 10 days, the VOCs are assumed to be out of the domain or consumed by chemical reactions, so that the sensitivity of the concentrations within the domain to the emissions is negligible.

In order to check these approximations, the concentration fields from the adjoint model, obtained from the contribution of the source ($\mathbf{H}\boldsymbol{\alpha}$ when $\boldsymbol{\alpha} = \mathbf{1}$), were compared with the concentration fields from the direct simulation (for $\mathbf{e} = \mathbf{e}^b$, the EMEP inventory first-guess, with clean air boundary and initial conditions). This is the so-called duality test [Davoine and Bocquet, 2007]. The correlation between both computations, for all the species, is 0.995. The average values of the direct model concentrations and adjoint model concentrations are $0.38\,\mu\mathrm{g/m^3}$ and $0.36\,\mu\mathrm{g/m^3}$, respectively. The normalised mean square error (NMSE) between the two sets of results is about 3×10^{-3}. Figure 5.2 shows the comparison between the source contribution estimated with the forward model and with the adjoint model. These results indicate that the adjoint model is accurate enough for our inverse modelling purpose.

5.3.4 Values of the hyperparameters

In the Gaussian case, the optimal values of the hyperparameters r_s and m_s are obtained by value screening of the pdf Eq. (5.11). Their optimal values are reported in Table 5.4 for each species. As an example, Fig. 5.3 displays $p(\boldsymbol{\mu}^s|r_s, m_s)$ for the species NBUT. The coordinates are normalised with respect to the values of the hyperparameters obtained from the fixed-point

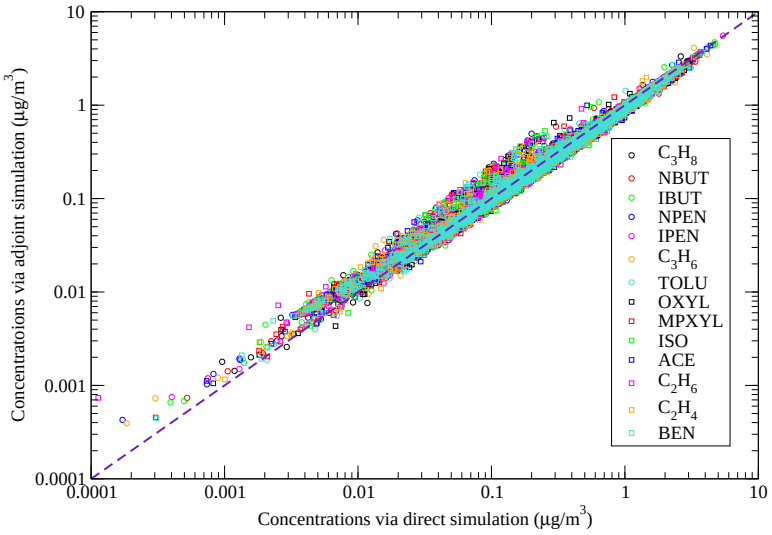


Figure 5.2: Comparison of the source contributions estimated with the direct model and the adjoint model.

solutions of Eq. (5.12). It can be seen that the optimal value of the likelihood Eq. (5.11) is equal to the result of the fixed-point method.

A comparison between the likelihoods Eq. (5.11) and Eq. (5.14) is also shown in Table 5.4. The optimal value of r_s^+ and m_s^+ computed with Eq. (5.14) are obviously different from r_s and m_s. Remarkably r_s^+ is always larger than r_s. Indeed, the Gaussian assumption inversion incorrectly interprets part of the noise within the observations as useful information, while the positivity constraint of the truncated Gaussian assumption offers less flexibility. Therefore, the inversion based on Gaussian assumptions underestimates the errors' magnitude. The comparison between m_s^+ and m_s is of less relevance because the two parameters do not represent the same statistical information within the Gaussian or truncated Gaussian assumptions.

As an example, Fig. 5.4 displays the likelihood of the hyperparameters for ethane in the truncated Gaussian case. Again, the variables are normalised with respect to the values obtained from the Desroziers method.

5.4 Inversion results

Four types of simulations conducted for the year 2005 are reported here:

- In case A, the VOC concentrations are simulated using the EMEP inventories, $\mathbf{e}_k = \mathbf{e}_k^b$ in Eq. (5.1). This is a free run serving as a reference since the observations are not assimilated.

- In cases B1 and B2, the concentrations are simulated using the emissions obtained from Eq. (5.4). The scaling factors $\boldsymbol{\alpha}^s$ used in this equation, are obtained by the minimisation of Eq. (5.7). The L-BFGS-B optimisation tool [Byrd et al., 1995] is used under the constraints that $\boldsymbol{\alpha}^s \geq \mathbf{0}$ (case B2). In case B1, we merely use the BLUE matrix formula,

Table 5.4: Estimated standard deviations of the observation error and background error, under the Gaussian likelihood and truncated-Gaussian likelihood. The units of r_s and r_s^+ are $\mu\mathrm{g}/\mathrm{m}^3$.

species	r_s	m_s	r_s^+	m_s^+
C_3H_8	0.48	16.04	0.50	4.03
NBUT	0.36	2.02	0.37	0.64
IBUT	0.24	35.17	0.28	2.72
NPEN	0.19	6.01	0.20	0.76
IPEN	0.30	8.16	0.32	1.63
C_3H_6	0.14	14.22	0.15	2.84
TOLU	0.34	3.16	0.35	0.63
OXYL	0.08	16.75	0.09	1.33
PXYL	0.16	6.52	0.17	0.52
MXYL	0.16	6.52	0.17	0.52
ISO	0.46	82.65	0.60	2.61
ACE	0.35	41.26	0.36	8.23
C_2H_6	0.55	17.40	0.57	4.37
C_2H_4	0.35	31.19	0.40	3.93
BEN	0.24	11.12	0.27	1.11

Eq. (5.8), for the estimator of $\boldsymbol{\alpha}^s$. However, it is assumed in both cases that the errors are essentially Gaussian, so that the hyperparameters used in the inversion are computed with the Gaussian likelihood following Section 5.2.5. Because the statistical assumptions are different in the estimation of the fluxes and the estimation of the hyperparameters, this may lead to inconsistencies. However, the fact that the r^+ and r hyperparameters are not too different proves that these inconsistencies are small. Yet, by construction, case B1 can lead to negative emission fluxes. This may be considered unphysical but it might have some potential use for air quality forecast.

We have also considered a third case B3, that takes the results of B2 and artificially sets all negative values to zero. However, we found that because it is not a minimum of Eq. (5.7), it leads to a poorly performing estimation. Therefore, case B3 uses ruled out and is not reported here.

- Case C is similar to case B2 except that the hyperparameters are obtained using the truncated Gaussian likelihood, following Section 5.2.5. In this case the estimation of the emission fluxes and the estimation of the hyperparameters are statistically fully consistent.

5.4.1 Analysis of the inversion results

5.4.1.1 A posteriori verification of the model linearisation

In the full CTM, which is used in the present study, the chemical kinetics of the reactions can be written as follows:

$$\mathcal{X}(c) = \left(\frac{\delta\mathcal{X}(c)}{\boldsymbol{X}(c)} + 1\right)\boldsymbol{X}(c) \tag{5.16}$$

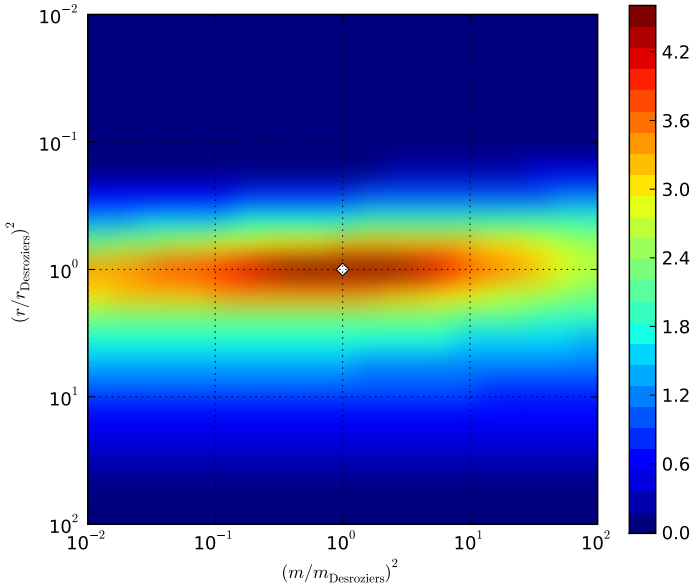


Figure 5.3: This density plot displays a monotonic transform of the likelihood of the hyperparameters for NBUT in the Gaussian case. The monotonic transform is used to obtain a better contrast in the density plot. The abscissa and ordinate are normalised according to the optimal parameters obtained from the fixed-point method.

where $\delta\mathcal{X}(c)$ denotes the variation of the VOC concentrations field with respect to the variation of the oxidant concentration field. Until now, we assumed that $\delta\mathcal{X}(c) \ll \boldsymbol{X}(c)$, and that the reduced model was linear. In order to check that hypothesis, the a priori and the a posteriori oxidant concentration fields are compared. Fig. 5.5 shows the comparison of the concentration fields of OH before and after data assimilation (case B2). Each point in the figure denotes the average concentration of OH over the spatial domain for a 2-hour period. The mean value of the concentrations is of 5.51×10^{-5} μg/m^3 for the a priori fields and 5.02×10^{-5} μg/m^3 for the a posteriori fields. For the species NO_3 and O_3, the average values of the a priori concentrations are 0.0147 μg/m^3 and 87.73 μg/m^3 respectively. They are 0.0131 μg/m^3 and 88.73 μg/m^3 for the a posteriori concentrations. The Pearson correlation between the a priori and a posteriori concentration fields is about 1.00. Furthermore, an examination of the a posteriori VOC concentration fields shows that the results obtained with the reduced linear model are very close to those obtained with the complete model. The relative bias between the two sets of concentrations for all of the VOC species and over the entire spatial domain is 1% for the year 2005 and the correlation is 1.00. Therefore, since the oxidant concentration fields are little affected by the VOC data assimilation, we consider that the hypothesis that the reduced model is about linear is verified.

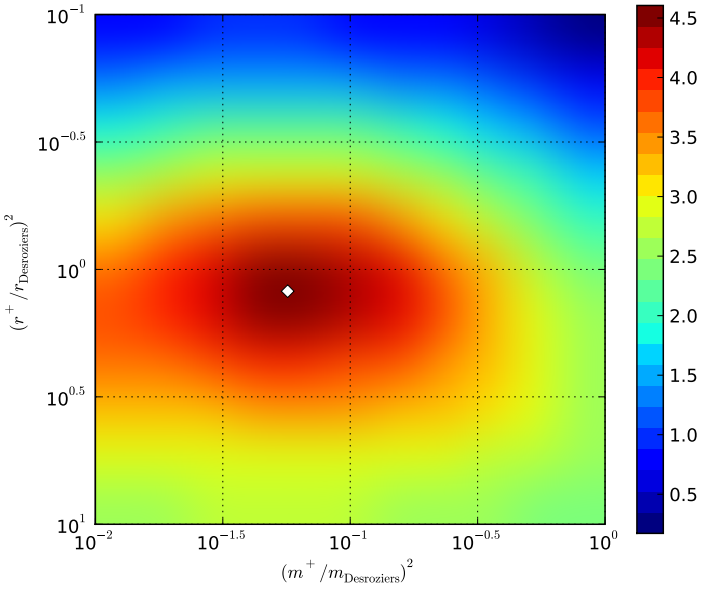


Figure 5.4: A monotonic transform of the likelihood of the hyperparameters for C_2H_6 in the truncated Gaussian case. The abscissa and ordinate are normalised according to the fixed-point method.

5.4.1.2 Comparison to the observations

The fours runs A, B1, B2 and C are compared with the observations of the analysis period (year 2005). Statistical indicators for this comparison are reported in Table 5.5. For most species the bias between the concentrations and the observations decreases with data assimilation, except when the mean of the simulation is already close to the measurement mean. The root mean square errors (RMSEs) and the normalised mean square errors (NMSEs) are systematically improved in the re-analysis runs, which is consistent with the fact that our inversion scheme minimises the quadratic error. For all species, except ACE, the Pearson correlation coefficients R, FA_2 and FA_5 are remarkably improved in the re-analyses. Note that FA_x is the fraction of the simulated concentrations within a factor x of the corresponding observations. In the very few cases where an indicator is not improved, other indicators are improved. The decrease of the correlation in the ACE case is due to a very large bias, which is compensated by a very significant improvement of the other indicators, starting with the bias.

Considering all species (14) and all statistical indicators (6) together, we have counted how often the forecast runs B1, B2 and C beat the free run A: 81 times out of 84 in case B1; 80 times out of 84 in case B2; and 79 times out of 84 in case C.

The fact that run B1 is slightly closer to the observations than B2 and C is consistent with the fact that optimisation on which B1 relies is less constrained than that of B2 or C (emission fluxes can be negative in case B1). Yet, it does not prove that method B1 is better than method B2 or C, since a comparison with the (already assimilated) observations is merely a check of

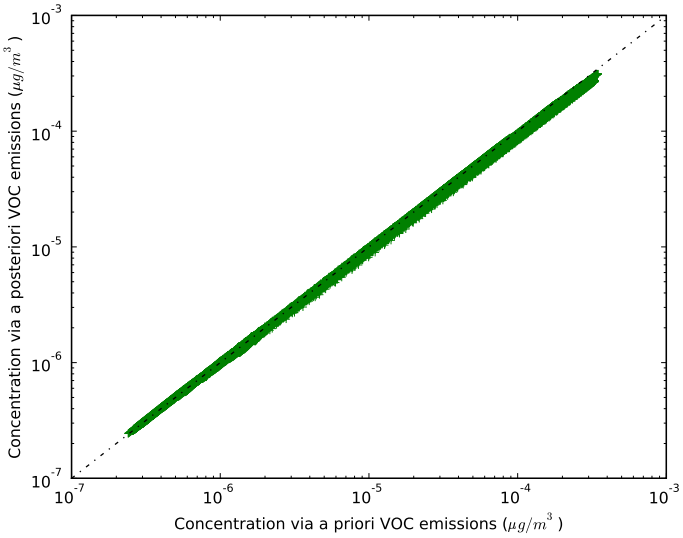


Figure 5.5: Comparison between the a priori and a posteriori OH concentration fields.

consistency, not a validation.

5.4.1.3 Estimated inventories

The total emitted mass of the EMEP inventories for the fifteen species are compared with the a posteriori emissions obtained from data assimilation following Gaussian assumption (B1/B2) and truncated Gaussian assumption (C). The results are reported in Table 5.6. Caution must be used on the interpretation of the results of the statistically consistent Gaussian case B1 since fluxes are allowed to be negative in this case (e.g. isoprene emissions).

The results of case B2 and case C indicate that the EMEP inventories may underestimate the true emissions for C_3H_8, IBUT, IPEN, C_3H_6, ACE, C_2H_6, C_2H_4, and BEN. They may overestimate the true emissions for the other VOC species. For all the species, the total mass obtained from the inversion based on the truncated Gaussian assumption (C case) is between the EMEP inventory and the total mass estimated from the Gaussian assumption (case B2). Comparison between cases A and C shows a strong correction for C_3H_8, ACE, C_2H_6, and C_2H_4. It is less than 20% for NBUT, NPEN, IPEN, TOLU, OXYL and MXYL. Figure 5.6 presents the ratio between the correction of the emission (i.e., the total posterior emission minus the total prior EMEP emissions) and the total prior emission for the Gaussian cases (B1 and B2) and the truncated Gaussian case (C).

5.4.1.4 Spatial distribution

The spatial extent of the corrections from the EMEP network depends on the nature of the species. As an example, Fig. C.1 displays the spatial ratio between the posterior emissions and the prior EMEP emissions for the species C_2H_6 (with an average lifetime of about 60 days),

Table 5.5: Scores from the comparison between the observations and the simulated concentrations for four simulations. For each species, the first line represents the scores for the simulations with the a priori fluxes (case A). The scores of the second line and third lines are related to the simulations with the a posteriori emissions from Gaussian hyper parameters estimation (case B1 and case B2 respectively). The scores of the fourth line are related to the simulations performed with the a posteriori emissions under the truncated Gaussian assumption (case C). The means and the RMSE are in $\mu g/m^3$. Bold numbers compared to the best agreement with the observations.

species	$\overline{O}$	case	$\overline{C}$	RMSE	NMSE	R	FA_2	FA_5
C_3H_8	1.13	A	0.54	0.83	1.14	0.77	0.40	0.946
		B1	**1.08**	**0.46**	**0.18**	**0.85**	**0.91**	**0.999**
		B2	1.07	0.49	0.20	0.83	**0.91**	0.998
		C	1.04	0.51	0.22	0.82	0.90	0.998
NBUT	0.62	A	**0.64**	0.41	0.43	0.68	0.80	0.989
		B1	0.59	**0.36**	**0.35**	**0.75**	**0.85**	**0.996**
		B2	0.59	**0.36**	0.36	0.74	**0.85**	0.995
		C	0.58	0.37	0.38	0.72	0.84	0.995
IBUT	0.33	A	0.18	0.37	2.21	0.55	0.59	0.960
		B1	**0.33**	**0.23**	**0.46**	**0.82**	**0.79**	0.972
		B2	**0.33**	0.28	0.69	0.71	0.78	0.978
		C	0.31	0.30	0.86	0.65	0.77	**0.984**
NPEN	0.30	A	0.25	0.25	0.86	0.48	0.57	0.939
		B1	**0.26**	**0.18**	**0.41**	**0.70**	**0.77**	**0.987**
		B2	0.25	0.20	0.51	0.65	0.71	0.974
		C	0.24	0.21	0.63	0.60	0.66	0.972
IPEN	0.51	A	0.45	0.42	1.39	0.42	0.44	0.875
		B1	**0.46**	**0.29**	**0.36**	**0.65**	**0.82**	**0.993**
		B2	0.45	0.31	0.42	0.60	0.78	0.989
		C	0.43	0.32	0.47	0.56	0.77	0.989
C_3H_6	0.18	A	0.06	0.21	4.37	0.45	0.27	0.653
		B1	**0.16**	**0.13**	**0.61**	**0.72**	**0.78**	**0.970**
		B2	0.15	0.15	0.77	0.66	0.71	0.964
		C	0.15	0.18	0.85	0.64	0.70	0.963
TOLU	0.46	A	**0.46**	0.39	0.73	0.47	0.65	0.953
		B1	0.43	**0.34**	**0.56**	**0.58**	**0.72**	0.957
		B2	0.43	0.35	0.60	0.56	0.70	0.959
		C	0.43	0.36	0.64	0.53	0.68	**0.960**
OXYL	0.09	A	0.055	0.11	2.21	0.37	0.49	0.879
		B1	**0.086**	**0.07**	**0.65**	**0.74**	**0.68**	**0.962**
		B1	0.083	0.09	0.97	0.60	0.63	0.950
		C	0.080	0.09	1.06	0.57	0.62	0.946
MPXYL	0.19	A	**0.17**	0.20	1.25	0.33	0.53	0.909
		B1	**0.17**	**0.15**	**0.71**	**0.59**	**0.65**	**0.956**
		B2	0.16	0.17	0.89	0.52	0.62	0.940
		C	0.16	0.17	1.01	0.47	0.57	0.919
ISO	0.31	A	0.18	0.86	13.47	0.63	0.35	0.708
		B1	**0.31**	**0.46**	**2.22**	**0.91**	0.45	0.716
		B2	0.36	0.59	3.18	0.82	**0.46**	0.789
		C	0.35	0.59	3.22	0.82	**0.46**	**0.795**
ACE	0.52	A	0.12	0.57	4.95	**0.69**	0.11	0.723
		B1	**0.49**	**0.34**	**0.45**	0.67	**0.72**	**0.995**
		B2	**0.49**	0.35	0.49	0.64	0.71	0.992
		C	0.47	0.38	0.59	0.57	0.69	0.989
C_2H_6	1.93	A	1.17	0.99	0.43	0.78	0.76	**1.000**
		B1	**1.83**	**0.53**	**0.08**	**0.86**	**0.99**	**1.000**
		B2	1.81	0.55	0.09	0.85	**0.99**	**1.000**
		C	1.76	0.59	0.10	0.83	**0.99**	**1.000**
C_2H_4	0.64	A	0.20	0.68	3.50	0.63	0.22	0.677
		B1	**0.61**	**0.33**	**0.28**	**0.85**	**0.81**	**0.987**
		B2	0.59	0.39	0.40	0.79	0.77	0.984
		C	0.58	0.43	0.49	0.74	0.76	0.983
BEN	0.47	A	0.42	0.31	0.47	0.69	0.74	**0.994**
		B1	**0.46**	**0.23**	**0.25**	**0.81**	**0.84**	0.977
		B2	**0.46**	0.26	0.31	0.77	0.82	0.992
		C	**0.46**	0.28	0.36	0.73	0.82	0.993

Table 5.6: For all species, the total emitted mass (in Gg) for the EMEP inventory run (case A), the a posteriori emissions under Gaussian assumption (cases B1 and B2) and the a posteriori emissions under the truncated Gaussian assumption (case C).

symbols	Case A	Case B1	Case B2	Case C
C_3H_8	7.4	18.3	18.4	15.1
NBUT	19.0	16.5	16.8	18.0
IBUT	4.7	10.3	9.1	6.1
NPEN	9.7	7.1	7.8	8.9
IPEN	9.2	11.4	10.8	10.1
C_3H_6	4.1	5.4	5.9	5.4
TOLU	12.7	11.1	11.4	12.1
OXYL	2.6	1.9	2.4	2.5
PXYL	5.7	3.5	4.4	5.4
MXYL	2.3	2.4	2.1	2.2
ISO	165.0	-925.5	124.9	143.4
ACE	2.1	10.6	9.9	4.8
C_2H_6	7.2	22.9	22.0	17.2
C_2H_4	7.9	19.1	20.1	14.5
BEN	5.6	7.1	6.7	5.8

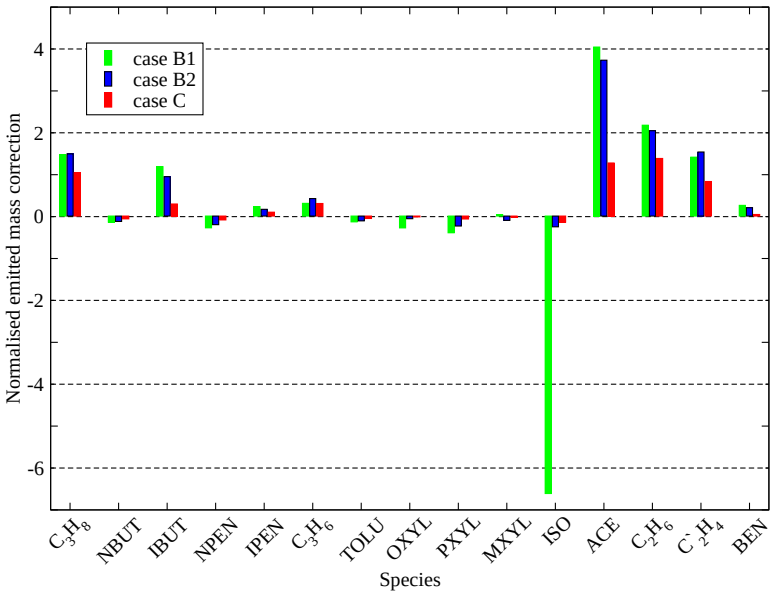


Figure 5.6: The total emitted mass correction, normalised with respect to the total emitted mass of the EMEP inventory for cases B1, B2 and C.

IPEN (with an average lifetime of 4 days), OXYL (with a lifetime of about 1.1 days), and ISO (with an average lifetime of 1.7 hours), respectively. Obviously, the corrections extend much farther from the monitoring stations for the long-lived species, such as C_2H_6 and IPEN than for the short-lived species.

As can be seen, with the B2 approach based on a Gaussian estimation of the prior errors, the magnitude of the corrections is significantly higher than when using the fully consistent non-Gaussian approach C. One way to put it is that, with the Gaussian assumption, part of the error (noise) is mis-interpreted as valuable information (signal), so that the Gaussian assumption leads to over-corrections. A similar phenomenon was put forward by Koohkan and Bocquet [2012] in the inversion of carbon monoxide emission fluxes: the proper identification of representativeness errors leads to smaller corrections of the emission fluxes.

Several studies have performed inverse modelling study of isoprene emissions over Europe. To do so, they assimilate satellite observations of formaldehyde [e.g. Curci et al., 2010; Dufour et al., 2009]. Specifically, they exploit the observations of SCIAMACHY (instrument operating onboard the sunsynchronous Envisat satellite) with a resolution of 30×60 km^2. The study of Curci et al. [2010] shows an increase of about 5% for the MEGAN [Guenther et al., 2006] emissions. Our results show that the emission of isoprene decreases by about 24% in case B2 and by 13% in case C for the emissions obtained from Simpson et al. [1999]. This discrepancy is explained in part by the fact that the emission inventories of isoprene by Simpson et al. [1999] and MEGAN differ significantly [Bessagnet et al., 2008; Sartelet et al., 2012] with the former leading to greater isoprene emissions on average over Europe by a factor of about 2.5 [Sartelet et al., 2012]. This discrepancy is also explained by the very short lifetime of isoprene (1-2 hours). Assimilation of in-situ observations are inoperative because the information is not spread far enough by the model, because its transport representation becomes of minor interest. Our results are only valid near the five stations measuring isoprene. On the contrary, satellite observations are well-suited for this short-lived species as remote sensing offers an (indirect) spatially well-resolved snapshot of the concentrations. In addition, the errors made by approximate adjoint model (or automatic adjoint up to some numerical precision) are larger for short-lived species [Bocquet, 2012]. As a consequence, the case of isoprene in this study is somehow singular because its lifetime is too short. Nevertheless, most results of isoprene are given in this study for the sake of comparison, and to document the inadequacy of the in situ observations for regional inverse modelling in such a case.

5.4.2 Forecast test

In order to test and possibly validate the corrected inventory obtained from inverse modelling, one needs observations that have not been assimilated. One stringent test is to perform a forecast using the corrected inventory over a period of time different from the data assimilation window, assuming some time persistence of the VOC inventories.

Four inventories are generated with Eq. (5.4), using the background EMEP emission $\mathbf{e}_b^s$, or $\boldsymbol{\alpha}_b^s = \mathbf{1}$ (case A), and the scaling factors of cases B1, B2 and C. In each case, a forecast is performed for the year 2006. These simulations are then compared with the independent observations of year 2006.

Note that because some of the fluxes retrieved in case B1 are negative, and because numerical schemes of CTMs often rely on the positivity of the concentrations, we had to circumvent the difficulty. One solution consists in decomposing the scaling factors into a positive part and a negative part: $\boldsymbol{\alpha}^s = \boldsymbol{\alpha}_+^s + \boldsymbol{\alpha}_-^s$, so that, invoking the linearity of physics, the concentrations are given by $\mathbf{H}^s(\boldsymbol{\alpha}^s) = \mathbf{H}^s(\boldsymbol{\alpha}_+^s) - \mathbf{H}^s(-\boldsymbol{\alpha}_-^s)$.

The statistical indicators are reported in Table 5.7. The results indicate that, for most species, the scores are improved using the retrieved scaling factors of case B2 and case C,

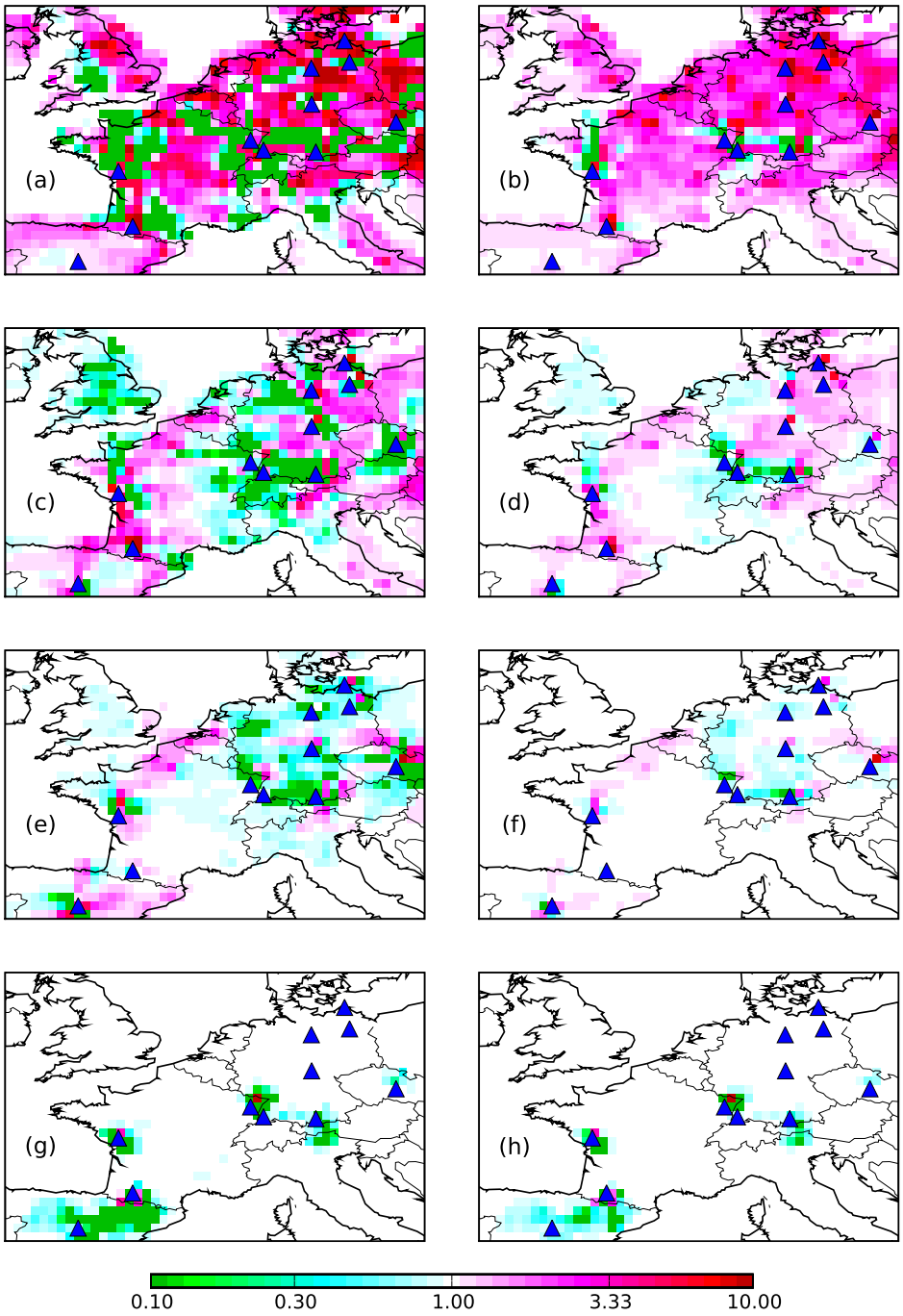


Figure 5.7: Gridded ratios of time-integrated retrieved flux to EMEP time-integrated flux for C_2H_6, (a,b), IPEN (c,d), OXYL (e,f) and ISO (g,h). Green and blue colours correspond to reduction of the emission fluxes, whereas red and pink colours correspond to increase of the emission fluxes. The left column (a,c,e,g) corresponds to case B2 and the right column (b,d,f,h) corresponds to case C. The species are ordered by decreasing lifetime.

whereas they are degraded in case B1. Considering all species and all statistical indicators together, we have counted how often the forecast runs B1, B2 and C beat the free run A: 41 times out of 84 in case B1; 69 times out of 84 in case B2; and 77 times out of 84 in case C.

The fully-consistent truncated Gaussian approach C performs best and beats the Gaussian-based but positively constrained approach B2. Both positively constrained approaches beats the fully consistently Gaussian approach B1. Since B1 partially leads to unphysical negative fluxes, this could have been expected. However, as was shown by Bocquet [2012], an unconstrained optimisation of parameters that are allowed to take unphysical values may sometimes lead to valuable better forecasts because it compensates for other sources of model error. Obviously, this is not the case here.

According to Table 5.7, an improvement of all the scores can be seen for C_3H_8, NBUT, NPEN, C_3H_6, C_2H_6, C_2H_4 and BEN using the optimal scaling factors. However, the bias is increased for OXYL (in case B2), MPXYL (in cases B2 and C) and ISO (in case B2), using the corrected emissions. Despite that degraded bias, the other indicators are improved. For IBUT, the RMSE is increased in case B2, and the correlation is decreased in cases B2 and C. The statistical indicators show that the RMSE increases for IPEN using the corrected emissions. For that species, the Pearson correlation coefficient is also deteriorated in case B2. For TOLU, the bias, NMSE and FA_5 are degraded using scaling factors. The RMSE is also increased in case B2. The scores also show that a decrease of the Pearson correlation coefficient for ACE, using the corrected emissions.

5.4.3 Cross-validation test

Since the EMEP Netherlands VOC station Kollumerwaard does not provide any observation for the year 2005, that station is not included in the list of stations used for the analysis of 2005. Yet, for the year 2006, eleven out of fourteen VOCs were measured at Kollumerwaard, excluding NBUT, ACE and C_2H_6. Kollumerwaard is located far from the other stations whose observations lead to the corrections in the emission inventories. Therefore, we do not expect the simulated concentrations of the VOC with a short lifetime, at this station, to be very sensitive to the correction of the emissions. Indeed, as shown in Table 5.8, the scores obtained from the comparison of the simulated concentrations in case A of ISO (lifetime of about 1.7 hours) with the measurements are similar to those obtained from the comparison of the concentrations in case B and C with the observations.

For species with a longer lifetime (about one day or two), for C_3H_6 (lifetime of 15 hours), OXYL (lifetime of about 25-26 hours), MPXYL (lifetime of about 14-25 hours) and TOLU (lifetime of 2.4 days), the corrections of the inventories are more spread over the domain. The scores indicate that the results in case C are in a better agreement with the observations. The results also show that for the species with a lifetime between 4 to 7 days (IBUT, NPEN, and IPEN), the scores are deteriorated. It is likely that the corrections performed close to the Kollumerwaard station, are not reliable enough. For the species with a lifetime longer than 10 days (C_3H_8 and BEN), the scores are remarkably improved. The emission inventories of C_3H_8 are corrected almost all over the domain.

Considering all species and all statistical indicators together, we have counted how often the forecast runs B2 and C beat the free run A at the Kollumerwaard station: 34 times out of 66 in case B2; and 43 times out of 66 in case C. The use of the corrected emission from case C has a positive impact on the forecast at the station, even though the station is far away from the other stations whose observations were assimilated.

Table 5.7: Scores of the forecast test (year 2006) from the comparison of the observations and the simulated concentrations for four simulations: case A, case B1, case B2 and case C.

species	$\overline{O}$	case	$\overline{C}$	RMSE	NMSE	R	FA_2	FA_5
C_3H_8	1.10	A	0.55	0.85	1.19	0.76	0.46	0.958
		B1	**1.14**	0.63	**0.32**	**0.80**	0.85	0.992
		B2	**1.06**	0.61	**0.32**	0.78	**0.89**	**0.998**
		C	1.00	**0.59**	**0.32**	0.79	0.88	**0.998**
NBUT	0.56	A	0.75	0.45	0.49	0.72	0.70	0.977
		B1	0.38	0.39	0.40	0.78	**0.83**	0.994
		B2	**0.62**	**0.36**	**0.38**	**0.76**	0.80	**0.994**
		C	0.65	0.37	**0.38**	**0.76**	0.79	**0.994**
IBUT	0.34	A	0.21	**0.33**	1.46	**0.61**	0.67	0.967
		B1	**0.36**	0.46	1.73	0.49	0.56	0.825
		B2	0.37	0.35	**0.93**	0.55	**0.75**	**0.975**
		C	**0.32**	**0.33**	0.98	0.54	0.74	**0.975**
NPEN	0.29	A	**0.31**	0.31	1.08	0.45	0.56	0.919
		B1	0.18	0.30	1.72	0.48	0.39	0.692
		B2	0.26	0.28	1.03	0.51	0.61	0.949
		C	**0.27**	**0.27**	**0.94**	**0.52**	**0.62**	**0.950**
IPEN	0.47	A	0.29	**0.37**	0.99	0.49	0.54	0.906
		B1	0.37	0.62	2.17	0.31	0.43	0.733
		B2	**0.46**	0.46	0.99	0.42	0.64	0.963
		C	0.42	0.38	**0.73**	**0.49**	**0.65**	**0.969**
C_3H_6	0.19	A	0.08	0.24	3.86	0.44	0.39	0.674
		B1	0.07	0.24	4.40	0.51	0.24	0.521
		B2	**0.13**	**0.20**	**1.58**	**0.60**	**0.50**	**0.877**
		C	**0.13**	**0.20**	1.62	**0.60**	0.49	0.873
TOLU	0.66	A	**0.57**	0.69	**1.28**	0.28	0.55	**0.876**
		B1	0.47	0.73	1.65	0.25	0.56	0.852
		B2	0.48	0.70	1.55	0.27	0.58	0.864
		C	0.50	**0.68**	1.42	**0.29**	**0.60**	0.875
OXYL	0.07	A	**0.07**	**0.07**	1.05	0.42	0.54	0.885
		B1	0.04	0.08	2.35	0.42	0.36	0.693
		B2	0.06	**0.07**	0.98	**0.52**	0.54	**0.918**
		C	**0.07**	**0.07**	**0.95**	0.50	**0.56**	0.910
MPXYL	0.21	A	**0.21**	0.25	1.45	0.32	0.51	0.856
		B1	0.10	0.23	2.73	0.49	0.38	0.720
		B2	0.16	**0.21**	1.38	**0.50**	**0.53**	**0.885**
		C	0.18	0.22	**1.33**	0.44	**0.53**	0.863
ISO	0.31	A	**0.35**	1.04	10.25	0.21	0.26	0.571
		B1	1.39	19.95	939.9	-0.18	0.16	0.287
		B2	0.25	**0.81**	8.59	**0.46**	**0.34**	**0.682**
		C	**0.27**	**0.81**	**8.05**	0.44	0.33	0.656
ACE	0.68	A	0.14	0.83	7.17	0.57	0.17	0.649
		B1	**0.49**	**0.58**	**1.00**	**0.60**	**0.63**	**0.966**
		B2	0.47	0.62	1.22	0.51	0.54	0.952
		C	0.36	0.70	1.98	0.39	0.52	0.837
C_2H_6	1.89	A	1.15	1.05	0.51	0.78	0.80	0.998
		B1	**1.96**	0.76	**0.16**	**0.81**	**0.98**	0.999
		B2	1.78	0.76	0.17	0.78	0.97	**1.000**
		C	1.71	**0.75**	0.17	0.79	0.97	0.999
C_2H_4	0.66	A	0.26	0.87	4.43	0.59	0.36	0.751
		B1	0.39	0.74	2.15	0.65	0.39	0.713
		B2	**0.61**	**0.62**	**0.97**	**0.73**	0.64	0.978
		C	0.56	0.66	1.18	0.69	**0.65**	**0.980**
BEN	0.47	A	0.44	0.38	0.69	0.65	0.77	0.995
		B1	0.57	0.42	0.65	0.70	0.71	0.966
		B2	**0.47**	**0.35**	**0.55**	**0.71**	**0.81**	0.994
		C	0.46	0.36	0.59	0.69	0.80	**0.997**

Table 5.8: Scores at the Kollumerwaard station (for the year 2006) from the comparison of the observations and the simulated concentrations for three simulations: case A, case B2 and case C.

species	case	$\overline{O}$	$\overline{C}$	RMSE	NMSE	R	FA_2	FA_5
	A		0.82	2.12	3.24	0.05	0.39	0.799
C_3H_8	B2	1.69	**0.97**	**1.81**	**2.08**	0.43	**0.63**	**0.936**
	C		0.81	1.87	2.56	**0.47**	0.56	0.932
	A		**0.44**	**0.33**	1.32	**0.54**	0.60	0.868
IBUT	B2	0.33	0.41	0.39	1.19	0.34	0.60	0.927
	C		0.36	0.34	**0.98**	0.37	**0.62**	**0.948**
	A		**0.28**	0.42	**1.53**	0.52	0.50	0.899
NPEN	B2	0.42	0.18	**0.48**	2.99	0.48	0.45	0.859
	C		0.23	0.44	2.00	**0.53**	**0.52**	**0.917**
	A		**0.43**	**0.59**	**1.32**	**0.20**	**0.48**	**0.868**
IPEN	B2	0.61	0.37	0.63	2.04	0.13	0.41	0.770
	C		0.33	0.65	1.75	0.12	0.36	0.814
	A		0.08	0.30	6.10	0.65	0.36	0.719
C_3H_6	B2	0.17	**0.14**	**0.26**	**2.79**	0.63	**0.44**	**0.780**
	C		0.13	**0.26**	2.97	**0.66**	0.39	0.734
	A		**0.52**	0.43	**0.68**	0.48	**0.67**	0.886
TOLU	B2	0.53	0.35	0.35	0.64	0.72	0.61	0.937
	C		0.39	**0.33**	0.52	**0.73**	**0.67**	**0.942**
	A		**0.06**	0.10	**1.51**	0.52	**0.39**	**0.847**
OXYL	B2	0.11	0.05	0.10	1.76	0.59	0.36	0.818
	C		**0.06**	**0.09**	1.58	**0.60**	0.38	0.829
	A		**0.15**	0.22	**1.44**	0.72	**0.39**	0.881
MPXYL	B2	0.22	0.11	0.23	2.26	**0.75**	0.36	0.875
	C		0.14	**0.21**	1.48	**0.75**	0.38	**0.888**
	A		**0.05**	**0.35**	14.3	**0.33**	0.19	0.445
ISO	B2	0.18	**0.05**	**0.35**	**13.29**	0.32	**0.20**	**0.463**
	C		**0.05**	**0.35**	**13.29**	0.32	**0.20**	**0.463**
	A		0.14	0.44	4.33	**0.57**	0.24	0.609
C_2H_4	B2	0.18	**0.17**	0.45	3.58	0.44	**0.33**	**0.691**
	C		0.20	**0.40**	**2.42**	0.55	0.31	0.662
	A		**0.54**	0.83	**2.92**	0.00	0.48	0.772
BEN	B2	0.44	0.24	0.67	4.26	0.23	0.43	0.816
	C		0.30	**0.65**	3.25	**0.27**	**0.52**	**0.869**

5.4.4 Information content and DFS

The degrees of freedom for the signal (DFS) is a metric that is representative of the fraction of the observations effectively used in the inversion to retrieve the source [Koohkan et al., 2012]. A better data assimilation system can either lead to an increase of the DFS when the observations are better used by the system, or lead to a decrease of the DFS if the system better diagnoses the errors and correctly identifies more noise in the observations [Koohkan and Bocquet, 2012]. Therefore, a comparison of DFS from two data assimilation systems relying on different assumptions is not straightforward. However, a comparison between DFS of different species for the same data assimilation system is of simpler interpretation. The DFS for each species is given by:

$$\mathrm{DFS}_s = \mathbb{E}\left[\left(\boldsymbol{\alpha}^s - \mathbf{1}\right)^{\mathrm{T}} \mathbf{B}_s^{-1} \left(\boldsymbol{\alpha}^s - \mathbf{1}\right)\right] . \tag{5.17}$$

The ratios between DFS_s and $\mathcal{L}_s$ (see Eq.(5.7)) are reported in Table 5.9 for cases B1, B2, and C. For all species, the ratio is greater for B1. Indeed, by permitting negative fluxes, the data assimilation system is incorrectly interpreting degrees of freedom for the noise as DFS. As for the positively constrained inversions, for almost all species except BEN, the ratio is greater in the statistically consistent truncated Gaussian system. The DFS, $4 \pm 2\%$ in the B2 case, and $7 \pm 2\%$, is consistent with the figures usually met in air pollution source inverse modelling systems [Koohkan et al., 2012, argue that it is usually between 5% and 15% for dispersion problems]. Isoprene is clearly identified as the species with the lower ratio in the B2 case.

Table 5.9: The ratio of DFS to the cost function for each species.

symbols	case B1	case B2	case C
C_3H_8	0.076	0.055	0.108
NBUT	0.040	0.033	0.035
IBUT	0.142	0.052	0.067
NPEN	0.088	0.050	0.067
IPEN	0.072	0.044	0.069
C_3H_6	0.067	0.040	0.091
TOLU	0.045	0.030	0.031
OXYL	0.119	0.037	0.052
MPXYL	0.141	0.034	0.045
ISO	0.081	0.015	0.042
ACE	0.074	0.055	0.069
C_2H_6	0.080	0.060	0.131
C_2H_4	0.126	0.056	0.109
BEN	0.081	0.040	0.033

5.5 Conclusion

The goal of this study was to estimate the emission inventories of fifteen VOC species using ground-based in situ measurements. The concentration observations at eleven stations from

the EMEP network over western Europe were assimilated to perform inverse modelling of the emission field for each one of the fifteen species, for the year 2005.

For that purpose, the Jacobian matrix, i.e., the source-receptor relationship, was built using the POLAIR3D CTM. To compute that matrix, a fast version of this CTM, as well as its validated approximate adjoint model have been developed. The chemistry module of this fast version only includes the chemical reactions between the VOC species and three oxidants (OH, NO_3 and O_3), the concentrations of which are pre-computed with the full CTM.

For each species and each grid cell, a scaling factor that multiplies the local EMEP emission flux, is computed. The uncertainty attached to the prior scaling factors and the covariance matrix of the observation errors, which are crucial statistical components of the inversion, are obtained using the maximum likelihood principle. The principle was implemented using two different assumptions: (1) the errors attached to the scaling factors follow a Gaussian pdf or (2) the scaling factor follows a truncated Gaussian pdf.

In the Gaussian case, the simulated concentrations for the year 2005 using the corrected emissions lead to a significant improvement in most statistical indicators. However the fact that the VOC fluxes are positive is not statistically accounted for, and this is shown to lead to a probable over-fitting to the observations, and to over-corrections of the EMEP emissions. Using a fully consistent truncated Gaussian assumption for the emission fluxes, including the use of a non-Gaussian likelihood for the estimation of the hyperparameters, the corrections are significantly smaller.

For short-lived species, it is shown that information cannot propagate far from the monitoring stations, so that the corrections are rather local to the stations. That is why we deem the isoprene inversion to be unreliable. That is a typical case where remote sensing assimilation is necessary to offer a satisfying coverage.

The corrected emissions have been partly validated thanks to a forecast conducted for the year 2006 using independent observations. The simulations using the corrected emissions often led to significant improvements in the statistical indicators. Considering all statistical indicators, the fully consistent truncated Gaussian approach emerged as the best approach from this test.

The 2006 forecasts have also been compared to the observations at the Kollumerwaard station, the Netherlands. The Kollumerwaard station is not part of the 11 stations used in the analysis of 2005. Even though this station is far away from the 2005 network, and its surroundings fluxes little affected by the 2005 analysis, some improvements are noticed for several long-lived VOC species using the statistically consistent positively constrained inversion.

Chapter 6

Summary and perspectives

6.1 Conclusion

This study introduces new applications of data assimilation in air quality modelling. It is divided in three different studies for which a brief description is presented hereafter.

In the first part of this PhD thesis, an adjoint model to the chemistry transport model, as well as a 4D-Var routine are developed and validated. The 4D-Var routine developed is used to invert the carbon monoxide emissions. The data assimilation approach to correct the emission fields leading to a mild improvement only in the bias and the correlation (since it does not handle the representativeness issue), a simple statistical subgrid model is introduced and coupled to the 4D-Var routine.

In a second part, data assimilation is used as a decision-making support in order to assess the ability of a monitoring network (as a matter of fact, the International Monitoring System network of radionuclides) to reconstruct accidental sources.

In the third and last part, the maximum likelihood method is used to assess the hyper-parameters of a cost function. In particular, that method is applied to invert the emissions of Volatile Organic Compounds (VOCs).

6.1.1 Adjoint of chemistry transport model

The adjoint of the chemistry transport model based on the POLAIR3D Eulerian code of POLYPHEMUS is validated in four steps, using the *duality* test, involving one term of the equation at a time (source terms of surface and volume emissions, initial and boundary conditions). The concentrations computed using the CTM and its adjoint model are in good agreement. The pearson coefficient is 99.8% when validating the surface and volume emission fields. It is of 98.7% and 93% when validating the initial conditions and boundary conditions, respectively. This shows that the developed adjoint model is reliable enough as the errors it yields are much smaller than the modelling errors (departure between the observation and the simulated concentrations).

6.1.2 4D-Var algorithm

The 4D-Var algorithm, using the newly developed adjoint model, is validated with the help of two kinds of unity tests. In that aim, the gradient of the cost function computed within the 4D-var subroutine is compared to the one computed with the finite difference method. The first test (perturbation) enabled to validate the gradient with respect to the emission fields and to the initial conditions. The second test validated the gradient with respect to the control parameters (periodical scale factors of the emissions).

6.1.3 Representativeness error and subgrid model

The stations of BDQA recording carbon monoxide concentrations are impacted by representativeness errors. Comparisons between the observations and simulated concentrations show that the variability of the observations is much stronger than that of the simulated concentrations. The scores of a preliminary assimilation run via the 4D-Var method, for carbon monoxide and during 8 weeks from January 1 2005, show that the consistency between the analysed concentrations and the observations is low, in spite of a Pearson correlation coefficient increasing from 0.16 to 0.36. The results also show that the 4D-Var method artificially over-estimates the total emitted mass in the first 8 weeks of the year 2005. In order to take the representativeness issue into account, which is an unresolved part of the model, a statistical subgrid model is developed and is coupled to the CTM. That new subgrid model provides a coefficient ξ_i to each station i. This statistical coefficient ξ_i links the average of the variation of the representativeness error to the variation of the emission inventories. The subgrid model in question, although simple removes an important part of the observation mismatch. The correlation between the observations and the simulated results of the coupled model , 4D-Var-ξ, increases to 0.73. The bias between the simulated concentrations and the measurements is also eliminated using the 4D-Var-ξ algorithm. The amount of the total emitted mass of CO is computed either with 4D-Var and 4D-Var-ξ. That value is independent on the station locations for the network in case using 4D-Var-ξ, while it is sensitive to the network when using 4D-Var. This results is consistent with the one obtained from CITEPA.

This study shows that the 4D-Var system is not able to assess the emission fluxes of CTM for the case in which in-situ measurements are impacted by the representativeness errors. That is why the estimation of representativeness errors is important for the assimilation of the measurements of the proximity stations.

6.1.4 Multiscale method data assimilation and application to network design

Beyond the detection capability of an observation network and the geostatistics approaches, it is crucial to evaluate the potential of a monitoring system for inverse modelling. To do so, multiscale adaptative grids of the IMS radionuclide network sources are built and optimised under the DFS criterion. The influence functions linked to the observations are generated using the Lagrangian transport model FLEXPART (driven by ECMWF meteorological fields) over the year 2009. The ratio DFS/d (d is the total number of observations), which is used to control the error of the a priori source and the modelling error (instrumental, representativeness and transport model) is carried out for the system performance level. At each level of performace, the adaptative grids are optimised. The results of the optimisation show that the IMS radionuclide network is not quite able to construct the source in the Intertropical Convergence Zone. In case of the large modelling errors, the system is not able to construct the sources far from the observation point. Considering the realistic case where DFS/d is about 10%, the information from many locations, such as African continent, cannot be sufficiently propageted.

The same test is performed for the noble gas network. The results of the realistic case show that the network does not perform well enough to detect the informations coming from the Intertropical Convergence Zone, from the African continent and the central Asia.

6.1.5 Emission flux estimation for Volatile Organic Compounds

In this study, the method of data assimilation is used in order to estimate the emission fluxes of fifteen Volatile Organic Compounds (five aromatics, six alkanes, two alkenes, one alkyne and one biogenic diene). To begin with, it can be notice that the results of the CTM simulation show

the discrepancies between the computed concentrations and the ground-based observations provided by the European Monitoring and Evaluation Programme (EMEP) for the year 2005. Furthermore, the RACM 2 (Regional Atmospheric Chemistry Mechanism, version 2) chemical kinetic mechanism used within this CTM and contains more than three hundreds reactions. The mentioned CTM is not fast enough. Therefore, a new version of that chemical kinetic mechanism is built which uses pre-computed fields of oxidants (OH,NO_3 and O_3). The adjoint of this fast CTM version is also developed and validated. As the concentrations of the computed VOCs change linearly with respect to the emission fluxes, the Jacobian matrix of the model can be built to be used in inversion problems. In the present data assimilation system, the number of the unknown variables (gridded emissions) is higher than the number of the observations. In order to reduce the dimension of the control space (the space of the gridded emission fluxes) in the inverse problem, the scaling factors attached to each grid cell are introduced for the emission fluxes. These are the very factors which would be estimated in the inversion system instead of the gridded emission fluxes. The parameters of the errors (the uncertainty of the control parameters and the observational error), used in the inversion system are obtained via a maximum likelihood method. Three types of distributions for the prior flux errors are assumed: Gaussian, Gaussian under positivity assumption and truncated Gaussian. The DA method leads to a significant reduction of the bias between the observations and the simulated concentrations. The statistical indicators show that the results of the model under the Gaussian assumption is in better agreement with the observations of the year 2005 than the results of the models under the Gaussian positive or under the truncated Gaussian assumptions. A forecast test is performed for the observations of the year 2006. The statistical indicators of the model under truncated Gaussian assumption outperform those of the Gaussian assumptions. The observations of the year 2006 done in a single station in the Netherlands is used for a validation test. It is also shown that the non-Gaussian approach provides the best scores.

The positivity assumption on the VOC emission fluxes is crucial for a successful inversion. The Gaussian assumption on the VOC emissions leads to unrealistic emission fields. During the analysis period, DA system without the positivity assumption on the emission fields more reduce the departure between the observations and model compared to the system taking the positivity assumption on the emission fields into account. However, the model is not physically admissible (since the emission flux value can be negative): therefore, it does not able to model a new set of observations. Although DA with Gaussian positive assumption on the emissions is physically allowable, but it is not performance enough. Because the error parameters ($\mathbf{R}_s$ and $\mathbf{B}_s$) used for DA system are not compatible with the positivity assumption. DA system with truncated Gaussian assumption on the emissions gives the best performance, not only due to the positivity of the posterior emissions, but also to the accurate estimation of error parameters used for DA system.

The test results also show that the efficiency in correcting the inventory depends on the lifetime of the VOCs, when using the in-situ observations. For instance, using a sparse monitoring network to estimate the emissions of isoprene is not suitable because its short chemical lifetime significantly limits the spatial radius of influence of the monitoring data. For species with longer lifetimes (a few days), emission corrections can reach regions hundreds of kilometres away from the stations.

6.2 Outlook

The new applications of data assimilation presented in this study can be generalised to the other pollutant species. For instance, the subgrid statistical model can be used for pollutants whose observations are impacted by representativeness errors. The multiscale method is an

advantageous method which helps to optimise the control space for inverse modelling. The hyper-parameters (or error parameters) estimation under the positivity constraint is useful to invert the initial condition, boundary condition, emission fluxes, etc.
Some perspectives for this thesis are presented in this section. The methodologies used in this thesis can be optimised and extended. The DA method can be also used to estimate other parameters of the model such as vertical diffusion and boundary conditions.

6.2.1 A more complex subgrid model

In chapter 3 a statistical subgrid model is introduced to take the representativeness issue into account. Although simple, that statistical model significantly helps to better estimate model errors. That subgrid model could also take the effects of the mesoscale wind direction into account. Let us assume that the θ^t angle defines the wind direction at time t in a grid cell where the station i is located. The subgrid model can be written as:

$$\xi_i f(\theta^t, \theta_i) \mathbf{\Pi}_{i,k} \mathbf{e} \tag{6.1}$$

where f is the effective wind function and the angle θ_i accounts for the effective wind direction. If $\theta^t = \theta_i$, then the impact of the nearby source on the station is maximum. A simple form for function f can be:

$$f(\theta^t, \theta_i) = \cos^2\left(\frac{\theta^t - \theta_i}{2}\right). \tag{6.2}$$

The use of a such function can help to produce a more accurate physical model. In addition to ξ_i, the unknown parameter θ_i is also estimated in the optimisation algorithm.

6.2.2 Network design through the minimisation of representativeness error

Comparison between model output and in-situ measurements are more or less impacted by representativeness errors. The accessibility to the background station observations is essential for inverse modelling [Koohkan and Bocquet, 2012] if one cannot precisely estimate representativeness errors. The mismatch between the model results and the measurements done in proximity stations (due to the local emission sources) decreases by increasing the resolution of the model. This can be expressed by the representativeness term, $\mathcal{H}\left(\mathbf{I} - \mathbf{\Gamma}^\star \mathbf{\Gamma}\right) s$, mentioned in chapter 3. If the mismatch between the observations and the simulation results remains unchanged by choosing finer or coarser grid resolutions, then it can be concluded that the measurements at the station of interest are not directly impacted by nearby sources. Therefore, in order to minimise the representativeness errors, the choice of the in-situ stations (used to compare their measurements with the computations) could be done by finding out the station locations for which the results of the CTM remain unchanged with respect to the grid resolution. This kind of study can be made possible if emission inventories with fine resolution are available (for example, using the EDGAR3 inventory with a resolution of 10km $\times$ 10km). To give an example, this methodology can be used to set up a background observation network of CO over France. Alternatively, one can pick up the stations with the lower ξ_i for a provisional monitoring network.

6.2.3 Multiscale data assimilation for VOC species

One of the challenges in inverse problem studies is finding an optimal control space. In other words, to decrease the cost of computation, instead of the real number of model parameters a smaller number of parameters called the effective parameters are used for the computations.

Of course, the computations run with the real and effective number of parameters must give almost the same results at each observational time and location. The choice of optimal space can be determined by multiscale data assimilation. The multiscale DA can be effective to invert the emissions of the VOC species with short lifetime. The spatial distribution (Fig. C.1) of these species show that the effective number of parameters which should be estimated is smaller than the dimension of the control space. Furtheremore, this type of study helps to better understand the quality of the assimilation performed in chapter 5. It also shows the performance of the EMEP monitoring network for inverse modelling of VOCs. The Jacobian matrices and the error parameters ($\mathbf{R}_s$ and $\mathbf{B}_s$) computed for inverse modelling of VOCs can be used for multiscale inverse modelling.

6.2.4 Inversion of the boundary conditions fields for long lifetime VOC species

In chapter 5, the emission fluxes of Volatile Organic Compounds are estimated using data assimilation. For long lifetime species such as C_3H_8, ACE, $C_2H_4$and C_2H_6, the emission fluxes are corrected even far from the in-situ stations and almost all over the domain (western Europe). For these species, the influence of the boundary conditions on the observations should not be neglected. To better estimate the simulated concentrations, a more accurate assessment of the boundary conditions is needed. One of the additional challenges could be to invert the boundary conditions field for long-range VOCs.

Bibliography

Abida, R. and Bocquet, M. (2009). Targeting of observations for accidental atmospheric release monitoring. *Atmospheric Environment*, 43. 26, 77

Abida, R., Bocquet, M., Vercauteren, N., and Isnard, O. (2008). Design of a monitoring network over France in case of a radiological accidental release. *Atmos. Env.*, 42:5205–5219. 78, 88

Arellano, A. and Hess, P. (2006). Sensitivity of top-down estimates of co sources to gctm transport. *J. Geophys. Res.*, 33:L21807. 50, 51

Becker, A., Wotawa, G., Ringbom, A., and Saey, P. (2010). Backtracking of noble gas measurements taken in the aftermath of the announced october 2006 event in North Korea by means of PTS methods in nuclear source estimation and reconstruction. *Pure and Applied Geophysics*, 167:581–599. 77

Bennett, A. (1992). *Inverse Methods in Physical Oceanography*. Cambridge Unversity Press. 27

Bessagnet, B., Menut, L., Curci, G., Hodzic, A., Guillaume, B., Liousse, C., Moukhtar, S., Pun, B., C., S., and Schulz, M. (2008). Regional modeling of carbonaceous aerosols over europe-focus on secondary organic aerosols. *Journal of Atmospheric Chemistry*, 61:175–202. 110

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. 77, 102

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. 77

Bocquet, M. (2007). High resolution reconstruction of a tracer dispersion event. *Q. J. Roy. Meteor. Soc.*, 133:1013–1026. 77

Bocquet, M. (2008). Inverse modelling of atmospheric tracers: Non-Gaussian methods and second-order sensitivity analysis. *Nonlin. Processes Geophys.*, 15:127–143. 30

Bocquet, M. (2009). Towards optimal choices of control space representation for geophysical data assimilation. *Mon. Wea. Rev.*, 137:2331–2348. 31, 34, 81, 83

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. 44, 77, 79, 110, 112

Bocquet, M., Pires, C. A., and Wu, L. (2010). Beyond Gaussian statistical modeling in geophysical data assimilation. *Mon. Wea. Rev.*, 138:2997–3023. 80

Bocquet, M. and Wu, L. (2011). Bayesian design of control space for optimal assimilation of observations. II: Asymptotics solution. *Q. J. Roy. Meteor. Soc.*, 0:0–0. in press. 81, 83

Bocquet, M., Wu, L., and Chevallier, F. (2011). Bayesian design of control space for optimal assimilation of observations. I: Consistent multiscale formalism. *Q. J. Roy. Meteor. Soc.*, 137:1340–1356. 31, 57, 58, 81, 82, 84

Boutahar, J., Lacour, S., Mallet, V., Musson-Genon, L., Quélo, D., Roustan, Y., and Sportisse, B. (2004). Development and validation of a fully modular platform for the numerical modeling of air pollution: POLAIR. *Int. J. of Environ. and Pollution*, 22:17–28. 20

Byrd, R. H., Lu, P., and Nocedal, J. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific and Statistical Computing*, 16:1190–1208. 103

Chao, W. and Chang, L. (1992). Development of a four dimensional variational analysis system using the adjoint method at at GLA. Part 1: Dynamics. *Mon. Wea. Rev.*, 120:1661–1674. 46

Chapnik, B., Desroziers, G., Rabier, F., and Talagrand, O. (2006). Properties and first application of an error-statistics tuning method in variational assimilation. *Q. J. Roy. Meteor. Soc.*, 130:2253–2275. 30, 99

Chevallier, F., Fisher, M., Peylin, P., Serrar, A., Bousquet, P., Bréon, F.-M., Chédin, A., and P., C. (2005). Inferring CO_2 sources and sinks from satellite observations: Method and application to TOVS data. *J. Geophys. Res.*, 110:D24309. 54

Constantinescu, E. M., Sandu, A., and Carmichael, G. R. (2008). Modeling atmospheric chemistry and transport with dynamic adaptive resolution. *Computational Geosciences*, 12(2):133–151. 81

Cressman, G. P. (1959). An operational objective analysis system. *American Meteorological Society*, 87(10):367–374. 38

Curci, G., Palmer, P. I., Kurosu, T. P., Chance, K., and Visconti, G. (2010). Estimating European volatile organic compound emissions using satellite observations of formaldehyde from the Ozone Monitoring Instrument. *Atmos. Chem. Phys.*, 10:11501–11517. 110

Davoine, X. and Bocquet, M. (2007). Inverse modelling-based reconstruction of the Chernobyl source term available for long-range transport. *Atmos. Chem. Phys.*, 7:1549–1564. 44, 77, 79, 84, 99, 102

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., and 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. 77

Desroziers, G. and Ivanov, S. (2001). Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Q. J. Roy. Meteor. Soc.*, 127:1433–1452. 29, 30, 56, 99

Dufour, G., Wittrock, F., Camredon, M., Beekmann, M., Richter, A., Aumont, B., and Burrows, J. P. (2009). SCIAMACHY formaldehyde observations: constraint for isoprene emission estimates over Europe? *Atmos. Chem. Phys.*, 9:1647–1664. 95, 110

Elbern, H., Strunk, A., Schmidt, H., and Talagrand, O. (2007). Emission rate and chemical state estimation by 4-dimensional variational inversion. *Atmos. Chem. Phys.*, 7:3749–3769. 50, 55, 95

Emmons, L. K., Walters, S., Hess, P. G., Lamarque, J.-F., Pfister, G. G., Fillmore, D., Granier, C., Guenther, A., Kinnison, D., Laepple, T., Orlando, J., Tie, X., Tyndall, G., Wiedinmyer, C., Baughcum, S. L., and Kloster, S. (2010). Description and evaluation of the model for Ozone and related chemical tracers, version 4 (MOZART-4). *Geosci. Model Dev.*, 3:43–67. 50, 51, 52

Fagerli, H., Gauss, M., Jonson, J., Nyíri, Á., Simpson, D., Tarrasón, L., Tsyro, S., and Wind, P. (2008). Transboundary acidification, euthrophication and ground level ozone in Europe. Part I: Unified EMEP model description. Technical report, EMEP. 101

Fisher, M. and Leny, D. J. (1995). Lagrangian four-dimensional variational data assimilation of chemical species. *Q. J. Roy. Meteor. Soc.*, 121:1681–1704. 50

Fortems-Cheiney, A., Chevallier, F., Pison, I., Bousquet, F., Carouge, C., Clerbaux, C., Coheur, P.-F., and George, M. (2009). On the capability of IASI measurements to inform about CO surface emissions. *Atmos. Chem. Phys.*, 9:8735–8743. 50, 56

Fortems-Cheiney, A., Chevallier, F., Pison, I., Bousquet, P., Szopa, S., Deeper, N. M., and Clerbaux, C. (2011). Ten years of CO emissions as seen from measurements of pollution in the troposphere (MOPITT). *J. Geophys. Res.*, 116:D05304. 50, 56

Fu, T.-M., Jacob, D., Palmer, P., Chance, K., Wang, Y., Barletta, B., Blake, D., Stanton, J., and Pilling, M. (2007). Space-based formaldehyde measurements as constraints on volatile organic compound emissions in east and south asia and implications for ozone. *J. Geophys. Res.*, 112:D06312. 95

Gandin, L. S. (1963). Objective analysis of meteorological fields. *Leningrad. Hydromet. Press*. Translated from Russian by Israel Program for Scientific Translations. Jerusalem 1965. 38

GENEMIS (1994). Generation of european emission data for episodes (GENEMIS) project. Technical report, EUROTRAC annual report 1993, Garmisch-Partenkirchen, Germany. 53, 101

Goliff, W. S. and Stockwell, W. R. (2008). The regional atmospheric chemistry mechanism, version 2, an update. International con- ference on Atmospheric Chemical Mechanisms, University of California, Davis. 96

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P., and Geron, C. (2006). Estimates of global terrestrial isoprene emissions using megan (Model of Emissions of Gases and Aerosols from Nature). *Atmos. Chem. Phys.*, 6:3181–3210. 53, 110

Guillas, S., Bao, J., Choi, Y., and Wang, Y. (2008). Statistical correction and downscaling of chemical transport model ozone forecasts over atlanta. *Atmos. Env.*, 42:1338–1348. 57

Hall, M. C. G. and Cacuci, D. G. (1983). Physical interpretation of the adjoint functions for sensitivity analysis of atmospheric models. *J. Atmos. Sci.*, 40:2537–2546. 38

Hansen, P. C. (2010). *Discrete Inverse Problems: Insight and Algorithms*. SIAM. 79

Harley, R. and Cass, G. (1995). Modeling the atmospheric concentrations of individual volatile organic compounds. *Atmos. Env.*, 29:905–922. 95

Henne, S., Brunner, D., Folini, D., Solberg, S., Klausen, J., and Buchmann, B. (2010). Assessment of parameters describing representativeness of air quality in-situ measurement sites. *Atmos. Chem. Phys.*, 10:3561–3581. 52

Ho, K., Ho, S., Cheng, Y., Lee, S., and Yu, J. (2007). Real-world emission factors of fifteen carbonyl compounds measured in a hong kong tunnel. *Atmos. Env.*, 41:1747–1758. 95

Hopkins, J. R., Evans, M. J., Lee, J. D., Lewis, A. C., H Marsham, J., McQuaid, J. B., Parker, D. J., Stewart, D., Reeves, C. E., and Purvis, R. M. (2009). Direct estimates of emissions from the meacity of lagos. *Atmos. Chem. Phys.*, 9:8471–8477. 95

Horowitz, L. W., Walters, S., Mauzerall, D. L., Emmons, L. K., Rasch, P. J., Granier, C., Tie, X., Lamarque, J. F., Schultz, M. G., Tyndall, G. S., Orlando, J. J., and Brasseur, G. P. (2003). A global simulation of tropospheric ozone and related tracers: Description and evaluation of MOZART, version 2. *J. Geophys. Res.*, 108:D24. 52, 101

Hourdin, F. and Issartel, J.-P. (2000). Sub-surface nuclear tests monitoring through the ctbt xenon network. *Geophys. Res. Lett.*, 27:2245–2248. 78

Issartel, J.-P. and Baverel, J. (2003). Inverse transport for the verification of the Comprehensive Nuclear Test Ban Treaty. *Atmos. Chem. Phys.*, 3:475–486. 77

Kaipio, J. and Somersalo, E. (2010). *Statistical and Computational Inverse Problems*. Springer. 27

Karamchandani, P., Lohman, K., and Seigneur, C. (2009). Using a sub-grid scale modeling approach to simulate the transport and fate of toxic air pollutants. *Environ. Fluid. Mech.*, 9:59–71. 57

Kim, Y., Sartelet, K., and Seigneur, C. (2009). Comparison of two gas-phase chemical kinetic mechanisms of ozone formation over europe. *Journal of Atmospheric Chemistry*, 62:89–119. 96

Kontarev, G. (1980). The adjoint equation technique applied to meteorological problems. Technical Report 21, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, UK. 38

Koohkan, M. and Bocquet, M. (2012). Accounting for representativeness errors in the inversion of atmospheric constituent emissions: Application to the retrieval of regional carbon monoxide fluxes. *Tellus-B*, 64. 30, 35, 95, 110, 115, 120

Koohkan, M. R., Bocquet, M., Wu, L., and Krysta, M. (2012). Potential of the international monitoring system radionuclide network for inverse modelling. *Atmos. Env.*, 54:557–567. 33, 35, 115

Kopacz, M., Jacob, D. J., Fisher, J. A., Logan, J. A., Zhang, L., Megretskaia, I. A., Yantosca, R. M., Singh, K., Henze, D. K., Burrows, J. P., Buchwitz, M., Khlystova, I., McMillan, W. W., Gille, J. C., Edwards, D. P., Eldering, A., Thouret, V., and Nedelec, P. (2010). Global estimates of CO sources with high resolution by adjoint inversion of multiple satellite datasets MOPITT, AIRS, SCIAMACHY, TES. *Atmos. Chem. Phys.*, 10:855–876. 50, 56

Kopacz, M., Jacob, D. J., Henze, D. K., Heald, C., Streets, D., and Zhang, Q. (2009). A comparison of analytical and adjoint bayesian inversion methods for constraining asian sources of co using satellite (mopitt) measurements of co columns. *J. Geophys. Res.*, 114:D04305. 50

Krysta, M. and Bocquet, M. (2007). Source reconstruction of an accidental radionuclide release at european scale. *Q. J. Roy. Meteor. Soc.*, 133. 38, 77, 80

Krysta, M., Bocquet, M., and Brandt, J. (2008). Probing ETEX-II data set with inverse modelling. *Atmos. Chem. Phys.*, 8:3963–3971. 77, 79, 80

Lahoz, W., Khattatov, B., Ménard, R., and (Eds.) (2010). *Data Assimilation: Making Sense of Observations*. Springer. 24

Larssen, S., Sluyter, R., and Helmis, C. (1999). Criteria for EUROAIRNET: The EEA air quality monitoring and information network. Technical report, European Environment Agency. 52

Le Dimet, F.-X. and Talagrand, O. (1986). Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus A*, 38:97–110. 27, 38, 50

Lorenc, A. (1986). Analysis methods for numerical weather prediction. *Q. J. Roy. Meteor. Soc.*, 112:1177–1194. 27

Mallet, V., Quélo, D., Sportisse, B., Ahmed de Biasi, M., Debry, É., Korsakissok, I., Wu, L., Roustan, Y., Sartelet, K., Tombette, M., and Foudhil, H. (2007). Technical note: The air quality modeling system polyphemus. *Atmos. Chem. Phys.*, 7:5479–5497. 95

Mallet, V. and Sportisse, B. (2006). Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations: An ensemble approach applied to ozone modeling. *J. Geophys. Res.*, 111:D01302. 23

Mason, L. R. and Bohlin, J. B. (1995). Network optimization of a radionuclide monitoring system for the comprehensive nuclear test ban treaty. Technical Report 2585. 25

Ménard, R., Cohn, S. E., Chang, L.-P., and Lyster, P. M. (2000). Assimilation of stratospheric chemical tracer observations using a kalman filter. Part I: Formulation. *Mon. Wea. Rev.*, 128:2654–2671. 30, 51, 55

Michalak, A., Hirsch, A., Bruhwiler, L., Gurney, K., Peters, W., and Tans, P. (2005). Maximum likelihood estimation of covariance parameters for bayesian atmospheric trace gas surface flux inversions. *J. Geophys. Res.*, 110:D24107. 30

Millet, D., Jacob, D., Boersma, K., Fu, T.-M. Kurosu, T., Chance, K., Heald, C., and Guenther, A. (2008). Spatial distribution of isoprene emissions from north america derived from formaldehyde column measurements by the omi satellite sensor. *J. Geophys. Res.*, 113. 95

Mulholland, M. and Seinfeld, J. (1995). Inverse air pollution modelling of urban-scale carbon monoxide emissions. *atmoenv*, 4:497–516. 50

Müller, J.-F. and Stavrakou, T. (2005). Inversion of CO abd NOx emissions using the adjoint of the Image model. *Atmos. Chem. Phys.*, 5:1157–1186. 51

Nappo, C. J., Caneill, J. Y., Furman, R. W., Gifford, F. A., Kaimal, J. C., Kramer, M. L., Lockhart, T. J., Pendergast, M. M., Pielke, R. A., Randerson, D., Shreffler, J. H., and Wyngaard, J. C. (1982). Workshop on the representativeness of meteorological observations. *B. Am. Meteorol.Soc.*, pages 761–764. 52

Nikodym, O. (1933). Sur une classe de fonctions considérée dans l'étude du problème de dirichlet. *Fund. Math.*, 21:129–150. 41

Nodop, K., Connolly, R., and Girardi, F. (1998). The field campaigns of the European Tracer Experiment (ETEX): Overview and results. *Atmos. Env.*, 32:4095–4108. 77

Penkett, S. A., J., B. N., Lightman, P., Marsh, A. R. W., Anwyl, P., and Butcher, G. (1993). The seasonal variation of nonmethane hydrocarbons in the free troposphere over the north atlantic ocean: Possible evidence for extensive reaction of hydrocarbons with the nitrate radical. *J. Geophys. Res.*, 98:2865–2885. 102

Pétron, G., Granier, C., Khattatov, B., Lamarque, J.-F., Yudin, V., Müller, J.-F., and Gille, J. (2002). Inverse modeling of carbon monoxide surface emissions using Climate Monitoring and Diagnostics Laboratory network observations. *J. Geophys. Res.*, 107:4761. 50, 56

Pétron, G., Granier, C., Khattatov, B., Yudin, V., Lamarque, J.-F., Emmons, L., Gille, J., and Edwards, D. P. (2004). Monthly CO surface sources inventory based on the 2000-2001 MOPITT satellite data. *Geophys. Res. Lett.*, 31:L21107. 56

Peylin, P., Bousquet, P., and Ciais, P. (2001). Inverse modeling of atmospheric carbon dioxid fluxes - response. *Science*, 294:2292–2292. 84

Politis, K. and Robertson, L. (2004). Bayesian updating of atmospheric dispersion after a nuclear accident. *Appl. Statist.*, 53:583–600. 77

Pudykiewicz, J. A. (1998). Application of adjoint transport tracer equations for evaluating source parameters. *Atmos. Env.*, 32:3039–3050. 77

Quélo, D., Krysta, M., Bocquet, M., Isnard, O., Minier, Y., and Sportisse, B. (2007). Validation of the Polyphemus platform on the ETEX, Chernobyl and Algeciras cases. *Atmos. Env.*, 41:5300–5315. 20, 52, 89

Quélo, D., Mallet, V., and Sportisse, B. (2005). Inverse modeling of nox emissions at regional scale over northern france: Preliminary investigation of the second-order sensitivity. *J. Geophys. Res.*, 110:D24310. 95

Rigby, M., Manning, A. J., and Prinn, R. G. (2011). Inversion of long-lived trace gas emissions using combined Eulerian and Lagrangian chemical transport models. *Atmos. Chem. Phys. Discuss.*, 11:14689–14717. 84

Ringbom, A. and Miley, H. (2009). *Radionuclide Monitoring. Science for Security: Verifying the Comprehensive Nuclear-Test-Ban Treaty*, pages 23–28. Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization, Vienna, Austria. 76, 78

Robertson, L. and Langner, J. (1998). Source function estimate by means of adjoint variational data assimilation applied to the ETEX-I tracer experiment. *Atmos. Env.*, 32(24):4219–4225. 77

Rodgers, C. (2000). *Inverse Methods for Atmospheric Sounding: Theory and Practice*. World Scientific, Series on Atmospheric, Oceanic and Planetary Physics: Volume 2. 27, 80, 81

Roustan, Y. and Bocquet, M. (2006a). Inverse modeling for mercury over europe. *Atmos. Chem. Phys.*, 6:3085–3098. 39, 44

Roustan, Y. and Bocquet, M. (2006b). Sensitivity analysis for mercury over europe. *J. Geophys. Res.*, 111:D14304. 97

Rudolph, J. and Ehhalt, D. (1981). Measurements of c2-c5 hydrocarbons over the north atlantic. *J. Geophys. Res.*, 86:11959–11964. 102

Rudolph, J. and Johnen, F. J. (1990). Measurements of light atmospheric hydrocarbons over the atlantic in regions of low biological activity. *J. Geophys. Res.*, 95:20583–20591. 102

Saide, P., Bocquet, M., Osses, A., and Gallardo, L. (2011). Constraining surface emissions of air pollutants using inverse modeling: method intercomparison and a new two-step multi-scale approach. *Tellus B*, 63:360–370. 50, 79

Sartelet, K., Couvidat, F., Seigneur, C., and Roustan, Y. (2012). Impact of biogenic emissions on air quality over europe and north america. *Atmos. Env.*, 53:131–141. 110

Sartelet, K., Debry, E., Fahey, K., Roustan, Y., Tombette, M., and Sportisse, B. (2007). Simulation of aerosols and gas-phase species over europe with the polyphemus system. part i: model-to-data comparison for 2001. *Atmos. Env.*, 41:6116–6131. 95

Sasaki, Y. (1958). An objective analysis based on the variational method. *J. Meteor. Soc. Japan*, pages 77–88. 27

Sawyer, R., Harley, R., Cadle, S., Norbeck, J., Slott, R., and Bravo, H. (2000). Mobile sources critical review: 1998 narsto assessment. *Atmos. Env.*, 34:2161–2181. 95

Seibert, P. and Stohl, A. (2000). Inverse modelling of the ETEX-1 release with a Lagrangian particle model. In *Proceedings of the Third GLOREAM Workshop, September 1999, Ischia, Italy*. 77

Shim, C., Wang, Y., Choi, Y., Palmer, P., Abbot, D., and Chance, K. (2005). Constraining global isoprene emissions with global ozone monitoring experiment (gome) formaldehyde column measurements. *J. Geophys. Res.*, 110:D24301. 95

Simpson, D., Winiwarter, W., Börjesson, G., Cinderby, S., Ferreiro, A., Guenther, A., Hewitt, C., Janson, R., Khalil, M., Owen, S., Pierce, T., Puxbaum, H., Shearer, M., Skiba, U., Steinbrecher, R., Tarrason, L., and Oquist, M. (1999). Inventorying emissions from nature in Europe. *J. Geophys. Res.*, 104(D7):8113–8152. 101, 110

Staehelin, J., Keller, C., Stahel, W., Schläpfer, K., and Wunderli, S. (1998). Emission factors from road traffic from a tunnel study (gubris tunnel, switzerland), part iii: results of organic compounds, so2 and speciation of organic exhaust emission. *Atmos. Env.*, 32:999–1009. 95

Stavrakou, T. and Müller, J.-F. (2006). Grid-based versus big region approach for inverting co emissions using measurement of pollution in the troposphere (mopitt) data. *jgr*, 111:D15304. 50

Stemmler, K., Bugmann, S., Buchmann, B., Reimann, S., and Staehelin, J. (2005). Large decrease of voc emissions of switzerland's car fleet during the past decade: results from a highway tunnel study. *Atmos. Env.*, 39:1009–1018. 95

Stohl, A., Forster, C., Frank, A., Seibert, P., and Wotawa, G. (2005). Technical note : The lagrangian particle dispersion model FLEXPART version 6.2. *Atmos. Chem. Phys.*, 5:2461–2474. 20, 22, 83

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. SIAM. 30

Tarrasón, L., Fagerli, H., Jonson, J., Simpson, D., Benedictow, A., Klein, H., and Vestreng, V. (2007). Transboundary acidification, euthrophication and ground level ozone in Europe in 2005. Technical report, EMEP. 101

Tikhonov, A. and Arsenin, V. (1977). *Solutions of Ill Posed Problems*. Winston, Washington DC. 28

Touaty, M. and Bonsang, B. (2000). Hydrocarbon emissions in a highway tunnel in the paris area. *Atmos. Env.*, 34:985–996. 95

United Nations (1996). *Protocol of the Comprehensive Nuclear-Test-Ban Treaty*. Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization, Vienna, Austria. 76

United Nations (2000). United nations: Sources and effects of ionizing radiation. United Nations scientific commitee on the effects of atomic radiation. Report to general assembly. Technical report, United Nations, New-York. 77

Vestreng, V., Myhre, G., Fagerli, H., Reis, S., and Terrasón, L. (2007). Twenty-five years of continuous sulphir dioxide emission reduction in europe. *Atmos. Chem. Phys.*, 7:3663–3681. 67

Vijayaraghavan, K., Snell, H., and Seigneur, C. (2008). Practical aspects of using satellite data in air quality modeling. *Environ. Sci. Technol.*, 42:8187–8192. 95

Vogel, C. R. (2002). *Computational Methods for Inverse Problems*. SIAM, Frontiers in Applied Mathematics. 79

Winiarek, V., M. Bocquet, M., Saunier, O., and 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.*, pages in–press. 99, 100

Winiarek, V., Vira, J., Bocquet, M., Sofiev, M., and 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. 31, 78, 79

Wotawa, G., De Geer, L.-E., Denier, P., Kalinowski, M., Toivonen, H., D'Amours, R., Desiato, F., Issartel, J.-P., Langer, M., Seibert, P., Frank, A., Sloan, C., and Yamazawa, H. (2003). Atmospheric transport modelling in support of CTBT verification–overview and basic concepts. *Atmos. Env.*, 37:2529–2537. 78

Wu, L. and Bocquet, M. (2011). Optimal redistribution of the background ozone monitoring stations over France. *Atmos. Env.*, 45:772–783. 78

Wu, L., Bocquet, M., Lauvaux, T., Chevallier, F., Rayner, P., and Davis, K. (2011). Optimal representation of source-sink fluxes for mesoscale carbon dioxide inversion with synthetic data. *J. Geophys. Res.*, 116:D21304. 33, 58, 80

Xiao, Y., Logan, J., Jacob, D., Hudman, R., Yantosca, R., and Blake, D. (2008). Global budget of ethane and regional constraints on u.s. sources. *J. Geophys. Res.*, 113:D21306. 95

Yee, E., Lien, F.-S., Keats, A., and 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. 77

Yumimotoa, K. and Uno, I. (2006). Adjoint inverse modeling of CO emissions over Eastern Asia using four-dimensional variational data assimilation. *Atmos. Env.*, 40:6836–6845. 50, 56

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., and Baklanov, A. (2012). Real-time air quality forecasting, part ii: State of the science, current research needs, and future prospects. *Atmos. Env.* 50, 71

Zou, X., Vandenberghe, F., Pondeca, M., and Kuo, Y.-H. (1997). Introduction to adjoint techniques and the MM5 adjoint modeling system. Technical report, NCAR. 46

Appendix A

A posteriori formalism of the cost function

As explained in the chapter of introduction, it is important to reformulate the cost function $J(\mathbf{x})$, with the help of statistical information on the a posteriori values of the model parameters.

The background term of the cost function $J_b(\mathbf{x})$ includes the statistical information about the first guess of the model parameters. First of all, the following cost function is assumed:

$$J(\mathbf{x}) = (\boldsymbol{\mu} - \mathbf{H}\mathbf{x})^\dagger \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}) + (\mathbf{x} - \mathbf{x}_b)^\dagger \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_b) \tag{A.1}$$

The first term on the right hand side of this function can be written as:

$$\begin{aligned} (\boldsymbol{\mu} - \mathbf{H}\mathbf{x})^\dagger \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}) &= (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b)^\dagger \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b) \\ &\quad + (\mathbf{x} - \mathbf{x}_b)^\dagger \mathbf{H}^\dagger \mathbf{R}^{-1} \mathbf{H} (\mathbf{x} - \mathbf{x}_b) \\ &\quad - (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b)^\dagger \mathbf{R}^{-1} \mathbf{H} (\mathbf{x} - \mathbf{x}_b) \\ &\quad - (\mathbf{x} - \mathbf{x}_b)^\dagger \mathbf{H}^\dagger \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b) \,. \end{aligned} \tag{A.2}$$

When using the inverse of the posterior covariance matrix of the model parameters, it comes:

$$\mathbf{P}_a^{-1} = \mathbf{B}^{-1} + \mathbf{H}^\dagger \mathbf{R}^{-1} \mathbf{H} \,. \tag{A.3}$$

Finally, one can write the cost function as follows:

$$\begin{aligned} J(\mathbf{x}) &= (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b)^\dagger \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b) + (\mathbf{x} - \mathbf{x}_b)^\dagger \mathbf{P}_a^{-1} (\mathbf{x} - \mathbf{x}_b) \\ &\quad - (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b)^\dagger \mathbf{R}^{-1} \mathbf{H} (\mathbf{x} - \mathbf{x}_b) - (\mathbf{x} - \mathbf{x}_b)^\dagger \mathbf{H}^\dagger \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b) \,. \end{aligned} \tag{A.4}$$

Now, using the BLUE analysis (Eq. (1.28)), the the second term on the right hand side of the equation (A.4) can be written as:

$$\begin{aligned} (\mathbf{x} - \mathbf{x}_b)^\dagger \mathbf{P}_a^{-1} (\mathbf{x} - \mathbf{x}_b) &= (\mathbf{x} - \mathbf{x}_a)^\dagger \mathbf{P}_a^{-1} (\mathbf{x} - \mathbf{x}_a) \\ &\quad + (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b)^\dagger \mathbf{K}^\dagger \mathbf{P}_a^{-1} \mathbf{K} (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b) \\ &\quad + (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b)^\dagger \mathbf{R}^{-1} \mathbf{H} (\mathbf{x} - \mathbf{x}_a) \\ &\quad + (\mathbf{x} - \mathbf{x}_a)^\dagger \mathbf{H}^\dagger \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b) \,. \end{aligned} \tag{A.5}$$

Therefore, it comes:

$$\begin{aligned} J(\mathbf{x}) &= (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b)^\dagger (\mathbf{R}^{-1} + \mathbf{K}^\dagger \mathbf{P}_a^{-1} \mathbf{K}) (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b) + (\mathbf{x} - \mathbf{x}_a)^\dagger \mathbf{P}_a^{-1} (\mathbf{x} - \mathbf{x}_a) \\ &\quad - (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b)^\dagger \mathbf{R}^{-1} \mathbf{H} (\mathbf{x}_b - \mathbf{x}_a) - (\mathbf{x}_b - \mathbf{x}_a)^\dagger \mathbf{H}^\dagger \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b) \,. \end{aligned} \tag{A.6}$$

The BLUE analysis can be still formulated as:

$$\mathbf{x}_b - \mathbf{x}_a = -\mathbf{K}(\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b)\,. \tag{A.7}$$

Finally, the following cost function is detailed, this time, using the BLUE analysis,

$$\begin{aligned} J(\mathbf{x}) \;=\; & (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b)^\dagger(\mathbf{R}^{-1} + \mathbf{K}^\dagger\mathbf{P}_a^{-1}\mathbf{K} - \mathbf{R}^{-1}\mathbf{H}\mathbf{K} - \mathbf{K}^\dagger\mathbf{H}^\dagger\mathbf{R}^{-1})(\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b) \\ & +(\mathbf{x} - \mathbf{x}_a)^\dagger\mathbf{P}_a^{-1}(\mathbf{x} - \mathbf{x}_a)\,. \end{aligned} \tag{A.8}$$

The gain matrix, $\mathbf{K}$, can be derived as a function of $\mathbf{B}$ and $\mathbf{R}$ (Eq. (1.29)). Then,

$$\mathbf{R}^{-1} + \mathbf{K}^\dagger\mathbf{P}_a^{-1}\mathbf{K} - \mathbf{R}^{-1}\mathbf{H}\mathbf{K} - \mathbf{K}^\dagger\mathbf{H}^\dagger\mathbf{R}^{-1} = (\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^\dagger)^{-1}. \tag{A.9}$$

This leads to the following from of the cost function.

$$J(\mathbf{x}) = (\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b)^\dagger(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^\dagger)^{-1}(\boldsymbol{\mu} - \mathbf{H}\mathbf{x}_b) + (\mathbf{x} - \mathbf{x}_a)^\dagger\mathbf{P}_a^{-1}(\mathbf{x} - \mathbf{x}_a)\,. \tag{A.10}$$

Appendix B

Carbon monoxide scores for each of the stations

B.1 Simulation scores

Table B.1: Comparison of the observations and the simulated concentrations.

	N	$\overline{C}$	$\overline{O}$	RMSE	R	FA_2	FA_5
HAYANGE (industrial)	1273	343	584.3	545.5	0.28	0.51	0.91
Brest 3 CDM (traffic)	954	256.5	649.6	621.6	-0.01	0.44	0.87
SAMONZET (urban)	780	222.8	673.4	562.8	-0.04	0.24	0.88
TOULON FOCH (urban)	1330	239.1	995.9	1123	0.07	0.25	0.72
ALEXIS CARREL ROUENG (urban)	1275	295	356.3	263.3	0.25	0.72	0.95
Jardin Lecoq CF (urban)	1137	236.4	283.9	478.9	-0.09	0.36	0.69
VICTOR HUGO (traffic)	1320	271.6	367.3	465.3	0.16	0.39	0.75
Roubaix/Serres (traffic)	1244	357.6	539.3	378.2	0.47	0.71	0.98
Cherbourg Paul Doume (urban)	1341	244	342.1	226.7	0	0.81	0.99
Station MARSANNAY (urban)	1285	280	437.1	282.4	0.16	0.73	0.99
Saint Denis (traffic)	1341	252	895.7	903.2	0.28	0.29	0.79
Mirabeau (traffic)	1340	276.7	754.5	667.3	0.08	0.35	0.89
Grenoble Foch (traffic)	1323	247	1064	1053	0.14	0.15	0.7
Muhl.ASPA (traffic)	1308	298.5	451.5	425.7	0.16	0.68	0.96
Hotel de ville (urban)	1296	296.8	428.9	571.5	0.09	0.81	0.99
MERIGNAC (traffic)	781	260.9	581.9	477	0.19	0.53	0.96
Nice Pellos (urban)	1242	233.8	2156	2500	0.04	0.06	0.34
Ecole Jules Ferry (urban)	578	273.7	520.2	410.5	-0.04	0.63	0.96
GONESSE (periurban)	1073	472.3	554.7	293.7	0.48	0.86	0.99
Liane Boulogne Sud (traffic)	1141	270.6	367.8	351.3	0.32	0.51	0.89
VAUCELLES (urban)	1243	262.1	585	549.3	0.04	0.45	0.89
COUBERTIN (periurban)	1331	243.3	312.9	239.7	0	0.7	0.95
Gambeta (traffic)	1333	286.1	715.9	635.4	0.04	0.39	0.92
ANGLET (urban)	710	228.5	602.7	589.5	-0.08	0.4	0.89
Bar-le-Duc (urban)	1231	285	521.4	450.7	-0.02	0.65	0.96
Aurillac Centre (traffic)	1327	228.3	698.1	833.2	-0.09	0.35	0.78
CTRE VILLE MEGEVAND (traffic)	1337	266.5	457.8	367.9	0.13	0.62	0.95
LA ROE (traffic)	1316	267	755.8	668.3	0.05	0.32	0.9
AVIGNON ROCADE (traffic)	1341	317.7	683	658.7	0.51	0.57	0.96

	N	$\overline{C}$	$\overline{O}$	RMSE	R	FA_2	FA_5
Epinal (urban)	1338	275.6	431.6	253.1	0.21	0.8	0.99
Macon Paul Bert (periurban)	1292	287.7	544.1	431.1	0.28	0.69	0.98
Rue de la Tour (urban)	1307	279.3	435.7	331.2	0.07	0.78	0.99
Strasbourg Clemenceau (traffic)	1164	308.3	816.4	700.5	0.1	0.36	0.93
Amiens Saint Leu ()	584	299.4	511.1	382.9	0.29	0.77	0.98
Forbach (12) (urban)	1319	319.6	354.8	394.5	0.29	0.57	0.85
Place du Marche (urban)	1319	265.3	560.4	494.2	-0.09	0.56	0.94
Mairie MALO (traffic)	975	387.3	461.5	498.9	-0.06	0.68	0.88
Halles centralles (traffic)	1344	272.4	609.8	585.2	0.02	0.49	0.9
GENERAL DE GAULLE (traffic)	1285	264.5	597.4	568.2	-0.07	0.5	0.92
Valence traffic (traffic)	1329	260.6	462.9	359.1	-0.02	0.65	0.96
place de VERDUN (urban)	102	195.7	522.6	741.8	0.45	0.62	0.89
Le Puy Fayolle (traffic)	1222	221.7	831.2	880.8	-0.12	0.24	0.78
Le Creusot Molette (periurban)	1295	263.6	457	454.4	0.1	0.72	0.96
Marignane Ville (urban)	1341	273.9	527.7	444	0.3	0.65	0.96
Hotel Districal (urban)	1330	293.6	562.7	436.8	-0.08	0.64	0.97
METZ-BORNY (urban)	786	343.1	248	243	0.22	0.49	0.86
Port de Bouc EDF (urban)	1312	294.1	459.4	378.1	0.44	0.61	0.98
PLOMBIERES (traffic)	1342	260.8	1463	1500	0.17	0.13	0.49
AIX CENTRE (traffic)	1338	281.4	595.5	490	0.24	0.55	0.95
Place Victor Basch (traffic)	1316	447.1	1642	1463	0.22	0.23	0.72
AUBERVILLIERS (urban)	1203	505.3	408	332.8	0.35	0.68	0.95
Avenue des Champs Elysees (traffic)	1331	445.6	933.6	739.3	0.03	0.53	0.94
Boulevard peripherique Auteuil (traffic)	1326	423.7	1392	1164	0.29	0.23	0.86
PARIS 1er Les Halles (urban)	1304	480.5	435.9	229.8	0.35	0.85	0.99
Autoroute A1 - Saint-Denis (traffic)	1313	481.4	1255	970.3	0.32	0.34	0.93
Rue Bonaparte (traffic)	1267	473.3	896.2	642.9	0.25	0.58	0.99
Quai des Celestins (traffic)	1333	480.4	1420	1193	0.29	0.29	0.86
LeHavre Republique (traffic)	1333	259.1	582.6	580.7	0.03	0.39	0.89
ESQUERCHIN A DOUAI (traffic)	1340	327.3	555.8	396.4	0.32	0.7	0.99
Rousillon (traffic)	1254	240.3	553.6	512.1	-0.07	0.39	0.89
Pres Arenes (urban)	1323	250.2	518.6	581.3	0.31	0.63	0.95
Planas (traffic)	1308	283.5	869.3	925.8	0.28	0.35	0.84
rue de la GRILLE (traffic)	1320	240.1	1022	1054	-0.1	0.21	0.68
FORT-MARDYCK (industrial)	990	385.1	374.6	422.7	-0.1	0.49	0.88
Petite Synthe (periurban)	1130	370.7	384.1	450.3	0.08	0.62	0.96
Calais Centre (traffic)	1330	347.3	478.3	307.6	0.46	0.77	0.99
LIBERTE (traffic)	1341	353.3	603	540.8	0.45	0.72	0.96
PASTEUR (traffic)	1338	355.1	494.7	289.9	0.49	0.81	1
LA BASSEE/CENTRE (traffic)	1337	327.4	413.1	281	0.49	0.76	0.98
Le Rondeau (traffic)	1315	245.9	793.7	707.5	0.02	0.23	0.84
LAENNEC (traffic)	605	277	679.1	634.5	0.08	0.5	0.9
PUITS GAILLOT (traffic)	1340	318.7	815.2	702.2	0.21	0.39	0.93
BERTHELOT (traffic)	1340	315.3	1004	970.2	0.36	0.36	0.82
GARIBALDI (traffic)	1339	321.7	1407	1426	0.22	0.18	0.67
LA MULATIERE (traffic)	1341	308.8	829.6	693.1	0.34	0.35	0.92
Batiment ELF-ATO (industrial)	1075	328.1	257.3	196	0.24	0.62	0.92
ANTIBES GUYNEMER (urban)	1275	233.4	1279	1363	0.11	0.05	0.54
Rouen Le Conquerant (traffic)	1340	292.6	503.6	382.7	0.25	0.62	0.97

	N	$\overline{C}$	$\overline{O}$	RMSE	R	FA_2	FA_5
Pasteur (urban)	1227	287.6	628.9	552.5	0.11	0.53	0.94
station DAIX (periurban)	1329	281.3	307.3	195.5	0.06	0.71	0.96
BETHUNE PROX AUTO (traffic)	1293	305.3	509.6	383.3	0.26	0.7	0.99
ST ETIENNE ROND PT (traffic)	1294	243.6	597.7	502.3	0.02	0.39	0.94
RIVE DE GIER (traffic)	1312	268.4	457	333.2	0.25	0.59	0.96
Luneville (urban)	1315	285	559.4	427.9	0.16	0.61	0.97
BORDEAUX-BASTIDE (traffic)	654	262.3	405.1	299.4	0.24	0.71	0.95
Chalon centre ville (traffic)	1265	277	691.9	616	0.11	0.39	0.93
Champforgeuil (periurban)	1154	276.5	301.3	230.4	0.25	0.63	0.83
Hilaire Chardonnet (periurban)	1185	277.8	406.1	240.6	0.14	0.83	0.99
Montceau-les-Mines (periurban)	1324	260.8	443.7	304.2	0.12	0.72	0.99

B.2 4D-Var scores

Table B.2: Comparison of the observations and the simulated concentrations diagnosed by the 4D-Var system.

	N	$\overline{C}$	$\overline{O}$	RMSE	R	FA_2	FA_5
HAYANGE (industrial)	1273	363.5	584.3	528.3	0.36	0.53	0.92
Brest 3 CDM (traffic)	954	257.2	649.6	620.8	-0.01	0.44	0.87
SAMONZET (urban)	780	215.1	673.4	569.1	-0.06	0.23	0.86
TOULON FOCH (urban)	1330	261.4	995.9	1108	0.07	0.28	0.74
ALEXIS CARREL ROUENG (urban)	1275	308.6	356.3	254.9	0.33	0.73	0.96
Jardin Lecoq CF (urban)	1137	239.5	283.9	474.2	0.01	0.35	0.69
VICTOR HUGO (traffic)	1320	277.3	367.3	458.9	0.25	0.4	0.76
Roubaix/Serres (traffic)	1244	440.8	539.3	295.4	0.69	0.83	0.98
Cherbourg Paul Doume (urban)	1341	245.7	342.1	225.1	0.02	0.81	0.99
Station MARSANNAY (urban)	1285	304.9	437.1	259.9	0.35	0.78	0.99
Saint Denis (traffic)	1341	271	895.7	873.4	0.48	0.3	0.85
Mirabeau (traffic)	1340	288.3	754.5	657.1	0.13	0.39	0.91
Grenoble Foch (traffic)	1323	273.3	1064	1023	0.35	0.16	0.78
Muhl.ASPA (traffic)	1308	304	451.5	423.9	0.16	0.69	0.96
Hotel de ville (urban)	1296	311.1	428.9	567.2	0.11	0.83	0.99
MERIGNAC (traffic)	781	269.6	581.9	466.7	0.29	0.55	0.96
Nice Pellos (urban)	1242	304.4	2156	2422	0.32	0.07	0.39
Ecole Jules Ferry (urban)	578	279	520.2	406.6	-0.03	0.64	0.96
GONESSE (periurban)	1073	749.3	554.7	317.1	0.71	0.82	0.98
Liane Boulogne Sud (traffic)	1141	283.5	367.8	340.7	0.4	0.53	0.9
VAUCELLES (urban)	1243	270.6	585	541.4	0.12	0.47	0.9
COUBERTIN (periurban)	1331	253.7	312.9	229.8	0.19	0.71	0.95
Gambeta (traffic)	1333	305.3	715.9	619.7	0.11	0.46	0.93
ANGLET (urban)	710	226.2	602.7	590.6	-0.07	0.4	0.89
Bar-le-Duc (urban)	1231	294.3	521.4	445.2	0	0.66	0.97
Aurillac Centre (traffic)	1327	240.2	698.1	824.7	-0.03	0.35	0.8
CTRE VILLE MEGEVAND (traffic)	1337	275.9	457.8	362.9	0.13	0.63	0.95
LA ROE (traffic)	1316	272.5	755.8	662.7	0.1	0.33	0.91
AVIGNON ROCADE (traffic)	1341	385.9	683	580.3	0.61	0.69	0.99
Epinal (urban)	1338	288.4	431.6	245.4	0.21	0.82	0.99
Macon Paul Bert (periurban)	1292	364.3	544.1	358.5	0.5	0.81	0.99
Rue de la Tour (urban)	1307	287.3	435.7	325.4	0.11	0.81	0.99

	N	$\overline{C}$	$\overline{O}$	RMSE	R	FA_2	FA_5
Strasbourg Clemenceau (traffic)	1164	324.8	816.4	684.7	0.19	0.4	0.94
Amiens Saint Leu ()	584	317.3	511.1	366.9	0.35	0.8	0.99
Forbach (12) (urban)	1319	329.3	354.8	393.8	0.29	0.56	0.85
Place du Marche (urban)	1319	274.3	560.4	487	-0.03	0.58	0.94
Mairie MALO (traffic)	975	417.8	461.5	491.3	0.01	0.68	0.89
Halles centralles (traffic)	1344	288.6	609.8	564.1	0.25	0.51	0.93
GENERAL DE GAULLE (traffic)	1285	270.3	597.4	563.9	-0.04	0.51	0.93
Valence traffic (traffic)	1329	305.1	462.9	341.3	-0.03	0.71	0.97
place de VERDUN (urban)	102	196.7	522.6	741.7	0.45	0.6	0.89
Le Puy Fayolle (traffic)	1222	227.8	831.2	873.4	0.01	0.24	0.79
Le Creusot Molette (periurban)	1295	279.5	457	442.6	0.21	0.75	0.97
Marignane Ville (urban)	1341	329.1	527.7	401.7	0.41	0.75	0.98
Hotel Districal (urban)	1330	304.3	562.7	426.2	0.01	0.67	0.97
METZ-BORNY (urban)	786	357.3	248	244.7	0.29	0.47	0.86
Port de Bouc EDF (urban)	1312	397.5	459.4	344.3	0.35	0.62	0.98
PLOMBIERES (traffic)	1342	297.9	1463	1461	0.3	0.15	0.57
AIX CENTRE (traffic)	1338	313.7	595.5	465.5	0.29	0.62	0.97
Place Victor Basch (traffic)	1316	879.1	1642	986.2	0.73	0.6	1
AUBERVILLIERS (urban)	1203	1041	408	760.3	0.57	0.28	0.83
Avenue des Champs Elysees (traffic)	1331	876.6	933.6	474.9	0.55	0.81	1
Boulevard peripherique Auteuil (traffic)	1326	791.5	1392	769.8	0.71	0.65	1
PARIS 1er Les Halles (urban)	1304	1003	435.9	703.1	0.63	0.41	0.96
Autoroute A1 - Saint-Denis (traffic)	1313	933.7	1255	532.9	0.73	0.85	1
Rue Bonaparte (traffic)	1267	970.8	896.2	380.3	0.72	0.91	1
Quai des Celestins (traffic)	1333	1002	1420	708.8	0.67	0.8	1
LeHavre Republique (traffic)	1333	265	582.6	576	0.07	0.44	0.9
ESQUERCHIN A DOUAI (traffic)	1340	361.9	555.8	367.8	0.41	0.74	0.99
Rousillon (traffic)	1254	243.3	553.6	507.9	0	0.39	0.9
Pres Arenes (urban)	1323	268.7	518.6	560	0.39	0.66	0.96
Planas (traffic)	1308	328.3	869.3	869	0.45	0.42	0.89
rue de la GRILLE (traffic)	1320	242.3	1022	1052	-0.07	0.21	0.69
FORT-MARDYCK (industrial)	990	414.2	374.6	424.7	-0.05	0.48	0.87
Petite Synthe (periurban)	1130	399.2	384.1	454.3	0.05	0.57	0.96
Calais Centre (traffic)	1330	410.5	478.3	254.9	0.63	0.85	0.99
LIBERTE (traffic)	1341	452.4	603	452	0.61	0.79	0.99
PASTEUR (traffic)	1338	462.8	494.7	202.3	0.72	0.93	1
LA BASSEE/CENTRE (traffic)	1337	365.4	413.1	248	0.62	0.79	0.98
Le Rondeau (traffic)	1315	274	793.7	676.8	0.29	0.26	0.89
LAENNEC (traffic)	605	293.2	679.1	613.4	0.29	0.53	0.92
PUITS GAILLOT (traffic)	1340	826.4	815.2	529.1	0.51	0.78	0.98
BERTHELOT (traffic)	1340	773.7	1004	589.1	0.64	0.77	0.99
GARIBALDI (traffic)	1339	860.1	1407	893	0.66	0.63	0.98
LA MULATIERE (traffic)	1341	690.5	829.6	435.4	0.57	0.85	0.99
Batiment ELF-ATO (industrial)	1075	338.3	257.3	199.4	0.25	0.62	0.92
ANTIBES GUYNEMER (urban)	1275	282.2	1279	1309	0.32	0.08	0.67
Rouen Le Conquerant (traffic)	1340	305.9	503.6	372	0.3	0.64	0.98
Pasteur (urban)	1227	316.3	628.9	535.8	0.1	0.57	0.96
station DAIX (periurban)	1329	306.5	307.3	183.7	0.28	0.73	0.95
BETHUNE PROX AUTO (traffic)	1293	325.4	509.6	364.3	0.35	0.73	0.99

	N	$\overline{C}$	$\overline{O}$	RMSE	R	FA_2	FA_5
ST ETIENNE ROND PT (traffic)	1294	255.3	597.7	490.6	0.14	0.46	0.96
RIVE DE GIER (traffic)	1312	312.2	457	300.5	0.35	0.67	0.97
Luneville (urban)	1315	296.1	559.4	420.8	0.16	0.63	0.98
BORDEAUX-BASTIDE (traffic)	654	274.6	405.1	288.4	0.33	0.74	0.96
Chalon centre ville (traffic)	1265	309.9	691.9	588.1	0.22	0.48	0.96
Champforgeuil (periurban)	1154	311.1	301.3	228.4	0.27	0.63	0.83
Hilaire Chardonnet (periurban)	1185	312.5	406.1	216.3	0.32	0.89	1
Montceau-les-Mines (periurban)	1324	274.1	443.7	292.2	0.24	0.75	0.99

B.3 4D-Var-ξ scores

Table B.3: Comparison of the observations and the simulated concentrations diagnosed by the 4D-Var-ξ system.

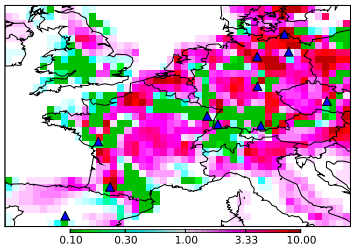
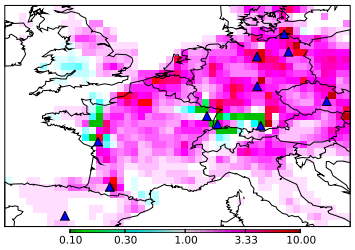
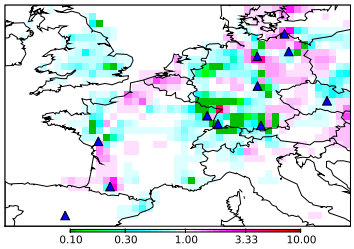
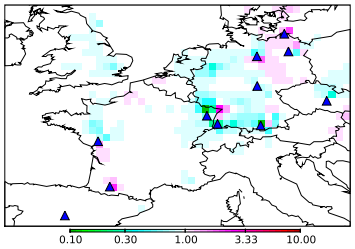
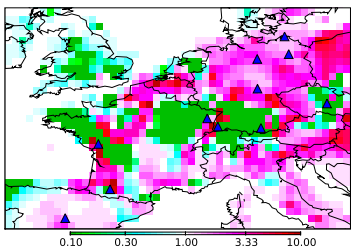
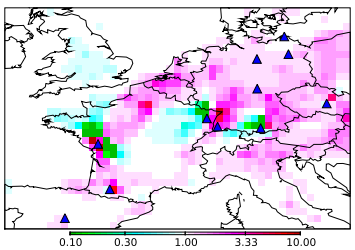
	N	$\overline{C}$	$\overline{O}$	RMSE	R	FA_2	FA_5
HAYANGE (industrial)	1273	593.3	584.3	449	0.49	0.59	0.89
Brest 3 CDM (traffic)	954	650.6	649.6	388.4	0.57	0.82	0.99
SAMONZET (urban)	780	645.3	673.4	261.2	0.63	0.94	0.99
TOULON FOCH (urban)	1330	988.6	995.9	695.2	0.55	0.69	0.95
ALEXIS CARREL ROUENG (urban)	1275	366.9	356.3	241.1	0.39	0.72	0.95
Jardin Lecoq CF (urban)	1137	321.8	283.9	462.8	0.23	0.31	0.67
VICTOR HUGO (traffic)	1320	438.7	367.3	396.3	0.57	0.42	0.74
Roubaix/Serres (traffic)	1244	545.3	539.3	273.2	0.69	0.84	0.97
Cherbourg Paul Doume (urban)	1341	356.6	342.1	186	0.36	0.9	1
Station MARSANNAY (urban)	1285	444.8	437.1	217.4	0.41	0.87	0.99
Saint Denis (traffic)	1341	888.4	895.7	536	0.57	0.78	1
Mirabeau (traffic)	1340	762.8	754.5	372.2	0.61	0.84	0.99
Grenoble Foch (traffic)	1323	1062	1064	599.3	0.45	0.82	1
Muhl.ASPA (traffic)	1308	477.2	451.5	367.6	0.42	0.74	0.96
Hotel de ville (urban)	1296	461.2	428.9	521.5	0.35	0.84	1
MERIGNAC (traffic)	781	545.6	581.9	278.7	0.64	0.9	1
Nice Pellos (urban)	1242	2126	2156	1176	0.68	0.82	0.99
Ecole Jules Ferry (urban)	578	402	520.2	323.4	0.35	0.84	1
GONESSE (periurban)	1073	538.1	554.7	240.7	0.66	0.9	0.99
Liane Boulogne Sud (traffic)	1141	401.5	367.8	309.2	0.57	0.55	0.85
VAUCELLES (urban)	1243	597.2	585	376.9	0.53	0.71	0.93
COUBERTIN (periurban)	1331	320.8	312.9	217.4	0.28	0.7	0.95
Gambeta (traffic)	1333	730.4	715.9	349.5	0.67	0.86	0.99
ANGLET (urban)	710	549	602.7	378	0.56	0.81	0.98
Bar-le-Duc (urban)	1231	530.6	521.4	339.5	0.45	0.79	0.99
Aurillac Centre (traffic)	1327	722	698.1	576.3	0.54	0.6	0.93
CTRE VILLE MEGEVAND (traffic)	1337	468	457.8	296.5	0.35	0.72	0.95
LA ROE (traffic)	1316	758	755.8	378.4	0.56	0.88	1
AVIGNON ROCADE (traffic)	1341	702.6	683	456.9	0.67	0.81	0.99
Epinal (urban)	1338	440.1	431.6	181.2	0.48	0.96	1
Macon Paul Bert (periurban)	1292	548	544.1	318.3	0.48	0.84	1
Rue de la Tour (urban)	1307	443.1	435.7	270.7	0.35	0.87	1
Strasbourg Clemenceau (traffic)	1164	813.7	816.4	382.3	0.62	0.9	1
Amiens Saint Leu ()	584	421.3	511.1	309.8	0.5	0.93	0.99

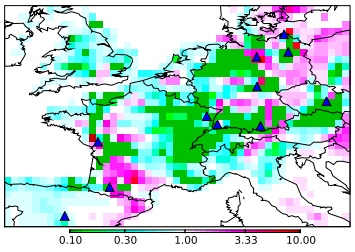
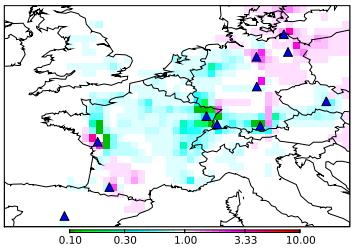
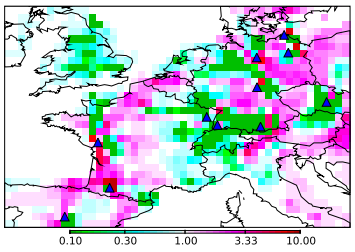
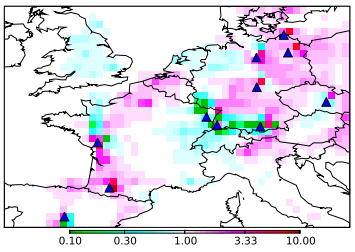
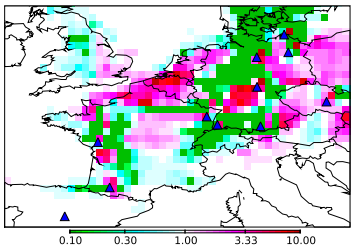
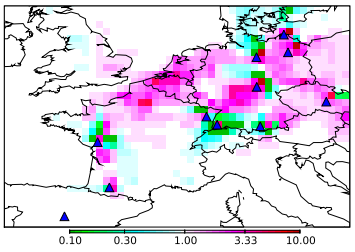
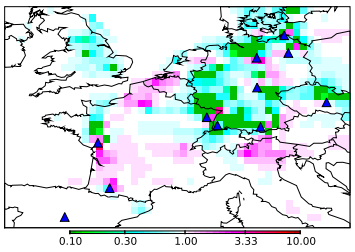
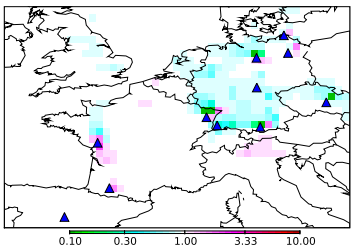
	N	$\overline{C}$	$\overline{O}$	RMSE	R	FA_2	FA_5
Forbach (12) (urban)	1319	379.9	354.8	387.1	0.35	0.52	0.85
Place du Marche (urban)	1319	569.8	560.4	352.2	0.42	0.8	0.99
Mairie MALO (traffic)	975	446.4	461.5	477	0.14	0.68	0.89
Halles centralles (traffic)	1344	668.4	609.8	385.8	0.6	0.75	0.97
GENERAL DE GAULLE (traffic)	1285	624.4	597.4	370.1	0.59	0.82	0.99
Valence traffic (traffic)	1329	472.8	462.9	283.5	0.24	0.79	0.99
place de VERDUN (urban)	102	221.3	522.6	725.7	0.55	0.66	0.9
Le Puy Fayolle (traffic)	1222	811	831.2	574.6	0.41	0.78	0.99
Le Creusot Molette (periurban)	1295	472.6	457	384.5	0.35	0.78	0.99
Marignane Ville (urban)	1341	572.4	527.7	335.4	0.49	0.74	1
Hotel Districal (urban)	1330	568.5	562.7	318.2	0.32	0.89	1
METZ-BORNY (urban)	786	340.3	248	237.4	0.29	0.49	0.86
Port de Bouc EDF (urban)	1312	504.2	459.4	321.2	0.5	0.64	0.95
PLOMBIERES (traffic)	1342	1464	1463	653.8	0.69	0.86	1
AIX CENTRE (traffic)	1338	620.6	595.5	333.1	0.51	0.81	0.99
Place Victor Basch (traffic)	1316	1618	1642	542.8	0.78	0.96	1
AUBERVILLIERS (urban)	1203	531.9	408	295.9	0.61	0.72	0.96
Avenue des Champs Elysees (traffic)	1331	953.1	933.6	432.5	0.6	0.88	1
Boulevard peripherique Auteuil (traffic)	1326	1374	1392	431.6	0.77	0.98	1
PARIS 1er Les Halles (urban)	1304	516.2	435.9	216.1	0.59	0.88	0.99
Autoroute A1 - Saint-Denis (traffic)	1313	1219	1255	448.3	0.69	0.96	1
Rue Bonaparte (traffic)	1267	897.3	896.2	363	0.69	0.93	1
Quai des Celestins (traffic)	1333	1416	1420	527.6	0.73	0.96	1
LeHavre Republique (traffic)	1333	619.2	582.6	369.6	0.65	0.72	0.95
ESQUERCHIN A DOUAI (traffic)	1340	573.6	555.8	277.5	0.59	0.89	0.99
Rousillon (traffic)	1254	556.6	553.6	362.2	0.39	0.68	0.95
Pres Arenes (urban)	1323	537.2	518.6	479.3	0.46	0.7	0.98
Planas (traffic)	1308	899.2	869.3	574.6	0.64	0.78	0.99
rue de la GRILLE (traffic)	1320	1012	1022	621.8	0.46	0.81	1
FORT-MARDYCK (industrial)	990	380.3	374.6	413.5	-0.01	0.5	0.88
Petite Synthe (periurban)	1130	376.3	384.1	448.2	0.1	0.6	0.96
Calais Centre (traffic)	1330	485.7	478.3	237.9	0.65	0.86	0.99
LIBERTE (traffic)	1341	618	603	404.8	0.64	0.75	0.99
PASTEUR (traffic)	1338	499.8	494.7	195.1	0.74	0.93	1
LA BASSEE/CENTRE (traffic)	1337	432.7	413.1	238.4	0.66	0.79	0.97
Le Rondeau (traffic)	1315	783	793.7	408	0.41	0.8	0.97
LAENNEC (traffic)	605	476.3	679.1	454.3	0.65	0.81	1
PUITS GAILLOT (traffic)	1340	806	815.2	452.1	0.45	0.81	0.98
BERTHELOT (traffic)	1340	988.5	1004	629.4	0.45	0.69	0.99
GARIBALDI (traffic)	1339	1447	1407	738.7	0.62	0.82	0.99
LA MULATIERE (traffic)	1341	827.5	829.6	424	0.45	0.81	1
Batiment ELF-ATO (industrial)	1075	330.1	257.3	192.7	0.31	0.63	0.92
ANTIBES GUYNEMER (urban)	1275	1243	1279	753.7	0.52	0.89	0.99
Rouen Le Conquerant (traffic)	1340	521	503.6	269.8	0.58	0.8	0.97
Pasteur (urban)	1227	638.6	628.9	372.7	0.53	0.79	0.99
station DAIX (periurban)	1329	314.6	307.3	185.9	0.24	0.73	0.95
BETHUNE PROX AUTO (traffic)	1293	525.6	509.6	281.6	0.56	0.83	1
ST ETIENNE ROND PT (traffic)	1294	615.1	597.7	293.8	0.56	0.89	1
RIVE DE GIER (traffic)	1312	490.3	457	250.7	0.48	0.76	0.95

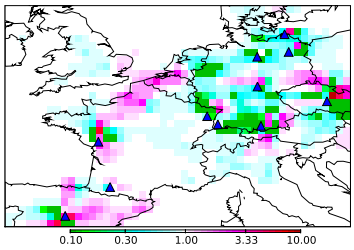
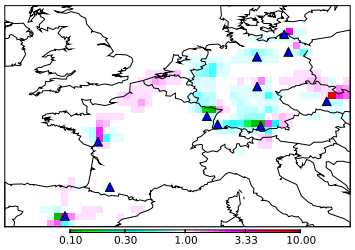
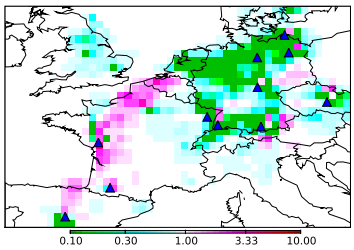
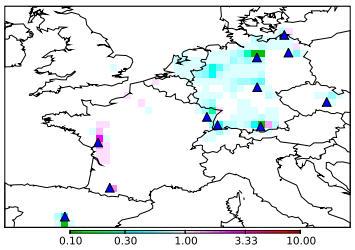
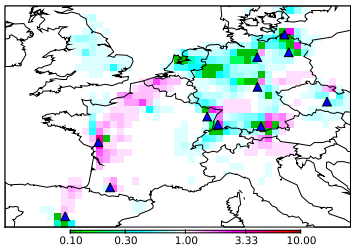
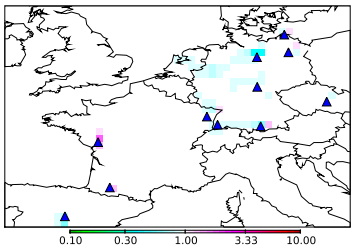
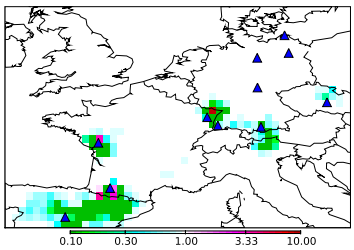
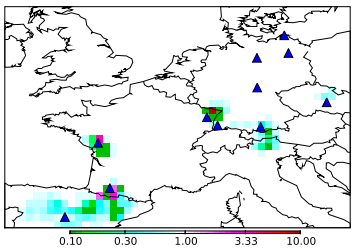
	N	$\overline{C}$	$\overline{O}$	RMSE	R	FA_2	FA_5
Luneville (urban)	1315	569.3	559.4	287.7	0.51	0.86	1
BORDEAUX-BASTIDE (traffic)	654	383.2	405.1	227.3	0.57	0.86	0.96
Chalon centre ville (traffic)	1265	693.4	691.9	390.6	0.52	0.81	0.99
Champforgeuil (periurban)	1154	317.2	301.3	222.7	0.36	0.64	0.82
Hilaire Chardonnet (periurban)	1185	400	406.1	190.2	0.35	0.92	1
Montceau-les-Mines (periurban)	1324	448.9	443.7	218.7	0.45	0.89	1

Appendix C

VOC emission scaling factor maps

case B2	case C
	
C_3H_8	
	
NBUT	
	
IBUT	

case B2	case C
	
NPEN	
	
IPEN	
	
C_3H_6	
	
TOLU	

case B2	case C
	
OXYL	
	
MXYL	
	
PXYL	
	
ISO	

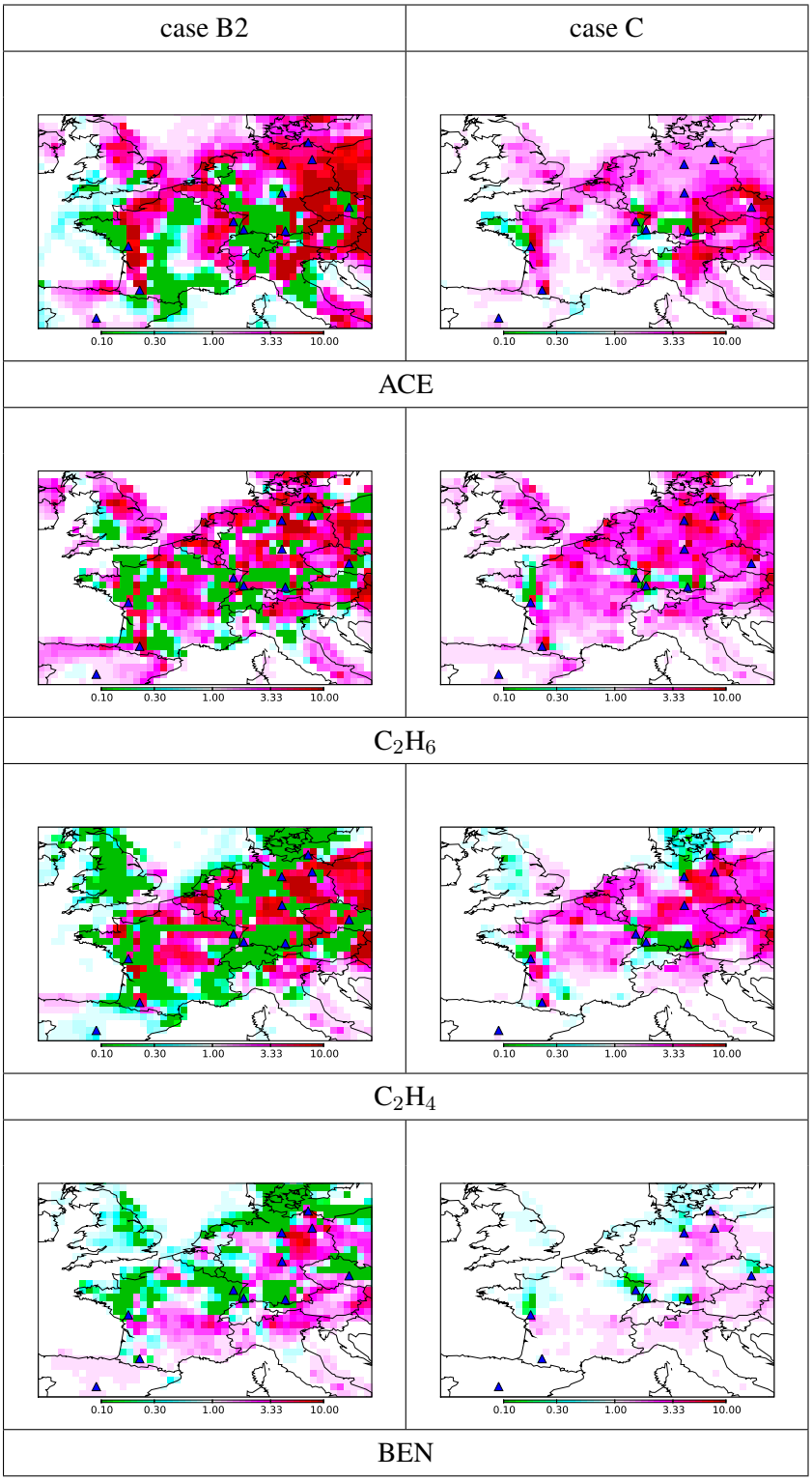
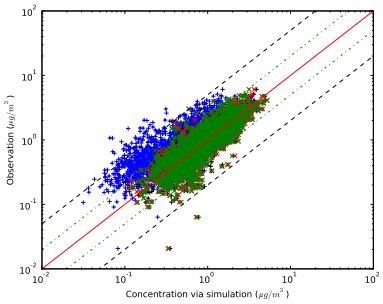


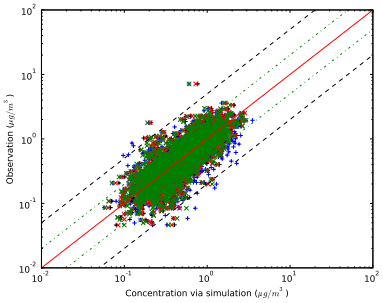
Figure C.1: Gridded ratios of time-integrated retrieved flux to EMEP time-integrated flux for VOC species

Appendix D

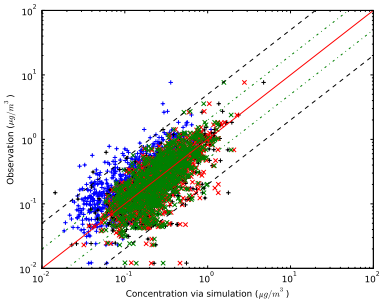
VOC scatterplots



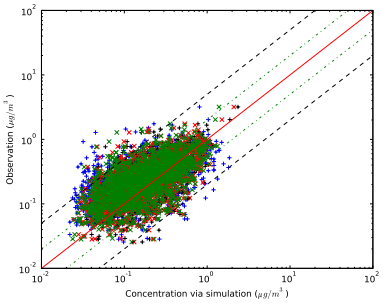
C_3H_8



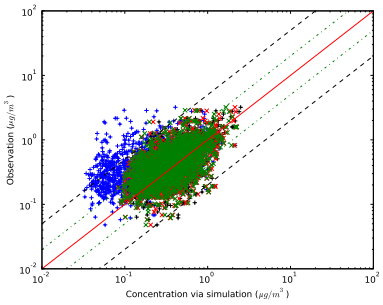
NBUT



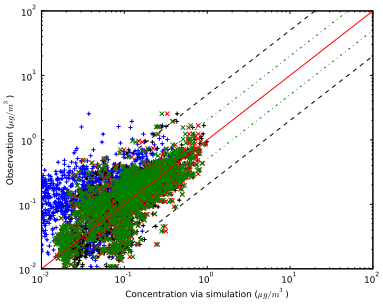
IBUT



NPEN



IPEN



C_3H_6

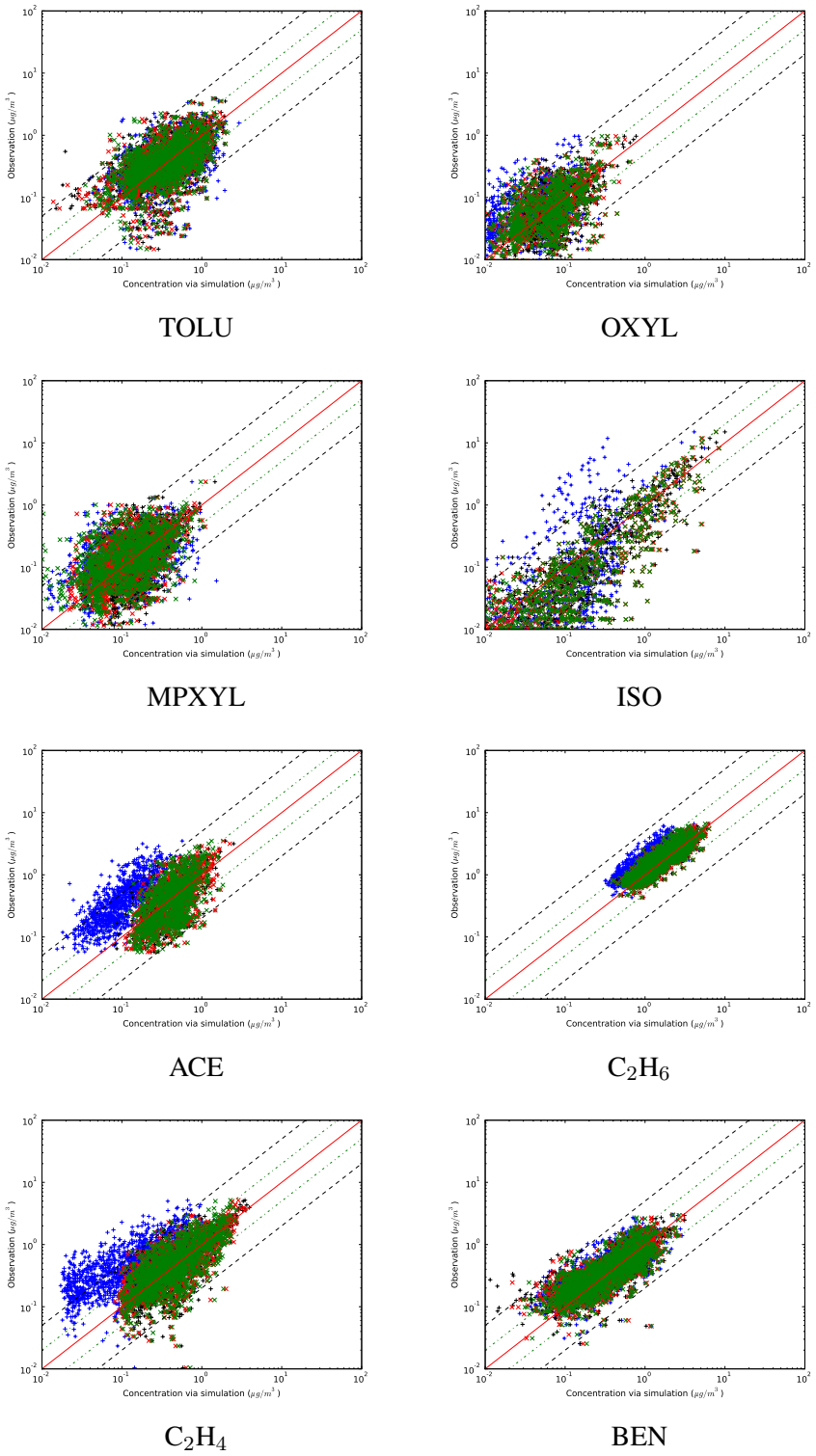


Figure D.1: comparison between the observations and the simulated concentrations (for the year 2005) in four cases: case A (blue), case B1 (black), case B2 (red), case C (green).