RMetS
Royal Meteorological Society

Bayesian design of control space for optimal assimilation of observations. Part I: Consistent multiscale formalism

M. Bocquet,[a,b]* L. Wu[a,b] and F. Chevallier[c]

[a] *Université Paris-Est, CEREA, Joint laboratory École des Ponts ParisTech and EDF R&D, Champs-sur-Marne, France*
[b] *INRIA, Paris Rocquencourt Research Centre, France*
[c] *Laboratoire des Sciences du Climat et de l'Environnement/IPSL, CEA-CNRS-UVSQ, Gif-sur-Yvette, France*
*Correspondence to: M. Bocquet, CEREA, École des Ponts ParisTech, 6–8 avenue Blaise Pascal,
Cité Descartes Champs-sur-Marne, 77455 Marne la Vallée Cedex, France. E-mail: bocquet@cerea.enpc.fr

In geophysical data assimilation, the control space is by definition the set of parameters which are estimated through the assimilation of observations. It has recently been proposed to design the discretizations of control space in order to assimilate observations optimally. The present paper describes the embedding of that formalism in a consistent Bayesian framework. General background errors are now accounted for. Scale-dependent errors, such as aggregation errors (that lead to representativeness errors) are consistently introduced. The optimal adaptive discretizations of control space minimize a criterion on a dictionary of grids. New criteria are proposed: degrees of freedom for the signal (DFS) built on the averaging kernel operator, and an observation-dependent criterion.

These concepts and results are applied to atmospheric transport of pollutants. The algorithms are tested on the European tracer experiment (ETEX), and on a prototype of CO_2 flux inversion over Europe using a simplified CarboEurope-IP network. New types of adaptive discretization of control space are tested such as quaternary trees or factorised trees. Quaternary trees are proven to be both economical, in terms of storage and CPU time, and efficient on the test cases. This sets the path for the application of this methodology to high-dimensional and noisy geophysical systems. Part II of this article will develop asymptotic solutions for the design of control space representations that are obtained analytically and are contenders to exact numerical optimizations. Copyright © 2011 Royal Meteorological Society

Key Words: multiscale data assimilation; atmospheric chemistry; reduction method; ETEX

Received 30 September 2010; Revised 1 April 2011; Accepted 6 April 2011; Published online in Wiley Online Library 17 May 2011

Citation: Bocquet M, Wu L, Chevallier F. 2011. Bayesian design of control space for optimal assimilation of observations. Part I: Consistent multiscale formalism. *Q. J. R. Meteorol. Soc.* **137**: 1340–1356. DOI:10.1002/qj.837

1. Introduction

1.1. *The resolution issue*

Researchers using inverse modelling techniques in atmospheric chemistry have faced the so-called 'resolution problem'.

A first example is given by the gridded emission inventories which are multidimensional fields and key components of the models. Unfortunately, the uncertainty of these fields is quite high (of the order of 40% for the ozone precursors in air quality at continental scale, for instance). Observations could help to constrain the emission fields through inverse modelling and reduce this uncertainty, e.g. Elbern *et al.* (2007) for an application to the precursors of ozone, or Davoine and Bocquet (2007) for an application to an accidental release of radionuclides. Both the model equations and the control space of the emission field need to be discretised at some predefined space and time resolution. The space and time resolutions of the discretised control space are not necessarily the same as those of state space. There is a non-trivial choice of resolution to be made. Furthermore, inventories are built at a given resolution, the model runs at another, and the data assimilation scheme injects the information of all observations into the system at still another scale depending on the nature of the instruments: ground-based, satellite, radar, lidar, etc. Therefore, the system should ideally be considered multiscale.

Another example pertains to the inverse modelling of greenhouse gases. Early carbon flux inversions relied on a partition of the globe (the control space of fluxes) into about 20 sub-domains representing several types of continental or ocean exchange with the atmosphere, with an annual or monthly time resolution (e.g. Fan *et al.*, 1998; Bousquet *et al.*, 2000). This was necessary because of the limited computational power together with a limited number of precise observations of CO_2 concentration. However, such gross partitioning led to severe aggregation errors (Trampert and Sneider, 1996; Kaminski *et al.*, 2001). Thus it is tempting to increase the space and time resolutions of control space. But the total number of variables could dramatically exceed the total number of observations. Besides, because of the nature of transport and dispersion, the inverse modelling problem is ill-posed. Therefore a regularisation is needed (Rödenbeck *et al.*, 2003), which can be written as a Tikhonov regularising term, as is usually done in geophysical data assimilation. This regularisation, which spatially and temporally correlates the errors, may stem from real physical correlations due, for instance, to similar ecosystems (Chevallier *et al.*, 2006). But it may also be artificial and correspond to a smooth aggregate of variables. Note that this distinction is not always made clear in the literature.

In both cases, there is a difficult choice to be made on the resolution of control space. To make the problem worse, Bocquet (2005) has shown that, for atmospheric dispersion problems, the source estimation of atmospheric pollutant from inversions using pointwise measurements depends strongly on the control space resolution, even when using a proper classical Tikhonov regularisation (background-error term of quadratic form in the cost function).

1.2. Multiscale approach

To partially solve this resolution issue, a multiscale framework for such inversions was proposed (Bocquet, 2009). It is at the crossroads between a coarse partitioning of control space subject to aggregation errors and a highly resolved control space where regularisation is decisive. The method consists of constructing an adaptive grid of control space (also called a *representation* of control space in the following). This adaptive grid is optimal in the sense that it is designed to optimally capture the information carried by the observations and inject into control space through a model and the assimilation system. This is achieved by maximizing an objective function that measures the reduction of uncertainty granted by the observations on a space of all potential adaptive grids (later called a *dictionary* or *class*).

The method quantifies how observational information is propagated into control space. It diagnoses poorly observed areas. It informs how space- and time-scales should be related for the problem at hand. Also, it has strong algorithmic implication. Indeed, the method shows how to devise adaptive grids of control space that have significantly fewer grid cells than the original finest regular grid, but which can still capture most of the information content of observations. Such an adaptive grid was built and tested on the European Tracer Experiment (ETEX; Nodop *et al.*, 1998). The inversion of the source term of this dispersion event was performed much faster with an optimization over about 100 times fewer independent variables in this adaptive grid, with results very similar to those obtained with a regular fine grid.

The method also offers a starting point for a general conceptual and mathematical framework for multiscale data assimilation in atmospheric chemistry, or in other areas of geophysics.

This two-part article aims to continue and improve the potential of this formalism and prepare for large-scale applications. The first part explores a few essential questions still unanswered, such as

- Can the Bayesian approach that is currently used in geophysical data assimilation be made consistent with the multiscale framework of the method?
- Can a non-diagonal background-error covariance matrix be taken into account in this formalism? Such matrices are often used in air quality, greenhouse gas flux inversions and, more generally, in data assimilation schemes for geophysical forecasting systems.
- Can scale-dependent errors be accounted for?
- Can one use other grid optimization objective functions, such as DFS, or observation-dependent criteria?
- Can one perform the optimization within a simpler or more economical dictionary of adaptive grids than the so-called *tiling* dictionary introduced by Bocquet (2009)?

The results are obtained with a view to applications in atmospheric chemistry and air quality, but most of the findings are more general and could be applied outside this scope whenever the choice of control space is complex and decisive.

1.3. Outline

The conceptual and mathematical framework will be presented in section 2. The multiscale description of control space is made consistent with the assimilation of observations using Bayesian principles.

Section 3 deals with errors which may enter the inversions, and which are scale-dependent. Of particular interest are the aggregation errors occurring when grid cells are merged. They lead to representativeness errors.

The construction of optimal representations of control space requires the definition of a criterion that ranks adaptive grids in a given dictionary of representations. In addition to the so-called Fisher criterion introduced by Bocquet (2009), we add two new criteria in section 4. One is based on the DFS which measures the theoretical information gain in the analysis. A third criterion is defined with an objective function that not only depends on the prior statistics but also on the observations themselves.

In section 5, most of the developments will be illustrated on two test cases: the ETEX-I dispersion event using real measurements and realistic physics (from a chemistry and transport model), and another demonstration case based on a simplified European CO_2 network (CarboEurope-IP).

In section 6, the dictionary of tilings is compared to a dictionary of quaternary tree structures (later called *qtrees*). Although suggested in Bocquet (2009), the quaternary tree structure was not tested and studied there.

Finally, in section 7, we summarise the results. We discuss its connection with other multiscale formalisms introduced very recently in data assimilation. We also discuss the scope of the method and its extension to nonlinear models. Elements that justify the need for Part II (Bocquet *et al.*, 2011) of this work are explained.

2. Multiscale modelling

This section extends the multiscale approach developed in Bocquet (2009). It goes farther on several points, and unifies the concepts using a Bayesian methodology.

2.1. *Data assimilation context*

A simplified typical data assimilation set-up is employed. For the data assimilation problem at hand, the control space, named after its domain Ω, is discretised into cells of a grid ω. This grid can be regular (grid cells of equal size in a given system of coordinates), or not. It may have several space dimensions, and possibly one time dimension. For instance, in atmospheric chemistry inverse modelling, the control space is often the vector space of emission gas fluxes from the ground, at any time. Therefore there are two space dimensions plus the time dimension (2D+T). A vector representing a discretised flux or source field is denoted $\boldsymbol{\sigma}$.

A clear distinction is made between control space and state space, which could be discretised at different resolutions, although they could share a subspace or even be identical.

We assume that the measurement vector $\boldsymbol{\mu}$ is related to the source $\boldsymbol{\sigma}$ through a Jacobian matrix $\mathbf{H}$, which stands for both the system's evolution model and the observation operator. It could result from the linearisation of models, but for simplicity we shall hypothesise that the models are linear. The equation that links the observation to the source via the models reads

$$\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\sigma} + \boldsymbol{\epsilon}\,, \qquad (1)$$

where $\boldsymbol{\epsilon}$ is the vector of errors (of any type). Note that space and time are not split up in this simplified equation, so that $\mathbf{H}$ links elements in space and time. We assume that $\boldsymbol{\sigma}$ follows a Gaussian prior probability density function (pdf): $\boldsymbol{\sigma} \sim \mathcal{N}(\boldsymbol{\sigma}_\mathrm{b}, \mathbf{B})$ where $\boldsymbol{\sigma}_\mathrm{b}$ is the first guess, and $\mathbf{B}$ the background-error covariance matrix. The errors are supposed to be unbiased and follow a normal pdf: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, where $\mathbf{R}$ is the observational-error covariance matrix. We shall designate by $\boldsymbol{\sigma}_\mathrm{a}$ and $\mathbf{P}_\mathrm{a}$ the analysed source and the analysis-error covariance matrix respectively. Both result from a data assimilation or inverse modelling analysis.

2.2. *Multiscale framework*

The control space discretization is discussed now, as well as the way to define a multiscale Jacobian.

2.2.1. *Multiscale mesh*

A multiscale mesh is defined first. It is assumed that the domain Ω is discretised into a fine-resolution regular grid, which represents the finest available discretization. The number of grid cells in the grid is N_fg. Grid cells at coarser scales will be obtained by *dyadic* coarse-grainings of cells in the finest grid, i.e. two grid cells that are adjacent along one direction can be merged into a coarser cell (hence the adjective *dyadic* which qualifies this binary grouping). The dyadic coarse-grainings can be performed in each space or time direction of the domain Ω. The number of coarse-grainings in each direction is limited by the number of accessible scales denoted by n_x, n_y, and n_t for a 2D+T domain (ETEX-I case), or n_x and n_y for a 2D domain (simplified CarboEurope-IP case). In the ETEX-I, and similarly in the simplified CarboEurope-IP, each coarse-grained cell has an intrinsic scale vector of integers

$$\mathbf{l} = (l_x, l_y, l_t)\,,$$

where $0 \le l_x < n_x$, $0 \le l_y < n_y$, and $0 \le l_t < n_t$. The scale levels are set by l_x, l_y and l_t. For each direction, label 0 corresponds to the finest scale. For instance, the cells in the finest regular grid all share the same scale vector $\mathbf{l} = (0, 0, 0)$.

2.2.2. *Multiscale Jacobian*

Correspondingly, the Jacobian defined in Eq. (1) is generalized to a multiscale Jacobian. $\mathbf{H}$ is usually computed in the finest regular grid, using either direct forward simulations or backward adjoint simulations. Then dyadic coarse-grainings of the Jacobian are performed by simple averaging. Note that the multiscale Jacobian could also be defined using several Jacobians obtained at different scales from different models, or different versions of the same core model.

2.2.3. *Adaptive representations*

Using this multiscale framework, one can build representations (adaptive grids) of Ω. A representation ω is a set of cells of many sizes (depending on the scale of the cell), that cover Ω. A set, or dictionary, of representations will generically be called $\mathcal{R}(\Omega)$. Besides, a representation will be called *admissible*, if it is a strict partition of Ω. That is, a single grid cell corresponds to each point in Ω.

Several kinds of multiscale structures were contemplated in Bocquet (2009). In each case, successive time coarse-grainings were represented by a binary tree. 2D space could be considered as the *tensor product* of two binary trees, one for each space direction. This means that the grid cells, or *tiles* of such a representation are the Kronecker products of two 1D elements of binary trees, one for each direction. This led to the so-called *tiling* representations.

In the case of two directions of space, one could use instead a quaternary tree, called *qtree* later. This means that each mother tile can be refined into four daughter tiles, instead of two. This reduces the space occupied by the multiscale Jacobian at the expense of a smaller (therefore less rich) dictionary $\mathcal{R}(\Omega)$. Note that the dictionary of qtrees is included in the dictionary of tilings (any qtree is a tiling). In Figure 1, a 2D tiling made of the *tensorial product* of two binary trees, one for each space direction,

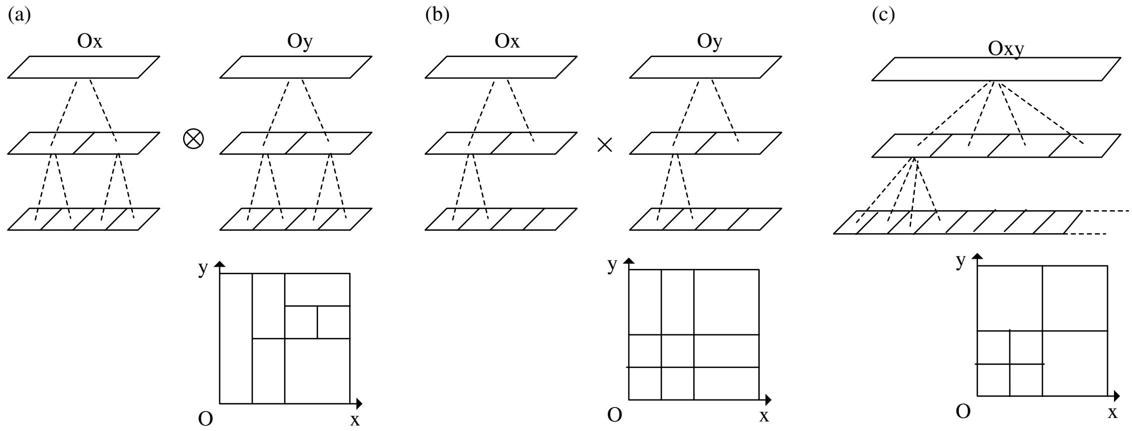

Figure 1. Schematic illustration of three types of structure, in two dimensions. (a) shows the *tensorial product* structure of two binary trees, one for O*x*, one for O*y*, to produce a collections of 2D tilings. (b) shows the direct product structure of two binary trees (or ftree), one for O*x*, one for O*y*. (c) shows a quaternary tree structure, or qtree. In each case, an example of a generated grid is drawn.

is plotted. An example of qtree is also shown. A third type of representation, which is the direct product of two binary trees (called *factorised tree* or *ftree* later) is also displayed.

2.3. Restriction and prolongation

To climb up or down the scales in the multiscale ladder, one needs to define a *restriction operator* that tells how a source is coarse-grained, and a *prolongation operator* that tells how the source is refined through the scales. Rodgers (2000) gives an in-depth discussion on the topic.

First, let us consider the restriction operator. Assume $\boldsymbol{\sigma}$ is a source vector which is known in the finest regular grid. Let ω be an adaptive representation of a dictionary $\mathcal{R}(\Omega)$. The coarse-graining of $\boldsymbol{\sigma}$ in ω is defined by $\boldsymbol{\sigma}_\omega = \boldsymbol{\Gamma}_\omega \boldsymbol{\sigma}$, where $\boldsymbol{\Gamma}_\omega : \mathbb{R}^{N_{\rm fg}} \to \mathbb{R}^N$ stands for the coarse-graining operator. This operator is supposed to be unambiguously defined. In most of the article, we suppose it identifies with simple averaging. But the formalism does not rule out more complex coarse-graining with associated prolongation operator given by a spline interpolation, or model-specific coarser Jacobians.

A source can also be refined thanks to a prolongation operator $\boldsymbol{\Gamma}^\star_\omega : \mathbb{R}^N \to \mathbb{R}^{N_{\rm fg}}$ which refines $\boldsymbol{\sigma}_\omega$ into $\boldsymbol{\sigma} = \boldsymbol{\Gamma}^\star_\omega \boldsymbol{\sigma}_\omega$. This operator is ambiguous, since additional information is needed to reconstruct a source at higher resolution. One possible choice, which we shall call the deterministic one, is to set $\boldsymbol{\Gamma}^\star_\omega = \boldsymbol{\Gamma}^{\rm T}_\omega$. A schematic of the use of the restriction and prolongation operators is displayed in Figure 2

However, in this data assimilation framework, one has prior information on the source that may be exploited. The pdf $q(\boldsymbol{\sigma})$ gives prior information on $\boldsymbol{\sigma}$. Following the statistical assumptions after Eq. (1), it is chosen to be Gaussian $q(\boldsymbol{\sigma}) \sim \mathcal{N}(\boldsymbol{\sigma}_{\rm b}, \mathbf{B})$. From this prior defined in the finest regular grid, one can infer, thanks to $\boldsymbol{\Gamma}_\omega$, the prior pdf of $\boldsymbol{\sigma}$ in representation ω

$$q_\omega(\boldsymbol{\sigma}_\omega) \sim \mathcal{N}(\boldsymbol{\sigma}^{\rm b}_\omega, \mathbf{B}_\omega), \tag{2}$$

with

$$\boldsymbol{\sigma}^{\rm b}_\omega = \boldsymbol{\Gamma}_\omega \boldsymbol{\sigma}_{\rm b}, \qquad \mathbf{B}_\omega = \boldsymbol{\Gamma}_\omega \mathbf{B} \boldsymbol{\Gamma}^{\rm T}_\omega. \tag{3}$$

Conversely, assume one knows $\boldsymbol{\sigma}_\omega$ in representation ω. Since the problem is underdetermined, then one could opt for the most likely refinement. It is given by the mode of $q(\boldsymbol{\sigma}|\boldsymbol{\sigma}_\omega)$. From Bayes' rule, it is clear that

$$q(\boldsymbol{\sigma}|\boldsymbol{\sigma}_\omega) = \frac{q(\boldsymbol{\sigma})}{q_\omega(\boldsymbol{\sigma}_\omega)} \delta\left(\boldsymbol{\sigma}_\omega - \boldsymbol{\Gamma}_\omega \boldsymbol{\sigma}\right), \tag{4}$$

where δ is the Dirac distribution. Then the mode of this posterior Gaussian distribution is given by

$$\boldsymbol{\sigma}^\star = \boldsymbol{\sigma}_{\rm b} + \mathbf{B}\boldsymbol{\Gamma}^{\rm T}_\omega \left(\boldsymbol{\Gamma}_\omega \mathbf{B} \boldsymbol{\Gamma}^{\rm T}_\omega\right)^{-1} (\boldsymbol{\sigma}_\omega - \boldsymbol{\Gamma}_\omega \boldsymbol{\sigma}_{\rm b}). \tag{5}$$

Thus $\boldsymbol{\Gamma}^\star_\omega$ would be an affine operator. We denote by $\boldsymbol{\Lambda}^\star_\omega$ its tangent linear component

$$\boldsymbol{\Lambda}^\star_\omega \equiv \mathbf{B}\boldsymbol{\Gamma}^{\rm T}_\omega \left(\boldsymbol{\Gamma}_\omega \mathbf{B} \boldsymbol{\Gamma}^{\rm T}_\omega\right)^{-1}. \tag{6}$$

Moreover, we define

$$\boldsymbol{\Pi}_\omega \equiv \boldsymbol{\Lambda}^\star_\omega \boldsymbol{\Gamma}_\omega = \mathbf{B}\boldsymbol{\Gamma}^{\rm T}_\omega \left(\boldsymbol{\Gamma}_\omega \mathbf{B} \boldsymbol{\Gamma}^{\rm T}_\omega\right)^{-1} \boldsymbol{\Gamma}_\omega, \tag{7}$$

so that we can choose as a prolongation operator

$$\boldsymbol{\Gamma}^\star_\omega \equiv (\mathbf{I}_{N_{\rm fg}} - \boldsymbol{\Pi}_\omega)\boldsymbol{\sigma}_{\rm b} + \boldsymbol{\Lambda}^\star_\omega, \tag{8}$$

where $\mathbf{I}_{N_{\rm fg}}$ is the identity operator. Since the refinement is now a probabilistic process, errors are attached to it. The corresponding error covariance matrix is

$$\begin{aligned}\mathbf{P}^\star_\omega &= \mathbf{B} - \mathbf{B}\boldsymbol{\Gamma}^{\rm T}_\omega \left(\boldsymbol{\Gamma}_\omega \mathbf{B} \boldsymbol{\Gamma}^{\rm T}_\omega\right)^{-1} \boldsymbol{\Gamma}_\omega \mathbf{B} \\ &= \left(\mathbf{I}_{N_{\rm fg}} - \boldsymbol{\Pi}_\omega\right) \mathbf{B}.\end{aligned} \tag{9}$$

As expected, if the representation ω is close to the finest grid, $\{N_{\rm fg} - {\rm Rank}(\boldsymbol{\Pi}_\omega)\}/N_{\rm fg} \ll 1$, the refinement error is negligible. If the representation is coarse, ${\rm Rank}(\boldsymbol{\Pi}_\omega)/N_{\rm fg} \ll 1$, the refinement error is limited by that of the background.

Those operators first satisfy

$$\boldsymbol{\Gamma}_\omega \boldsymbol{\Gamma}^\star_\omega = \mathbf{I}_N, \tag{10}$$

which is a consistency identity. Any reasonable prolongation operator should satisfy it. Then, one verifies that

$$\boldsymbol{\Gamma}^\star_\omega \boldsymbol{\Gamma}_\omega = \left(\mathbf{I}_{N_{\rm fg}} - \boldsymbol{\Pi}_\omega\right) \boldsymbol{\sigma}_{\rm b} + \boldsymbol{\Pi}_\omega. \tag{11}$$

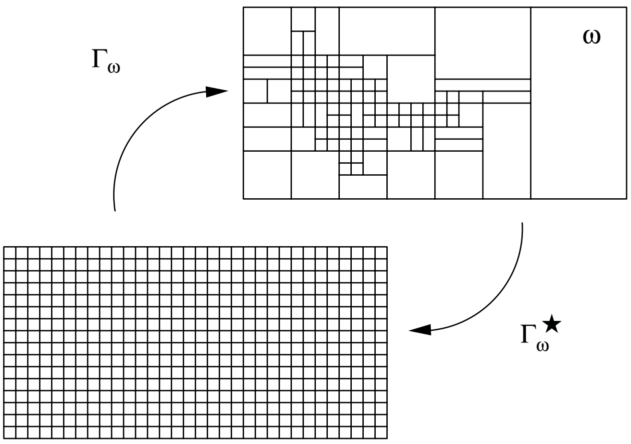

Figure 2. Schematic of the restriction and prolongation operators from the finest regular grid to a representation (adaptive grid) ω, and vice versa.

The linear operator $\mathbf{\Pi}_\omega$ is a projector since it can be checked that $\mathbf{\Pi}_\omega^2 = \mathbf{\Pi}_\omega$. Besides, it is $\mathbf{B}^{-1}$-symmetric since

$$\begin{aligned}\langle \mathbf{x}, \mathbf{\Pi}_\omega \mathbf{y}\rangle_{\mathbf{B}^{-1}} &= \mathbf{x}^{\mathrm{T}}\mathbf{B}^{-1}\mathbf{B}\mathbf{\Gamma}_\omega^{\mathrm{T}}\left(\mathbf{\Gamma}_\omega\mathbf{B}\mathbf{\Gamma}_\omega^{\mathrm{T}}\right)^{-1}\mathbf{\Gamma}_\omega\mathbf{y}\\ &= \mathbf{x}^{\mathrm{T}}\mathbf{\Pi}_\omega^{\mathrm{T}}\mathbf{B}^{-1}\mathbf{y}\\ &= \langle \mathbf{\Pi}_\omega^{\mathrm{T}}\mathbf{x}, \mathbf{y}\rangle_{\mathbf{B}^{-1}}\,, \end{aligned}\qquad (12)$$

where $\langle , \rangle_{\mathbf{B}^{-1}}$ is the scalar product built on $\mathbf{B}^{-1}$. In matrix form, this is equivalent to

$$\mathbf{\Pi}_\omega\mathbf{B} = \mathbf{B}\mathbf{\Pi}_\omega^{\mathrm{T}}\,. \qquad (13)$$

$\mathbf{\Pi}_\omega$ cannot be the identity because the coarse-graining implies a loss of information that, in general, cannot be fully recovered. The approach will be called the Bayesian or probabilistic prescription of $\mathbf{\Gamma}^\star_\omega$.

2.4. *Observation equation in any representation*

The mathematical formalism being laid, the observation equation Eq. (1) can be written in any representation ω of $\mathcal{R}(\Omega)$. The Jacobian $\mathbf{H}$ becomes $\boldsymbol{H}_\omega = \mathbf{H}\mathbf{\Gamma}^\star_\omega$. Inheriting from $\mathbf{\Gamma}^\star_\omega$, $\boldsymbol{H}_\omega$ is an affine operator. The observation equation reads

$$\begin{aligned}\boldsymbol{\mu} &= \boldsymbol{H}_\omega\boldsymbol{\sigma}_\omega + \boldsymbol{\epsilon}_\omega = \mathbf{H}\mathbf{\Gamma}^\star_\omega\mathbf{\Gamma}_\omega\boldsymbol{\sigma} + \boldsymbol{\epsilon}_\omega\\ &= \mathbf{H}\boldsymbol{\sigma}_{\mathrm{b}} + \mathbf{H}\mathbf{\Pi}_\omega(\boldsymbol{\sigma} - \boldsymbol{\sigma}_{\mathrm{b}}) + \boldsymbol{\epsilon}_\omega\,.\end{aligned}\qquad (14)$$

The error $\boldsymbol{\epsilon}_\omega$ has been made scale-dependent, because several sources of errors depend on the scale, such as the aggregation errors, or the errors in model subgrid physical parametrisations.

2.5. *Reduction of the correlated case*

When $\mathbf{B}$ is not diagonal, correlation between errors of values defined on different tiles will occur. Non-zero covariances in $\mathbf{B}$ may come from true physical correlation in the errors. They may also come from imposed correlations between variables, a form of variable aggregation, or coarse-graining. This second case is discarded here because our coarsening scheme already copes with this explicitly.

Off-diagonal terms in $\mathbf{B}$, which induce correlations between tiles, complicate the optimization scheme considerably. In particular the calculation of $\mathbf{\Gamma}^\star_\omega$ entails the representation-dependent computation of the inverse of $\mathbf{B}_\omega$.

One way out of this is to redefine the original coarsening scheme $\mathbf{\Gamma}_\omega$ so that $\mathbf{B}$ induces no error cross-correlations between coarse-grained tiles. To do so, one defines a new coarse-graining operator, a substitute for $\mathbf{\Gamma}_\omega$:

$$\widetilde{\mathbf{\Gamma}}_\omega = \mathbf{\Gamma}_\omega\mathbf{B}^{-1/2}\,. \qquad (15)$$

This implies that the adaptive grid cells no longer represent a partition of the control space domain. Instead, they are (*a priori*) statistically independent linear combinations of the former cells. Coarse-graining is now applied to these combinations, maintaining the property of statistical independence, rather than to the original grid cells.

As a result, redefining $\mathbf{\Gamma}_\omega$ into $\widetilde{\mathbf{\Gamma}}_\omega$, the background-error covariance matrix in representation ω becomes

$$\mathbf{B}_\omega = \widetilde{\mathbf{\Gamma}}_\omega\mathbf{B}\widetilde{\mathbf{\Gamma}}_\omega^{\mathrm{T}} = \mathbf{\Gamma}_\omega\mathbf{\Gamma}_\omega^{\mathrm{T}}\,. \qquad (16)$$

Also the prolongation operator Eq. (8) changes according to

$$\begin{aligned}\widetilde{\boldsymbol{\sigma}}^\star_\omega &= \boldsymbol{\sigma}_{\mathrm{b}} + \mathbf{B}\widetilde{\mathbf{\Gamma}}_\omega^{\mathrm{T}}\left(\widetilde{\mathbf{\Gamma}}_\omega\mathbf{B}\widetilde{\mathbf{\Gamma}}_\omega^{\mathrm{T}}\right)^{-1}\left(\widetilde{\boldsymbol{\sigma}}_\omega - \widetilde{\mathbf{\Gamma}}_\omega\boldsymbol{\sigma}_{\mathrm{b}}\right)\\ &= \boldsymbol{\sigma}_{\mathrm{b}} + \mathbf{B}^{1/2}\mathbf{\Gamma}_\omega^{\mathrm{T}}\left(\mathbf{\Gamma}_\omega\mathbf{\Gamma}_\omega^{\mathrm{T}}\right)^{-1}\left(\widetilde{\boldsymbol{\sigma}}_\omega - \mathbf{\Gamma}_\omega\mathbf{B}^{-1/2}\boldsymbol{\sigma}_{\mathrm{b}}\right).\end{aligned}\qquad (17)$$

As a consequence, one obtains

$$\widetilde{\mathbf{\Gamma}}_\omega\widetilde{\mathbf{\Gamma}}^\star_\omega = \mathbf{I}_N\,, \qquad (18)$$

$$\widetilde{\mathbf{\Lambda}}^\star_\omega\widetilde{\mathbf{\Gamma}}_\omega = \mathbf{B}^{1/2}\mathbf{\Gamma}_\omega^{\mathrm{T}}(\mathbf{\Gamma}_\omega\mathbf{\Gamma}_\omega^{\mathrm{T}})^{-1}\mathbf{\Gamma}_\omega\mathbf{B}^{-1/2} = \widetilde{\mathbf{\Pi}}_\omega\,. \qquad (19)$$

One checks also that $\widetilde{\mathbf{\Pi}}_\omega$ is $\mathbf{B}^{-1}$-symmetric.

3. Accounting for scale-dependent errors

The observations in $\boldsymbol{\mu}$ are representative of some scale. This scale may not be accessible to modellers. Vector $\boldsymbol{\mu}$ is related to $\boldsymbol{\sigma}$ at the finest scale, but also $\boldsymbol{\sigma}_\omega$ at a coarser scale through Eq. (1) and Eq. (14):

$$\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\sigma} + \boldsymbol{\epsilon} = \boldsymbol{H}_\omega\boldsymbol{\sigma}_\omega + \boldsymbol{\epsilon}_\omega\,. \qquad (20)$$

Then consistency would impose that the errors are scale-dependent (hence the notation $\boldsymbol{\epsilon}_\omega$), because the numerical model is.

3.1. *Scale-free errors*

In Bocquet (2009), only scale-independent errors $\boldsymbol{\epsilon}$ were considered. It means that these errors are attached to the observations themselves (instrumental errors), or pertain to model errors that are scale-free. For the sake of consistency, the measurements themselves are to be scale-dependent in this case:

$$\boldsymbol{\mu}_\omega = \boldsymbol{H}_\omega\boldsymbol{\sigma}_\omega + \boldsymbol{\epsilon}\,. \qquad (21)$$

This is a natural standpoint in a synthetic data assimilation experiment performed at several scales. In this context, each Jacobian at different scales is assumed to be derived from a perfect model, so that discretization errors are discarded.

The synthetic measurements of such an experiment are

$$\boldsymbol{\mu}^{\mathrm{t}}_{\omega} = \boldsymbol{H}_{\omega}\boldsymbol{\sigma}^{\mathrm{t}}_{\omega}\,, \tag{22}$$

where $\boldsymbol{\sigma}^{\mathrm{t}}_{\omega} = \boldsymbol{\Gamma}_{\omega}\boldsymbol{\sigma}^{\mathrm{t}}$. These synthetic measurements are possibly made noisy. This is the point of view adopted by Bocquet (2005) and Saide *et al.* (2011). $\boldsymbol{H}_{\omega}$ could either be obtained by coarsening of $\mathbf{H}$ or by several models at several resolutions that are assumed perfect.

Since scale-dependent errors are discarded, this type of study is ideal to assess the signal in the observations without bothering about scale-dependent biases in the model, especially representativeness errors.

3.2. *Errors due to aggregation only*

Let us assume that errors are specified in the finest grid level, $\boldsymbol{\epsilon} = \boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}$, and that they may originate from many sources. Then, errors at larger scale $\boldsymbol{\epsilon}_{\omega} = \boldsymbol{\mu} - \boldsymbol{H}_{\omega}\boldsymbol{\sigma}_{\omega}$ are supposed to be solely due to this original error, plus errors entirely due to coarsening, or aggregation error that leads to representativeness error. In that case, the model scaling is entirely explained by the coarsening $\boldsymbol{H}_{\omega} = \mathbf{H}\boldsymbol{\Gamma}^{\star}_{\omega}$. Since $\boldsymbol{\mu} = \mathbf{H}\boldsymbol{\sigma} + \boldsymbol{\epsilon} = \mathbf{H}\boldsymbol{\sigma}_{\mathrm{b}} + \mathbf{H}\boldsymbol{\Pi}_{\omega}\left(\boldsymbol{\sigma} - \boldsymbol{\sigma}_{\mathrm{b}}\right) + \boldsymbol{\epsilon}_{\omega}$, the aggregation error, or *scale-covariant* error, can be identified:

$$\boldsymbol{\epsilon}_{\omega} = \boldsymbol{\epsilon} + \mathbf{H}\left(\mathbf{I}_{N_{\mathrm{fg}}} - \boldsymbol{\Pi}_{\omega}\right)\left(\boldsymbol{\sigma} - \boldsymbol{\sigma}_{\mathrm{b}}\right)\,. \tag{23}$$

Assuming independence of the error and source error priors, the computation of the covariance matrix of these errors yields

$$\mathbf{R}_{\omega} = \mathbf{R} + \mathbf{H}\left(\mathbf{I}_{N_{\mathrm{fg}}} - \boldsymbol{\Pi}_{\omega}\right)\mathbf{B}\mathbf{H}^{\mathrm{T}}\,. \tag{24}$$

The fact that $\boldsymbol{\Pi}_{\omega}$ is $\mathbf{B}^{-1}$-symmetric has been used in the derivation. Since $\mathbf{H}\left(\mathbf{I}_{N_{\mathrm{fg}}} - \boldsymbol{\Pi}_{\omega}\right)\mathbf{B}\mathbf{H}^{\mathrm{T}}$ is a positive matrix, the mean variance of the errors always increases because of the aggregation.

Intuitively, the statistics of the innovation vector $\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_{\mathrm{b}}$ should not depend on the scale. However, when written in terms of errors, the innovation depends formally on the representation ω:

$$\begin{aligned}\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_{\mathrm{b}} &= \boldsymbol{\mu} - \boldsymbol{H}_{\omega}\boldsymbol{\sigma}_{\omega} + \boldsymbol{H}_{\omega}\boldsymbol{\sigma}_{\omega} - \mathbf{H}\boldsymbol{\sigma}_{\mathrm{b}}\\ &= \boldsymbol{\epsilon}_{\omega} + \boldsymbol{H}_{\omega}\boldsymbol{\sigma}_{\omega} - \boldsymbol{H}_{\omega}\boldsymbol{\sigma}^{\mathrm{b}}_{\omega}\\ &= \boldsymbol{\epsilon}_{\omega} + \boldsymbol{H}_{\omega}\left(\boldsymbol{\sigma}_{\omega} - \boldsymbol{\sigma}^{\mathrm{b}}_{\omega}\right)\,.\end{aligned} \tag{25}$$

We have used the fact that:

$$\begin{aligned}\boldsymbol{\mu}_{\mathrm{b}} &= \boldsymbol{H}_{\omega}\boldsymbol{\sigma}^{\mathrm{b}}_{\omega} = \mathbf{H}\boldsymbol{\Gamma}^{*}_{\omega}\boldsymbol{\Gamma}_{\omega}\boldsymbol{\sigma}_{\mathrm{b}}\\ &= \mathbf{H}\boldsymbol{\sigma}_{\mathrm{b}} + \mathbf{H}\boldsymbol{\Pi}_{\omega}(\boldsymbol{\sigma}_{\mathrm{b}} - \boldsymbol{\sigma}_{\mathrm{b}}) = \mathbf{H}\boldsymbol{\sigma}_{\mathrm{b}}\,.\end{aligned} \tag{26}$$

This paradox is only superficial since one can check that the statistics of the innovation are truly scale-independent:

$$\begin{aligned}&\mathbf{R}_{\omega} + \mathbf{H}_{\omega}\mathbf{B}_{\omega}\mathbf{H}^{\mathrm{T}}_{\omega}\\ &\quad= \mathbf{R} + \mathbf{H}(\mathbf{I}_{N_{\mathrm{fg}}} - \boldsymbol{\Pi}_{\omega})\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{H}\boldsymbol{\Pi}_{\omega}\mathbf{B}\boldsymbol{\Pi}^{\mathrm{T}}_{\omega}\mathbf{H}^{\mathrm{T}}\\ &\quad= \mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\,.\end{aligned} \tag{27}$$

More generally, an analysis performed in the representation ω is obtained by coarsening the analysis at the finest scale. Hence, in this case, the multiscale formalism has no theoretical benefit compared to performing data assimilation in the finest grid (although there are major practical advantages). This can be understood by applying Bayes' rule directly, using Gaussian statistics,

$$\begin{aligned}q(\boldsymbol{\sigma}_{\omega}|\boldsymbol{\mu}) &= \frac{q(\boldsymbol{\mu}|\boldsymbol{\sigma}_{\omega})\,q(\boldsymbol{\sigma}_{\omega})}{q(\boldsymbol{\mu})}\\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\sigma}_{\omega} - \boldsymbol{\sigma}^{\mathrm{b}}_{\omega})^{\mathrm{T}}\mathbf{B}^{-1}_{\omega}(\boldsymbol{\sigma}_{\omega} - \boldsymbol{\sigma}^{\mathrm{b}}_{\omega})\right.\\ &\qquad\left.-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{H}_{\omega}\boldsymbol{\sigma}_{\omega})^{\mathrm{T}}\mathbf{R}^{-1}_{\omega}(\boldsymbol{\mu} - \boldsymbol{H}_{\omega}\boldsymbol{\sigma}_{\omega})\right\}\,.\end{aligned} \tag{28}$$

This leads to the estimate

$$\begin{aligned}\boldsymbol{\sigma}^{\mathrm{a}}_{\omega} &= \boldsymbol{\sigma}^{\mathrm{b}}_{\omega} + \mathbf{B}_{\omega}\mathbf{H}^{\mathrm{T}}_{\omega}\left(\mathbf{R}_{\omega} + \mathbf{H}_{\omega}\mathbf{B}_{\omega}\mathbf{H}^{\mathrm{T}}_{\omega}\right)^{-1}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_{\mathrm{b}})\\ &= \boldsymbol{\Gamma}_{\omega}\left\{\boldsymbol{\sigma}_{\mathrm{b}} + \mathbf{B}\mathbf{H}^{\mathrm{T}}\left(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right)^{-1}(\boldsymbol{\mu} - \mathbf{H}\boldsymbol{\sigma}_{\mathrm{b}})\right\}\\ &= \boldsymbol{\Gamma}_{\omega}\boldsymbol{\sigma}_{\mathrm{a}}\,,\end{aligned} \tag{29}$$

with $\boldsymbol{\sigma}_{\mathrm{a}}$ the emission estimation in the finest grid. The analysis-error covariance matrix transforms similarly according to

$$\begin{aligned}\mathbf{P}^{\mathrm{a}}_{\omega} &= \boldsymbol{\Gamma}_{\omega}\left\{\mathbf{B} - \mathbf{B}\mathbf{H}^{\mathrm{T}}\left(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{H}\mathbf{B}\right\}\boldsymbol{\Gamma}^{\mathrm{T}}_{\omega}\\ &= \boldsymbol{\Gamma}_{\omega}\mathbf{P}^{\mathrm{a}}\boldsymbol{\Gamma}^{\mathrm{T}}_{\omega}\,,\end{aligned} \tag{30}$$

where $\mathbf{P}^{\mathrm{a}}$ is the analysis-error covariance matrix in the finest grid. This can also be consistently obtained, through the finest scale:

$$\begin{aligned}q(\boldsymbol{\sigma}_{\omega}|\boldsymbol{\mu}) &= \int\mathrm{d}\boldsymbol{\sigma}\;q(\boldsymbol{\sigma}_{\omega}|\boldsymbol{\sigma},\boldsymbol{\mu})\,q(\boldsymbol{\sigma}|\boldsymbol{\mu})\\ &= \int\mathrm{d}\boldsymbol{\sigma}\;\delta\left(\boldsymbol{\sigma}_{\omega} - \boldsymbol{\Gamma}_{\omega}\boldsymbol{\sigma}\right)\,q(\boldsymbol{\sigma}|\boldsymbol{\mu})\,,\end{aligned} \tag{31}$$

which yields Eqs (29) and (30), by a simple convolution of Gaussian pdfs.

3.3. *Scale-dependent model errors*

As a first step, the errors were assumed to be scale-free $\boldsymbol{\epsilon}_{\omega} \equiv \boldsymbol{\epsilon}$, for instance coming from the observation: instrumental errors. Then, in addition, aggregation errors were taken into account by coarse-graining at fine resolution: $\boldsymbol{\epsilon}_{\omega} \equiv \boldsymbol{\epsilon} + \boldsymbol{\epsilon}^{\mathrm{c}}_{\omega}$, where $\boldsymbol{\epsilon}^{\mathrm{c}}_{\omega} = \mathbf{H}\left(\mathbf{I}_{N_{\mathrm{fg}}} - \boldsymbol{\Pi}_{\omega}\right)(\boldsymbol{\sigma} - \boldsymbol{\sigma}_{\mathrm{b}})$.

A third decomposition could involve
(i) the scale-independent observation error $\boldsymbol{\epsilon}^{\mathrm{o}}$ which would also include model error that could be scale-independent,
(ii) an error due to discretization $\boldsymbol{\epsilon}^{\mathrm{c}}_{\omega}$ (coarse-graining), and a model error that would be scale-dependent $\boldsymbol{\epsilon}^{\mathrm{m}}_{\omega}$:

$$\boldsymbol{\epsilon}_{\omega} = \boldsymbol{\epsilon}^{\mathrm{o}} + \boldsymbol{\epsilon}^{\mathrm{c}}_{\omega} + \boldsymbol{\epsilon}^{\mathrm{m}}_{\omega}\,. \tag{32}$$

On the one hand $\mathrm{Tr}\left(\mathrm{E}\left[\boldsymbol{\epsilon}^{\mathrm{c}}_{\omega}(\boldsymbol{\epsilon}^{\mathrm{c}}_{\omega})^{\mathrm{T}}\right]\right)$ would be decreasing as the resolution increases. On the other hand, $\mathrm{Tr}\left(\mathrm{E}\left[\boldsymbol{\epsilon}^{\mathrm{m}}_{\omega}(\boldsymbol{\epsilon}^{\mathrm{m}}_{\omega})^{\mathrm{T}}\right]\right)$ may have various behaviours depending on how the physics of the problem is parametrised and how the errors of the parametrisations depend on scale.

For instance, and for the latter source of errors, a large error increase is observed in atmospheric dispersion, when increasing the resolution of the atmospheric dispersion model beyond the reliable resolution of the meteorological fields used to drive the simulations.

In the rest of the article, we shall assume that the errors that are modelled account for scale-independent errors of all kinds, plus the scale-covariant aggregation errors. Additional scale-dependent model errors will not be considered.

4. Optimality criteria and optimization

4.1. *Three optimality criteria*

In addition to a multiscale formalism, the dependence of errors on the scale has been studied. Now, the optimal design of the representation of control space can be introduced. Three possible criteria of optimality are tested.

4.1.1. *The Fisher criterion*

Given our original incentive, which is to construct an adaptive grid of control space, optimal for data assimilation, the optimality criterion must be a measure of the quality of the analysis. In Bocquet (2009), the following criterion was chosen

$$\mathcal{J} = \mathrm{Tr}\left(\mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right) . \tag{33}$$

It is inspired by the Fisher information matrix, normalised by the background-error covariance matrix, so that the criterion is invariant by a change of coordinate in control space (for a given grid). Specifically, it measures the reduction of uncertainty granted by the observations.

In a representation ω, the criterion reads

$$\mathcal{J}_\omega = \mathrm{Tr}\left(\mathbf{B}_\omega\mathbf{H}_\omega^{\mathrm{T}}\mathbf{R}_\omega^{-1}\mathbf{H}_\omega\right) . \tag{34}$$

The operator $\mathbf{H}_\omega = \mathbf{H}\mathbf{\Lambda}_\omega^\star$ is the tangent linear operator of the affine operator $\boldsymbol{H}_\omega$ (which explains the difference of notation). Because only the linear part of $\boldsymbol{H}_\omega$ survives when averaging over the errors to obtain second-order moments, $\mathbf{H}_\omega$ appears in the criterion rather than $\boldsymbol{H}_\omega$.

If one assumes that the errors are essentially scale-independent, then $\mathbf{R}_\omega \simeq \mathbf{R}$. In that case, $\mathcal{J}_\omega$ can be written in terms of $\mathbf{\Pi}_\omega$ using the machinery developed earlier:

$$\begin{aligned}\mathcal{J}_\omega &= \mathrm{Tr}\left[\mathbf{\Gamma}_\omega\mathbf{B}\mathbf{\Gamma}_\omega^{\mathrm{T}}\left(\mathbf{\Lambda}_\omega^\star\right)^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{\Lambda}_\omega^\star\right]\\ &= \mathrm{Tr}\left[\mathbf{\Pi}_\omega\mathbf{B}\mathbf{\Pi}_\omega^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right] .\end{aligned} \tag{35}$$

Using the Bayesian prolongation operator $\mathbf{\Gamma}_\omega^\star$ that makes use of the prior, one obtains further

$$\mathcal{J}_\omega = \mathrm{Tr}\left(\mathbf{\Pi}_\omega\mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right) , \tag{36}$$

owning to the $\mathbf{B}^{-1}$-symmetry of $\mathbf{\Pi}_\omega$.

But, if the errors are scale-covariant following Eq. (23), the Fisher criterion Eq. (33) reads

$$\begin{aligned}\mathcal{J}_\omega &= \mathrm{Tr}\left[\mathbf{B}_\omega\mathbf{H}_\omega^{\mathrm{T}}\mathbf{R}_\omega^{-1}\mathbf{H}_\omega\right]\\ &= \mathrm{Tr}\left[\mathbf{\Pi}_\omega\mathbf{B}\mathbf{H}^{\mathrm{T}}\left\{\mathbf{R}+\mathbf{H}(\mathbf{I}_N-\mathbf{\Pi}_\omega)\mathbf{B}\mathbf{H}^{\mathrm{T}}\right\}^{-1}\mathbf{H}\right],\end{aligned} \tag{37}$$

which is more difficult to optimize because of the nonlinear dependence of $\mathcal{J}_\omega$ in $\mathbf{\Pi}_\omega$. The additional term is expected to increase the trust in the finest grid descriptions rather than the coarser ones.

4.1.2. *Degrees of freedom for the signal*

The dependence in $\mathbf{\Pi}_\omega$ is actually simpler if the criterion (to be maximized) is chosen to be

$$\begin{aligned}\mathcal{J}_\omega &= -\mathrm{Tr}\left[\mathbf{B}_\omega^{-1}\mathbf{P}_\omega^{\mathrm{a}} - \mathbf{I}_N\right]\\ &= \mathrm{Tr}\left[\mathbf{H}_\omega\mathbf{B}_\omega\mathbf{H}_\omega^{\mathrm{T}}\left(\mathbf{R}_\omega + \mathbf{H}_\omega\mathbf{B}_\omega\mathbf{H}_\omega^{\mathrm{T}}\right)^{-1}\right]\\ &= \mathrm{Tr}\left[\mathbf{\Pi}_\omega\mathbf{B}\mathbf{H}^{\mathrm{T}}\left(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{H}\right],\end{aligned} \tag{38}$$

using the innovation statistics scaling, Eq. (27).

This criterion $\mathcal{J}_\omega = \mathrm{Tr}\left(\mathbf{I}_N - \mathbf{B}_\omega^{-1}\mathbf{P}_\omega^{\mathrm{a}}\right)$ is known to measure the number of DFS, i.e. the information load that helps resolve the parameter space. It is actually more common in data assimilation literature than the cost function (Eq. (36)). In the absence of any source of errors, the DFS are equal to the number of scalar observations that are assimilated (p here). In the presence of errors, the DFS ranges between 0 and the number of observations p, because the information of the observations is also used to resolve the noise (Rodgers, 2000). So the maximization of $\mathcal{J}_\omega$ entails maximizing these degrees of freedom, which seems very natural. Note that criterion Eq. (36) is the limiting case of this DFS criterion when $\mathbf{R}$ is inflated or when $\mathbf{B}$ vanishes.

In this vein, given an admissible representation ω,

$$\epsilon_{\mathbf{l},k} = \frac{\mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}\mathbf{B}\mathbf{H}^{\mathrm{T}}\left(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right)^{-1}\mathbf{H}\mathbf{v}_{\mathbf{l},k}}{\mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}\mathbf{H}\mathbf{v}_{\mathbf{l},k}} \tag{39}$$

would represent the number of degrees of freedom per grid cell or tile. It is an objective measure of the data density (Rodgers, 2000) in parameter space.

4.1.3. *Data-dependent criterion*

One could consider the relative entropy, that is to say a gain in information, attached to the reconstructed parameters of control space (such as source variables). When the inference leading to the reconstructed source is Bayesian, and when the statistics are Gaussian, this information gain is (Kleeman, 2002)

$$\begin{aligned}\mathcal{K}_{\boldsymbol{\sigma}}^{\mathrm{Bayes}} =& \frac{1}{2}\ln\left|\mathbf{B}\left(\mathbf{B}^{-1}+\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right)\right|\\ &+\frac{1}{2}\mathrm{Tr}\left[\left(\mathbf{B}^{-1}+\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{B}^{-1}-\mathbf{I}_N\right]\\ &+\frac{1}{2}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}^{\mathrm{b}})^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}\\ &\qquad\times\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}^{\mathrm{b}}),\end{aligned} \tag{40}$$

whereas in a maximum entropy inference context, only the third term of Eq. (40) appears (Bocquet, 2008):

$$\begin{aligned}\mathcal{K}_{\boldsymbol{\sigma}} =& \frac{1}{2}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}^{\mathrm{b}})^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}\\ &\times\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}(\boldsymbol{\mu}-\mathbf{H}\boldsymbol{\sigma}^{\mathrm{b}}) .\end{aligned} \tag{41}$$

This term is a measure of the gain of information on the estimate of the source, whereas the additional terms in Eq. (40) focus on the gain in the knowledge of the uncertainty of this estimate. The former measures the information gain on the first-order moment, while the latter measures the information gain on the second-order moments. The average of the Bayesian result over all potential $\boldsymbol{\mu}$ is

$$\mathrm{E}_{\boldsymbol{\mu}}\left[\mathcal{K}_{\boldsymbol{\sigma}}^{\mathrm{Bayes}}\right] = \frac{1}{2}\ln\left|\mathbf{I}+\mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right| , \tag{42}$$

whereas in the maximum entropy case, it is

$$\mathrm{E}_{\boldsymbol{\mu}}\left[\mathcal{K}_{\boldsymbol{\sigma}}\right] = \frac{1}{2}\mathrm{Tr}\left[(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}}\right] , \tag{43}$$

which is half of the DFS.

Therefore, Eq. (41) could be used as a criterion for its simplicity and its physical interpretation. Applied to a representation ω, and defining $\mathcal{J}_\omega \equiv \mathcal{K}_{\boldsymbol{\sigma}}^{\omega}$, it reads

$$\begin{aligned}\mathcal{J}_\omega =& \mathrm{Tr}\left[\mathbf{B}_\omega\mathbf{H}_\omega^{\mathrm{T}}(\mathbf{R}_\omega+\mathbf{H}_\omega\mathbf{B}_\omega\mathbf{H}_\omega^{\mathrm{T}})^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_{\mathrm{b}})\right.\\ &\left.\times(\boldsymbol{\mu}-\boldsymbol{\mu}_{\mathrm{b}})^{\mathrm{T}}(\mathbf{R}_\omega+\mathbf{H}_\omega\mathbf{B}_\omega\mathbf{H}_\omega^{\mathrm{T}})^{-1}\mathbf{H}_\omega\right] ,\end{aligned} \tag{44}$$

where $\boldsymbol{\mu}_{\mathrm{b}} = \mathbf{H}\boldsymbol{\sigma}_{\mathrm{b}} = \boldsymbol{H}_\omega\boldsymbol{\sigma}_\omega^{\mathrm{b}}$ is scale-independent. The choice of the scale-covariant error Eq. (23) leads to

$$\begin{aligned}\mathcal{J}_\omega =& \mathrm{Tr}\left[\boldsymbol{\Pi}_\omega\mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_{\mathrm{b}})\right.\\ &\left.\times\ (\boldsymbol{\mu}-\boldsymbol{\mu}_{\mathrm{b}})^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}\mathbf{H}\right] ,\end{aligned} \tag{45}$$

where the cyclic property of the trace operator has been used. Contrary to the Fisher and DFS criteria, this criterion depends on the observation vector $\boldsymbol{\mu}$. By Eq. (43), when averaged over all possible sources and errors following the prior statistics, it yields half of the DFS criterion.

The total gain of information both on the source and on the errors in the maximum entropy inference is

$$\begin{aligned}\mathcal{K}_{\boldsymbol{\sigma},\boldsymbol{\epsilon}} =& \mathcal{K}_{\boldsymbol{\sigma}} + \mathcal{K}_{\boldsymbol{\epsilon}}\\ =& \frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_{\mathrm{b}})^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_{\mathrm{b}}) .\end{aligned} \tag{46}$$

Using a scale-covariant error Eq. (23) implies that $\mathcal{K}_{\boldsymbol{\sigma},\boldsymbol{\epsilon}}$ is scale-invariant. However, $\mathcal{K}_{\boldsymbol{\sigma}}$ is not. Therefore the information is distributed differently depending on the scale, or more generally the representation ω.

4.2. *Reduction of the criteria in the correlated case*

When $\mathbf{B}$ is not necessarily diagonal, a redefinition of the original restriction operator $\boldsymbol{\Gamma}_\omega$ into $\widetilde{\boldsymbol{\Gamma}}_\omega = \boldsymbol{\Gamma}_\omega\mathbf{B}^{-1/2}$ was advocated. Let us take the example of the Fisher criterion. With this redefinition, the optimality criterion becomes

$$\begin{aligned}\mathcal{J}_\omega &= \mathrm{Tr}\left(\widetilde{\boldsymbol{\Pi}}_\omega\mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right)\\ &= \mathrm{Tr}\left(\boldsymbol{\Pi}_\omega\mathbf{B}^{1/2}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{B}^{1/2}\right) ,\end{aligned} \tag{47}$$

where $\boldsymbol{\Pi}_\omega$ is now reduced to

$$\boldsymbol{\Pi}_\omega = \boldsymbol{\Gamma}_\omega^{\mathrm{T}}\left(\boldsymbol{\Gamma}_\omega\boldsymbol{\Gamma}_\omega^{\mathrm{T}}\right)^{-1}\boldsymbol{\Gamma}_\omega , \tag{48}$$

where $\boldsymbol{\Gamma}_\omega$ is the original coarse-graining restriction operator obtained by simple averaging. Similar results can be obtained for the other two criteria.

In the following, we shall assume that either $\mathbf{B}$ is proportional to the identity, or one applies the above redefinition to $\boldsymbol{\Gamma}_\omega$. Although this reduction of the correlated case is to be used in future work and needed to be addressed in this methodological article, it will not be directly used in the following test cases.

4.3. *Algebraic formalism*

Since the main goal is to optimize the representations of control space, we need to transform this abstract description of the multiscale structure and errors into numerical mathematics. For each tile at scales $\mathbf{l}$, a vector $\mathbf{v}_{\mathbf{l},k}$ in $\mathbb{R}^{N_{\mathrm{fg}}}$ is defined. Here $\mathbf{l} = (l_x, l_y, l_t)$ represents the scales of the tile, k is the tile index in the set of tiles of the same type (i.e. of the same scales $\mathbf{l}$). Recall that the finest regular grid is made of the tiles of scales $\mathbf{l} = (0, 0, 0)$. By construction, these tiles are in one-to-one correspondence with the canonical vectors $\{\mathbf{e}_{i,j,h}\}$ of $\mathbb{R}^{N_{\mathrm{fg}}}$. At a coarser scale, a tile at scales $\mathbf{l}$ can be partitioned into finer grid cells of the finest regular grid. Correspondingly, its vector $\mathbf{v}_{\mathbf{l},k}$ is defined as the sum of the canonical vectors $\{\mathbf{e}_{i,j,h}\}$ representing the finest grid cells that compose the tile:

$$\mathbf{v}_{\mathbf{l},k} = \sum_{\delta i=1}^{2^{l_x}}\sum_{\delta j=1}^{2^{l_y}}\sum_{\delta h=1}^{2^{l_t}}\mathbf{e}_{i_k+\delta i-1,\, j_k+\delta j-1,\, h_k+\delta h-1} , \tag{49}$$

where i_k, j_k and h_k are the smallest indices of the finest tiles composing tile $(\mathbf{l}, k)$.

From Eq. (7), and using the fact that $\mathbf{B}$ is proportional to the identity by definition or after the above redefinition Eq. (15), one obtains an explicit formula for $\boldsymbol{\Pi}_\omega$:

$$\boldsymbol{\Pi}_\omega = \sum_{\mathbf{l}}\sum_{k=1}^{n_{\mathbf{l}}}\alpha_{\mathbf{l},k}^{\omega}\frac{\mathbf{v}_{\mathbf{l},k}\mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}}{\mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}\mathbf{v}_{\mathbf{l},k}} , \tag{50}$$

where $n_{\mathbf{l}}$ is the number of tiles in the set of tiles with scale vector $\mathbf{l}$, which runs on all predefined scales $0 \le l_x < n_x$, $0 \le l_y < n_y$ and $0 \le l_t < n_t$. The coefficients $\alpha_{\mathbf{l},k}^{\omega}$ define representation ω: $\alpha_{\mathbf{l},k}^{\omega}$ is 1 when tile $(\mathbf{l}, k)$ belongs to the representation ω and is zero when it does not. Equation (50) can be checked by applying projector Eq. (7) on any vector $\mathbf{v}_{\mathbf{l},k}$. From now on, the superscript ω on $\alpha_{\mathbf{l},k}^{\omega}$ will be dropped to simplify the notation.

Since $\mathbf{B}$ is (truly or effectively) diagonal, then for any two vectors, $\mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}\mathbf{B}\mathbf{v}_{\mathbf{l}',k'}$ is non-zero, if and only if the two vectors correspond to overlapping tiles. If they belong to an admissible representation (the tiles form partition of Ω), then the matrix element is non-zero only if $(\mathbf{l}, k) = (\mathbf{l}', k')$.

Then, inserting Eq. (50) into Eq. (47), the cost function reads

$$\mathcal{J}_\omega = \sum_{\mathbf{l}}\sum_{k=1}^{n_{\mathbf{l}}}\alpha_{\mathbf{l},k}\epsilon_{\mathbf{l},k} , \tag{51}$$

where $\epsilon_{\mathbf{l},k} = \mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}\mathbf{W}\mathbf{v}_{\mathbf{l},k}/\mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}\mathbf{v}_{\mathbf{l},k}$. For instance, in the case of the Fisher criterion one has $\mathbf{W} = \mathbf{B}^{1/2}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{B}^{1/2}$. The local *energy* $\epsilon_{\mathbf{l},k}$ is a local measure of the contribution of the cell to the cost function.

In the following subsection we assume that $\mathcal{J}_\omega$ is of this form.

4.4. *Solving for optimal representations*

The goal is to optimize the functional Eq. (51) on all admissible representations. In order to lift the constraint of admissibility (the tiles cannot overlap), one introduces a Lagrangian. A fixed number of tiles is imposed thanks to a single multiplier ζ. The one point:one tile requirement is imposed thanks to a vector $\boldsymbol{\lambda}$ of N_{fg} multipliers. Each multiplier is associated with one grid cell of the finest regular grid. The Lagrangian reads

$$\begin{aligned}\mathcal{L}(\omega) = &\sum_{\mathbf{l}}\sum_{k=1}^{n_{\mathbf{l}}}\alpha_{\mathbf{l},k}\,\epsilon_{\mathbf{l},k} + \sum_{k=1}^{N_{\mathrm{fg}}}\lambda_k\left(\sum_{\mathbf{l}}\alpha_{\mathbf{l},\widetilde{k}} - 1\right)\\ &+ \zeta\left(\sum_{\mathbf{l}}\sum_{k=1}^{n_{\mathbf{l}}}\alpha_{\mathbf{l},k} - N\right).\end{aligned} \tag{52}$$

The sum on $k = 1, \ldots, N_{\mathrm{fg}}$ runs on all cells of the finest grid. In this sum, $\alpha_{\mathbf{l},\widetilde{k}}$ is the coefficient attached to the tile at scale $\mathbf{l}$ that covers cell k from the finest grid. This tile has index $\widetilde{k}$ among the $n_{\mathbf{l}}$ tiles related to scale $\mathbf{l}$. The Lagrangian can also be written as

$$\mathcal{L}(\omega) = \sum_{\mathbf{l}}\sum_{k=1}^{n_{\mathbf{l}}}\left(\epsilon_{\mathbf{l},k} + \mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}\boldsymbol{\lambda} + \zeta\right)\alpha_{\mathbf{l},k} - \sum_{k=1}^{N_{\mathrm{fg}}}\lambda_k - \zeta N\,. \tag{53}$$

Then the maximum can formally be taken on all representations, admissible or not, with any number of tiles in $[N_{\mathrm{cg}}, N_{\mathrm{fg}}]$, where N_{cg} is the number of grid cells in the coarsest regular grid. As a first step, the optimization is performed 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 one obtains an effective cost function of the Lagrange parameters:

$$\begin{aligned}\widehat{\mathcal{L}}(\boldsymbol{\lambda}, \zeta) &= \max_{\omega\in\mathcal{R}(\Omega)}\mathcal{L}(\omega) = \max_{\alpha_{\mathbf{l},k}}\mathcal{L}(\omega)\\ &= \sum_{\mathbf{l}}\sum_{k=1}^{n_{\mathbf{l}}}\max\left(0, \epsilon_{\mathbf{l},k} + \mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}\boldsymbol{\lambda} + \zeta\right)\\ &\quad - \sum_{k=1}^{N_{\mathrm{fg}}}\lambda_k - \zeta N\,.\end{aligned} \tag{54}$$

Because this cost function is dual to $\mathcal{L}(\omega)$, it needs to be minimized, not maximized (Borwein and Lewis, 2000). Note that the cost function is not smooth since it is non-differentiable on the edges of a polytope. Hence, optimization on the Lagrange parameters cannot make direct use of gradient-based minimization techniques. Besides, this functional may not be convex, nor is it guaranteed that it has a single minimum. To overcome these potential problems, a regularisation of this effective cost function is needed.

A statistical mechanics analogy was used earlier by Bocquet (2009) to solve this problem. We develop here an equivalent analytical approach through information theory. We look for the least committed representation, described by a pdf $q(\boldsymbol{\alpha})$ in the vector $\boldsymbol{\alpha}$, given that all constraints are satisfied *on average*. At finite temperature β^{-1}, the optimal pdf is the one that maximizes the criterion with a weight β, plus the relative entropy of the representation pdf relative to the (non-admissible) geometry where all tiles are equiprobable.

$$\begin{aligned}\mathcal{J}(q) = &-\sum_{\boldsymbol{\alpha}}q(\boldsymbol{\alpha})\ln q(\boldsymbol{\alpha})\\ &+ \beta\sum_{\mathbf{l}}\sum_{k=1}^{n_{\mathbf{l}}}\sum_{\boldsymbol{\alpha}}q(\boldsymbol{\alpha})\alpha_{\mathbf{l},k}\,\epsilon_{\mathbf{l},k}\\ &+ \sum_{k=1}^{N_{\mathrm{fg}}}\lambda_k\sum_{\alpha}q(\boldsymbol{\alpha})\left(\sum_{\mathbf{l}}\alpha_{\mathbf{l},\widetilde{k}} - 1\right)\\ &+ \zeta\left(\sum_{\mathbf{l}}\sum_{k=1}^{n_{\mathbf{l}}}\sum_{\boldsymbol{\alpha}}q(\boldsymbol{\alpha})\,\alpha_{\mathbf{l},k} - N\right)\\ = &-\sum_{\boldsymbol{\alpha}}q(\boldsymbol{\alpha})\ln q(\boldsymbol{\alpha})\\ &+ \sum_{\mathbf{l}}\sum_{k=1}^{n_{\mathbf{l}}}q(\boldsymbol{\alpha})\left(\beta\,\epsilon_{\mathbf{l},k} + \mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}\boldsymbol{\lambda} + \zeta\right)\alpha_{\mathbf{l},k}\\ &- \sum_{k=1}^{N_{\mathrm{fg}}}\lambda_k - \zeta N\,.\end{aligned} \tag{55}$$

A first optimization on q leads to

$$q(\boldsymbol{\alpha}) \propto \exp\left\{\sum_{\mathbf{l},k}\alpha_{\mathbf{l},k}\left(\beta\,\epsilon_{\mathbf{l},k} + \mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}\boldsymbol{\lambda} + \zeta\right)\right\}. \tag{56}$$

$\sum_{\mathbf{l},k}$ is shorthand for $\sum_{\mathbf{l}}\sum_{k=1}^{n_{\mathbf{l}}}$. The substitution of q given by Eq. (56) into Eq. (55) leads to a dual Lagrangian

$$\widehat{\mathcal{J}}_\beta(\boldsymbol{\lambda}, \zeta) = \ln Z_\beta(\boldsymbol{\lambda}, \zeta) - \sum_{k=1}^{N_{\mathrm{fg}}}\lambda_k - \zeta N\,, \tag{57}$$

where the partition function Z_β, is given (after factorisation) by

$$Z_\beta(\boldsymbol{\lambda}, \zeta) = \prod_{\mathbf{l},k}\left\{1 + \exp\left(\beta\epsilon_{\mathbf{l},k} + \mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}\boldsymbol{\lambda} + \zeta\right)\right\}. \tag{58}$$

This leads to the dual Lagrangian, function of the Lagrange parameters

$$\begin{aligned}\widehat{\mathcal{J}}_\beta(\boldsymbol{\lambda}, \zeta) = &\sum_{\mathbf{l},k}\ln\left\{1 + \exp\left(\beta\,\epsilon_{\mathbf{l},k} + \mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}\boldsymbol{\lambda} + \zeta\right)\right\}\\ &- \sum_{k=1}^{N_{\mathrm{fg}}}\lambda_k - \zeta N\,.\end{aligned} \tag{59}$$

This cost function is the one that was obtained in Bocquet (2009) using the statistical mechanics analogy. From the minimization of this free energy, yielding $\boldsymbol{\lambda}^*$ and ζ^*, one obtains the filling factor

$$\alpha_{\mathbf{l},k}^* = \frac{1}{\beta}\frac{\partial\ln Z_\beta}{\partial\epsilon_{\mathbf{l},k}} = \frac{1}{1+\exp\left(-\beta\,\epsilon_{\mathbf{l},k} - \mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}\boldsymbol{\lambda}^* - \zeta^*\right)}. \tag{60}$$

When β goes to infinity, the filling factors $\alpha_{\mathbf{l},k}^*$ converge to either 0 or 1. An alternate statistical regularisation is proposed in the Appendix.

5. Illustrations

The formalism described in the previous sections will be illustrated on two examples related to the transport and fate of atmospheric constituents.

5.1. *Simplified CarboEurope-IP network*

5.1.1. *Set-up*

The CarboEurope-IP network routinely measures CO_2 concentrations over Europe at a precision of 0.1 ppm, and is part of the global monitoring network of greenhouses gases. The observations from this network of 22 stations can be used to perform inverse modelling of CO_2 sources and sinks (http://www.carboeurope.org). Here we will use a much simpler prototype to apply the above formalism to this issue. Firstly, we shall use only one annual-mean observation for each station (for a total of 22). Secondly, we will use a drastically simpler model to construct the Jacobian $\mathbf{H}$, made of the influence functions for each of those observations. Each influence function c^* attached to an observation i is assumed to be an average power law

$$c_i^*(r) \propto \frac{1}{r^\alpha}, \qquad (61)$$

where r is the great-circle distance separating the observation location and the point where this sensitivity is being computed. The exponent $\alpha \simeq 2.4$ is chosen heuristically following Roustan and Bocquet (2006). As an average midlatitude footprint, it bears some realism. The Jacobian entries are given by $[\mathbf{H}]_{ik} = [\mathbf{c}_i^*]_k$, where $\mathbf{c}_i^*$ is the discretised influence function. One is then looking for an optimal stationary adaptive representation.

A multiscale structure of six levels for each direction is defined. The domain Ω of control space is 22°W–42°E, 34–66°N. Its finest regular grid has dimensions $N_x = 128$ and $N_y = 64$, with grid-cell sizes $\Delta_x = \Delta_y = 0.50°$. The total number of cells in this grid is therefore $N_{\rm fg} = 8192$.

With such simple assumptions, this example only represents a prototype of the kind of results that could be achieved with a more realistic physical model and observation set. It allows us to test the ideas presented in this article, as well as sketch a future computationally demanding full-scale application.

5.1.2. *DFS criterion for the simplified CarboEurope-IP network case*

It is first assumed that the model and observations are perfect (the error covariance matrix $\mathbf{R} \simeq \mathbf{0}$ is negligible). The background-error covariance matrix is taken diagonal. For simplicity it is assumed that $\boldsymbol{\sigma}_{\rm b} = \mathbf{0}$. Figure 3(a) shows the optimal adaptive grid with $N = 512$, which represents 6% of the number of cells in the finest grid. The DFS obtained is 21.514, as compared to $p = 22$ observations. This means that this representation is able to capture all the degrees of freedom that could have been obtained in the finest grid. The densification of the grid close to Scandinavia is due to two outlying stations: Pallas and Zeppelin, while the densification in the Atlantic is due to the outlying station of Ivittuut, Greenland. These three stations are used in the optimization, but do not lie in the part of the domain that is shown here.

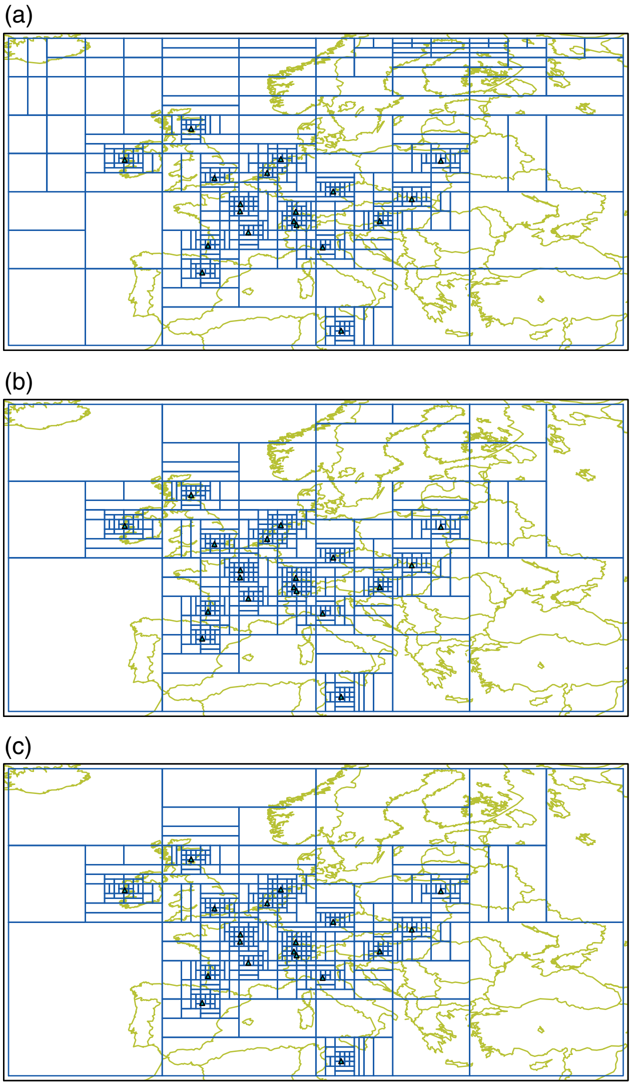

Figure 3. Optimal adaptive grids over the tiling dictionary for the CarboEurope-IP prototype with $N = 512$. (a) corresponds to an optimization using the DFS criterion with negligible errors $\mathbf{R} \simeq \mathbf{0}$. (b) corresponds to an optimization using the DFS criterion with errors. (c) corresponds to an optimization using criterion Eq. (36).

Then we assume a diagonal non-null error covariance matrix $\mathbf{R}$, such that the theoretical maximum DFS is 5.89, much lower than the available $p = 22$ observations, which is quite realistic. The optimal grid that is obtained with $N = 512$ reaches the DFS 5.88. The result is displayed in Figure 3(b). The main difference is that the grid is even more peaked around the stations. Indeed, in Eq. (38), $\mathbf{R}$ acts as a threshold below which the propagation of information from control space to the observations, represented by $\mathbf{HBH}^{\rm T}$, becomes less relevant (the denominator is dominated by $\mathbf{R}$ rather than by $\mathbf{HBH}^{\rm T}$). As the errors represented by $\mathbf{R}$ increase, the information is propagated at shorter distances. The criterion Eq. (36) is the limiting case of Eq. (38) when observation/model errors dominate the background errors: $\mathbf{R}$ is far superior to $\mathbf{HBH}^{\rm T}$ for any reasonable norm. It consistently leads to the grid design displayed in Figure 3(c), which is even slightly more peaked than the grid of Figure 3(b).

5.2. *ETEX-I dispersion experiment*

5.2.1. *Set-up*

The second example is the European Tracer Experiment (ETEX), and in particular its first campaign, ETEX-I. Organised by the Joint Research Centre at Ispra, Italy, it dates back to 12 October 1994, 1600 UTC, when 340 kg of perfluoromethylcyclohexane were released uniformly over 12 h, at Monterfil, in Brittany, France. 168 stations of the World Meteorological Organisation (WMO) monitored the subsequent plume throughout Europe. The weather conditions (low pressure over Scotland) were selected so that the plume would be advected eastward toward the stations.

The measurements were intensively used to benchmark chemistry and transport models (Nodop *et al.*, 1998), but also more recently for the tests of inverse modelling methodologies (Krysta *et al.*, 2008). In particular, it was shown that, with a considerable reduction of the grid-cell numbers, the optimal tiling leads to inversions very similar to the one obtained with a fine regular grid.

A multiscale structure of five levels for each direction is defined. The finest regular grid is 20.8125°W–15.1875°E, 36.5625–54.5625°N, with $N_x = 64$, $N_y = 32$, and $N_t = 160$. The number of cells of size $\Delta_x = \Delta_y = 0.5625°$ and $\Delta_t = 1$ h in the finest grid is $N = 327680$.

Contrary to the example of CarboEurope-IP, $\mathbf{H}$ is obtained from a realistic Eulerian chemistry and transport model (Bocquet, 2007, gives modelling details). Also the adaptive grid will be dynamic: it will be optimized on the ground and in time.

5.2.2. *Comparing designs with ETEX-I*

The differences between the data-dependent criterion and the DFS, data-free, criterion are illustrated on the ETEX-I case. For both criteria, a scale-covariant error is assumed. The data-free criterion is therefore Eq. (38) while the data-dependent criterion is Eq. (45). A limited dataset of 201 real observations of tracer concentration is used. The same set was employed by Bocquet (2009), but with criterion Eq. (36).

We seek optimal grids of the same size $N = 402$ as in Bocquet (2009). The optimal 2D+T grid obtained from the data-free criterion is displayed in Figure 4, while the optimal 2D+T grid obtained from the data-dependent criterion is displayed in Figure 5. $N = 402$ offers a tight compromise for the data-free criterion optimization in terms of high DFS and significant reduction of the tile numbers (0.1% of the total number of grid cells in the finest grid: $N = 327\,680$). The resulting DFS is 75.70, while the maximum achievable DFS in the finest grid is 157.5. Since error statistics have been taken into account, it is lower than the perfect model case of 201 DFS. It is also the maximum of criterion Eq. (45), that is the maximum of the achievable information gain via a maximum entropy on the mean inference.

The main difference is seen over Ireland. Indeed the value of the concentrations at the stations in Brittany are high and do not rule out a source upwind near Ireland or in the Atlantic, so that the grid is refined there. On the contrary, the data-free criterion accounts for any set of values compatible with the prior. The true observation set used in the data-dependent criterion is only one specific set, so that a refinement near the monitoring network is preferred at the expense of a refinement over Ireland and the Atlantic.

5.2.3. *Inverse modelling with ETEX-I*

Inverse modelling is performed using several adaptive grids and the results are reported in Table I. The details of the set-up of the inverse modelling are the same as those reported by Bocquet (2009), and they are not repeated here.

Two types of inversion are considered: Gaussian and non-Gaussian. The Gaussian type is based on Gaussian background errors, such as those assumed in this article, while the non-Gaussian type is based on non-Gaussian background errors (following Bocquet *et al.*, 2010, and references therein) that ensures positiveness of the source. The total retrieved mass and the mass retrieved near the location of the release site are reported in the table. Scalar m_0 is a mass scale that parametrises the background-error term, whereas χ is the prior observation-error standard deviation. The results obtained for the DFS criterion are similar to those obtained with the Fisher criterion, and the remarks of Bocquet (2009) are still valid. However, for the grid obtained from the data-dependent criterion, the inversions lead to a very good localization of the source (not shown here but it can be inferred from the figures of the table) and an underestimation of the retrieved mass. The better localization is due to a stronger refinement of the grid close to the release site. On the downside, it probably strengthens the importance of the measurements performed on nearby sites, known to be largely overestimated by Eulerian dispersion models on ETEX-I, leading to an underestimation by the data assimilation scheme because of this model error. The inversions with adaptive grids are also compared to those performed with the regular grid of resolution 2.25° × 2.25° × 1h in Table I.

5.3. *An optimal number N of tiles?*

When one adds one more tile to an optimal adaptive grid, there is a marginal gain in the objective function which is defined mathematically by $\partial_N \mathcal{J}_{\omega_N^*}$. It can numerically be accessed through the parameter $-\zeta^*$, conjugate to the number of tiles, because

$$\begin{aligned}\frac{\mathrm{d}\mathcal{J}_{\omega_N^*}}{\mathrm{d}N} &= \frac{\mathrm{d}\widehat{\mathcal{L}}_N}{\mathrm{d}N}(\boldsymbol{\lambda}^*, \zeta^*) \\ &= \frac{\partial \widehat{\mathcal{L}}_N}{\partial N} + \left(\frac{\partial \boldsymbol{\lambda}^*}{\partial N}\right)^{\mathrm{T}} \nabla_{\boldsymbol{\lambda}} \widehat{\mathcal{L}}_N + \left(\frac{\partial \zeta^*}{\partial N}\right) \partial_\zeta \widehat{\mathcal{L}}_N \\ &= -\,\zeta^*,\end{aligned} \tag{62}$$

which can be checked by differentiating Eq. (54) with respect to N. If there is an optimal number $N^\star$ of tiles which is non-trivial (i.e. strictly $N_{\mathrm{cg}} < N^\star < N_{\mathrm{fg}}$), then $\zeta^*_{N^\star}$ is zero. To obtain such a grid, with an optimal $N^\star$, it is sufficient to get rid of the tile number constraint in the optimization of Eq. (54). Unfortunately, the existence of such a non-trivial $N^\star$ is not a simple issue.

Consider the generic objective function $\mathcal{J}_\omega = \mathrm{Tr}\,(\boldsymbol{\Pi}_\omega \boldsymbol{\Omega})$, where $\boldsymbol{\Pi}_\omega$ is the projector onto representation ω and $\boldsymbol{\Omega}$ is a positive definite matrix. If ω_N^* is the optimal representation for this criterion with N tiles, then $\mathcal{J}_{\omega_N^*}$ is an increasing function of N. Suppose ω_N^* has been determined for

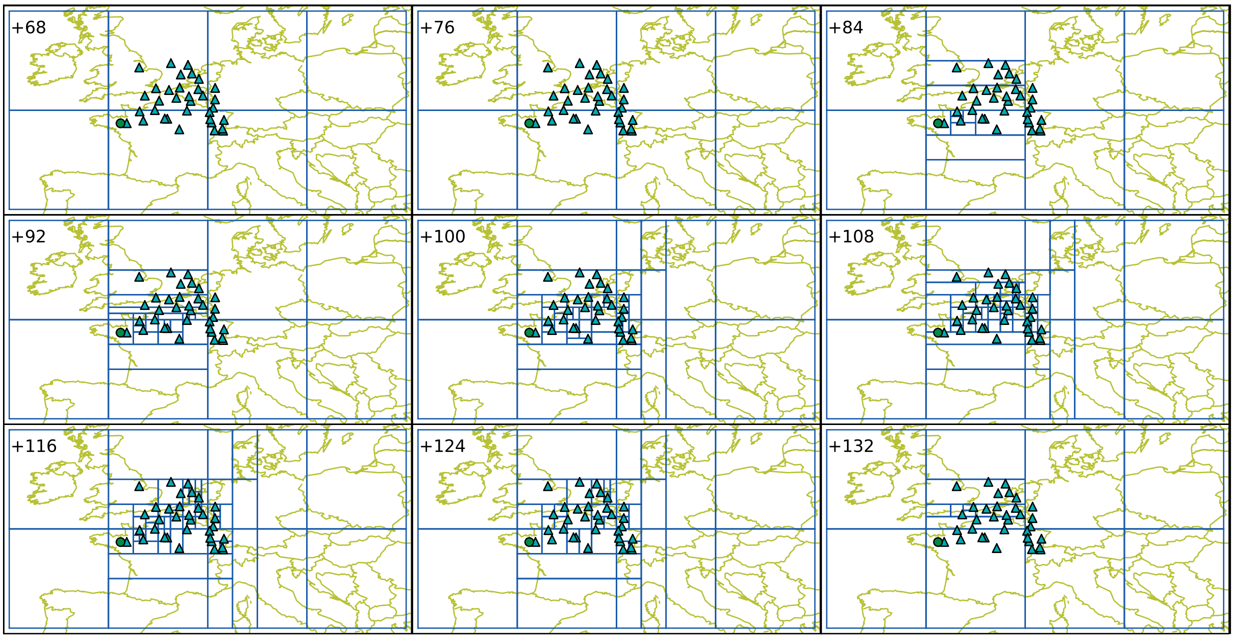

Figure 4. Snapshots of the 2D+T optimal adaptive grid with $N = 402$ tiles for a selection of 201 concentration observations of the ETEX-I dispersion event. The criterion is given by the data-independent cost function Eq. (38). Time is indicated in the top left-hand corner of each panel. The triangles indicate the WMO stations that reported at least one of these 201 observations. The disk indicates the true source location of ETEX-I.

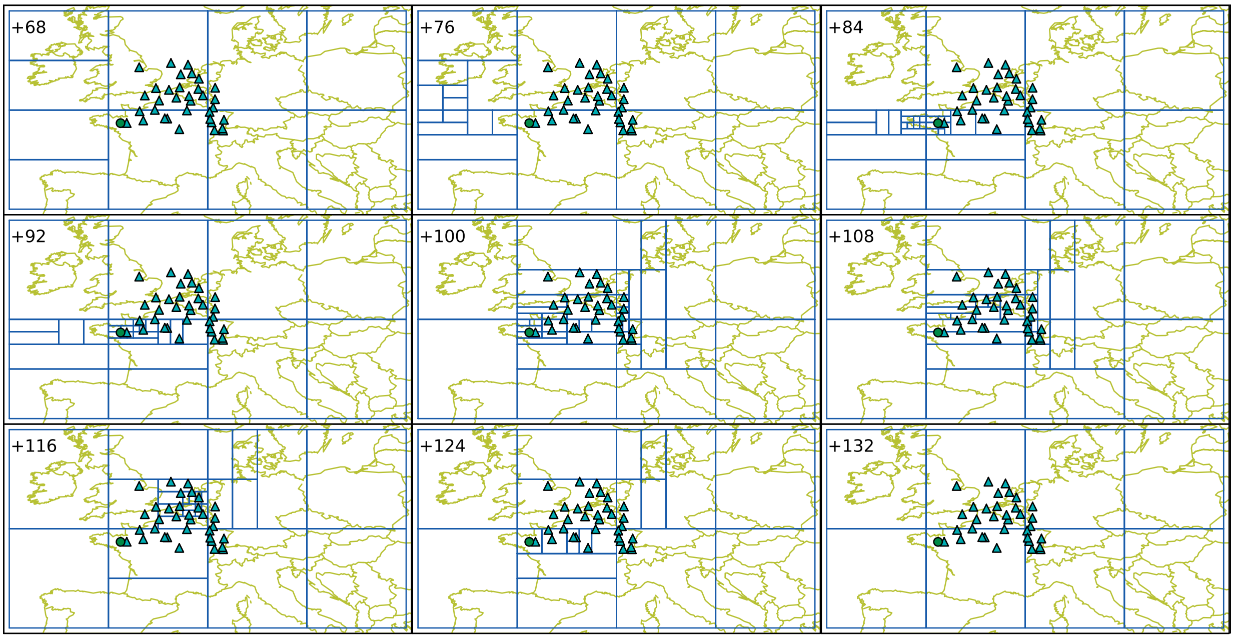

Figure 5. As Figure 4, but the optimality criterion is now given by the data-dependent cost function Eq. (45).

Table I. Results of source inverse modelling experiments on ETEX-I, using several types of regular or adaptive grids built from the criteria introduced in this article. The total mass of tracer released during ETEX-I was 340 kg at a point location.

Grid type	Criterion type	N	Inversion type	m_0 (kg)	χ (ng m^{-3})	Local mass (kg)	Total mass (kg)
Regular		20480	Gaussian	0.025	0.25	234	680
Regular		20480	Non-Gaussian	5	0.25	220	327
Tiling	Fisher	402	Gaussian	0.025	0.25	270	1005
Tiling	Fisher	402	Non-Gaussian	5	0.25	205	238
Tiling	DFS	402	Gaussian	0.025	0.25	268	1136
Tiling	DFS	402	Non-Gaussian	5	0.25	200	252
Tiling	Data-dependent	402	Gaussian	0.025	0.25	173	599
Tiling	Data-dependent	402	Non-Gaussian	5	0.25	134	195

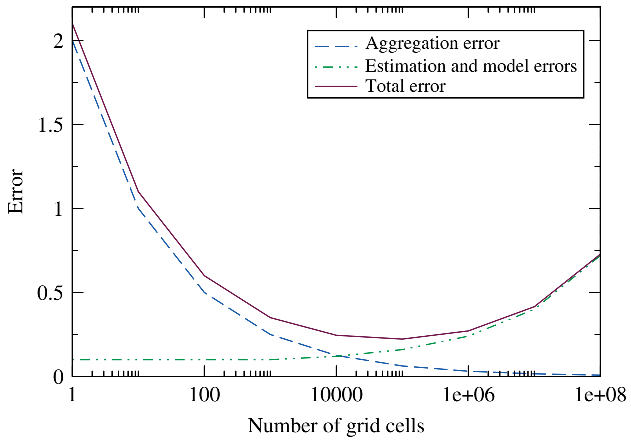

Figure 6. Schematic of the posterior error (arbitrary units), or of a reasonable criterion resulting from the aggregation, model and estimation errors, as a function of the resolution.

$N \le N_{\rm fg} - 1$ and let us look for a better representation ω_{N+1} with $N+1$ tiles. Take any tile of ω_N^* and split it into two sub-tiles. This leads to a representation ω_{N+1} that does not have to be optimal. If the eigensystem of $\boldsymbol{\Omega}$ is $\{\mathbf{v}_i, \zeta_i\}_{i=1,\cdots,N_{\rm fg}}$, then

$$\mathcal{J}_\omega = \mathrm{Tr}\left(\boldsymbol{\Pi}_\omega \boldsymbol{\Omega}\right) = \sum_{i=1}^{N_{\rm fg}} \zeta_i \sum_{\mathbf{l},k} \alpha_{\mathbf{l},k} \frac{\left(\mathbf{v}_{\mathbf{l},k}^{\rm T}\mathbf{v}_i\right)^2}{\mathbf{v}_{\mathbf{l},k}^{\rm T}\mathbf{v}_{\mathbf{l},k}}\,. \tag{63}$$

It is not difficult to show that the sum of the quantity $(\mathbf{v}_{\mathbf{l},k}^{\rm T}\mathbf{v}_i)^2/\mathbf{v}_{\mathbf{l},k}^{\rm T}\mathbf{v}_{\mathbf{l},k}$ of two sub-tiles is greater or equal to the same quantity for the mother tile. Since the ζ_i are all positive, one concludes that $\mathcal{J}_{\omega_N^*} \le \mathcal{J}_{\omega_{N+1}}$, so that $\mathcal{J}_{\omega_N^*} \le \mathcal{J}_{\omega_{N+1}^*}$.

As a consequence, criteria Eqs (36) and (38) are monotonic functions of the optimal representations ω_N^*, when N increases. The maximum of the objective function is reached in the finest regular grid. This was numerically checked by Bocquet (2009) in the case of the first objective function. It will be checked in section 6 for the second objective function based on the DFS.

This also applies to the data-dependent objective function Eq. (45), because in this case $\boldsymbol{\Omega}$ is of the form $\boldsymbol{\Omega} = \mathbf{u}\mathbf{u}^{\rm T}$, where $\mathbf{u}$ is

$$\mathbf{u} = \mathbf{B}^{1/2}\mathbf{H}^{\rm T}\mathbf{w} \quad \text{with} \quad \left(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\rm T}\right)\mathbf{w} = \boldsymbol{\mu} - \boldsymbol{\mu}_{\rm b}\,. \tag{64}$$

However, such monotonic behaviour may not be satisfied for an arbitrary objective function. The DFS and data-dependent objective functions used in this article account for aggregation errors that decrease with N, and for estimation errors that increase with N (for a given dataset). The net result is an error reduction with increasing N. In an even more realistic context, one should also take into account scale-dependent model errors $\boldsymbol{\epsilon}_\omega^{\rm m}$, that are not of aggregation type, as discussed in section 3. Then there may be an optimal N^*, as illustrated in Figure 6. This paradigm has been established in the greenhouse gas inversion community (e.g. Peylin *et al.*, 2001).

Such a non-trivial optimum should also exist when the errors are scale-free (as discussed in section 3). For instance, in the case of the data-dependent cost function, yet without taking into account aggregation (scale-covariant) errors, it was shown by Bocquet (2005) that the objective function vanishes when N goes to infinity. For a finite resolution limit (large but finite $N_{\rm fg}$), the objective function is expected to ultimately decrease to a finite limiting value imposed by the finest accessible resolution. Taking into account aggregation errors counteracts this increase in information gain, because fields on coarser grids are not as trusted as fields defined in the finest grid. Again, this trust in the finest grid is likely to be mitigated by taking into account realistic scale-dependent model errors, yielding a non-trivial $N^\star$.

6. General tilings versus qtrees and ftrees

In the previous sections, a multiscale framework has been defined and a data assimilation system was made consistent with it, including scale-covariant aggregation errors. This allowed optimal representations of control space for the assimilation of observations to be built. Up to this point, the adaptive grids were optimized on a dictionary of general tilings. For a 2D+T parameter field, and when employing a dyadic multiscale structure, storing the multiscale Jacobian in memory requires up to eight times the size of the Jacobian of the finest grid. It is thus of practical concern to use a smaller, but still efficient enough, dictionary of representations.

6.1. *Qtrees*

If one adopts a quaternary tree structure (qtree) for the spatial part instead of the tensor product of two dyadic structures while keeping a binary tree multiscale structure for time, then storing the multiscale Jacobian in memory requires at most 8/3 times the size of the Jacobian of the finest grid. Note that the set of qtrees built on the same domain is a subset of the tilings.

In order to compare the results, we use again the ETEX-I example to visually illustrate the qtree representations. Figure 7 displays an optimal representation with the same assumption and for the same criterion as for the example of Figure 5. The corresponding tiling and qtree representations are consistently refined at the same space and time spots.

Figure 8 displays the DFS of optimal tilings, optimal qtrees, and regular grids, for a wide range of N. The optimal tilings and optimal qtrees are far superior to regular grids: much more information is captured with the same number of cells in an optimal adaptive grid. Besides, for a fixed N, the optimal qtree captures fewer DFS than the corresponding optimal tiling. This must be so since qtrees form a subset of tilings. Nevertheless, the drop in performance is very moderate. Moreover the optimization times for these computations were roughly two times shorter for the qtrees than for the tilings. Their respective numerical efficiencies will discussed in Part II of this article. Therefore, we believe that optimizing on qtrees is a good substitute for an optimization on tilings.

6.2. *Ftrees*

A *factorised tree*, or *ftree* is defined as the *direct product* of binary trees. In the 2D case, the ftree is the direct product of two binary trees, one for each of the two directions. An example of such an adaptive grid is displayed in Figure 1(b). It is similar to the grid used by global numerical weather prediction models or chemical transport models that require zooming onto some region, such as

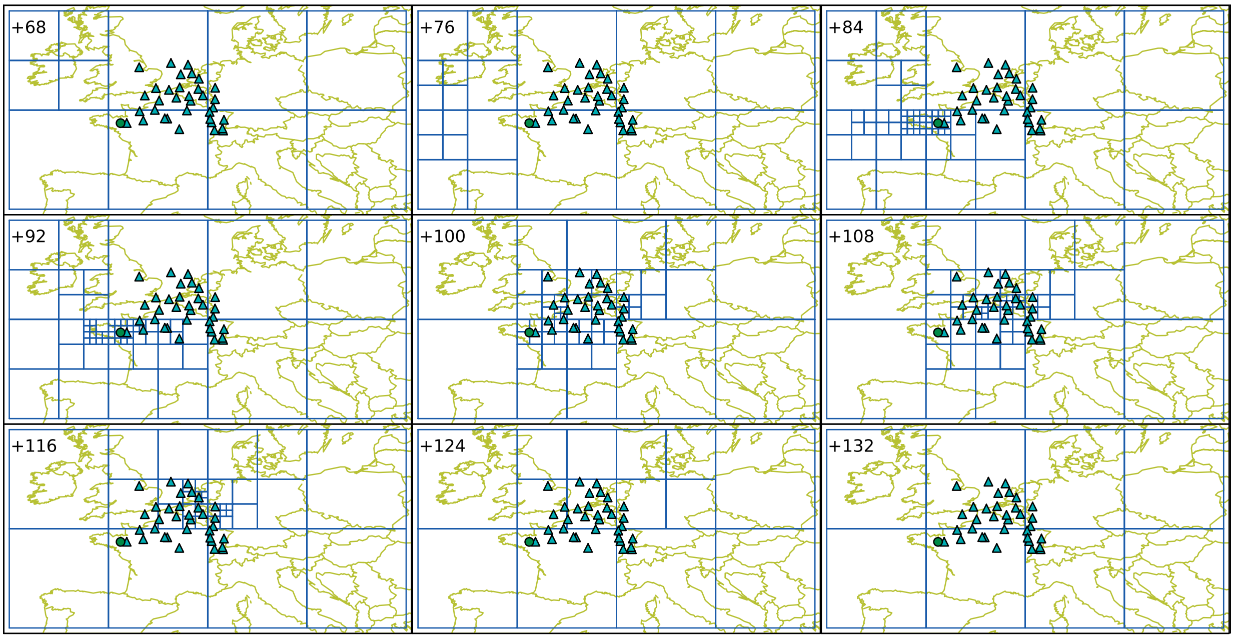

Figure 7. As Figure 4, but the optimal representation is searched in the qtree set.

Arpège (Action de Recherche Petite Echelle Grande Echelle) by Météo-France or LMDZ (Laboratoire de Météorologie Dynamique 'Zoom').

Contrary to the qtree dictionary, the generation of all ftrees requires computation of the value of the Jacobian for any tile, so the same amount of memory would be required as that of the dictionary of general tilings.

This dictionary of ftrees has considerably fewer degrees of freedom than both the dictionary of tilings and the dictionary of qtrees. Moreover, the optimization algorithm over the set of ftrees requires an adaptation of the general algorithm used for the tilings and for the qtrees. Two vectors of filling factors, say $\boldsymbol{\alpha}^x$ and $\boldsymbol{\alpha}^y$, one for each direction, are required. The global filling factor, at scales (l_x, l_y) and position (k_x, k_y) could thus be a product of the two directional ones (other choices are possible):

$$\alpha_{(l_x,l_y),(k_x,k_y)} = \alpha^x_{l_x,k_x}\alpha^y_{l_y,k_y} . \qquad (65)$$

The application of our optimization algorithm leads to the computation of the partition function Eq. (58). However, its computation is less simple for the ftrees because, on the one hand, $\boldsymbol{\alpha}$ is factorised into two contributions (one for each direction), and on the other hand, the energies $\epsilon_{(l_x,l_y),(k_x,k_y)}$ cannot be factorised. So there is no trivial factorisation of the partition function according to the two directions.

We have opted for solving this optimization problem iteratively. At first, one of the directions (say Ox) is frozen and the vector $\boldsymbol{\alpha}_x$ is fixed. Then, one solves for $\boldsymbol{\alpha}_y$, using our algorithm applied to a 1D problem. Then, in turn, direction Oy is frozen, and the newly obtained $\boldsymbol{\alpha}_y$ is fixed. Then one solves for a new estimation of $\boldsymbol{\alpha}_x$, and so on, until convergence. We have contemplated a variant of the algorithm, where one imposes a fixed number of tiles for each direction, $\mathcal{N}_x$ and $\mathcal{N}_y$, the global number of tiles N being equal to $N = \mathcal{N}_x\mathcal{N}_y$.

This geometry is tested on the ETEX-I example and its performance is compared to the skills of the tilings and qtrees in Figure 8. So, one obtains results which are significantly inferior to the performance of the qtrees, with a substantial complication in the optimization. That is why we recommend qtrees over ftrees in this context.

7. Summary, discussion and future work

7.1. Summary

In this article, we have developed a consistent Bayesian framework for the optimal design of control space in geophysical data assimilation. Prior information on the parameters of control space, including correlation of errors, is now accounted for and embedded in a multiscale framework. Prior information is also consistently used in the prolongation operator, so that every bit of available information is used when moving up and down the scale ladder. Note that, since the control space parameters can depend on both space and time, this framework accounts for space and time together.

Observation errors originating from aggregation were also explicitly considered in this framework. These scale-covariant errors consistently yield scale-invariant innovation statistics. The impact of observation errors on the optimal design of the representation was illustrated in a CO_2 flux inversion context using a simplified CarboEurope-IP monitoring network. More general scale-dependent errors, such as complex model errors, could not be studied here since they are case-specific.

New objective functions to rank adaptive grids of a dictionary of representation of control space have been defined. The first one is a normalised measure of the uncertainty $\mathrm{Tr}\left(\mathbf{B}^{-1}\mathbf{P}^{\mathrm{a}}\right)$, which is similar to the criterion at the heart of the Best Linear Unbiased Estimator (BLUE) approach used in most current data assimilation schemes. It is equal to the degrees of freedom for the signal (DFS). This DFS measure, together with scale-covariant errors, leads to an elegant criterion which is easier to optimize.

However, this DFS criterion is an implicit statistical average over all potential observation sets prescribed by the prior. That is why an observation-dependent criterion

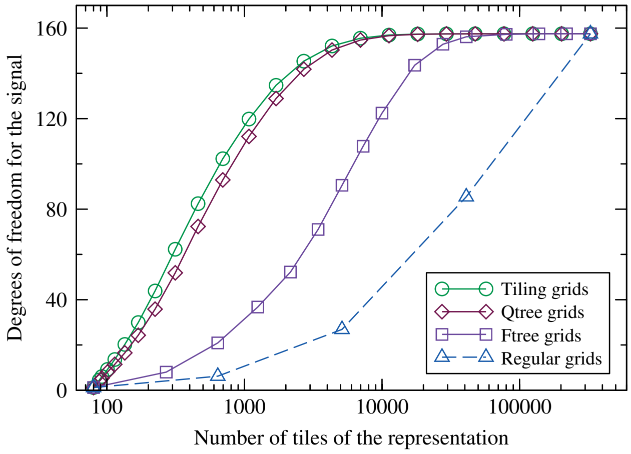

Figure 8. Degrees of freedom for the signal of optimal tilings, optimal qtrees, optimal ftrees and regular grids *versus* the number of grid cells in the representation (ETEX-I example).

has been defined, which corresponds to a gain of information in the inference. Application to the real tracer dispersion campaign ETEX-I, has shown that the optimal grid obtained from this new criterion is not only refined around the observation site and upwind of those stations, but also in areas where an inversion of these observations might indicate. However, one may object that an *inversion crime* is committed when using such a data-dependent cost function, since the adaptive grid that is used to perform Bayesian inverse modelling using a set of observations has been constructed with the help of the same set of observations. A solution to this subtle issue is left for future work.

The existence of an optimal number of tiles N was also discussed. All the well-controlled examples given here lead to the choice of the largest (numerically) possible N. But it was shown that taking into account a more complex model error may lead to a finite optimal N. So this issue remains very dependent on the physical context and on the specification of the model through the various scales.

The choice of the representation dictionary on which these criteria are optimized is another issue of practical concern. General tilings, where grid cells are defined as the Kronecker product of leaves of 1D binary tree structures, offer a rich set, but the numerical optimization scheme can be computationally demanding. As an alternative, we have implemented and tested a qtree structure where spatial tiles belong to a quaternary tree structure. Qtrees form a subset of the dictionary of tilings. It is more economical and faster to optimize in that subset. Furthermore, it has been shown on the ETEX-I example that optimal qtrees could be almost as efficient as optimal tilings. This is not so for another class of representations, the ftrees, whose skills are significantly inferior with a greater complexity in the optimization algorithm.

7.2. Connection with other multicale data assimilation approaches

The introduction of consistent multiscale formalisms is very recent in data assimilation, even though the inner and outer loops of 4D-Var can be seen as a precursor methodology (Courtier, 1994). Exploiting the framework developed by Willsky (2002), Zhou *et al.* (2008) have introduced a multiscale tree structure. A model operating at a different scale is assigned to each level of the tree. Using conditional probabilities and Bayes' rule, the information carried by the observations is propagated up and down the tree. This formalism is meant to be efficient with ensemble Kalman filtering. In a variational context, a fully consistent 4D-Var scheme has been developed on top of a two-way nested model (Simon *et al.*, 2011). It has been used to propagate information back and forth between the coarser and the finer grids. Multigrid methods used in the numerical solution of partial differential equations are percolating into data assimilation, although making them consistent with a data assimilation method is a challenge. As a preliminary step, Neveu *et al.* (2010) have tested such a scheme on a Burgers equation. The main advantage of the multigrid methods is the acceleration of convergence of the data assimilation scheme. In particular, it is shown to outperform the inner and outer loops scheme.

These formalisms, as well as the one of this article, are derived from first Bayesian principles. Therefore, to a large extent, they should be equivalent. However, they most naturally apply to different data assimilation schemes: Kalman filters, 4D-Var, or BLUE matrix equations in our case. That is why making connections between them cannot be a simple task.

7.3. Extension of the formalism to general data assimilation problems

The formalism developed in this article is expected to suit environmental problems whose monitoring network is known *a priori*, so that a grid optimization of control space can be performed prior to any inference. Yet the observation sites are not necessarily fixed, since the optimal representations can be dynamical. It is expected to be of primary interest for systems with sparse and inhomogeneous observations, and for data assimilation systems where the observational information is not propragated far, or anisotropically. At the very least, the methodology can help assess the areas which are poorly resolved (by the conjunction of models and observations).

The formalism of this study has been developed using the Jacobian of the system. In geophysics, the computation of the Jacobian is not always affordable, especially when the evolution model is nonlinear. To generalize our methodology to the nonlinear forecasting context of meteorology or oceanography, one needs to optimize the representations and compute the required pieces of the Jacobian when needed, similar to a standard 4D-Var. The main difficulty is that the optimal grids require to access second-order sensitivities, which are related to the Hessian $\mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$. We anticipate that this might be achieved using randomization techniques of Desroziers *et al.* (2005) or a stochastic gradient algorithm. Using ensemble Kalman filter methods, the error covariance matrix can be more easily accessed since it is given by the empirical statistics of the ensemble. As opposed to 4D-Var, control or state space representations can only be optimized in space, and not in time. After adequate inflation and localization, the methodology could used to optimize the representations of the state space and of control space. Furthermore, the methodology might be seen as a substitute for the localization of the raw empirical error covariance matrix. Insteading of choosing an adequate localization length, one chooses an adequate number of grid cells for the representation. The selection of an optimal representation would project the raw covariance matrix onto the active (real) degrees of freedom in the problem, curing any rank deficiency. The methodology is adaptive and can capture regions of the error covariance matrix where there is structure that might be smoothed out using standard uniform localization methods.

7.4. *Towards computationally efficient designs*

An optimization on tilings or qtrees could still be quite time-consuming when the number of grid cells in the finest grid reaches several hundred thousands, and when the hierarchical structure is deep. The application of the theory to large-dimensional systems, such as those contemplated earlier, may therefore be computationally challenging. As a short cut, an analytical approach based on the asymptotic properties of the optimal grids has been developed to offer an approximate but quick solution to the Bayesian design of control space. This will be reported in Part II of this article.

Acknowledgements

The authors would like to thank the associate editor, Dan Cornford, and two anonymous reviewers for their constructive and extensive comments and suggestions. The authors are grateful to Michel Ramonet for providing the CarboEurope-IP network characteristics. They thank Peter Rayner for stimulating discussions, and Monika Krysta for a careful reading of the manuscript. This article is a contribution to the MultiScale Data Assimilation in Geophysics (MSDAG) project supported by the Agence Nationale de la Recherche, grant ANR-08-SYSC-014.

Appendix

Alternate statistical regularisation

Before enforcing the tile number and the one tile–one point constraints, a tile can either be selected or not in the representation, with a probability that depends on its energy $\varepsilon_{\mathbf{l},k}$. Following this idea, one is led to a derivation similar to that of subsection 4.4, but with a more physical touch. It is assumed that the prior distribution (before imposing the constraints) of the tiles follows a Bernoulli law: tile $(\mathbf{l},k)$ is *a priori* selected with probability

$$\gamma_{\mathbf{l},k} = \frac{1}{1+\mathrm{e}^{-\beta\varepsilon_{\mathbf{l},k}}}\,, \tag{A.1}$$

which is the standard distribution factor of systems following Fermi–Dirac statistics. The prior law is then

$$\nu(\boldsymbol{\alpha}) = \prod_{\mathbf{l},k}\left\{(1-\gamma_{\mathbf{l},k})\,\delta_{\alpha_{\mathbf{l},k},0} + \gamma_{\mathbf{l},k}\,\delta_{\alpha_{\mathbf{l},k},1}\right\}\,. \tag{A.2}$$

One should then minimize the gain of information (maximize the entropy) from the prior distribution to the equilibrium distribution of tiles that satisfies the constraints. The information gain is measured by the Kullback–Leibler divergence (Kullback, 1959)

$$\mathcal{K}(p,\nu) = \sum_{\boldsymbol{\alpha}} q(\boldsymbol{\alpha})\ln\frac{q(\boldsymbol{\alpha})}{\nu(\boldsymbol{\alpha})}\,. \tag{A.3}$$

The resulting cost function to be maximized is

$$\begin{aligned}\widetilde{\mathcal{J}}(p) = &-\sum_{\boldsymbol{\alpha}} q(\boldsymbol{\alpha})\ln\frac{q(\boldsymbol{\alpha})}{\nu(\boldsymbol{\alpha})}\\ &+\sum_{\mathbf{l}}\sum_{k=1}^{n_{\mathbf{l}}} q(\boldsymbol{\alpha})\left(\mathbf{v}_{\mathbf{l},k}^{\mathrm{T}}\boldsymbol{\lambda}+\zeta\right)\alpha_{\mathbf{l},k}\\ &-\sum_{k=1}^{N_{\mathrm{fg}}}\lambda_k - \zeta N\,.\end{aligned} \tag{A.4}$$

The rest of the derivation is then unchanged, with the same intermediate and final results.

References

Bocquet M. 2005. Grid-resolution dependence in the reconstruction of an atmospheric tracer source. *Nonlin. Proc. Geophys.* **12**: 219–233.

Bocquet M. 2007. High-resolution reconstruction of a tracer dispersion event. *Q. J. R. Meteorol. Soc.* **133**: 1013–1026.

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

Bocquet M. 2009. Towards optimal choices of control space representation for geophysical data assimilation. *Mon. Weather Rev.* **137**: 2331–2348.

Bocquet M, Pires CA, Wu L. 2010. Beyond Gaussian statistical modeling in geophysical data assimilation. *Mon. Weather Rev.* **138**: 2997–3023.

Bocquet M, Wu L, Chevallier F. Bayesian design of control space for optimal assimilation of observations. Part II: Asymptotic solutions. 2011. *Q. J. R. Meteorol. Soc.* **137**: DOI: 10.1002/qj.841.

Borwein JM, Lewis AS. 2000. *Convex analysis and nonlinear optimization: theory and examples*. 273, Springer.

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

Chevallier F, Viovy N, Reichstein M, Ciais P. 2006. On the assignment of prior errors in Bayesian inversion of $\mathrm{CO_2}$ surface fluxes. *Geophys. Res. Lett.* **33**: L13802.

Courtier P. 1994. A strategy for operational implementation of 4D-Var using an incremental approach. *Q. J. R. Meteorol. Soc.* **120**: 1367–1387.

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

Desroziers G, Brousseau P, Chapnik B. 2005. Use of randomization to diagnose the impact of observations on analyses and forecasts. *Q. J. R. Meteorol. Soc.* **131**: 2821–2837.

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

Fan S-M, Gloor M, Mahlman J, Pacala S, Sarmiento JL, Takahashi T, Tans PP. 1998. Atmopsheric and oceanic $\mathrm{CO_2}$ data and models imply a large terrestrial carbon sink in North America. *Science* **282**: 442–444.

Kaminski T, Rayner PJ, Heimann M, Enting IG. 2001. On aggregation errors in atmospheric transport inversions. *J. Geophys. Res.* **106**: 4703–4715.

Kleeman R. 2002. Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.* **59**: 2057–2072.

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

Kullback S. 1959. *Information theory and statistics.* John Wiley & Sons: New York.

Neveu E, Debreu L, Le Dimet F-X. 2010. 'Multigrid methods and data assimilation applied to a linear advection equation'. Pp. 181–188 in *Proceedings of the 20th African Conference on Research in Computer Science and Applied Mathematics*, Yamoussoukro, Ivory Coast.

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

Peylin P, Bousquet P, Ciais P. 2001. Inverse modeling of atmospheric carbon dioxin fluxes – Response. *Science* **294**: 2292–2292.

Rödenbeck C, Houweling S, Gloor M, Heimann M. 2003. CO_2 flux history 1982–2001 inferred from atmospheric data using a global inversion of atmospheric transport. *Atmos. Chem. Phys.* **3**: 1919–1964.

Rodgers CD. 2000. *Inverse methods for atmospheric sounding. Series on Atmospheric, Oceanic and Planetary Physics*. World Scientific.

Roustan Y, Bocquet M. 2006. Sensitivity analysis for mercury over Europe. *J. Geophys. Res.* **111**: D14304.

Saide P, Bocquet M, Osses A, Gallardo L. 2011. Constraining surface emissions of air pollutants using inverse modeling: Method intercomparison and a new two-step multiscale approach. *Tellus B* in press.

Simon E, Debreu L, Blayo E. 2011. 4D variational data assimilation for locally nested models: Complementary theoretical aspects and application to a 2D shallow-water model. *Int. J. Num. Methods Fluids* in press.

Trampert J, Snieder R. 1996. Model estimations biased by truncated expansions: Possible artifacts in seismic tomography. *Science* **271**: 1257–1260.

Willsky AS. 2002. Multiresolution Markov models for signal and image processing. *IEEE Proc.* **90**: 1396–1458.

Zhou Y, McLaughlin A, Entekhabi D, Crystal Ng G-H. 2008. An ensemble multiscale filter for large nonlinear data assimilation problems. *Mon. Weather Rev.* **136**: 678–698.