Toward Optimal Choices of Control Space Representation for Geophysical Data Assimilation

MARC BOCQUET

Université Paris-Est, CEREA Joint Laboratory école des Ponts ParisTech, and EDF R&D, Champs-sur-Marne, and INRIA, Paris Rocquencourt Research Center, Le Chesnay, France

(Manuscript received 9 September 2008, in final form 22 January 2009)

ABSTRACT

In geophysical data assimilation, observations shed light on a control parameter space through a model, a statistical prior, and an optimal combination of these sources of information. This control space can be a set of discrete parameters, or, more often in geophysics, part of the state space, which is distributed in space and time. When the control space is continuous, it must be discretized for numerical modeling. This discretization, in this paper called a representation of this distributed parameter space, is always fixed a priori.

In this paper, the representation of the control space is considered a degree of freedom on its own. The goal of the paper is to demonstrate that one could optimize it to perform data assimilation in optimal conditions. The optimal representation is then chosen over a large dictionary of adaptive grid representations involving several space and time scales.

First, to motivate the importance of the representation choice, this paper discusses the impact of a change of representation on the posterior analysis of data assimilation and its connection to the reduction of uncertainty. It is stressed that in some circumstances (atmospheric chemistry, in particular) the choice of a proper representation of the control space is essential to set the data assimilation statistical framework properly.

A possible mathematical framework is then proposed for multiscale data assimilation. To keep the developments simple, a measure of the reduction of uncertainty is chosen as a very simple optimality criterion. Using this criterion, a cost function is built to select the optimal representation. It is a function of the control space representation itself. A regularization of this cost function, based on a statistical mechanical analogy, guarantees the existence of a solution.

This allows numerical optimization to be performed on the representation of control space. The formalism is then applied to the inverse modeling of an accidental release of an atmospheric contaminant at European scale, using real data.

1. Introduction

Data assimilation is a set of mathematical techniques that aims at optimally combining several sources of information: data of an experimental nature that come from observation of the system, statistical information that comes from a prior knowledge of the system, and a numerical model that relates the space of observation to the space of the system state. Modern data assimilation has been carried out in meteorological operational centers or in oceanographic research centers over the last 15 yr with success and a significant improvement in the forecast skills (Le Dimet and Talagrand 1986; Lorenc 1986; Courtier and Talagrand 1990; Ghil and Malanotte-Rizzoli 1991; Courtier et al. 1994—to mention just a few of the seminal works on the topic). The ideas and methods have also percolated in atmospheric chemistry over the last 10 yr on a research basis (the example of the methodological development of this paper pertains to this field).

One characteristic of geophysical data assimilation is the huge domain in which it is meant to be applied and the very inhomogeneous distribution of fields. For numerical purposes, the physical model and its equations need to be discretized on this domain. The number of spatial variables is then close to a billion (currently). Thus, compromises must always be made in terms of the sophistication of the mathematical methods, the amount of data needed, and the control parameters (which shall

Corresponding author address: Marc Bocquet, CEREA, école des Ponts ParisTech, 6-8 avenue Blaise Pascal, Cité Descartes, Champs-sur-Marne, 77455 Marne la Vallée, CEDEX 2, France.
E-mail: bocquet@cerea.enpc.fr

DOI: 10.1175/2009MWR2789.1

also be interchangeably called variables) that can be handled.

At the confluence of these data assimilation ingredients is the choice of the resolution of the control variables space. The resolution choice has a direct impact on the number of control variables. For instance, in the context of inverse modeling, a finer resolution may lead to underdetermination. In geophysics, changing the resolution (or the scale at which the phenomena are examined) also has physical modeling consequences. Indeed, the physics that are modeled are usually scale dependent; the same process may be represented differently at different scales. Setting the resolution of a geophysical model is an intricate choice, which is full of implications. As a consequence, this choice is usually made very early in the modeling study. The implementation of (complex enough) data assimilation methods usually follows this early decision. From time to time, operational centers would change the resolution of their main model and a fresh study would be conducted on the basis of the fixed targeted resolution. For variational data assimilation purposes, two grids can be used but are fixed nevertheless; for instance, one for the inner loop and another for the outer loop in four-dimensional variational data assimilation (4DVAR).

In this paper, we would like to reverse the point of view and consider the resolution a free parameter in the implementation of data assimilation techniques.

a. Data assimilation and the representation of control space

In this paper, we are not directly interested in the nature of the control fields or in a careful choice of uncorrelated control fields. This choice is assumed to already have been made. Because the physics are modeled through numerics, they have to be discretized and the choice of a resolution should be made. This resolution problem generalizes to the question of the representation of the control fields.

Note that in numerical simulation (no assimilated observations) the issue of choosing a proper representation (e.g., an adaptive grid) is well studied. It allows speeding up the simulations and refines the computations in which the numerical error could be large (see, e.g., Saad 2003). An example in geophysics, and in particular air quality, can be found in Constantinescu et al. (2008).

However, in data assimilation, control variables and observations are equally important components. Because they do not share the same representation space but are nevertheless combined in the data assimilation process, the optimal representation problem is much less simple there.

Note that the control space representation can be different from the forward model (state) space discretization, even if the control space is a subspace of the state space. For instance, their resolutions may differ resulting from use of regular meshes of the same underlying domain.

b. Can the representation be chosen?

A fundamental issue is to decide how control fields should be discretized. In particular, can we choose, in a reasonably nonarbitrary fashion, the resolution of the control space in data assimilation? Hints could come from information theory. Provided adequate tools exist, one would compare the information content of the observations and the capacity of the control space grid to contain this information filtered out by the data assimilation analysis. In addition to the Bayesian inference principle used in data assimilation, a (typically variational) side principle would help construct an optimal representation. Such a cost function could establish how fine the grid resolution should be if it has to accommodate for the set of available data. Such a choice may stem from the balance found between the information content of the observation (but also the background) and the amount of information that the control variables are able to hold.

In this paper, a simple criterion of optimality is chosen based on the trace of the inverse analysis error covariance matrix, with known links to an information content analysis to be discussed and recalled.

c. Regularization, continuum limit, and scaling divergence

In a data assimilation system, what if the control space is discretized in too fine a manner? This question is actually central to inverse problems. Without regularization, the problem is underdetermined and therefore ill-posed. Regularization has been devised to circumvent this difficulty. In the data assimilation literature, regularization is performed through background implementation with a clear physical interpretation. In many cases, common regularization [such as Tikhonov; i.e., a least square constraint between the system state and a first guess (Rodgers 2000)] is sufficient to accommodate a finer resolution of control space (Bocquet 2005a). As the resolution increases, no new information coming from the observation is strengthening the inference or data assimilation analysis. Obviously, the computation load is heavier. However, the solution (the analysis) is only marginally changed.

Yet, it has been shown recently that, in some circumstances (e.g., the identification of an atmospheric pollutant accidental-release source), this regularization

may not be enough to support the increase in the resolution (Bocquet 2005a; Furbish et al. 2008). As the resolution increases, the solution ignores the benefit of the observations more and more. With a Tikhonov regularization scheme, it may converge to the first guess. The analysis is mathematically sound, and the degeneracy is understood on physical grounds (Bocquet 2005a). However, the solution is no longer impacted by the observations except in the vicinity of the observation sites. In this context, it is therefore mandatory to make an educated choice on a (if not the) proper scale of the representation.

Building on Bocquet (2005a), general results will be obtained on this issue in section 2.

d. Greenhouse gas fluxes estimation: An optimal representation problem

The driving force of global warming is very likely to be greenhouse gases and their relative balance (Pachauri and Reisinger 2007). The first one, in terms of impact, is carbon dioxide. From climate theory, it is therefore crucial to evaluate precisely the fluxes and exchange of carbon on earth. Therefore, biogeochemists have built up inventories of fluxes that are still uncertain, especially biogenic emissions and uptakes. A complementary approach is to use partial flux data and carbon dioxide concentration measurements and, through inverse modeling and because of prior hypotheses and a global dispersion model, estimate carbon fluxes on the globe. This is the so-called top-down approach. However, until the arrival of carbon satellite data, the pointwise concentration and flux measurements are sparse on the globe, albeit very reliable. Their information content is insufficient to accommodate the global-scale high-resolution map of fluxes that the carbon community is hoping for. So, it was decided very early to split the world into a few areas, leading to clearly identified aggregation errors (Fan et al. 1998; Bousquet et al. 2000). With more data available, the community was then inclined toward high-resolution reconstruction of the fields (Peylin et al. 2005). But then the information content of the observations does not match this too-fine control space. So here, too, the choice of the representation of the control space is thought to be crucial, and quantitative prescription of the right scale would be decisive. The formalism developed in this paper should help in this respect.

e. Objectives and outline

The aim of this paper is to demonstrate that the representation of the control parameter space can be chosen optimally to better match the available sources of information introduced in the data assimilation system.

In section 2, we reinterpret old data assimilation results by studying, in general terms, how the data assimilation analysis, as a measure of the inference quality, evolves through changes in the representation of control space defined on the same domain. This justifies how important an optimal choice of representation can be. In section 3, a simple optimality criterion is chosen to select a representation. A mathematical framework is then developed to make the idea explicit and implementable. In section 4, the formalism is applied in the field of atmospheric dispersion, in which an optimal representation is obtained given a fixed number of grid cells. Using this optimal representation, inverse modeling of a tracer source is then performed and a comparison is made with earlier results on regular grids. A summary and perspectives are given in section 5.

2. Posterior analysis and change of representation

In this section, basic results of data assimilation will be scrutinized, keeping in mind a possible change of the control space grid structure. This will justify the importance of an optimal choice of representation.

a. Reduction of error and analysis error covariance matrix

In the framework of the best linear unbiased estimator (BLUE) analysis, the posterior error covariance matrix reads as follows:

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

where $\mathbf{B}$ is the background covariance matrix (which will be assumed full rank in the following), $\mathbf{R}$ is the observation covariance matrix, and $\mathbf{H}$ is, broadly speaking, the observation operator. Here, T denotes the usual vector or matrix transposition. Alternatively, the inverse of the error covariance matrix, or the confidence matrix, is the following:

$$(\mathbf{P}_a)^{-1} = \mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}. \tag{2}$$

This matrix is also called the Fisher information matrix (for a general definition in a geophysical context, see Rodgers 2000, p. 32), which is a measure of the information conveyed by the observations (typically a vector of measurements $\boldsymbol{\mu} \in \mathbb{R}^d$, where d is the number of observations) to the control variables (a vector $\boldsymbol{\sigma} \in \mathbb{R}^N$, where N is the number of variables) through the data assimilation optimality system, when $\boldsymbol{\mu}$ is seen as a random vector.

It is assumed that the control space has an underlying spatial and temporal structure, with a discretization parameter r and a spacing that could be spatial, temporal,

or both. This structure has a continuum limit ($r \to 0$), which is considered as the "truth" in geophysics. For instance, one may be attempting to retrieve a ground emission field (two spatial dimensions and time; 2D + T). Then $\mathbf{H}$, as a Jacobian matrix, relates the concentration measurements to the space–time emissions. It is computed because of numerical schemes solving consistently, for some discretization r' (possibly different from r), a set of partial differential equations.

As is well known, the analysis systematically reduces the uncertainty because $\mathbf{P}_a \prec \mathbf{B}$, where $\prec$ denotes partial ordering between semidefinite positive matrices. Alternatively, the confidence is systematically increased: $\mathbf{P}_a^{-1} \succ \mathbf{B}^{-1}$. A similar analysis based on the same dataset is contemplated but the resolution, or more generally the representation of the control space, is now changed. When both observations and model are perfect (i.e., $\mathbf{R} = 0$), then $\mathbf{P}_a = \mathbf{B} - \mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}\mathbf{H}\mathbf{B}$. A limiting case is reached when a representation of the control space has d cells, which is the number of observations. As a consequence, $\mathbf{H}$ may become invertible, at least in an infinite precision numerical context, so that in this specific case $\mathbf{P}_a = 0$. Still assuming $\mathbf{R} = 0$, suppose that the number of variables is greater than the number of observations $N > d$. This obviously happens when $r \to 0$. Then the uncertainty is not null anymore but its reduction is controlled by $\mathbf{B}^{-1/2}\mathbf{P}_a\mathbf{B}^{-1/2} = \mathbf{I}_N - \underline{\mathbf{H}}^{\mathrm{T}}(\underline{\mathbf{H}}\underline{\mathbf{H}}^{\mathrm{T}})^{-1}\underline{\mathbf{H}}$, where $\underline{\mathbf{H}} = \mathbf{H}\mathbf{B}^{1/2}$ and $\mathbf{I}_N$ is the identity matrix in $\mathbb{R}^N$. Therefore, the posterior uncertainty is essentially controlled by the ability of the averaging kernel $\underline{\mathbf{H}}^{\mathrm{T}}(\underline{\mathbf{H}}\underline{\mathbf{H}}^{\mathrm{T}})^{-1}\underline{\mathbf{H}}$ to reproduce the identity. [The averaging kernel is the sensitivity matrix of the analysis state with the true state; see Rodgers (2000) for a clear definition and properties.]

One invariant measure of the total variance of the error is the trace of $\mathbf{P}_a$, which serves as the optimality criterion for establishing BLUE. The expression $\mathbf{B}^{-1/2}\mathbf{P}_a\mathbf{B}^{-1/2}$ reads the posterior covariance matrix in the basis where all parameters, viewed as random variables, are independent with variance unity. Therefore, it actually measures the reduction of uncertainty with respect to the uncertainty attached to the independent degrees of freedom of the prior. Then $\mathrm{Tr}(\mathbf{I}_N - \mathbf{P}_a\mathbf{B}^{-1}) = d$ measures the number of degrees of freedom that are elucidated in the analysis. In an errorless context ($\mathbf{R} = 0$), this equals the number of (independent) observations, as noted by Rodgers (2000). Similarly, $\mathrm{Tr}(\mathbf{P}_a\mathbf{B}^{-1}) = N - d$ rigorously states the obvious: whereas the degrees of freedom are increasing with increasing mesh resolution, the reduction of uncertainty from observation remains the same.

Note that $\mathbf{P}_a\mathbf{B}^{-1}$ also appears in the measure of the average gain of information (average reduction of entropy) in the BLUE analysis (Fisher 2003):

$$E_\sigma[\mathcal{K}_\sigma] = -\frac{1}{2}\ln|\mathbf{P}_a\mathbf{B}^{-1}| = -\frac{1}{2}\ln|\mathbf{I}_N - \mathbf{A}|, \qquad (3)$$

where $|\mathbf{M}|$ is the determinant of matrix $\mathbf{M}$ and $\mathbf{A} = \mathbf{I}_N - \mathbf{P}_a\mathbf{B}^{-1}$ is the averaging kernel matrix. Here, $\mathcal{K}_\sigma$ is the relative entropy or Kullback–Leibler information measure (Cover and Thomas 1991), using Gaussian hypotheses. The reduction of uncertainty and the reduction of entropy are closely related and coincide in the limit where the reduction of entropy (gain of information) is small with respect to the prior, up to a conventional factor measuring the information unit: $E_\sigma[\mathcal{K}_\sigma] = -\frac{1}{2}\ln|\mathbf{I}_N - \mathbf{A}| \simeq \frac{1}{2}\mathrm{Tr}(\mathbf{A})$. In the same limit, a closely related expansion of the information measure can also be performed:

$$\begin{aligned} E_\sigma[\mathcal{K}_\sigma] &= \frac{1}{2}\ln|\mathbf{B}\mathbf{P}_a^{-1}| \\ &= \frac{1}{2}\ln|\mathbf{I}_N + (\mathbf{B}\mathbf{P}_a^{-1} - \mathbf{I}_N)| \simeq \frac{1}{2}\mathrm{Tr}(\mathbf{B}\mathbf{P}_a^{-1} - \mathbf{I}_N). \end{aligned} \qquad (4)$$

This expression will be used in section 3 as an optimality criterion.

These elementary data assimilation statements show that increasing the sharpness of the representation (e.g., the resolution by decreasing r), and hence increasing the number of unknown variables N beyond the available observational information (d observations), could be detrimental to the quality of the analysis. The background acts as a mathematical regularization in the data assimilation system when the information content of the d observations cannot account for the N unknown variables. However, the background is usually inducing extra confidence that is often criticized, especially within the atmospheric chemistry applications where emission inventories are very uncertain (Elbern et al. 2007). So, it may often come as a surprise that increasing the resolution in data assimilation decreases the performance of the analysis.

This may be considered less surprising when data assimilation is regarded as an inverse problem. Then, this quest for information (Tarantola and Valette 1982) amounts to matching just enough unknown variables for the available information content. This is obviously a trivial task when inverting an algebraic linear system because there is only one flux of information (the individual equations of the system). In data assimilation, the difficulty comes from quantifying the fluxes of information because the sources are multiple (e.g., background, model, and observations) and of distinct nature.

Note that the background modeling is also dependent on representation (or scale) because variances of control parameters usually vary with the scale. For instance,

considering intensive variables, statistically independent subelements have a stronger variance than the variance attached to their aggregated mother element. However, taking into account the proper scaling does not necessarily prevent the degradation of the analysis. If we do not assume that the variables are statistically independent, then a correlated proper background might remove the degrading effect. But this popular trick actually amounts to reducing the effective number of independent degrees of freedom, though in a smoother manner.

b. Continuum limit for the posterior analysis

We have proved in the previous section that increasing the resolution beyond some threshold determined by the information content of the observations is always detrimental to the analysis. Let us now show that it may even lead to a severe failure of the analysis.

As mentioned earlier, $\mathbf{H}$ could be a Jacobian matrix that directly relates the observations to the control space. Again, let us change the representation of the control space (e.g., the resolution) and consider its consequence on the analysis. In the continuum limit, at least two kinds of behavior can actually occur, which have been described in Bocquet (2005a). The actual regime is decided by the behavior of the innovation statistics matrix $\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}$ in the continuum limit. A possible divergence was first pointed out by Issartel and Baverel (2003). To stress the significance of a proper representation of control space, let us first exhibit a simplified case relevant to physical applications in which such divergence may occur.

1) ANOMALOUS REGIME: EXAMPLE OF DIVERGENCE

Essentially following Bocquet (2005a), let us give a simple example of the singular regime. A Fickian diffusion equation is considered. A field of pollutant c obeys the following diffusion equation on the space–time domain Ω:

$$\frac{\partial c}{\partial t} = \mathrm{div}(\boldsymbol{\kappa} \cdot \nabla c) + \sigma, \tag{5}$$

where σ is the source field of the pollutant and $\boldsymbol{\kappa}$ is a diagonal diffusivity tensor. The choice of a Fickian operator may look overly simplified, but it does capture qualitative features of atmospheric transport, such as convective diffusion and mixing for a sufficiently long time, that are relevant to data assimilation and inverse modeling applications. To directly relate a pointwise concentration measurement μ_i to the source field, the adjoint equations can be introduced, one for each observation:

$$-\frac{\partial c_i^*}{\partial t} = \mathrm{div}(\boldsymbol{\kappa} \cdot \nabla c_i^*) + \pi_i, \tag{6}$$

where π_i is a retro source, which, in the case of a pointwise and instantaneous concentration measurement at $(\mathbf{x}_i, t_i)$, reads $\pi_i(\mathbf{x}, t) = \delta(\mathbf{x} - \mathbf{x}_i, t - t_i)$ so that

$$\begin{aligned}\mu_i &= \int_\Omega d\mathbf{x}\, dt\, \pi_i(\mathbf{x}, t) c(\mathbf{x}, t) + \epsilon_i \\ &= \int_\Omega d\mathbf{x}\, dt\, c_i^*(\mathbf{x}, t) \sigma(\mathbf{x}, t) + \epsilon_i,\end{aligned} \tag{7}$$

where errors ϵ_i can also be accounted for.

In the discrete case, the vector of measurements $\boldsymbol{\mu} \in \mathbb{R}^d$ is related to the source or flux components $\boldsymbol{\sigma} \in \mathbb{R}^N$ through the Jacobian matrix $\mathbf{H} \in \mathbb{R}^{d \times N}$, whose lines are approximated by the discretization $\mathbf{c}_i^*$ of the adjoint solutions c_i^*.

Assume that the source or emission field has a prior model formalized by a simple background, a Tikhonov regularization, given by

$$\nu(\boldsymbol{\sigma}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{\sigma}^{\mathrm{T}}\mathbf{B}^{-1}\boldsymbol{\sigma}\right)}{\sqrt{(2\pi)^N |\mathbf{B}|}}, \tag{8}$$

where $\mathbf{B} = m^2 \mathbf{I}_N$ and m is a scalar, homogeneous to a root-mean-square error. Then, in this example, the innovation statistics diverge in the continuum limit. In particular, one has $\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} \simeq m^2 \mathbf{H}\mathbf{H}^{\mathrm{T}}$. Indeed the elements of this Gram matrix are given by scalar products of adjoint solutions. For noncoinciding observations, the scalar product of two elementary diffusion solutions has a proper continuum limit: the two singularities of the integral coincide with the two observations that are integrable because the species concentrations are locally integrable by mass conservation. However, this is not the case for coinciding observations, so that the diagonal of $\mathbf{H}\mathbf{H}^{\mathrm{T}}$ diverges. In particular, it was shown (Bocquet 2005a) that in the continuum limit one typical element of the diagonal either reaches a finite limit or behaves like $g \sim m^2 \int_{\Omega_r} d\mathbf{x}\, dt\, [c_i^*(\mathbf{x}, t)]^2 \sim r^{-\alpha}$, where Ω_r is the full space and time domain, punctured by a ball with radius r around the related observation site. The problem-dependent exponent $\alpha > 0$ qualifies the rate of divergence. Note that this divergence can also characterize a wide range of non-Fickian diffusion operators and maybe other types of evolution operators that are relevant to many geophysical situations.

Keeping in mind this possible divergence in the continuum limit, we now look at the consequences for the analysis and the space–time spread of the observational information in quite general terms.

2) Consequences for the analysis

Following the previous example, $\mathbf{H}$ is assumed to relate the vector of fluxes or source elements $\boldsymbol{\sigma}$ in $\mathbb{R}^N$ to the vector of observations $\boldsymbol{\mu}$ in $\mathbb{R}^d$ so that $\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\sigma} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is the errors vector. However, this model equation is more general, even though it assumes a linear model. It was shown that the innovation statistics may diverge in some well-founded geophysical applications when r vanishes. Let us analyze the implications on the BLUE analysis.

Here, $\mathbf{H}$ has a singular value decomposition $\mathbf{H} = \mathbf{UDV}^{\mathrm{T}}$, where $\mathbf{U}$ is an orthogonal matrix of size $d \times d$, $\mathbf{D}$ is a diagonal $d \times d$ matrix made of the singular values λ_i (which implicitly depend on the spacing r), and $\mathbf{V}$ is a $N \times d$ matrix satisfying $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_d$.

In the following, for the sake of simplicity, the background and observation error covariance matrices will be assumed diagonal of the form $\mathbf{B} = m^2\mathbf{I}_N$ and $\mathbf{R} = \chi\mathbf{I}_d$. As long as physical correlations are short range, the conclusions of the following developments will remain general. In particular, it is possible to perform similar calculations in the so-called standardized form, in which the singular value decomposition is applied to $\mathbf{R}^{-1/2}\mathbf{HB}^{1/2}$ instead of $\mathbf{H}$ (for details on the standardized form, see Rodgers 2000). Depending on the exact definition of the coarse-graining operation, m may or may not be scale dependent. Also, χ may depend on the scale because of the representativeness error, but this effect is neglected here.

If $\mathbf{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d)$, then the averaging kernel $\mathbf{A} = (\mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ can be written as follows:

$$\mathbf{A} = \left(m^{-2}\mathbf{I}_N + \sum_{i=1}^{d}\frac{\lambda_i^2}{\chi}\boldsymbol{v}_i\boldsymbol{v}_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{d}\frac{\lambda_i^2}{\chi}\boldsymbol{v}_i\boldsymbol{v}_i^{\mathrm{T}} = \sum_{i=1}^{d}\frac{\lambda_i^2}{m^{-2}\chi + \lambda_i^2}\boldsymbol{v}_i\boldsymbol{v}_i^{\mathrm{T}}, \tag{9}$$

where $\mathbf{A}$ is a projector on the $\{\boldsymbol{v}_i\}_{i=1,\ldots,d}$ followed by a contraction when errors are taken into account. In the singular regime, the vectors $\boldsymbol{v}_i$ converge to $\mathbf{c}_i^*/\|\mathbf{c}_i^*\|$, whereas the singular values λ_i diverge. Therefore, in the continuum limit $r \to 0$,

$$\mathbf{A} \underset{r\to 0}{\sim} \sum_{i=1}^{d}\frac{(\mathbf{c}_i^*)(\mathbf{c}_i^*)^{\mathrm{T}}}{(\mathbf{c}_i^*)^{\mathrm{T}}(\mathbf{c}_i^*)}, \tag{10}$$

provided m^{-2}/χ does not diverge faster than λ_i^2 (this issue is settled by the exponent α mentioned above and is therefore case dependent). In this limit, the averaging kernel is a mere projector onto the vector space generated by the adjoint solutions. If m^{-2}/χ competes with one λ_i (or more) in this limit, then this projection is followed by a contraction.

As for the gain matrix $\mathbf{K} = (\mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}$, if $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$, then

$$\mathbf{K} = \left(m^{-2}\mathbf{I}_N + \sum_{i=1}^{d}\frac{\lambda_i^2}{\chi}\boldsymbol{v}_i\boldsymbol{v}_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{d}\frac{\lambda_i}{\chi}\boldsymbol{v}_i\mathbf{u}_i^{\mathrm{T}} = \sum_{i=1}^{d}\frac{\lambda_i}{m^{-2}\chi + \lambda_i^2}\boldsymbol{v}_i\mathbf{u}_i^{\mathrm{T}}. \tag{11}$$

In the singular regime, the vectors $\mathbf{u}_i$ converge to $\mathbf{e}_i$, where $(\mathbf{e}_1, \ldots, \mathbf{e}_d)$ is the canonical basis of the physical space. Therefore, in the continuum limit,

$$\mathbf{K} \underset{r\to 0}{\sim} \sum_{i=1}^{d}\frac{1}{\lambda_i}\frac{(\mathbf{c}_i^*)(\mathbf{e}_i)^{\mathrm{T}}}{\|\mathbf{c}_i^*\|}, \tag{12}$$

so that the gain $\mathbf{K}$ goes to zero in the singular limit case. To estimate the performance of data assimilation within a single event context (only one set of measurements $\boldsymbol{\mu}$), an objective root-mean-square-like score (Bocquet 2005a; Krysta and Bocquet 2007) is given by

$$\rho = \frac{\boldsymbol{\mu}^{\mathrm{T}}(\mathbf{R} + \mathbf{HBH}^{\mathrm{T}})^{-1}\boldsymbol{\mu}}{(\boldsymbol{\sigma}_t - \boldsymbol{\sigma}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\boldsymbol{\sigma}_t - \boldsymbol{\sigma}_b) + \boldsymbol{\epsilon}_t^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{\epsilon}_t}, \tag{13}$$

where $\boldsymbol{\sigma}_t$ is the true control state, $\boldsymbol{\sigma}_b$ is the first guess, and $\boldsymbol{\epsilon}_t$ is the true errors set. It can be shown that $0 \le \rho \le 1$ (Bocquet 2005a). The more resolved the control state is, the weaker the image of $\boldsymbol{\sigma}_t - \boldsymbol{\sigma}_b$ through the averaging kernel, unless all the mass of the state vector concentrates on the observation sites. In the singular limit case, for a randomly chosen $\boldsymbol{\sigma}_t$, one almost surely has the following:

$$\rho = \frac{1}{\|\boldsymbol{\sigma}_t - \boldsymbol{\sigma}_b\|^2 + m^{-2}\chi\|\boldsymbol{\epsilon}_t\|^2}\sum_{i=1}^{d}\frac{1}{m^{-2}\chi + \lambda_i^2}(\mathbf{u}_i^{\mathrm{T}}\boldsymbol{\mu})^2 \underset{r\to 0}{\sim} \frac{1}{\|\boldsymbol{\sigma}_t - \boldsymbol{\sigma}_b\|^2 + m^{-2}\chi\|\boldsymbol{\epsilon}_t\|^2}\sum_{i=1}^{d}\frac{\mu_i^2}{\lambda_i^2} \underset{r\to 0}{\to} 0. \tag{14}$$

In the control space, the information matrix $\mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ should also hint at this singular behavior. The question is whether the confidence that is attached to $\mathbf{R}^{-1}$ is or is not propagating through the action of $\mathbf{H}$. This is clearly the case in the absence of singularity as follows:

$$\mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H} = m^{-2}\mathbf{I}_N + \sum_{i=1}^{d}\frac{\lambda_i^2}{\chi}\boldsymbol{v}_i\boldsymbol{v}_i^{\mathrm{T}} \underset{r\to 0}{\sim} m^{-2}\mathbf{I}_N + \sum_{i=1}^{d}\frac{\lambda_i^2}{\chi}\frac{(\mathbf{c}_i^*)(\mathbf{c}_i^*)^{\mathrm{T}}}{(\mathbf{c}_i^*)^{\mathrm{T}}(\mathbf{c}_i^*)}. \tag{15}$$

The singular regime leads to a clear divergence in the confidence. As r goes to zero, the adjoint solution is

more and more peaked near the observation sites so that the propagation is prominent in an increasingly close neighborhood. The average gain of information in the analysis is supporting this:

$$E_\sigma[\mathcal{K}_\sigma] = \frac{1}{2}\sum_{i=1}^{d} \ln(1 + m^2\chi^{-1}\lambda_i^2) \underset{r\to 0}{\sim} \sum_{i=1}^{d} \ln\lambda_i. \quad (16)$$

Thus, paradoxically, the gain of information diverges in the singular regime. A likely reason is that when the sources are localized in the vicinity of the observation sites, the inference becomes very informative compared to the prior assumption of an omnipotent emission field. The averaging kernel Eq. (10) tells what one can really hope for in this limit: it converges to a projector on the observation site support. This projector has no spread in the space–time domain. The normalized information attached to an observation will not propagate to the distributed control space. It will only resolve the vicinity of the observation sites.

Although it depends on scale, the analysis is not flawed. However, it does depend on the magnification chosen to investigate the control space, which, depending on the analysis purpose, may be inappropriate. If the objective is to tell anything about the whole distributed control space, then this situation is not desirable and the high-resolution limit should be avoided. To a lesser extent, this conclusion also applies to data assimilation systems of the nonsingular regime.

3) BLUE ESTIMATORS IN THE TWO REGIMES

The generic BLUE results depend on two typical behaviors. First, suppose that the limit of the innovation statistics matrix $\mathbf{HBH}^{\mathrm{T}}$ is well defined. Then both the analysis error covariance matrix and the estimator tend to a proper nontrivial limit. The continuum limit of the analysis covariance matrix is still given by Eq. (1), and the estimator has the following usual form:

$$\boldsymbol{\sigma}_a = \boldsymbol{\sigma}_b + \mathbf{BH}^{\mathrm{T}}(\mathbf{R} + \mathbf{HBH}^{\mathrm{T}})^{-1}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_b). \quad (17)$$

Even though there is a finite continuum limit for the data assimilation problem, the analysis performance may degrade when the resolution is decreased, down to a lower nonzero threshold. This has been checked in Bocquet (2005a) and Furbish et al. (2008). In the later paper model, error is also taken into account. On the contrary, when $\mathbf{HBH}^{\mathrm{T}}$ is divergent, the posterior matrix then merely tends to the background covariance; that is, $\mathbf{B}^{1/2}\mathbf{H}^{\mathrm{T}}(\mathbf{R} + \mathbf{HBH}^{\mathrm{T}})^{-1}\mathbf{HB}^{1/2}$ tends to a less and less relevant operator in $\mathbb{R}^{N\times N}$, along with the BLUE gain matrix. Thus, almost surely,

$$\boldsymbol{\sigma}_a = \boldsymbol{\sigma}_b + \mathbf{BH}^{\mathrm{T}}(\mathbf{R} + \mathbf{HBH}^{\mathrm{T}})^{-1}\mathbf{H}(\boldsymbol{\sigma}_t - \boldsymbol{\sigma}_b) \underset{r\to 0}{\to} \boldsymbol{\sigma}_b, \quad (18)$$

where $\boldsymbol{\sigma}_t$ is a randomly chosen true state of the geophysical system.

3. Choosing a representation for the control space

The results of section 2 demonstrate that the choice of the representation of the control space can potentially lead to more efficient data assimilation. In this section, a mathematical formalism is proposed to determine a good representation. Beyond the simpler choice of the resolution of a regular grid on control space Ω, this may involve choosing an adaptive and multiscale grid for later data assimilation purposes. To achieve this, one must first adopt a criterion that defines optimality in a dictionary of possible representations.

a. Simple criterion, simpler problem

We shall first drastically simplify the matter by fixing the number of parameters to be controlled. It is chosen on the order of the number of observations: a few control variables for one observation at most. Matching the resolution of a regular grid on Ω to the fixed number of cells is too rigid an approach, especially for distributed geophysical systems with very inhomogeneous control parameters. If the control space is multidimensional (three-dimensional for this paper application), a few hundred control parameters to match a few hundred observations is likely to be insufficient to describe the multidimensional distributed system through a regular grid.

Hence, the number of control parameters is fixed but not the partition of the control space that defines them. Call ω this representation, which is a partition of the space–time domain Ω. Here, ω is a set of (unnecessarily regular) cells that fully covers the space–time control domain (which is a 2D + T manifold for the typical application here). Besides, a point in the control domain Ω belongs to one and only one cell of the representation.

The criterion that we could choose to select a particular representation would consist in maximizing the total posterior confidence [i.e., the trace of the Fisher information matrix $\mathrm{Tr}(\mathbf{P}_a^{-1})$] at a fixed number of control parameters. To make the criterion dimensionless and coordinate free and following the discussion of section 2, we prefer the closely related $\mathcal{J} = \mathrm{Tr}(\mathbf{BP}_a^{-1} - \mathbf{I}_N)$. It measures the gain in confidence obtained from observation relatively to the initial confidence in the background. It also reads $\mathcal{J} = \mathrm{Tr}(\mathbf{BH}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})$. The physical interpretation is clear; confidence in the measurements $\mathbf{R}^{-1}$ is propagated by the model $\mathbf{H}$ and measured in the

basis for which the background confidence $\mathbf{B}^{-1}$ is standardized. It is also connected to the related information gain via Eq. (4).

b. Space–time multiscale grid and Jacobian matrix

To perform multiscale data assimilation, the Jacobian matrix $\mathbf{H}$ needs to be defined at several scales. The reference scale will be the scale that characterizes the finest available regular grid of size (N_x, N_y, N_t), whose total number of space and time cells is $N_{\text{fg}} \equiv N = N_x N_y N_t$. In this paper, physics at other scales will be derived from this reference scale by simple arithmetic averaging. However, the formalism developed in this paper could also accommodate physics that is based on more complex coarse-graining approaches with dedicated subgrid parameterizations that depend on the scale.

Recall that the physics of the problem is encoded in $\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\sigma} + \boldsymbol{\epsilon}$, which is written in the reference regular grid. This equation could be written at a different scale or more generally with components of $\boldsymbol{\sigma}$ that could be defined at different scales. In our context, one can think about coarse-graining components of $\boldsymbol{\sigma}$ in $\mathbb{R}^{N_{\text{fg}}}$ at the reference scale into a coarser $\boldsymbol{\sigma}_\omega$ in $\mathbb{R}^N (N \leq N_{\text{fg}})$, whose components are unions of size-varying blocks of components of $\boldsymbol{\sigma}$.

We have chosen to precalculate coarse-grained versions of the finest Jacobian obtained from numerical models on the reference regular grid of the control domain Ω (other choices are possible, such as computing them on the fly). As a result, a multiscale Jacobian made of coarse-grained versions of the original Jacobian on the finest regular grid is stored in computer memory. This precomputation saves time in the optimizations. However, the size of this multiscale Jacobian matrix is a requirement to consider in practice, so that an estimation of its maximum occupancy will be given below.

The multiscale extension consists in recursive dyadic coarse grainings of the original finest regular grid. By a basic coarse graining in one direction, two adjacent grid cells are gathered into one and the physical quantities (Jacobian components) defined on them are averaged accordingly. For each direction, one considers a predefined number of scales: n_x, n_y, and n_t for the application ahead. For each vector of scales $\mathbf{l} = (l_x, l_y, l_t)$ such that $0 \leq l_x < n_x$, $0 \leq l_y < n_y$, and $0 \leq l_t < n_t$, there is a corresponding coarse-grained version of the finest grid, with $N_{\text{fg}}\, 2^{-lx-ly-lt}$ coarse cells (also called tiles) and a related coarse-grained Jacobian matrix (we assume N_x, N_y, and N_t are multiples of 2^{n_x}, 2^{n_y}, and 2^{n_t}, respectively). The full multiscale grid on Ω is the union of all these coarser grids. The total number of tiles of the multiscale grid is $8N_{\text{fg}}(1 - 2^{-n_x})(1 - 2^{-n_y})(1 - 2^{-n_t})$. Obviously, that is also the dimension of the space that the multiscale Jacobian matrix acts on. At worst, the multiscale grid is 8 times larger than the finest regular grid, which could be costly in terms of memory. A graphical representation of the memory occupancy is pictured in Fig. 1a.

c. Adaptive tiling

Optimizations on the dictionary $\mathcal{R}(\Omega)$ of representations require that this set be handled mathematically and computationally. The following formalism is based on the multiscale grid detailed in the previous section.

In the representations ahead, a distinction is being made between space and time dimensions. Following the dyadic structure of the multiscale grid, the partition of time line is implemented by a binary tree $\mathcal{T}$ with n_t levels.

In a first attempt and following a 2D + T example, the partition of the 2D space domain $\mathcal{S}_{xy}$ was built up because of a ''quaternary tree'' with n_s levels. It means that each 2D space cell could be divided into four subcells, down to a depth of n_s levels (see Knuth 1997; Fig. 1b). A tensorial product of the time and space structures was then peformed. Two natural distinct sets of representations can be obtained: $\mathcal{T} \otimes \mathcal{S}_{xy}$ or $\mathcal{S}_{xy} \otimes \mathcal{T}$. As far as computer memory saving is concerned, these sets are rather economical. A (2D + T) emission field would be represented with $8/3N_{\text{fg}}(1 - 4^{-n_s})(1 - 2^{-n_t})$ scalars, in both multiscale structures. In the worst case (full downscaling), this involves a factor 8/3 as compared to the finest regular grid representation. However these representations are quite demanding in terms of coding. In particular, parsing efficiently all tree configurations in an optimization process can be very time-consuming.

That is why, in this paper, the set of representations that encompass all space–time ''tilings'' of a domain, based on the multiscale grid and Jacobian matrix defined earlier, is chosen instead. This set is the tensorial product of dyadic (binary) trees on time direction, O_t, and the spatial directions, O_x and O_y. It is a richer space than the two previous double-tree structures, but it is also much more memory consuming because the total number of cells to be run through is $8N_{\text{fg}}(1 - 2^{-n_x})(1 - 2^{-n_y})(1 - 2^{-n_t})$. It corresponds to all grid cells, or tiles in this context, of the multiscale grid. At worst, it is 8 times larger than the finest regular grid. Again, this could be avoided if the Jacobian $\mathbf{H}$ is not stored in memory, or as a compromise a quaternary tree representation is chosen instead.

A schematic of a double-tree partition is displayed in Fig. 1b, and a schematic of a tiling partition (triple tree) is displayed in Fig. 1c. Note that tensorial products of space and time structures are preferred to Cartesian products because the resulting dictionary of representations is much richer. Cartesian products would have

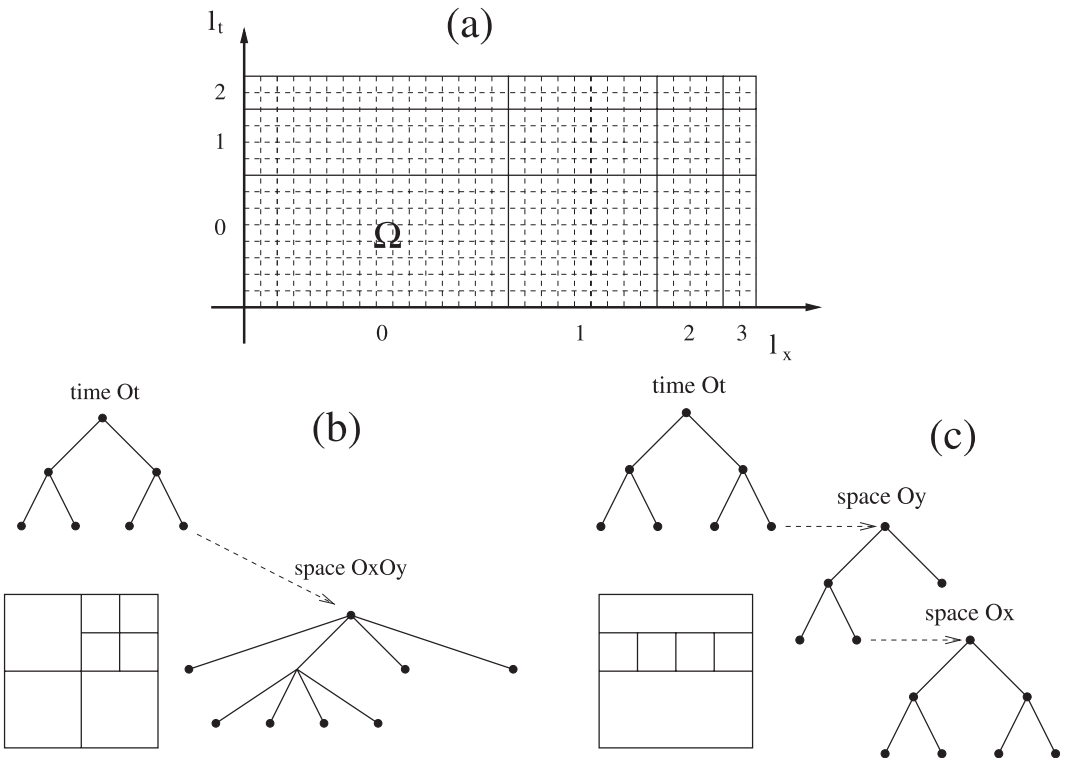

FIG. 1. (a) A representation of a multiscale structure on a 1D + T grid, with $n_x = 4$ and $n_t = 3$. Also depicted is how the multiscale Jacobian is stored in computer memory. The finest regular grid on Ω is on the left-bottom-hand corner. The other thick boxes represent coarse-grained versions of the original finest grid. They are coarse-grained copies of Ω. Each box (building block) is designated by two scale levels (l_x, l_t). (b) An example of a double-tree configuration that leads to a partition (quaternary space tree) of the space–time domain Ω. The resulting space partition is depicted at its side (at one fixed date). (c) An example of a triple binary tree configuration that describes a tiling partition of the space–time domain Ω. The resulting tiling is depicted at its side (at one fixed date corresponding to one single node in the time tree).

implied separation of the space and time degrees of freedom, which is too poor an assumption.

d. Vector space mathematical representation

The dictionary $\mathcal{R}(\Omega)$ of representations has been described geometrically. We shall now describe it algebraically.

Let us consider the vector space $\mathbb{R}^{N_{\text{fg}}}$ attached to the finest grid of dimension N_{fg}. To each tile is attached a vector $\boldsymbol{v}_{\mathbf{l},k}$ in this space, in which $\mathbf{l} = (l_x, l_y, l_t)$ is the set of scales of the tile and k is the index of this tile in the set of all tiles of the same scale $\mathbf{l}$ (one coarse version of the original grid). For the finest grid, one has $\mathbf{l} = (0, 0, 0)$ and $k = 1, \ldots, N_{\text{fg}}$. The cells of the finest grid (smallest tiles) are represented by the canonical basis of this vector space. A coarser tile, made of these finest tiles, has vector $\boldsymbol{v}_{\mathbf{l},k} \in \mathbb{R}^{N_{\text{fg}}}$, defined as the sum of the vectors of the canonical vectors $\{\mathbf{e}_{i,j,h}\}$ in $\mathbb{R}^{N_{\text{fg}}}$ with $1 \le i \le N_x$, $1 \le j \le N_y$, and $1 \le h \le N_t$ that are in one-to-one correspondence with the finest tiles that make up this tile: $\boldsymbol{v}_{\mathbf{l},k} = \Sigma_{\delta i=1}^{2^{l_x}} \Sigma_{\delta j=1}^{2^{l_y}} \Sigma_{\delta h=1}^{2^{l_t}} \mathbf{e}_{i_k+\delta i-1, j_k+\delta j-1, h_k+\delta h-1}$, where (i_k, j_k, h_k) are the coordinates on the finest grid of the tile indexed by k in the coarse-grained representation of Ω at level $\mathbf{l}$.

A tiling ω of Ω is a partition, in the strict mathematical sense, of the space and time control space with tiles. For each tiling ω in $\mathcal{R}(\Omega)$, we define an operator $\boldsymbol{\Pi}_\omega : \mathbb{R}^{N_{\text{fg}}} \to \mathbb{R}^{N_{\text{fg}}}$ built with the vectors attached to the tiles as follows:

$$\boldsymbol{\Pi}_\omega = \sum_{\mathbf{l}} \sum_{k=1}^{n_{\mathbf{l}}} \alpha_{\mathbf{l},k}^{\omega} \frac{\boldsymbol{v}_{\mathbf{l},k} \boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}}}{\boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \boldsymbol{v}_{\mathbf{l},k}}, \tag{19}$$

where $n_{\mathbf{l}}$ is the number of tiles in the set of tiles with scale vector $\mathbf{l}$; $\mathbf{l}$ runs on all predefined scales $0 \le l_x < n_x$, $0 \le l_y < n_y$, and $0 \le l_t < n_t$; and $\alpha_{\mathbf{l},k}^{\omega}$ are coefficients that define the representation ω. From now on, the superscript ω on $\alpha_{\mathbf{l},k}^{\omega}$ will be dropped to simplify the notation.

In a multigrid approach, a location in space and time is represented several times. On the contrary, in a single

representation of Ω, such a point is accounted for once, according to the strict definition of a partition. In this paper context, a point is a cell of the finest reference grid. A representation corresponding to a strict partition will be qualified as admissible. To obtain an admissible representation of the control space, one requires that $\alpha_{\mathbf{l},k}$ be 0 or 1 for all $(\mathbf{l}, k)$ and that any point in the domain be represented by one and only one tile:

$$\sum_{\mathbf{l}} \sum_{k=1}^{n_l} \alpha_{\mathbf{l},k} \boldsymbol{v}_{\mathbf{l},k} = (1, \ldots, 1)^{\mathrm{T}}. \tag{20}$$

Furthermore, we impose that the number N of variables (hence tiles) be set as follows:

$$\sum_{\mathbf{l}} \sum_{k=1}^{n_l} \alpha_{\mathbf{l},k} = N. \tag{21}$$

Obviously, for an admissible representation, $N_{\mathrm{cg}} \leq N \leq N_{\mathrm{fg}}$, where $N_{\mathrm{cg}} = N_{\mathrm{fg}} 2^{-n_x - n_y - n_t}$. On these conditions, it is not difficult to check that $\boldsymbol{\Pi}_\omega = \boldsymbol{\Pi}_\omega^{\mathrm{T}}$ and $(\boldsymbol{\Pi}_\omega)^2 = \boldsymbol{\Pi}_\omega$ so that $\boldsymbol{\Pi}_\omega$ is an orthogonal projector.

e. Restriction and prolongation

For a given representation ω, $\boldsymbol{\sigma}$ is coarse-grained into $\boldsymbol{\sigma}_\omega = \mathbf{l}_\omega \boldsymbol{\sigma}$, where $\mathbf{l}_\omega : \mathbb{R}^{N_{\mathrm{fg}}} \to \mathbb{R}^N$ is a restriction map that defines the coarse-graining operation. Although its precise definition depends on the context, it is always unambiguously defined. The prolongation operator $\mathbf{l}_\omega^* : \mathbb{R}^N \to \mathbb{R}^{N_{\mathrm{fg}}}$ relates $\boldsymbol{\sigma}_\omega$ back to $\boldsymbol{\sigma}$. Contrary to the restriction operator, several consistent definitions can be used for the prolongation operator. However, it should consistently satisfy $\mathbf{l}_\omega \mathbf{l}_\omega^* = \mathbf{I}_N$ (for a discussion of these operators, see Rodgers 2000). In addition, we impose that $\mathbf{l}_\omega^* \mathbf{l}_\omega = \boldsymbol{\Pi}_\omega$. This identity means that the net effect of coarse graining followed by an interpolation back in the finest grid is a local averaging on the area covered by each tile. Note that what follows does not depend on the precise definition of $\mathbf{l}_\omega^*$. Then $\mathbf{H}$ is coarse-grained into $\mathbf{H}_\omega = \mathbf{H}\mathbf{l}_\omega^*$, and the coarse-grained observation equation $\boldsymbol{\mu} = \mathbf{H}_\omega \boldsymbol{\sigma}_\omega + \boldsymbol{\epsilon}$ is made explicit in the finest grid vector space, because of the projection operator $\boldsymbol{\Pi}_\omega$:

$$\boldsymbol{\mu} = \mathbf{H}_\omega \boldsymbol{\sigma}_\omega + \boldsymbol{\epsilon} = \mathbf{H}\mathbf{l}_\omega^* \mathbf{l}_\omega \boldsymbol{\sigma} + \boldsymbol{\epsilon} = \mathbf{H}\boldsymbol{\Pi}_\omega \boldsymbol{\sigma} + \boldsymbol{\epsilon}. \tag{22}$$

In principle, $\boldsymbol{\epsilon}$ should be made dependent on ω through the representativeness and model errors. Besides, whereas $\mathbf{H}_\omega$ could be obtained from scale-dependent models at several scales, it is a mere coarse graining of a finescale Jacobian here. These important issues are beyond the scope of this paper, but the formalism developed ahead should accommodate them.

This allows to define a mathematically explicit multiscale criterion that depends on the choice of the representation $\omega \in \mathcal{R}(\Omega)$:

$$\mathcal{J}_\omega = \mathrm{Tr}_{\mathbb{R}^N}(\mathbf{B}_\omega \mathbf{H}_\omega^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H}_\omega) = \mathrm{Tr}_{\mathbb{R}^{N_{\mathrm{fg}}}}(\boldsymbol{\Pi}_\omega \mathbf{B} \boldsymbol{\Pi}_\omega \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H}). \tag{23}$$

The background error covariance matrix $\mathbf{B}$ has been coarse-grained into $\mathbf{B}_\omega = \mathbf{l}_\omega \mathbf{B} \mathbf{l}_\omega^{\mathrm{T}}$. In the following, $\mathbf{B}$ will be again assumed of the form $m^2 \mathbf{I}_{N_{\mathrm{fg}}}$ so that it commutes with $\boldsymbol{\Pi}_\omega$. As a consequence, $\mathcal{J}_\omega = \mathrm{Tr}_{\mathbb{R}^{N_{\mathrm{fg}}}}(\boldsymbol{\Pi}_\omega \mathbf{B} \mathbf{H}^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{H})$. This formula can be generalized to any positive definite $\mathbf{B}$, provided that $\boldsymbol{\Pi}_\omega$ is generalized to a projector which is $\mathbf{B}^{-1}$ orthogonal. This is not reported here because the application of section 4 does not require it.

The objective is to maximize the average reduction of uncertainty in the analysis $\mathcal{J}_\omega$ on all admissible tilings (i.e., on the $\alpha_{\mathbf{l},k}$ that satisfy the above conditions).

f. Lagrangian approach

To optimize $\mathcal{J}_\omega$ on ω, a Lagrangian is written with the above conditions implemented because of Lagrange multipliers. A fixed number of tiles is imposed from a single multiplier ζ. The one point–one tile requirement is imposed because of a vector $\boldsymbol{\lambda}$ of N_{fg} multipliers. The Lagrangian reads as follows:

$$\mathcal{L}(\omega) = \sum_{\mathbf{l}} \sum_{k=1}^{n_l} \alpha_{\mathbf{l},k} \frac{\boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \mathbf{W} \boldsymbol{v}_{\mathbf{l},k}}{\boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \boldsymbol{v}_{\mathbf{l},k}} + \sum_{k=1}^{n_0} \lambda_k \left(\sum_{\mathbf{l}} \alpha_{\mathbf{l},\tilde{k}} - 1 \right) + \zeta \left(\sum_{\mathbf{l}} \sum_{k=1}^{n_l} \alpha_{\mathbf{l},k} - N \right). \tag{24}$$

Here, $\mathbf{W}$ stands for $\mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ or any other appropriate criterion. The sum on $k = 1, \ldots, n_0 = \mathrm{N_{fg}}$ runs on all cells of the finest grid. In this sum, $\alpha_{\mathbf{l},\tilde{k}}$ is the coefficient attached to the tile at scale $\mathbf{l}$ that covers cell k from the finest grid. This tile has rank $\tilde{k}$ among the $n_{\mathbf{l}}$ tiles related to scale $\mathbf{l}$. The Lagrangian can also be written as follows:

$$\mathcal{L}(\omega) = \sum_{\mathbf{l}} \sum_{k=1}^{n_l} \left(\frac{\boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \mathbf{W} \boldsymbol{v}_{\mathbf{l},k}}{\boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \boldsymbol{v}_{\mathbf{l},k}} + \boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \boldsymbol{\lambda} + \zeta \right) \alpha_{\mathbf{l},k} - \sum_{k=1}^{n_0} \lambda_k - \zeta N. \tag{25}$$

Then, the maximum can formally be taken on all representations, admissible or not, with any number of tiles in $[N_{\mathrm{cg}}, N_{\mathrm{fg}}]$, on the set of coefficients $\alpha_{\mathbf{l},k}$ that have been freed from the constraints through the multipliers. This is made easier (to a limited extent) by the fact that

$\alpha_{\mathbf{l},k}$ can only be 0 or 1. Then, if $\widehat{\mathcal{L}}(\boldsymbol{\lambda},\zeta) = \max_{\omega\in\mathcal{R}(\Omega)} \mathcal{L}(\omega) = \max_{\alpha_{\mathbf{l},k}} \mathcal{L}(\omega)$,

$$\widehat{\mathcal{L}}(\boldsymbol{\lambda},\zeta) = \sum_{\mathbf{l}} \sum_{k=1}^{n_l} \max\left(0, \frac{\boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \mathbf{W} \boldsymbol{v}_{\mathbf{l},k}}{\boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \boldsymbol{v}_{\mathbf{l},k}} + \boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \boldsymbol{\lambda} + \zeta\right) - \sum_{k=1}^{n_{\boldsymbol{\theta}}} \lambda_k - \zeta N. \tag{26}$$

Because this cost function is dual to $\mathcal{L}(\omega)$, it needs to be minimized, not maximized. Because it depends on $\boldsymbol{\lambda}$ and ζ only, its evaluation is not too time consuming. Its complexity scales like an optimization on the finest gridcell variables. However, the cost function is not smooth because it is nonderivable on edges of a polytope. Hence, optimization on the Lagrange parameters cannot make direct use of gradient-based minimization techniques. Besides, it is unlikely that this function is convex; nor is it guaranteed that it has a single minimum. To overcome these potential problems, a regularization of this effective cost function is proposed.

g. Statistical physics interpretation

From now on, $\boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \mathbf{W} \boldsymbol{v}_{\mathbf{l},k} / (\boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \boldsymbol{v}_{\mathbf{l},k})$ is denoted by $\boldsymbol{\epsilon}_{\mathbf{l},k}$. If $\mathbf{W} = \mathbf{B}\mathbf{P}_a^{-1} - \mathbf{I}_{N_{\mathrm{fg}}}$, then $\boldsymbol{\epsilon}_{\mathbf{l},k}$ is a measure of the extra confidence in the posterior estimate of the scalar defined on tile $(\mathbf{l}, k)$ resulting from the observations. Then the criterion can be written as $\mathrm{Tr}(\boldsymbol{\Pi}_\omega \mathbf{W}) = \Sigma_{\mathbf{l},k} \alpha_{\mathbf{l},k} \boldsymbol{\epsilon}_{\mathbf{l},k}$, where $\Sigma_{\mathbf{l},k}$ is shorthand for $\Sigma_{\mathbf{l}} \Sigma_{k=1}^{n_l}$. Building on a statistical mechanics analogy introduced by Jaynes (1957a,b), we interpret the $\boldsymbol{\epsilon}_{\mathbf{l},k}$ as individual energies, one for each tile. The system is thermalized at inverse temperature β, which is the regularization parameter. The average number of cells of the system is set to N. Also, any point must be covered by an average of one, and only one, tile. From statistical mechanics, the partition function of this system is

$$Z_\beta(\boldsymbol{\lambda},\zeta) = \sum_{\boldsymbol{\alpha}} \exp\left[\sum_{\mathbf{l},k} (\beta \alpha_{\mathbf{l},k} \boldsymbol{\epsilon}_{\mathbf{l},k} + \alpha_{\mathbf{l},k} \boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \boldsymbol{\lambda} + \alpha_{\mathbf{l},k} \zeta)\right] = \prod_{\mathbf{l},k} [1 + \exp(\beta \boldsymbol{\epsilon}_{\mathbf{l},k} + \boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \boldsymbol{\lambda} + \zeta)], \tag{27}$$

where $\boldsymbol{\lambda}$ and ζ are conjugate parameters to the occupation constraint and to the tiles number. Because the number of tiles is fixed and because a point is covered only once, the right statistical potential is the free energy functional that derives from the partition function:

$$\widehat{\mathcal{L}}_\beta(\boldsymbol{\lambda},\zeta) = \sum_{\mathbf{l},k} \ln[1 + \exp(\beta \boldsymbol{\epsilon}_{\mathbf{l},k} + \boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \boldsymbol{\lambda} + \zeta)] - \sum_{k=1}^{n_{\boldsymbol{\theta}}} \lambda_k - \zeta N. \tag{28}$$

By construction, this free energy function is strictly convex, which guarantees the existence of a single minimum, provided the constraints can be satisfied by at least one configuration ω in $\mathcal{R}(\Omega)$. In particular, N needs to satisfy $N_{\mathrm{cg}} \le N \le N_{\mathrm{fg}}$. This allows for the use of classical optimization tools.

The minimization of the free energy $\widehat{\mathcal{L}}_\beta$ yields $\overline{\boldsymbol{\lambda}}$ and $\overline{\zeta}$. Then, it is possible to compute the average filling factors:

$$\overline{\alpha}_{\mathbf{l},k} = \frac{1}{\beta} \frac{\partial \ln Z_\beta}{\partial \boldsymbol{\epsilon}_{\mathbf{l},k}} = \frac{1}{1 + \exp(-\beta \boldsymbol{\epsilon}_{\mathbf{l},k} - \boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \overline{\boldsymbol{\lambda}} - \overline{\zeta})}, \tag{29}$$

which is essentially the main result of the optimization. The total reduction of uncertainty,

$$\overline{\mathcal{J}} = \frac{\partial \ln Z_\beta}{\partial \beta} = \sum_{\mathbf{l},k} \boldsymbol{\epsilon}_{\mathbf{l},k} \overline{\alpha}_{\mathbf{l},k}, \tag{30}$$

is given by the minimum of the free energy with an average reduction of $\boldsymbol{\epsilon}_{\mathbf{l},k} \overline{\alpha}_{\mathbf{l},k}$ per tile.

This approach has been checked to be a consistent regularization of our main optimization problem. In the low temperature limit, $\beta \to +\infty$, the variables can be rescaled by β and the cost function becomes

$$\widehat{\mathcal{L}}_\infty(\boldsymbol{\lambda},\zeta) = \beta \left[\sum_{\mathbf{l},k} \max(0, \boldsymbol{\epsilon}_{\mathbf{l},k} + \boldsymbol{v}_{\mathbf{l},k}^{\mathrm{T}} \boldsymbol{\lambda} + \zeta) - \sum_{k=1}^{n_{\boldsymbol{\theta}}} \lambda_k - \zeta N\right], \tag{31}$$

and the Lagrangian Eq. (26) is recovered, up to β. Hence, this statistical approach is a way to regularize our problem at a small but finite temperature. Any other regularization is possible but this one offers a statistical interpretation and a guarantee of convexity.

A similar Lagrangian to Eq. (26) has been obtained in the case of binary–quaternary trees representations, as well as an analog formalism for this set of representations. This will be reported elsewhere because no application is given for it next.

The optimal structure obtained from this technique is also the least committed representation given that all the constraints are satisfied on average. At finite temperature, the optimal representation is a compromise between the criterion and the relative entropy of the representation as compared to the (nonadmissible) geometry in which all tiles are equiprobable. The proof is not essential here and will be reported elsewhere.

h. Properties

A cost function Eq. (28) has been introduced to qualify representations. Several aspects of its minimization will be discussed now.

1) Representation solution and optimal admissible representation

The main drawback of a regularization, including the one above, is that it is by definition a continuous deformation of the nonsmooth optimization problem. A consequence is that, for a finite regularization parameter, the average $\overline{\alpha}_{\mathbf{l},k}$ is a scalar in the interval [0, 1] and not necessarily 0 or 1. So the representation that derives from the average $\overline{\alpha}_{\mathbf{l},k}$ is not usually an admissible one. As β goes to ∞, the values of $\overline{\alpha}_{\mathbf{l},k}$ gradually approach either 0 or 1. From the statistical point of view, the solution is still an average of admissible partitions of the domain for which the values of $\alpha_{\mathbf{l},k}$ are strictly 0 or 1, most of them being near optimal.

However, an admissible partition is needed to perform explicit data assimilation in the optimal representation. One solution is to optimize at very high β and, if possible, with high numerical precision. Unfortunately, this leads to unavoidable numerical instabilities (the regularization was used to overcome them in the first place). Besides, there is no guarantee that the solution does have a limit as β diverges, even though $\widehat{\mathcal{L}_\beta}$ is consistently deformed.

Therefore, a suboptimal admissible representation is constructed. It is based on the average filling factors computed for each tile by Eq. (29). Then, a representation is built by choosing for each point the tile with the largest average filling factor. This process yields a complete representation (every point is covered), but it may not be admissible because a point can be covered by several tiles. Therefore, this constraint is respected by somehow arbitrarily removing excess tiles. Note that the final number of tiles of the resulting admissible representation may be different from N, though close to it.

2) Links with the concept of resolution

In data assimilation, a concept of resolution quantifies the ability of the observations to locally resolve the control parameter space. It is notably applied in the assimilation of radiances for the reconstruction of profiles in the atmosphere (Purser and Huang 1993; Rodgers 2000). However, its quantification is empirical. As seen in section 2, the trace of the averaging kernel $\mathbf{A}$ is a measure of the gain of information because of the observations. The diagonal elements of the averaging kernel are a local gain of information within the control space. The inverse of each diagonal element $[\mathbf{A}]_{kk}^{-1}$ is then an empirical measure of the local resolution, or precision, in control space.

A similar but nonempirical concept can be derived from our formalism. For each point k, $1 \le k \le N_{\mathrm{fg}}$ (cell in the finest reference grid), a mean size of the local tile $\overline{S}_k$ can be obtained, along with a measure of the average resolution $\overline{\varrho}_k$:

$$\overline{S}_k = \sum_{\mathbf{l}} \overline{\alpha}_{\mathbf{l},\tilde{k}}\, \mathcal{A}(l_x, l_y, l_t) \quad \text{and} \quad \overline{\varrho}_k = \sum_{\mathbf{l}} \overline{\alpha}_{\mathbf{l},\tilde{k}}\, \mathcal{A}^{-1}(l_x, l_y, l_t), \tag{32}$$

where $\mathcal{A}$ is the volume of the tile, which would scale as $2^{l_x+l_y+l_t}$ in our dyadic representation in the units of cells of the finest grid. (For the sake of simplicity, the volumes are not converted in distance and time units.) The definition of $\overline{S}_k$ may be unpractical because the size of the tile grows geometrically in our multiscale coarse-graining picture. Therefore, either $\ln(\overline{S}_k)$ or another average, a "mean logarithmic size," can be used: $\overline{s}_k = \Sigma_{\mathbf{l}} \overline{\alpha}_{\mathbf{l},\tilde{k}} (l_x + l_y + l_t)$. This definition of the resolution implies that $0 \le \overline{\varrho}_k \le 1$. The best achievable resolution $\overline{\varrho}_k = 1$ means that the observations allow us to determine the parameter attached to the local cell in the finest reference grid. Similarly, one has $0 \le \overline{s}_k \le n_x + n_y + n_t - 3$ and $\mathcal{A}(0,0,0) \le \overline{S}_k \le \mathcal{A}(n_x - 1, n_y - 1, n_z - 1)$, with the lower bounds corresponding to the finer scale.

3) Numerics and parallelization

The optimization on tilings has been parallelized. Indeed, one practical advantage of the statistical approach through the partition function is that it enumerates all configurations. This sum within the computation of the regularized cost function Eq. (28) can easily be parallelized. Only a summation of the partial sums needs to be performed after all threads are finished. The search of the optimal representations in the example of section 4 requires a few seconds to a few hours, using an eight-core computer (double quad-core Intel Xeon). Because of the dual approach, the complexity of the algorithm depends not only on the number of grid cells of the finest level N_{fg} but also on the regularization parameter β (the higher the slower) and the targeted number of tiles N. The optimization algorithm is performed by the quasi-Newton limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS-B) minimizer (Byrd et al. 1995).

4. Application to atmospheric dispersion

In this section, the methodology of section 3 is applied to an experiment drawn from the air pollution context, using real data. It is taken from Bocquet (2005a), in which changes of regular inversion grids were tested on synthetic data. After determining an optimal representation for a fixed N, a data assimilation experiment will be performed on this representation.

a. Finding an optimal representation

This application is taken from (Bocquet 2005a,b,c, 2007) and is focused on source inversion of the European Tracer Experiment (ETEX). This European Joint Researcher Centre 1994 campaign (Girardi et al. 1998) consisted of a pointwise 12-h-long release into the atmosphere of an inert tracer at continental scale. In the papers mentioned earlier, several three-dimensional inversions of the source were attempted successfully on either the synthetic or the real concentration measurements from the campaign. However, these inversions were based on a regular discretization of the 2D + T source space. These meshes comprised from a few thousand to one million control variables (from 2.25° × 2.25° to 0.5625° × 0.5625°).

The aim is to construct an optimal adaptive representation (a tiling) for inversions based on a much more limited number of cells, using the mathematical formalism developed in section 3.

As proposed in section 3, the gain in confidence $\mathcal{J}_\omega = \mathrm{Tr}(\mathbf{B}_\omega \mathbf{H}_\omega^\mathrm{T} \mathbf{R}^{-1} \mathbf{H}_\omega)$ (or reduction in uncertainty) will be optimized on a dictionary of tilings, with the following multiscale structure: The finest grid has dimensions $N_x = 64$, $N_y = 32$, and $N_t = 160$ for a spacing of $\Delta_x = \Delta_y = 0.5625°$ and $\Delta_t = 1$ h, whereas $n_x = 5$ scales are used for O_x, $n_y = 5$ scales are used for O_y, and $n_t = 5$ scales are used for O_t. The inverse temperature is set to $\beta = 10^5$.

In the following, a small set of 201 observations is used. It was diagnosed to have a high information content for the inversion (Bocquet 2008). The number of required tiles (402) is twice the number of observations, which is very limited when compared to the number of cells of regular grids.

The resulting mean logarithmic tile size $\overline{s}_k$ is displayed in Fig. 2. A suboptimal admissible tiling of 403 tiles is displayed in Fig. 3. Its reduction of uncertainty is $\mathcal{J}_\omega = 802.98$, whereas the optimal reduction of uncertainty is $\overline{\mathcal{J}} = 803.89$. For the coarsest grid in this multiscale structure, $\mathcal{J}_\omega = 23.42$, whereas for the finest grid, $\mathcal{J}_\omega = 1324.74$. The latter is the maximum over all configurations for all N, which is due to the scaling divergence described in section 2.

Note how small the number of tiles $N = 402$ is compared to $N_\mathrm{fg} = 327\,680$. Nonetheless, 60% of the average gain of confidence would potentially be captured using this optimal representation compared to the choice of the finest grid. The average reduction of uncertainty for optimal representations with a tile number in the range [80, 327 680] is plotted on Fig. 4. This graph emphasizes how well performing an adaptive representation of a very limited number of cells can be for the purpose of data assimilation, especially in comparison to regular grid representations.

b. Inverse modeling using the optimal representation

Inverse modeling will now be performed on the basis of the optimal representation (adaptive grid) previously obtained (Fig. 3). We follow the methodology of Bocquet (2007) and Krysta et al. (2008), which was based on a nonoptimal regular grid. The difference is that the inversion is now based on the adaptive optimal grid for $N = 402$.

A standard Gaussian prior for the source is chosen from the generic form, Eq. (8). However, it is now defined on an adaptive grid $\omega : \nu_\omega(\boldsymbol{\sigma}_\omega) = \exp(-\tfrac{1}{2}\boldsymbol{\sigma}_\omega^\mathrm{T} \mathbf{B}_\omega^{-1} \boldsymbol{\sigma}_\omega)/\sqrt{(2\pi)^N |\mathbf{B}_\omega|}$. As mentioned earlier, $\mathbf{B}_\omega = E_{\nu_\omega}[\boldsymbol{\sigma}_\omega \boldsymbol{\sigma}_\omega^\mathrm{T}]$ is related to $\mathbf{B}$ through $\mathbf{B}_\omega = \mathbf{I}_\omega \mathbf{B} \mathbf{I}_\omega^\mathrm{T}$. We assume that $\mathbf{B} = m^2 \mathbf{I}_{N_\mathrm{fg}}$, where m sets the release rate magnitude. Here, $m = 0.025M/12$ is chosen arbitrarily, where $M = 340$ kg is the true released mass of tracer in ETEX-I and 12 is the number of time steps of the true release in the finest grid. The errors are also chosen from the Gaussian a priori: $\nu_\epsilon(\boldsymbol{\epsilon}) = \exp(-\tfrac{1}{2}\boldsymbol{\epsilon}^\mathrm{T} \mathbf{R}^{-1} \boldsymbol{\epsilon})/\sqrt{(2\pi)^d |\mathbf{R}|}$, where the error covariance matrix is diagonal and homoscedastic $\mathbf{R} = \chi \mathbf{I}_d$, where χ is the only degree of freedom left to weight the departure from the observations and the departure from the prior.

The cost function, the by-product of the Bayesian inference, is then as follows:

$$\mathcal{L}_\omega(\boldsymbol{\sigma}_\omega) = \frac{1}{2}\boldsymbol{\sigma}_\omega^\mathrm{T} \mathbf{B}_\omega^{-1} \boldsymbol{\sigma}_\omega + \frac{1}{2}(\boldsymbol{\mu} - \mathbf{H}_\omega \boldsymbol{\sigma}_\omega)^\mathrm{T} \mathbf{R}^{-1} (\boldsymbol{\mu} - \mathbf{H}_\omega \boldsymbol{\sigma}_\omega), \tag{33}$$

which leads to the usual normal equations and estimators of the BLUE analysis. Here, ω is the conditional representation that has been obtained earlier. The value of χ is selected from an L-curve analysis following Davoine and Bocquet (2007) and Krysta et al. (2008): $\sqrt{\chi} = 0.3$ ng m^{-3}, which is equal to the value obtained from the regular grid ($2.25° \times 2.25° \times 1h$).

The integrated map of the reconstructed source is plotted on Fig. 5, as well as the retrieved profile in the vicinity of the true released site. The total retrieved mass of tracer is 879 kg, as compared to 684 kg obtained on a regular grid (Krysta et al. 2008). It has been diagnosed that 242 kg of tracer have been released in the vicinity of the true release site, as compared to 241 kg in Krysta et al. (2008). Therefore, the inversion result is very similar to the inversion on the regular grid but at a much lower cost. A difference is the excess in the total mass estimate, which can be attributed to observation–unconstrained aggregation error resulting from large tiles over the Atlantic and Ireland. Note that in Krysta

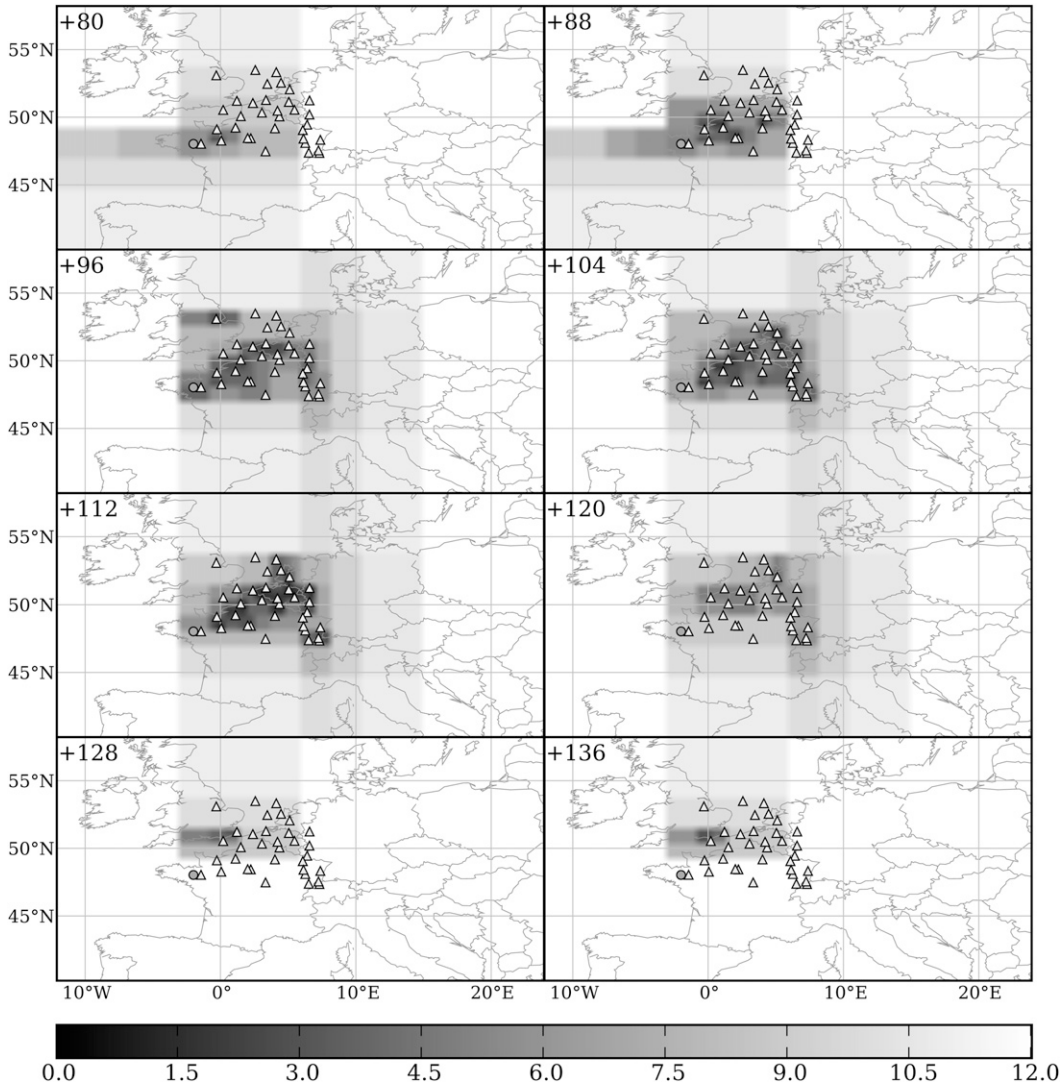

FIG. 2. Sequence of maps of mean tile logarithmic sizes $\overline{s}_k$ at different dates $t = 80, 88, 96, 104, 112, 120, 128$, and 136 h. The triangles indicate the measurement stations that are used. The circle points to the true release site of the ETEX campaign.

et al. (2008), 81 920 and 20 480 control variables are solved for, as compared to 403 tiles here. On the practical side, optimizations are significantly accelerated.

We have also tested this representation within a non-Gaussian prior framework. A Bernoulli prior, which ensures the positivity of the source, has been selected. The two parameters that enter the inversion have been selected using an iterative L-curve approach. As reported in Krysta et al. (2008), on a regular grid, this prior leads to a better estimate of the source than the Gaussian prior. However the profile is less satisfying, with significant fluctuations within the true release period.

The results obtained on the optimal adaptive grid are displayed in Fig. 5. The profile is smoother than with a regular grid. This is because the optimal representation determines the proper balance between the space and time scales. Given $N = 402$, the optimal time step is about 3–4 h.

5. Summary and perspectives

In geophysical data assimilation, the choice that is made in the discretization of the spatially distributed control space is not innocuous. It has a direct impact on the quality of the assimilation as measured by the reduction of uncertainty or by the gain in information through, for instance, the Fisher information matrix.

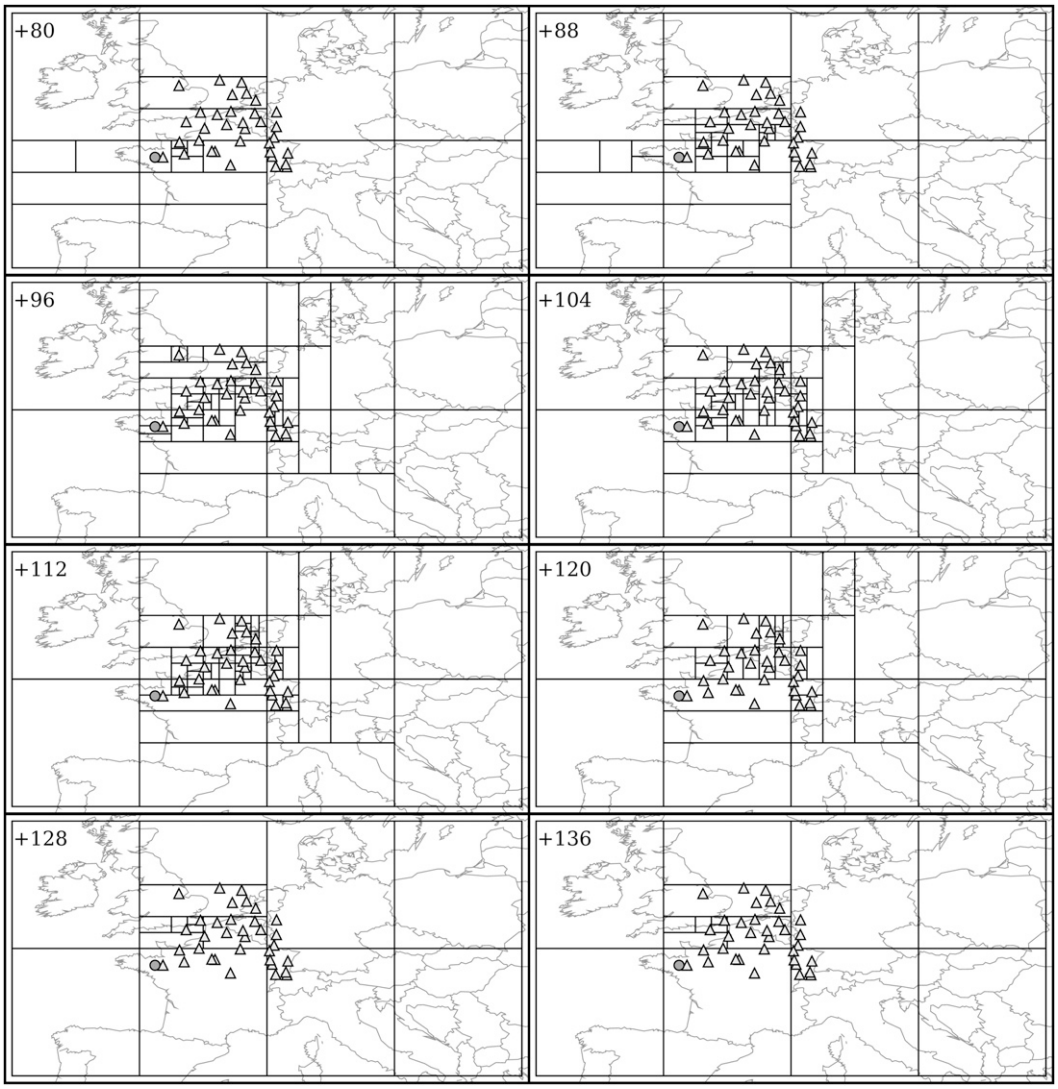

FIG. 3. Conditional tiling built from the average filling factor solution and displayed at several dates $t = 80, 88, 96, 104, 112, 120, 128$, and 136 h.

When the resolution of the control space discretization goes to zero, the input observational information remains the same while the uncertainty dramatically increases. Inverse modeling of atmospheric compounds is highly sensitive to this effect. Even though the information gain diverges, it is not spreading to the whole parameter space. Thus, in this limit, the estimator will not depart from the prior (outside the observation sites support). Therefore, there must be an optimal resolution for a regular space grid whose knowledge is even more needed in the latter context.

Hence, an idea is to consider the discretization of control space a degree of freedom. Instead of several regular grids at different scales, a richer set of representations to optimize on is given by adaptive grids. They have been defined as partitions (in the mathematical sense) of control space. Considering a three-dimensional control space 2D + T (typical of an atmospheric chemistry emission inverse problem), a product of binary trees was chosen, representing a tiling of the domain with nonoverlapping boxes.

To keep mathematical developments as simple as possible, the number of tiles of the partitions was fixed for all potential representations. The selected optimality criterion is the gain in information or confidence $\mathcal{J} = \mathrm{Tr}(\mathbf{B}\mathbf{P}_a^{-1} - \mathbf{I}_N)$ in the basis where control variables are independent with unit variance for the background error covariance matrix $\mathbf{B}$. This can also be understood as a measure of the average reduction of uncertainly for data assimilation analysis. The question addressed was

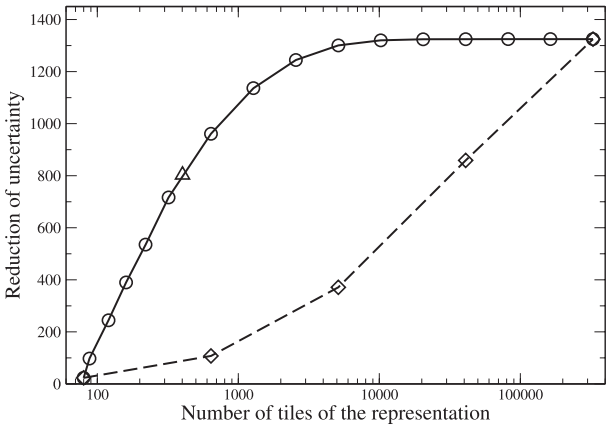

FIG. 4. Average gain of confidence (or reduction of uncertainty) for data assimilation as a function of the number of tiles of optimal adaptive representations. Note that the O_x scale is logarithmic. This shows that, at least in the context of this example and for a given monitoring network, a small optimally chosen adaptive representation can capture most of the information needed by data assimilation schemes. The triangle points to the value $N = 402$ that has been chosen for the inverse modeling experiment. Circles indicate the results of the optimizations that have actually been performed to plot this curve. Diamonds indicate the average reduction of uncertainty for regular grid representations at five different scales, which is much lower than its adaptive counterpart.

how to spatially arrange a fixed number of degrees of freedom in the control space, so as to maximize the quality of the subsequent data assimilation.

Once the dictionary of all representations has been described mathematically, a Lagrangian has been proposed to optimize the cost function while momentarily getting rid of the constraints (fixed number of tiles; one point–one tile). Because the cost function that is obtained appeared as a nonsmooth, possibly nonconvex, programming problem, it was regularized using a statistical mechanics analogy. The regularized cost function is convex, which guarantees the existence of a minimum.

This leads to rigorous concepts of local resolution, or local size of grid cells. The result from the regularized optimization consists of a superposition of admissible representations, which may not be admissible. Although it cannot be used straightforwardly in a data assimilation setup, it can be used to construct a suboptimal admissible representation.

In the last section, this mathematical formalism is applied on a realistic atmospheric tracer dispersion event: the European Tracer Experiment. The representation is optimized. A tiling of 403 cells that captures 60% of the average reduction of uncertainty is obtained. An inverse modeling experiment is then carried out on this conditional tiling. It leads to similar results as those obtained with a regular grid but with a much smaller number of grid cells. Depending on the target number of tiles, the computation may be much quicker. These tiles are supposed to better match the information load that can be delivered from the observations through the model and data assimilation analysis.

Preserving the continuity of this paper and confident in the ability of this formalism, we believe many developments need to be undertaken. We would like to mention five of them.

The formalism of this paper allowed for a fixed number of observations. Therefore, the optimization of the representation was focused on adaptive grids that would optimally dispatch the information from the observations onto the control space. From the mathematical side, this constraint can easily be relieved by getting rid of the ζ multiplier. However, the optimality criterion should be able to accommodate a free number of tiles. This is not the case for the criterion retained here. Indeed, in the singular regime case, the information from Fisher and also Kullback–Leibler has been shown to diverge in the high-resolution limit. This is a fundamental issue: how many degrees of freedom can be resolved by only the priors and the observations transported by the model? This aspect of the problem was kept low by choosing a fixed number of tiles. Further progress should clarify this question.

The effort of bringing data assimilation to a multiscale formalism along the proposed route is not complete. The data assimilation process was decomposed into the search of an optimal representation followed by the data assimilation inference on the parameters. Yet, what could be desired is the joint optimization of both the partition of control space and the actual values of the variables defined on this partition. In addition, the scale dependence of the model and representativeness errors were neglected, whereas we believe they should be accounted for in a fully consistent framework.

We would like to emphasize that this optimization could be useful not only in the singular regime, such as in air quality, but also in most data assimilation applications. Accordingly, the formalism should be extended to systems where the Jacobian is computed on the fly, which covers the majority of current data assimilation systems (sequential or variational).

In geophysical data assimilation, the description of the optimality system in terms of information exchanges remains conceptual. The information flow is not discriminated by individual components in the system (except for a distinction between the flows to the background and observation departures). This paper should also be understood as an effort to design individual components of the control space to match the flow of information coming from the observations and the priors. It shares conceptual similarities with the tracking of information

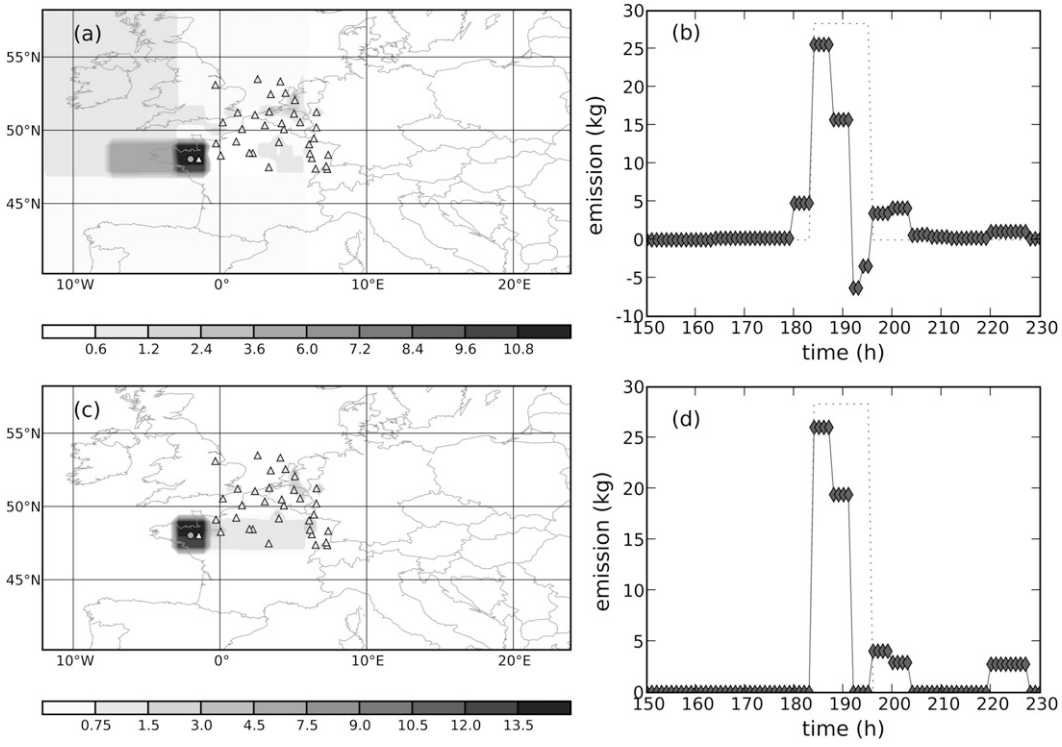

FIG. 5. (a),(c) Density plots of the time-integrated released mass retrieved by data assimilation. The mass unit is the kilogram. (b),(d) The reconstructed profile at the true release site with the true profile (dotted line). (a),(b) correspond to a Gaussian inversion, whereas (c),(d) are related to a non-Gaussian inversion.

exchanges between components of dynamical systems (Liang and Kleeman 2005).

Eventually, as far as applications are concerned, we believe that this analysis may be relevant to the inversion of greenhouse gases from pointwise concentrations or flux measurements. This representation issue is often put forward in the carbon inversion community, even though not necessarily in these terms. The hope is that such method would allow the determination of exactly how many and which degrees of freedom (fluxes) can be resolved unambiguously by the observations through the transport model.

Acknowledgments. The author is grateful to the organizers of the Banff International Research Station workshop ''Mathematical Advancement in Geophysical Data Assimilation,'' February 2008, which has fostered many exchanges and this special issue: K. Ide, P. Gauthier, C. K. R. T. Jones, and K. R. Thompson. The author would like to thank L. Wu for a careful reading of the manuscript and three anonymous reviewers for their suggestions. He acknowledges interesting discussions on related topics with P. Rayner and A. S. Denning. This paper is a contribution to the MSDAG project supported by the Agence Nationale de la Recherche, Grant ANR-08-SYSC-014.

REFERENCES

Bocquet, M., 2005a: Grid resolution dependence in the reconstruction of an atmospheric tracer source. *Nonlinear Processes Geophys.,* **12,** 219–234.

——, 2005b: Reconstruction of an atmospheric tracer source using the principle of maximum entropy. I: Theory. *Quart. J. Roy. Meteor. Soc.,* **131,** 2191–2208.

——, 2005c: Reconstruction of an atmospheric tracer source using the principle of maximum entropy. II: Applications. *Quart. J. Roy. Meteor. Soc.,* **131,** 2209–2223.

——, 2007: High resolution reconstruction of a tracer dispersion event. *Quart. J. Roy. Meteor. Soc.,* **133,** 1013–1026.

——, 2008: Inverse modelling of atmospheric tracers: Non-Gaussian methods and second-order sensitivity analysis. *Nonlinear Processes Geophys.,* **15,** 127–143.

Bousquet, P., P. Peylin, P. Ciais, P. Le Quéré, P. Friedlingstein, and P. P. Tans, 2000: Regional changes in carbon dioxide fluxes of land and oceans since 1980. *Science,* **290,** 1342–1346.

Byrd, R. H., P. Lu, and J. Nocedal, 1995: A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.,* **16,** 1190–1208.

Constantinescu, E., A. Sandu, and G. R. Carmichael, 2008: Modeling atmospheric chemistry and transport with dynamic adaptive resolution. *Comput. Geosci.,* **12,** 133–151.

Courtier, P., and O. Talagrand, 1990: Variational assimilation of meteorological observations with the direct and adjoint shallow-water equations. *Tellus,* **42A,** 531–549.

——, J.-N. Thépaut and A. Hollingsworth, 1994: A strategy for operational implementation of 4D-Var, using an incremental approach. *Quart. J. Roy. Meteor. Soc.,* **120,** 1367–1387.

Cover, T. M., and J. A. Thomas, 1991: *Elements of Information Theory*. Wiley, 542 pp.

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

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

Fan, S.-M., M. Gloor, J. Mahlman, S. Pacala, J. L. Sarmiento, T. Takahashi, and P. P. Tans, 1998: Atmospheric and oceanic CO_2 data and models imply a large terrestrial carbon sink in North America. *Science,* **282,** 442–444.

Fisher, M., 2003: Estimation of entropy reduction and degrees of freedom for signal for large variational analysis systems. ECMWF Tech. Rep. 397, 20 pp. [Available online at http://www.ecmwf.int/publications/library/ecpublications/_pdf/tm/301-400/tm397.pdf.]

Furbish, D., M. Y. Hussaini, F. -X. Le Dimet, P. Ngnepieba, and Y. Wu, 2008: On discretization error and its control in variational data assimilation. *Tellus,* **60A,** 979–991.

Ghil, M., and P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography. *Advances in Geophysics,* Vol. 33, Academic Press, 141–266.

Girardi, F., and Coauthors, 1998: The European Tracer Experiment. Office for Official Publications of the European Communities Tech. Rep. EUR 18143 EN.

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

Jaynes, E. T., 1957a: Information statistics and statistical mechanics. *Phys. Rev.,* **106,** 620–630.

——, 1957b: Information statistics and statistical mechanics II. *Phys. Rev.,* **108,** 171–190.

Knuth, D., 1997: *Fundamental Algorithms*. Vol 1, *The Art of Computer Programming,* 3rd ed. Addison-Wesley, 650 pp.

Krysta, M., and M. Bocquet, 2007: Source reconstruction of an accidental radionuclide release at European scale. *Quart. J. Roy. Meteor. Soc.,* **133,** 529–544.

——, ——, and J. Brandt, 2008: Probing ETEX-II data set with inverse modelling. *Atmos. Chem. Phys.,* **8,** 3963–3971.

Le Dimet, F.-X., and O. Talagrand, 1986: Variational algotrithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus,* **28A,** 97–110.

Liang, X. S., and R. Kleeman, 2005: Information transfer between dynamical system components. *Phys. Rev. Lett.,* **95,** 244101, doi:10.1103/PhysRevLett.95.244101.

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

Pachauri, R. K., and A. Reisinger, Eds., 2007: *Contribution of Working Groups I, II, and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* IPCC, 104 pp.

Peylin, P., and Coauthors, 2005: Daily CO_2 flux estimates over Europe from continuous atmospheric measurements: 1, Inverse methodology. *Atmos. Chem. Phys. Discuss.,* **5,** 1647–1678.

Purser, R., and H. -L. Huang, 1993: Estimating effective data density in a satellite retrieval or an objective analysis. *J. Appl. Meteor.,* **32,** 1092–1107.

Rodgers, C. D., 2000: *Inverse Methods for Atmospheric Sounding: Theory and Practice*. Series on Atmopsheric, Oceanic and Planetary Physics, Vol. 2, World Scientific, 240 pp.

Saad, Y., 2003: *Iterative Methods for Sparse Linear Systems*. 2nd ed. SIAM, 528 pp.

Tarantola, A., and B. Valette, 1982: Inverse problems = Quest for information. *J. Geophys. Res.,* **50,** 159–170.