University of Reading

Estimating model evidence using data assimilation

Article

Accepted Version

Carrassi, A. ORCID: https://orcid.org/0000-0003-0722-5600, Bocquet, M., Hannart, A. and Ghil, M. (2017) Estimating model evidence using data assimilation. Quarterly Journal of the Royal Meteorological Society, 143 (703). pp. 866-880. ISSN 1477-870X doi: https://doi.org/10.1002/qj.2972 Available at https://centaur.reading.ac.uk/90356/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.
Published version at: http://dx.doi.org/10.1002/qj.2972
To link to this article DOI: http://dx.doi.org/10.1002/qj.2972

Publisher: Royal Meteorological Society

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

University of Reading

Reading's research outputs online

Estimating model evidence using data assimilation

A. Carrassi[a*], M. Bocquet[b], A. Hannart[c] and M. Ghil[d,e]

[a] *Nansen Environmental and Remote Sensing Center, Bergen, Norway.*

[b] *CEREA, joint laboratory École des Ponts ParisTech and EDF R&D, Université Paris-Est, Champs-sur-Marne, France.*

[c] *IFAECI, CNRS-CONICET-UBA, Buenos Aires, Argentina.*

[d] *École Normale Supérieure, Paris, France.*

[e] *University of California, Los Angeles, USA*

[*]Correspondence to: Alberto Carrassi, e-mail: alberto.carrassi@nersc.no

We review the field of data assimilation (DA) from a Bayesian perspective and show that, in addition to its by now common application to state estimation, DA may be used for model selection. An important special case of the latter is the discrimination between a factual model — which corresponds, to the best of the modeler's knowledge, to the situation in the actual world in which a sequence of events has occurred — and a counterfactual model, in which a particular forcing or process might be absent or just quantitatively different from the actual world. Three different ensemble-DA methods are reviewed for this purpose: the ensemble Kalman filter (EnKF), the ensemble four-dimensional variational smoother (En-4D-Var), and the iterative ensemble Kalman smoother (IEnKS). An original contextual formulation of model evidence (CME) is introduced. It is shown how to apply these three methods to compute CME, using the approximated time-dependent probability distribution functions (pdfs) each of them provide in the process of state estimation. The theoretical formulae so derived are applied to two simplified nonlinear and chaotic models: (i) the Lorenz three-variable convection model (L63), and (ii) the Lorenz 40-variable mid-latitude atmospheric dynamics model (L95). The numerical results of these three DA-based methods and those of an integration based on importance sampling are compared. It is found that better CME estimates are obtained by using DA, and the IEnKS method appears to be best among the DA methods. Differences among the performance of the three DA-based methods are discussed as a function of model properties. Finally, the methodology is implemented for parameter estimation and for event attribution.

Key Words: model evidence; data assimilation; marginal likelihood; model selection; detection and attribution; ensemble Kalman filter; iterative ensemble Kalman smoother; ensemble 4DVar

Received . . .

1. Introduction and motivation

High-dimensional and nonlinear state-space evolution models arise in numerous and diverse fields, *e.g.,* numerical weather prediction (NWP) (Ghil and Malanotte-Rizzoli 1991), air quality forecasting (Zhang *et al.* 2012), subsurface flow modeling (Elsheikh *et al.* 2014a,b), oceanography (Balmaseda *et al.* 2009), climatic projection and reconstruction (Bhend *et al.* 2012), and signal processing (Crisan and Doucet 2002), to mention just a

© 2016 Royal Meteorological Society

few. Most of these applications require one to solve the problem of optimally estimating the partially observed state that evolves in time. This problem is often tackled by combining the observed data with a numerical model that is based on a comprehensive theoretical description of the processes at play. Designing methods and numerical algorithms that address this problem in a way that is theoretically sound and yet tailored to the specifics of the context at hand is an active research field in computational statistics in general, and in the aforementioned fields in particular.

This paper focuses on a related yet distinct problem, which has received less attention thus far. It consists in quantifying the resulting performance of the state inference by estimating the so-called *marginal likelihood of the observations* — also referred to as *model evidence* – which quantifies the "goodness-of-fit" between the data and the chosen state-space model (Baum *et al.* 1970). Solving the model evidence estimation problem has attracted, as we shall see, less research effort than the state estimation one. Still, estimating model evidence is just as important as state estimation.

Indeed, the model evidence can be used as a general metric for model comparison and selection. Areas of application include calibrating a given state-space model's parameters based on the observed data by maximizing the model evidence; comparing the skill of several candidate models (or model settings, or boundary conditions, or numerical schemes) in representing a given observed phenomenon or class of phenomena and thereby selecting the most appropriate candidate; quantifying the evidence that supports one of several, potentially conflicting, theoretical hypotheses for the physics of a phenomenon; or providing evidence for either the existence or the nonexistence of a causal relationship between a hypothetical forcing external to a system and an observed response of the system.

It is, in fact, the latter situation that largely motivated the present work, in the context of the climate system and of its increasingly significant anthropogenic forcing, cf. Hannart *et al.* (2016) and references therein. More generally speaking, model evidence may help address fundamental questions whenever confronting theory to observations in an attempt to improve the former, and prove to be a useful tool in extracting quantitative information from such a model–data confrontation.

Mathematically speaking, deriving model evidence is quite difficult as it requires, by definition, to integrate out the state vector. Although viable solutions exist in the case of Gaussian errors and linear model (see e.g. Winiarek *et al.* 2011), this daunting task is usually intractable, especially for the class of high-dimensional and non-linear models considered here. To circumvent this difficulty, different approaches were proposed in the literature to yield an accurate and more easily computable estimate of the desired value (Hürzeleri and Künsch 2001; Pitt 2002; Kantas *et al.* 2009).

Recently, marginal likelihoods have been used for the purposes of Bayesian inference and model selection. An efficient approach to deal with the former, based on Monte Carlo sampling, has been proposed by Elsheikh *et al.* (2014a,b) in the context of models of subsurface flows. Carson *et al.* (2015) have described an application to model selection; the method is again based on an advanced Monte Carlo technique and has proven to efficiently discriminate between phenomenological models of the glacial-interglacial cycle using a single dataset. Reich and Cotter (2015, Chapter 9.1) offer an exposition on model evidence, its use for model selection and how to use particle filters for its computation.

The present study focuses on the estimation of model evidence using data assimilation (DA) methods designed to deal with large numerical models and datasets, subject to partially Gaussian assumptions. DA methods were in fact initially developed for high-dimensional state estimation in the NWP context, in order to initialize an atmospheric model by estimating its state variables based on meteorological observations that are incomplete, diverse, unevenly distributed in space and time and are contaminated by measurement error (Bengtsson *et al.* 1981; Kalnay 2002, and references therein). These methods outgrew their original application field over the past decades, to reach a wide variety of fields in the geosciences and elsewhere. They thus present the advantage of being used operationally in many different contexts and by many practitioners, and to be the focus of a significant research effort (Blayo *et al.* 2015). In the recent past, model evidence estimates have started to appear in the DA context (Winiarek *et al.* 2011, 2012; Tandeo *et al.* 2015; Hannart *et al.* 2016).

© 2016 Royal Meteorological Society

The purpose of the present paper is to extend these studies by proposing several approaches for estimating the marginal likelihood (as model evidence). Section 2 recalls key aspects of Bayesian DA that are essential for the discussion; see Bocquet *et al.* (2010) and Reich and Cotter (2015) for an extensive treatment of the subject. Section 3 introduces the concept of *contextual* model evidence which is of particular interest in the DA setting. Section 4 lays out four different computational schemes to estimate contextual model evidence. Section 5 implements these schemes for two low-dimensional nonlinear models and compares their performance. Section 6 briefly illustrates the application of the proposed schemes to parameter estimation and causal attribution of weather events. Section 7 finally summarizes our results, provides conclusions and future directions.

2. Data assimilation from a Bayesian perspective: brief overview

Let us assume that a model of the physical process of interest is given as a discrete dynamical system in an M-dimensional Euclidean space $\mathbb{R}^M$,

$$\mathbf{x}_k = \mathcal{M}_{k:k-1}(\mathbf{x}_{k-1}) + \boldsymbol{\eta}_k. \tag{1}$$

Here $\mathbf{x}_k \in \mathbb{R}^M$ is the state vector, $\mathcal{M}_{k:k-1} : \mathbb{R}^M \to \mathbb{R}^M$ is usually a nonlinear, possibly chaotic, map and $\boldsymbol{\eta}_k \in \mathbb{R}^M$ stands for model error, represented as a stochastic additive term.

Noisy observations of $\mathbf{x}$ are available at discrete times and are represented as components of the observation vector $\mathbf{y} \in \mathbb{R}^d$. The relation between $\mathbf{y}$ and the model state $\mathbf{x}$ is given by

$$\mathbf{y}_k = \mathcal{H}_k(\mathbf{x}_k) + \boldsymbol{\epsilon}_k. \tag{2}$$

In Eq. (2), $\mathcal{H} : \mathbb{R}^M \to \mathbb{R}^d$ is the, possibly nonlinear, observation operator that maps the model solution to the observation space $\mathbb{R}^d$; typically $d \ll M$ and $\mathcal{H}$ may involve spatial interpolations in finite-difference models or spectral-to-physical space transformation in spectral models. Transformations based on physical laws for indirect measurements, such as radiative fluxes used to infer temperatures, can also be represented in this way (e.g., Kalnay 2002). The observational error $\boldsymbol{\epsilon}_k$ is also represented as a stochastic additive term.

Both random sequences $\{\boldsymbol{\eta}_k : k = 0, \ldots, K\}$ and $\{\boldsymbol{\epsilon}_k : k = 0, \ldots, K\}$ are assumed to be white in time, mutually independent and distributed according to the probability density functions (pdfs) $p_{\boldsymbol{\eta}}$ and $p_{\boldsymbol{\epsilon}}$, respectively. These pdfs represent the transitional kernel for the probabilistic transition from $\mathbf{x}_{k-1}$ to $\mathbf{x}_k$, and the likelihood of the observations $\mathbf{y}_k$ conditioned on the state $\mathbf{x}_k$, respectively,

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}) = p_{\boldsymbol{\eta}}[\mathbf{x}_k - \mathcal{M}_k(\mathbf{x}_{k-1})], \tag{3}$$

$$p(\mathbf{y}_k|\mathbf{x}_k) = p_{\boldsymbol{\epsilon}}[\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)]. \tag{4}$$

The output of the estimation process is the posterior pdf $p(\mathbf{x}|\mathbf{y})$ of the process $\mathbf{x}$ conditioned on the data $\mathbf{y}$,

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}\,. \tag{5}$$

In this straightforward application of the Bayes formula, $p(\mathbf{x})$ is the prior distribution that encodes all the knowledge about the process before assimilating the new observations, and $p(\mathbf{y})$ is the observation likelihood. The latter is usually independent from time when the estimation is performed and plays the role of a normalization coefficient. We will see in sections 3 and 4 how this view can be reversed when solving the DA problem for the purpose of evaluating model evidence, cf. Hannart *et al.* (2016).

Once the sequences of system states and observations are collected into $\mathbf{x}_{k:0} = \{\mathbf{x}_k, \mathbf{x}_{k-1}, ..., \mathbf{x}_0\}$ and $\mathbf{y}_{k:0} = \{\mathbf{y}_k, \mathbf{y}_{k-1}, ..., \mathbf{y}_0\}$, it is possible to define three estimation problems, depending on the time period where observations are distributed and the time when we want to estimate the state (Wiener 1949):

1. *Prediction*: Estimate $p(\mathbf{x}_l|\mathbf{y}_{k:0})$ with $l > k$.
2. *Filtering*: Estimate $p(\mathbf{x}_k|\mathbf{y}_{k:0})$.
3. *Smoothing*: Estimate $p(\mathbf{x}_{k:0}|\mathbf{y}_{k:0})$.

The prediction problem is formally addressed by solving the corresponding Chapman-Kolmogorov equation for the

© 2016 Royal Meteorological Society

propagation of a pdf under the model dynamics

$$p(\mathbf{x}_l|\mathbf{y}_{k:0}) = \int \mathrm{d}\mathbf{x}_k \, p_{\boldsymbol{\eta}}[\mathbf{x}_l - \mathcal{M}_{l:k}(\mathbf{x}_k)] p(\mathbf{x}_k|\mathbf{y}_{k:0}). \quad (6)$$

The filtering problem is the most common one in geophysical applications, and it is characterized by sequential processing, in which measurements are utilized as they become available (Jazwinski 1970; Bengtsson *et al.* 1981): a so-called analysis step, in which the conditional pdf $p(\mathbf{x}_k|\mathbf{y}_{k:0})$ is updated using the latest observation, $\mathbf{y}_k$, alternates with a forecast step in which this pdf is propagated forward until the time of a new observation. The analysis is based on the application of the Bayes formula (5), which becomes

$$p(\mathbf{x}_k|\mathbf{y}_{k:0}) = \frac{p_{\boldsymbol{\epsilon}}[\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)] p(\mathbf{x}_k|\mathbf{y}_{k-1:0})}{\int \mathrm{d}\mathbf{x}_k \, p_{\boldsymbol{\epsilon}}[\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)] p(\mathbf{x}_k|\mathbf{y}_{k-1:0})}, \quad (7)$$

while in the prediction step one integrates the Chapman-Kolmogorov equation (6), with $l \to k$ and $k \to (k-1)$.

Finally, using recursively the Bayes formula in the interval $\{t_0 \le t \le t_k\}$, the smoothing problem can be written as

$$p(\mathbf{x}_{k:0}|\mathbf{y}_{k:0}) \propto p(\mathbf{x}_0) \prod_{l=1}^{k} p_{\boldsymbol{\epsilon}}[\mathbf{y}_l - \mathcal{H}_l(\mathbf{x}_l)] p_{\boldsymbol{\eta}}[\mathbf{x}_l - \mathcal{M}_l(\mathbf{x}_{l-1})]. \quad (8)$$

The faithful numerical implementation of Eqs. (3)–(8) is impossible in realistic geophysical and other high-dimensional applications of DA because the huge size of the discrete models, $\mathcal{M}_k$, and of the observation vector, $\mathbf{y}_k$, renders the accurate representation of the relevant pdfs prohibitive. This problem is usually overcome by assuming that the error statistics are Gaussian. This assumption — along with some form of linearization of $\mathcal{M}_k$ and of $\mathcal{H}_k$ — allows one to fully characterize the pdfs in the equations by their first two moments only, *i.e.*, the means and covariances. Such a characterization results in an enormous simplification when applied to high-dimensional systems, a simplification that is in fact the basis of many successful practical DA algorithms. Some of these practical issues in applying the Bayesian perspective to DA will resurface in section 4 in computing model evidence.

3. Model evidence: a contextual formulation

Let us generalize the definition of an observation sequence given in section 2 to an arbitrary time interval $k:m$, with $k \ge m$, so that

$$\mathbf{y}_{k:m} = \{\mathbf{y}_k, \mathbf{y}_{k-1}, \cdots, \mathbf{y}_{m+1}, \mathbf{y}_m\}. \quad (9)$$

In particular, for $m = -\infty$ and $k = 0$, the sequence (9) contains all the observations from the far past up to the present time $k = 0$. The likelihood of the observations given $\mathcal{M}$ can be written as

$$p(\mathbf{y}_{k:}|\mathcal{M}) = \int \mathrm{d}\mathbf{x} \, p(\mathbf{y}_{k:}|\mathbf{x}, \mathcal{M}) p(\mathbf{x}|\mathcal{M}) \quad (10)$$

where we used the abbreviated notation $\mathbf{y}_{k:-\infty} = \mathbf{y}_{k:}$.

The likelihood $p(\mathbf{y}_{k:}|\,\mathcal{M})$ is referred to as *model evidence* and it is often used in model selection (e.g., Carson *et al.* 2015, and references therein). This likelihood depends on the underlying dynamics $\mathcal{M}$, but it can be formulated as being dependent on any hypothesis under scrutiny. When the model dynamics is ergodic, $p(\mathbf{x}|\mathcal{M})$ is the invariant distribution on the attractor and the likelihood $p(\mathbf{y}_{k:})$ will capture it with an accuracy that depends on the observation model in Eq. (2). The model evidence $p(\mathbf{y}_{k:}|\mathcal{M})$ is obtained by integrating over $\mathbf{x}$ given $\mathbf{y}_{k:}$ and it represents the probability that the data are actually observed under the hypothesis that the model $\mathcal{M}$ is the correct one. To simplify the notation hereafter, the explicit dependence on $\mathcal{M}$ is dropped. For instance, the marginal probability of the data is denoted by $p(\mathbf{y}_{k:}) = p(\mathbf{y}_{k:}|\mathcal{M})$.

The distribution $p(\mathbf{x})$ in Eq. (10) plays the role of a prior and it thus allows one to introduce additional information about the system. The choice of the prior is usually arbitrary and one can in principle use any distribution that suits a study's specific purposes. In many practical circumstances, however, the search for a good informative prior is not straightforward, although its choice may strongly, and sometimes negatively, affect the results of the study.

The use of the climatological invariant distribution for $p(\mathbf{x})$ in Eq. (10) has the advantage of characterizing the system globally, but it is not very informative about its specific current conditions. This is particularly true when the underlying dynamics is out of equilibrium, when it possesses multiple stationary points, or when

© 2016 Royal Meteorological Society

it is subject to large deviations. Moreover, in the case of the large-dimensional systems used in NWP and in climate prediction, as considered in this study, a proper estimate of $p(\mathbf{x})$ is complicated even further by the limitations of the computational resources.

Finally, if the system is subject to climate change driven by a time-dependent forcing — such as anthropogenic changes in greenhouse gas and aerosol concentrations — its dynamics has to be described self-consistently as non-autonomous, and the mere existence of a time-independent invariant set on which $p(\mathbf{x})$ is defined is questionable. In this case, one has to rely on the concept of a pullback or random attractor that is invariant under the system's dynamics but does depend on time, *i.e.* both p and $\mathcal{M}$ in Eq. (10) are explicit functions of t (e.g., Ghil *et al.* 2008; Chekroun *et al.* 2011; Dijkstra 2013). This non-autonomous situation is left for later investigation.

For the purpose of a time-dependent evaluation of the model evidence in the autonomous case — in which $\mathcal{M}$ only depends on time due to its non-linearity, *i.e.*, through its dependence on $\mathbf{x}$ — we are interested in a definition narrowed to the present moment. To this end, let us condition the observational likelihood from the present $t = t_0$ to some future time $t = t_k$, on the observation sequence up to the present, so that

$$p(\mathbf{y}_{k:}) = p(\mathbf{y}_{k:1}|\mathbf{y}_{0:})p(\mathbf{y}_{0:}). \tag{11}$$

The assumption here is to use $p(\mathbf{y}_{k:1}|\mathbf{y}_{0:})$ instead of $p(\mathbf{y}_{k:})$, and the conditional pdf $p(\mathbf{y}_{k:1}|\mathbf{y}_{0:})$ will be called the *contextual model evidence* (CME). Implicit in Eq. (11) is the idea that the informational content from past observations is propagated forward by conditioning on $\mathbf{y}_{0:}$. While $\mathbf{y}_{k:1}$ is still used to diagnose the evidence in the time interval from t_1 to t_k, $\mathbf{y}_{0:}$ allows us now to specify the *context*.

By marginalizing with respect to $\mathbf{x}_0$, one can write the CME as

$$\begin{aligned} p(\mathbf{y}_{k:1}|\mathbf{y}_{0:}) &= \int \mathrm{d}\mathbf{x}_0\, p(\mathbf{y}_{k:1}|\mathbf{x}_0, \mathbf{y}_{0:})p(\mathbf{x}_0|\mathbf{y}_{0:}) \\ &= \int \mathrm{d}\mathbf{x}_0\, p(\mathbf{y}_{k:1}|\mathbf{x}_0)p(\mathbf{x}_0|\mathbf{y}_{0:}). \end{aligned} \tag{12}$$

Equation (12) shows that the CME depends on two factors, on the conditional pdf $p(\mathbf{x}_0|\mathbf{y}_{0:})$, which plays the role of the prior and substitutes $p(\mathbf{x})$ in Eq. (10), and $p(\mathbf{y}_{k:1}|\mathbf{x}_0)$, the likelihood of the observational sequence $\mathbf{y}_{k:1}$ conditioned on the system's state.

The key point here is that the new prior, $p(\mathbf{x}_0|\mathbf{y}_{0:})$, is easier to compute than the invariant measure $p(\mathbf{x})$. Indeed, we saw in section 2 that the conditional pdf $p(\mathbf{x}_0|\mathbf{y}_{0:})$ is the posterior density in a Bayesian inference process designed to estimate $\mathbf{x}_0$ based on $\mathbf{y}_{0:}$ and that this posterior pdf is the standard, albeit approximate, outcome of applying a DA algorithm to the model $\mathcal{M}$ and the data $\mathbf{y}_{k:1}$. When such a forecast–assimilation cycle as described in section 2 is routinely running — as is the case in an operational NWP center — an approximation of $p(\mathbf{x}_0|\mathbf{y}_{0:})$ is already at hand. Given the model and the observational network, the level of accuracy of the DA-based approximation for $p(\mathbf{x}_0|\mathbf{y}_{0:})$ is related to the degree of sophistication of the DA scheme adopted. We shall show in section 4 that DA can also be used to estimate the other term in Eq. (12), namely $p(\mathbf{y}_{k:1}|\mathbf{x}_0)$, and thus to fully accomplish the task of estimating the CME, consistently and routinely.

4. Data assimilation for model evidence

Estimating the CME amounts to computing the integral in Eq. (12). Analytic solutions can only be obtained for elementary cases, and the problem becomes rapidly intractable as one moves toward realistic situations with practical relevance. Numerical methods are thus necessary in practice and their degree of complexity grows with the dimension M of the model and that of the data set, d. For high-dimensional systems with large M, Monte Carlo methods using importance sampling are a viable approach, but their convergence as the sample size increases is usually very slow even when M is only moderately large. An approximate Monte Carlo suitable for NWP applications is described in Sect. 4.2. Large dimensionality d of the data calls for Laplace method, in which the integrand is approximated as a Gaussian and a solution for the integral can be found as a function of the mode and the covariances of this normal distribution. We will make use of the Laplace approximation in combination with smoothers in Sect. 4.4.

In the geosciences, one often encounters both conditions, with M and d up to $O(10^9)$ and $O(10^7)$, respectively, and computing the integral in Eq. (12) is a very challenging task that requires a trade-off between accuracy and computational efficiency. In this

© 2016 Royal Meteorological Society

section, we present a hierarchy of methods based on DA practice that allow one to evaluate the CME integral in Eq. (12) within a good approximation, and that are suitable for the large systems and big data sets typical of the environmental sciences. To a certain extent, the accuracy of the methods described below can be ranked according to the degree of sophistication of the DA approach on which they are based.

4.1. *General setting*

We start by providing an iterative formula that will be used later, and that allows one to decompose the contextual evidence pdf as

$$\begin{aligned} p(\mathbf{y}_{K:1}|\mathbf{y}_{0:}) &= p(\mathbf{y}_{K:2}|\mathbf{y}_{1:})p(\mathbf{y}_1|\mathbf{y}_{0:}) \\ &= p(\mathbf{y}_{K:3}|\mathbf{y}_{2:})p(\mathbf{y}_2|\mathbf{y}_{1:})p(\mathbf{y}_1|\mathbf{y}_{0:}) \end{aligned} \tag{13}$$

and so on, up to time t_K:

$$p(\mathbf{y}_{K:1}|\mathbf{y}_{0:}) = \prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{y}_{k-1:}). \tag{14}$$

Hence the contextual evidence of the sequence $\mathbf{y}_{K:0}$ can be written as the product of single contextual evidences, one for each $\mathbf{y}_k$. Moreover, the individual contextual evidence $p(\mathbf{y}_k|\mathbf{y}_{k-1:})$ is often a tractable output of a DA scheme. After marginalizing over the state vector $\mathbf{x}_k$, we get

$$p(\mathbf{y}_k|\mathbf{y}_{k-1:}) = \int \mathrm{d}\mathbf{x}_k \, p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{y}_{k-1:}) \,, \tag{15}$$

where $p(\mathbf{y}_k|\mathbf{x}_k)$ is the observation likelihood and $p(\mathbf{x}_k|\mathbf{y}_{k-1:})$ is the forecast state pdf at t_k. Carson *et al.* (2015) recently used the identity in Eq. (14), whereas Del Moral (2004) provides an alternative proof of its validity.

We assume that, at any arbitrary time t_0, an estimate of the posterior density $p(\mathbf{x}_0|\mathbf{y}_{0:})$ is available as the outcome of a forecast–assimilation cycle up to t_0. This posterior is then used as a prior in the estimation of the CME in Eq. (12). Unless otherwise stated, a DA method that uses first- and second-order error moments is adopted herein, so that the posterior pdf is approximated as a Gaussian, $p(\mathbf{x}_0|\mathbf{y}_{0:}) \approx p^{\mathrm{DA}} = \mathcal{N}(\mathbf{x}_0, \mathbf{P}^{\mathrm{a}})$, with mean $\mathbf{x}_0$ and covariance matrix $\mathbf{P}^{\mathrm{a}}$, where the superscript stands for analysis. Similarly, let us define $\mathbf{P}^{\mathrm{f}}$ and $\mathbf{R}$, to be used in the following sections, as the forecast and observation error covariance, respectively.

The *evidencing window* is defined as the interval $[t_0, t_K]$, and we aim at estimating the CME

$$p(\mathbf{y}_{K:1}|\mathbf{y}_{0:}) = \int \mathrm{d}\mathbf{x}_0 \, p(\mathbf{y}_{K:1}|\mathbf{x}_0)p(\mathbf{x}_0|\mathbf{y}_{0:}). \tag{16}$$

The forward model, Eq. (1), is assumed here to be deterministic and perfect, so that $\boldsymbol{\eta}_k = 0$. The implications of this choice are discussed in section 7 and a follow-on study will relax these assumptions. See also (Reich and Cotter 2015, Chapter 9.1) for a definition of model evidence in the case of a stochastic forward model.

4.2. *Monte Carlo and importance sampling*

The CME integral, Eq. (16), can be estimated using a Monte Carlo approach with importance sampling. Samples are drawn from the Gaussian pdf $p^{\mathrm{DA}} = \mathcal{N}(\mathbf{x}_0, \mathbf{P}^{\mathrm{a}})$, outcome of a Gaussian DA scheme at time t_0, used as proposal density for the unknown actual $p(\mathbf{x}_0|\mathbf{y}_{0:})$.

Let us suppose that, at t_0, a sample of N members, $\{\mathbf{x}_0^i : i = 1, \ldots, N\}$, is drawn from $p(\mathbf{x}_0|\mathbf{y}_{0:})$, and its members are used as initial conditions for a forward integration over the evidencing window. The CME can then be estimated as

$$p(\mathbf{y}_{K:1}|\mathbf{y}_{0:}) \approx \frac{1}{N}\sum_{i=1}^{N} p(\mathbf{y}_{K:1}|\mathbf{x}_0^i). \tag{17}$$

The approximation in Eq. (17) gets progressively better by increasing N, but the computational resources usually set a bound to the size of the sample. We will use this Monte Carlo estimate, Eq. (17), with $N = 10^6$, for some cases in Sect. 5.2. Equation (17), but for imperfect models, is at the basis of a model evidence computing algorithm described in Reich and Cotter (2015, Algorithm 9.2). To prevent filter degeneracy and enhance efficiency with an affordable number of members/particles, they suggest the use of sequential Monte Carlo methods with resampling.

We are interested here in a straightforward use of the approximation Eq. (17) in a high-dimensional context where the sample's size is unavoidably very small. In the geosciences, one

© 2016 Royal Meteorological Society

often uses an *ensemble Kalman filter* (EnKF, Evensen 2009) with N members to assimilate data up to the beginning t_0 of the evidencing window. In any realistic setup, $N \ll M$ and these N members provide a reduced-order, but dynamically consistent, representation of $p(\mathbf{x}_0|\mathbf{y}_{0:})$ that can be used in Eq. (17) in place of a random draw. This approach, conveniently referred here to as importance sampling (IS), albeit not very accurate, is computationally affordable in NWP and climate prediction centers. We will use it here in comparison with some Gaussian ensemble-DA methods in the numerical comparison described in Sect. 5.

We turn now to the description of several methods to compute the CME, each of which is based on a distinct Gaussian ensemble-DA algorithm.

4.3. *Filtering: Kalman filter and ensemble Kalman filter*

Kalman Filter (KF). Assuming the evolution and observation models in Eqs. (1) and (2) are both linear, and that the observation errors and the initial errors are Gaussian, the Kalman filter (KF: Kalman 1960; Ghil and Malanotte-Rizzoli 1991) is optimal. In the logarithmic formulation of the KF, the observation likelihood pdfs are

$$\ln p(\mathbf{y}_k|\mathbf{x}_k) = -\frac{1}{2}\left\|\mathbf{y}_k - \mathbf{H}_k\mathbf{x}_k\right\|^2_{\mathbf{R}_k} - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\mathbf{R}_k\right|, \quad k = 1,\ldots,K, \tag{18}$$

and the forecast pdfs are of the form

$$\ln p(\mathbf{x}_k|\mathbf{y}_{k-1:}) = -\frac{1}{2}\left\|\mathbf{x}_k - \mathbf{x}^{\mathrm{f}}_k\right\|^2_{\mathbf{P}^{\mathrm{f}}_k} - \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\mathbf{P}^{\mathrm{f}}_k\right|, \quad k = 1,\ldots,K; \tag{19}$$

here $\mathbf{H}_k$ is the linearized observation operator at time t_k. The weighted Euclidean norm $\|\mathbf{x}\|^2_{\mathbf{A}} = \mathbf{x}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{x}$ is used, while $|\mathbf{A}|$ indicates the determinant of $\mathbf{A}$.

The contextual evidence

$$p(\mathbf{y}_k|\mathbf{y}_{k-1:}) = \int \mathrm{d}\mathbf{x}_k p(\mathbf{y}_k|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{y}_{k-1:}) \tag{20}$$

is, in this case, the product of two Gaussian pdfs, and hence it is itself Gaussian. Besides, the two statistical moments that suffice to characterize it are given by those of the innovations: $\mathbf{H}_k\mathbf{x}^{\mathrm{f}}_k$ for the mean and $\mathbf{R}_k + \mathbf{H}_k\mathbf{P}^{\mathrm{f}}_k\mathbf{H}^{\mathrm{T}}_k$ for the covariance matrix. As a result

$$\ln p(\mathbf{y}_k|\mathbf{y}_{k-1:}) = -\frac{1}{2}\left\|\mathbf{y}_k - \mathbf{H}_k\mathbf{x}^{\mathrm{f}}_k\right\|^2_{\mathbf{R}_k+\mathbf{H}_k\mathbf{P}^{\mathrm{f}}_k\mathbf{H}^{\mathrm{T}}_k} - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\mathbf{R}_k + \mathbf{H}_k\mathbf{P}^{\mathrm{f}}_k\mathbf{H}^{\mathrm{T}}_k\right|, \quad k = 1,\ldots,K, \tag{21}$$

so that the factorization formula in Eq. (14) now reads

$$p(\mathbf{y}_{K:1}|\mathbf{y}_{0:}) = \prod_{k=1}^{K} \frac{\exp\left(-\frac{1}{2}\left\|\mathbf{y}_k - \mathbf{H}_k\mathbf{x}^{\mathrm{f}}_k\right\|^2_{\mathbf{R}_k+\mathbf{H}_k\mathbf{P}^{\mathrm{f}}_k\mathbf{H}^{\mathrm{T}}_k}\right)}{\sqrt{(2\pi)^d\left|\mathbf{R}_k + \mathbf{H}_k\mathbf{P}^{\mathrm{f}}_k\mathbf{H}^{\mathrm{T}}_k\right|}}. \tag{22}$$

An alternative, less direct, proof of Eq. (22) can be found in Hannart *et al.* (2016).

Ensemble Kalman Filter (EnKF). If the forward model $\mathcal{M}(\mathbf{x}_k)$ in Eq. (1) is nonlinear, and even when both initial errors $\mathbf{P}^{\mathrm{f}}_0$ and observation errors $\mathbf{R}_k$ are Gaussian, the evolution of the forecast and analysis pdfs will not remain Gaussian, with very few exceptions. There are many ways of approximating their evolution in time, including the *extended KF* (EKF: Jazwinski 1970; Miller *et al.* 1994), the EnKF (Evensen 2009) and the *unscented KF* (UKF: Grewal and Andrews 2001). Of these, the EKF is probably most widely used in engineering applications, while the EnKF is very widely used in the geosciences where the typical problem size is much higher.

To approximate the CME in nonlinear cases, we consider here the EnKF and introduce the matrix $\mathbf{E}_k = [\mathbf{x}_1,\ldots,\mathbf{x}_N]$, whose columns contain the ensemble members, as well as the corresponding normalized forecast anomalies with respect to the ensemble mean $\mathbf{X}_k = \mathbf{E}_k\left(\mathbf{I}_N - \mathbf{1}\mathbf{1}^T/N\right)/\sqrt{N-1}$; here $\mathbf{1} \in \mathbb{R}^N$ is the column vector of ones, and $\mathbf{I}_N$ is the $N \times N$ identity matrix. We obtain

$$p(\mathbf{y}_{K:1}|\mathbf{y}_{0:}) \simeq \prod_{k=1}^{K} \frac{\exp\left(-\frac{1}{2}\left\|\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}^{\mathrm{f}}_k)\right\|^2_{\mathbf{R}_k+\mathbf{Y}_k\mathbf{Y}^{\mathrm{T}}_k}\right)}{\sqrt{(2\pi)^d\left|\mathbf{R}_k + \mathbf{Y}_k\mathbf{Y}^{\mathrm{T}}_k\right|}}, \tag{23}$$

© 2016 Royal Meteorological Society

where $\mathbf{Y}_k = \mathbf{H}_k \mathbf{X}_k$ is the matrix of normalized observation anomalies at t_k, usually estimated as $\mathbf{Y}_k = \mathcal{H}(\mathbf{E}_k)\left(\mathbf{I}_N - \mathbf{1}\mathbf{1}^T/N\right)/\sqrt{N-1}$.

The estimate in Eq. (23) is exact, and coincides with that in Eq. (22), when the model is linear, the initial condition and observation errors are Gaussian and when the initial ensemble anomalies span the full range of uncertainty.

4.4. *Smoothing: ensemble-4D-Var and iterative ensemble Kalman smoother*

We will use here the Laplace approximation to estimate the CME integral Eq. (16). As mentioned in Sect. 4, in the Laplace method, the integrand is approximated as a Gaussian and the integral is a function of the mode and the covariance of this normal distribution. In the asymptotic limit $d \to \infty$, most of the contributions to the integral come from the vicinity of the maximum and the Laplace approximation — whose accuracy scales as the inverse of the variance of the approximating Gaussian — gets progressively more accurate. The reader is referred to Evans and Swartz (1995) for a review of various integration methods used in statistical inference. Reich and Cotter (2015, Example 9.4) suggest to use the Laplace approximation in the computation of model evidence.

In the following we describe two approaches, the ensemble 4D-Var and the iterative ensemble Kalman smoother, that can be used to compute the best estimator, the mode, and the associated uncertainty, the covariance. This latter is estimated using the Hessian of the corresponding cost-functions, in a way that is made clear later in this section.

Ensemble 4D-Var (En-4D-Var).

A four-dimensional variational (4D-Var) analysis (Talagrand and Courtier 1987; Ghil and Malanotte-Rizzoli 1991) is used with background error covariances formed from the ensemble forecast at t_0. According to the nomenclature recommendations for hybrid ensemble–variational methods (point 7 of Lorenc 2013), we refer to this approach as En-4D-Var. Here, the estimation and minimization processes are carried out in ensemble space, meaning that the control variable is expressed in terms of the ensemble members with corresponding ensemble coefficients, similarly to the iterative ensemble Kalman smoother (IEnKS) (Bocquet and Sakov 2013, 2014; Bocquet 2016). However, the minimization could be performed as well with the adjoint models if available.

In this approach, the ensemble coefficient $\mathbf{w}$ parameterizes the ensemble space spanned by the perturbations $\mathbf{X}_0$ at t_0, so that $\mathbf{x}_0 = \overline{\mathbf{x}}_0 + \mathbf{X}_0\mathbf{w}$ at t_0, with $\overline{\mathbf{x}}_0$ being the ensemble mean at t_0 and, in computing $p(\mathbf{y}_{K:1}|\mathbf{y}_{0:})$, we marginalize over $\mathbf{w}$,

$$p(\mathbf{y}_{K:1}|\mathbf{y}_{0:}) = \int \mathrm{d}\mathbf{w}\, p(\mathbf{y}_{K:1}|\mathbf{w}) p(\mathbf{w}|\mathbf{y}_{0:})\,. \tag{24}$$

From the theory of the IEnKS written in ensemble space (Bocquet and Sakov 2014), we get the first factor in the integrand of Eq. (24):

$$\ln p(\mathbf{y}_{K:1}|\mathbf{w}) = -\frac{1}{2}\sum_{k=1}^{K} \|\mathbf{y}_k - \mathcal{H}_k \circ \mathcal{M}_{k:0}(\overline{\mathbf{x}}_0 + \mathbf{X}_0\mathbf{w})\|^2_{\mathbf{R}_k} - \frac{Kd}{2}\ln(2\pi) - \frac{1}{2}\sum_{k=1}^{K}\ln|\mathbf{R}_k|\,, \tag{25}$$

with the $\circ$ symbol representing the composition of operators. Fixing the gauge in $\mathbf{w}$, *i.e.* accounting for the redundant degrees of freedom in it, yields the second factor:

$$\ln p(\mathbf{w}|\mathbf{y}_{0:}) = -\frac{1}{2}\|\mathbf{w}\|^2 - \frac{N}{2}\ln(2\pi)\,. \tag{26}$$

The sum of these two log-likelihoods yields the IEnKS cost function

$$\mathcal{J}(\mathbf{w}) = \ln p(\mathbf{y}_{K:1}|\mathbf{w}) + \ln p(\mathbf{w}|\mathbf{y}_{0:})\,. \tag{27}$$

This outcome of the IEnKS variational analysis provides, without using the explicit adjoint of the model, the argument of the minimum $\mathbf{w}^\star$, along with the associated state vector $\mathbf{x}_0^\star$ and approximate Hessian $\mathbf{I}_N + \sum_{k=1}^{K}(\mathbf{Y}_k^\star)^{\mathrm{T}}\mathbf{R}_k^{-1}\mathbf{Y}_k^\star$ of the cost function (*i.e.* the covariance) that are required in the Laplace approximation; here $\mathbf{Y}_k^\star = [\mathcal{H}_k \circ \mathcal{M}_{k:0}]'_{\mathbf{x}_0^\star}\mathbf{X}_0$, and $[\mathcal{H}_k \circ \mathcal{M}_{k:0}]'_{\mathbf{x}_0^\star}$ is the linearization of the nonlinear operator in the square

© 2016 Royal Meteorological Society

brackets at t_0. Hence, we obtain the approximation

$$\begin{aligned}\ln p(\mathbf{y}_{K:1}|\mathbf{y}_{0:}) \simeq & \mathcal{J}(\mathbf{w}^\star) + \frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\partial^2\mathcal{J}_{|\mathbf{w}^\star}\right| \\ \simeq & -\frac{1}{2}\sum_{k=1}^{K}\left\|\mathbf{y}_k - \mathcal{H}_k\circ\mathcal{M}_{k:0}(\mathbf{x}_0^\star)\right\|^2_{\mathbf{R}_k} \\ & -\frac{1}{2}\left\|\mathbf{w}^\star\right\|^2 - \frac{Kd}{2}\ln(2\pi) - \frac{1}{2}\sum_{k=1}^{K}\ln|\mathbf{R}_k| \\ & -\frac{1}{2}\ln\left|\mathbf{I}_N + \sum_{k=1}^{K}\left(\mathbf{Y}_k^\star\right)^{\mathrm{T}}\mathbf{R}_k^{-1}\mathbf{Y}_k^\star\right| \end{aligned} \tag{28}$$

or, exponentiating the log-likelihood,

$$p(\mathbf{y}_{K:1}|\mathbf{y}_{0:}) \simeq \frac{\exp\left(-\frac{1}{2}\sum_{k=1}^{K}\left\|\mathbf{y}_k - \mathcal{H}_k\circ\mathcal{M}_{k:0}(\mathbf{x}_0^\star)\right\|^2_{\mathbf{R}_k} - \frac{1}{2}\left\|\mathbf{w}^\star\right\|^2\right)}{\sqrt{(2\pi)^{Kd}\prod_{k=1}^{K}|\mathbf{R}_k|\left|\mathbf{I}_N + \sum_{k=1}^{K}\left(\mathbf{Y}_k^\star\right)^{\mathrm{T}}\mathbf{R}_k^{-1}\mathbf{Y}_k^\star\right|}}\,. \tag{29}$$

Note that, in computing the CME with the En-4D-Var, Eq. (29), the initial conditions at the beginning of the evidencing window are changed at each iteration of the minimization process, and the innovations at each observation time $\{t_k : k = 1,\ldots,K\}$ are recomputed based on the corresponding trajectory started with the new data.

Here, the IEnKS is merely a convenient means to the solution of En-4D-Var; it is not representative of what the IEnKS has to offer, as will be seen in the following.

Iterative ensemble Kalman smoother (IEnKS).

The IEnKS (Bocquet and Sakov 2014) allows the observations to be assimilated sequentially, one time t_k after another, rather than all of them together as in the En-4D-Var. The initial condition and the ensemble of anomalies at t_0 are sequentially updated by assimilating $\mathbf{y}_1$ in the first step, then $\mathbf{y}_2$ in the second step and so on until $\mathbf{y}_K$. The outcomes of the analysis at step k are the state $\mathbf{x}_k^\star$ and the normalized anomaly matrix $\mathbf{X}_k^\star$, both defined at t_0, and both then used in the subsequent step. This procedure corresponds to the *quasi-static* IEnKS, as advocated in Bocquet and Sakov (2014), but it will be simply referred to as IEnKS hereafter. The IEnKS makes possible to implement Eq. (14), and each single contextual evidence $p(\mathbf{y}_k|\mathbf{y}_{k-1:})$ corresponds to the k-th analysis of the IEnKS.

At step k, the ensemble coefficient vector $\mathbf{w}_k$ is now used in the computation of $p(\mathbf{y}_k|\mathbf{y}_{k-1:})$ to parameterize the ensemble space spanned by the anomalies $\mathbf{x}_0 = \mathbf{x}_{k-1}^\star + \mathbf{X}_{k-1}^\star\mathbf{w}_k$ at t_0, with the definition $\mathbf{x}_0^\star \equiv \overline{\mathbf{x}}_0$ and $\mathbf{X}_0^\star \equiv \mathbf{X}_0$. Marginalizing over $\mathbf{w}_k$, we get

$$p(\mathbf{y}_k|\mathbf{y}_{k-1:}) = \int \mathrm{d}\mathbf{w}_k\, p(\mathbf{y}_k|\mathbf{w}_k)p(\mathbf{w}_k|\mathbf{y}_{k-1:})\,. \tag{30}$$

The theory of the IEnKS written in ensemble space yields for $k = 1,\ldots,K$:

$$\begin{aligned}\ln p(\mathbf{y}_k|\mathbf{w}_k) = & -\frac{1}{2}\left\|\mathbf{y}_k - \mathcal{H}_k\circ\mathcal{M}_{k:0}(\mathbf{x}_{k-1}^\star + \mathbf{X}_{k-1}^\star\mathbf{w}_k)\right\|^2_{\mathbf{R}_k} \\ & -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{R}_k|\,,\end{aligned} \tag{31}$$

and, fixing the gauge in $\mathbf{w}_k$,

$$\ln p(\mathbf{w}_k|\mathbf{y}_{k-1:}) = -\frac{1}{2}\left\|\mathbf{w}_k\right\|^2 - \frac{N}{2}\ln(2\pi)\,. \tag{32}$$

The sum of these two log-likelihoods yields the IEnKS cost function, as in Eq. (27). Then the integral in Eq. (30) can be estimated by the Laplace method. We take advantage of the outcome of the IEnKS variational analysis to use the argument $\mathbf{w}_k^\star$ of the minimum and the approximate Hessian $\mathbf{I}_N + (\mathbf{Y}_k^\star)^{\mathrm{T}}\mathbf{R}_k^{-1}\mathbf{Y}_k^\star$ of the cost function required in the Laplace approximation, where $\mathbf{Y}_k^\star = [\mathcal{H}_k\circ\mathcal{M}_{k:0}]'_{\mathbf{x}_k^\star}\mathbf{X}_{k-1}^\star$. One thus obtains the approximation: for $k = 1,\ldots,K$,

$$\begin{aligned}\ln p(\mathbf{y}_k|\mathbf{y}_{k-1:}) \simeq & -\frac{1}{2}\left\|\mathbf{y}_k - \mathcal{H}_k\circ\mathcal{M}_{k:0}(\mathbf{x}_k^\star)\right\|^2_{\mathbf{R}_k} - \frac{1}{2}\left\|\mathbf{w}_k^\star\right\|^2 \\ & -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{R}_k| \\ & -\frac{1}{2}\ln\left|\mathbf{I}_N + \left(\mathbf{Y}_k^\star\right)^{\mathrm{T}}\mathbf{R}_k^{-1}\mathbf{Y}_k^\star\right|\,.\end{aligned} \tag{33}$$

This can be re-arranged differently by using the value of $\mathbf{w}_k^\star$,

$$\begin{aligned}\mathbf{w}_k^\star = & \left\{\mathbf{I}_N + (\mathbf{Y}_k^\star)^{\mathrm{T}}\mathbf{R}_k^{-1}\mathbf{Y}_k^\star\right\}^{-1}\mathbf{R}_k^{-1}\left\{\mathbf{y}_k - \mathcal{H}_k\circ\mathcal{M}_{k:0}(\mathbf{x}_k^\star)\right\} \\ = & (\mathbf{Y}_k^\star)^{\mathrm{T}}\left\{\mathbf{R}_k + \mathbf{Y}_k^\star(\mathbf{Y}_k^\star)^{\mathrm{T}}\right\}^{-1}\left\{\mathbf{y}_k - \mathcal{H}_k\circ\mathcal{M}_{k:0}(\mathbf{x}_k^\star)\right\}\,,\end{aligned} \tag{34}$$

© 2016 Royal Meteorological Society

and gathering the first two terms to yield

$$\ln p(\mathbf{y}_k|\mathbf{y}_{k-1:}) = -\frac{1}{2}\left\|\mathbf{y}_k - \mathcal{H}_k \circ \mathcal{M}_{k:0}(\mathbf{x}_k^\star)\right\|^2_{\mathbf{R}_k+\mathbf{Y}_k^\star(\mathbf{Y}_k^\star)^{\mathrm{T}}} - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\mathbf{R}_k + \mathbf{Y}_k^\star\left(\mathbf{Y}_k^\star\right)^{\mathrm{T}}\right| . \quad (35)$$

The factorization formula in Eq. (14) reads therewith

$$p(\mathbf{y}_{K:1}|\mathbf{y}_{0:}) \simeq \prod_{k=1}^{K} \frac{\exp\left(-\frac{1}{2}\left\|\mathbf{y}_k - \mathcal{H}_k \circ \mathcal{M}_{k:0}(\mathbf{x}_k^\star)\right\|^2_{\mathbf{R}_k+\mathbf{Y}_k^\star(\mathbf{Y}_k^\star)^{\mathrm{T}}}\right)}{\sqrt{(2\pi)^d\left|\mathbf{R}_k + \mathbf{Y}_k^\star\left(\mathbf{Y}_k^\star\right)^{\mathrm{T}}\right|}} . \quad (36)$$

If the models are linear, this formula coincides with Eq. (23) in the limit in which the DA window vanishes. Equation (36) differs from Eq. (29) by the observations being assimilated sequentially rather than all in one batch.

5. Numerical results

This section presents the numerical results of the following four approaches to computing the CME

1. importance sampling (IS), Eq. (17);
2. ensemble Kalman filter (EnKF), Eq. (23);
3. ensemble 4D-Var (En-4D-Var), Eq. (29); and
4. iterative ensemble Kalman smoother (IEnKS), Eq. (36).

5.1. *Experimental Set-up*

Design and performance evaluation. We evaluate the relative performance of the four methodologies by applying them to two prototypical low-order dynamical systems often used in theoretical NWP and climate studies. Given the potential uses of CME mentioned in section 1, we wish to evaluate the accuracy and robustness of the proposed schemes when observations are consistent with the model (*i.e.* perfect model), but also when they are not, insofar as CME is intended for the purpose of model evaluation and selection. It is, therefore, particularly relevant to estimate CME when the observations are not consistent with the model.

The contextual model evidence will thus be systematically computed for two models: the *correct* and the *incorrect* one. To be consistent with the intended applications to detection and attribution (Hannart *et al.* 2016, and references therein), we use the notation $\mathcal{M}_1$ for the true model and $\mathcal{M}_0$ for the other one. This notation agrees with the use, in causality theory (Pearl 2000), of index '1' for the *factual* world and '0' for the *counterfactual* one. In both cases, the CME is estimated based on the same sequence of observations, which by definition are generated by the true model. Statistically speaking, the CME of observations is expected to be higher when using the true model than when using another one, as will be discussed further on.

In order to decide which of the four approximations performs best in each of the situations being tested, we have to compare their outcome with that of alternative computational methods that are able to evaluate the CME with high accuracy. We have therefore used a massive Monte Carlo (MC, Eq. (17)) integration with 10^6 particles and a high-degree integration method, the Gauss-Hermite quadrature (GHQ, e.g., Liu and Pierce 1994). The results are reported in subsection "Performance of the four methods" of Sect. 5.2.

Gauss-Hermite quadrature is a suitable option in the present context given that the kernel distribution, which is the posterior pdf of the underlying EnKF running with the true model (see "Models and implementation" subsection below), is assumed Gaussian although the unknown actual posterior distribution may well not be so. An analytic approximation of the integral can be obtained as a function of the roots of the corresponding Gauss-Hermite polynomials (see Appendix for details). Both the accuracy of the method and the associated computational cost increase with the degree of the Gauss-Hermite polynomials. Similarly to the kernel for GHQ, the ensemble-based Gaussian posterior pdf provided by the EnKF at the beginning of the evidencing window is used as the importance density from which the 10^6 members of the MC are randomly drawn.

Models and implementation. The experimental setup is chosen to mimic the situation encountered in an operational NWP or climate prediction center, in which DA is routinely carried out to update the model state based on observations. We suppose that an EnKF is used for this purpose and that it assimilates noisy observations whose error is assumed to be an unbiased random Gaussian noise, so that $\boldsymbol{\epsilon}_k$ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{R}_k)$.

© 2016 Royal Meteorological Society

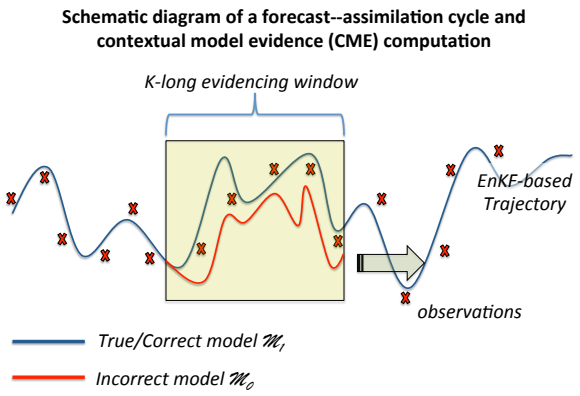

Figure 1. Schematic diagram of a routine forecast–assimilation cycle that includes contextual model evidence (CME) computation run alongside the EnKF-based trajectory of the correct model $\mathcal{M}_1$ (blue) and the incorrect one $\mathcal{M}_0$ (red).

We adopt a standard identical-twin experiment configuration (Bengtsson *et al.* 1981), in which a solution of the correct model $\mathcal{M}_1$ is taken to represent the truth and is sampled to generate synthetic observations. The control trajectory, into which the observations are assimilated, is generated by randomly perturbing the initial condition of the true trajectory. This control trajectory then evolves using the correct model, and the synthetic observations are assimilated into it using the EnKF with the setup described below.

In parallel, we carry out a sequence of model evidence diagnosis, at each observation time t_k and over an evidencing window comprising K observation vectors. The CME is then computed for the correct and incorrect models, $\mathcal{M}_1$ and $\mathcal{M}_0$, using the four methods. A schematic illustration of the experimental setup is given in Fig. 1.

Experiments are conducted using two low-order nonlinear chaotic models widely used in the predictability and DA literature: (i) the Lorenz 3-variable convection model (L63: Lorenz 1963); and (ii) the Lorenz 40-variable mid-latitude atmospheric dynamics model (L95: Lorenz and Emanuel 1998).

The original L63 model is modified to include an additional time-constant forcing, cf. Palmer (1999), as follows:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \sigma(y-x) + \lambda_i \cos\theta\,, \tag{37}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \rho x - y - xz + \lambda_i \sin\theta\,, \qquad i = 0, 1 \tag{38}$$

$$\frac{\mathrm{d}z}{\mathrm{d}t} = xy - \beta z\,. \tag{39}$$

The canonical values $(\sigma, \rho, \beta) = (10, 28, 8/3)$ for the standard coefficients are used, while the parameter λ modulates the strength of the external forcing; it is chosen to be $\lambda_1 = 0$ for the correct model, and $-8 \le \lambda_0 \ne 0 \le 8$ for the incorrect one. The coefficient θ gives the angle of the forcing and is set to $\theta = 7\pi/9$ as in Palmer (1999).

The equations of L95 read (Lorenz and Emanuel 1998): for $j = 1, \ldots, M$,

$$\frac{\mathrm{d}x_j}{\mathrm{d}t} = x_{j-1}\left(x_{j+1} - x_{j-2}\right) - x_j + F_i\,, \qquad i = 0, 1 \tag{40}$$

with $M = 40$ and periodic boundary conditions, $x_0 = x_M$, $x_{-1} = x_{M-1}$ and $x_{M+1} = x_1$. The standard value $F_1 = 8$ is used for the correct configuration, while $5 \le F_0 \ne 8 \le 11$ for the incorrect one. For both models, L63 and L95, the range of values chosen for the forcing in the incorrect configuration is such that the overall asymptotic stability properties of the system do not differ substantially from the true configuration. In all the situations we have examined, the true and perturbed models are both chaotic, although with a different spectrum of Lyapunov exponents.

The EnKF is used to assimilate the observations into the basic trajectory of the true model, shown in blue in Fig. 1. We use an ensemble square-root Kalman filter implementation, the ensemble transform Kalman filter (ETKF: Hunt *et al.* 2007). The numerical setup for the models and the EnKF is as follows.

- Number of ensemble members: $N = 4$ for L63; $N = 20$ for L95.
- Observation distribution: both models are fully observed, *i.e.* $d = 3$ for L63 and $d = 40$ for L95. The time interval between updated is $t_{k+1} - t_k = 0.10$ for L63 and $t_{k+1} - t_k = 0.05$ for L95.
- Observation error: unbiased Gaussian white noise with covariance $\mathbf{R} = \sigma^2_{\mathrm{obs}}\mathbf{I}_d$ with $\sigma_{\mathrm{obs}} = 2$ for L63, and $\sigma_{\mathrm{obs}} = 1$ for L95.
- Forecast error covariance inflation factor: $\mathbf{P}^{\mathrm{f}} \to \alpha^2\mathbf{P}^{\mathrm{f}}$, with $\alpha = 1.03$ for both models.

The experiments are ran after a 2,000 time step–long spin-up that is not taken into account in computing the statistics. The CME values are computed over K-long evidencing windows,

© 2016 Royal Meteorological Society

starting at each observation time t_k, and using the four methods under comparison: IS, EnKF, En-4D-Var and IEnKS, for both the "factual" scenario, with $\mathcal{M}_1$, and the "counterfactual" one, with $\mathcal{M}_0$ (see Fig. 1). In both cases, the same set of observations, sampled from the underlying true evolution, are used. A consequence of this latter assumption is that the correct model CME is generally larger than the incorrect one: observations are more likely to belong to the true rather than to the perturbed world. In the following, we will compute and show the logarithm of the CME, *i.e.* the logarithm of Eq. (16), but will indistinguishably refer to it as CME.

5.2. *Comparing the methods and CME values*

Performance of the four methods. In this subsection, we are interested in the accuracy of the four methods. To this end, they are compared here with MC and GHQ (see Sect. 5.1). The experiments last 200 time steps (after the 2,000 long spin-up), the evidencing window is $K = 10$, and the CMEs are computed at each time step. The forcing strength in the incorrect model is $\lambda_0 = 8$ and $F_0 = 11$ for L63 and L95, respectively.

To reduce the computational burden, GHQ is used only for L63 but with polynomial degrees as high as 32, while MC is used for both L63 and L95. MC and GHQ will not necessarily return the same estimate of the actual CME and, when necessary, we will assume GHQ to be the most accurate. To clarify this issue, we have computed the MC estimates as a function of N in the range $[10^2, 10^6]$, along with the best-fit to a power law of the form, $y(x) = a + bx^c$. This allows to extrapolate the asymptotic limiting value for $N \to \infty$, MC$^\infty$, by taking the limit $y^\infty = \lim_{x\to\infty} y(x)$. The results of this analysis are reported in Table 1. The root-mean-square-error (RMSE) of the best-fit is $O(10^{-2})$ for the correct L63 and $O(10^{-1})$ for both correct and incorrect L95. The incorrect L63 model is unsurprisingly the most intricate to treat and the RMSE is $O(10^0)$. In all cases the best-fit coefficient c is negative, so that the asymptotic values, MC$^\infty$, are retrieved from the best-fit coefficient a. The extrapolated asymptotic values can be compared with the estimates given by GHQ and MC (with $N = 10^6$) for L63, and with MC ($N = 10^6$) for L95. We see that for the L63 correct case, MC$^\infty$, MC (N=10^6) and GHQ all coincide and are thus indistinguishably good estimates of the target CME value to compare with the DA-based estimates. The numerical results for L63 incorrect model suggest that convergence has not yet been reached, as it is also reflected by the lower accuracy of its best-fit. The estimated limiting value, MC$^\infty$, is slightly closer to GHQ than MC with $N = 10^6$, and GHQ is taken as the reference here. For the L95, both correct and incorrect models, the asymptotic estimates, MC$^\infty$, are very close to those for $N = 10^6$, particularly for the correct model case as expected. We are thus confident that, in this case, MC with $N = 10^6$ provides reasonably good estimates of the actual CME to compare with the DA-based computations.

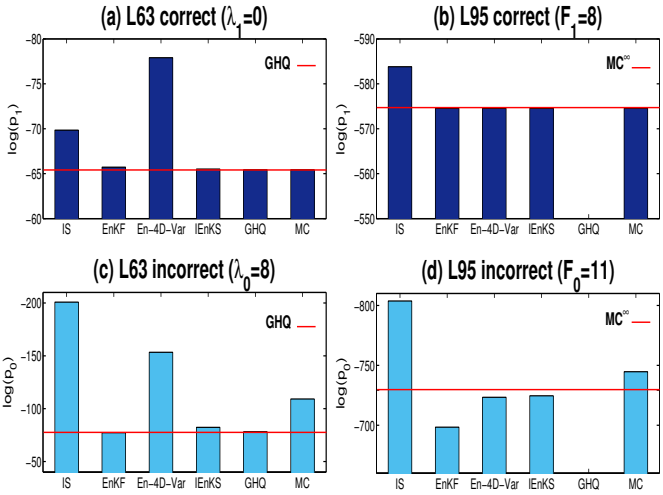

Figure 2. Average CME value for the correct and incorrect model, *i.e.* $\log(p_1)$ vs. $\log(p_0)$, estimated using our four methods — IS, EnKF, En-4D-Var and IEnKS — as well as the two high-accuracy validation methods, Monte Carlo (MC) with 10^6 particles, and Gauss-Hermite quadrature (GHQ) with degree 32. The horizontal line depicts the reference value given by the GHQ or MC$^\infty$ (see Table 1).

Results for the mean CME using the four methods along with GHQ and MC ($N = 10^6$) are displayed in Fig. 2; the red horizontal line depicts the target actual CME value.

For the L63 correct model, the EnKF and IEnKS are remarkably providing almost the same estimate, the closest to the GHQ target, followed by IS. In contrast, En-4D-Var is the farthest from the target, a result that is possibly related to the effect of a non-quadratic cost function and the resulting presence of multiple minima that makes difficult the convergence to the global minimum of the cost function. The problem experienced by the En-4D-Var smoother seems to be successfully overcome by the IEnKS, by virtue of the quasi-static formulation adopted in this study (see Pires *et al.* 1996; Bocquet and Sakov 2014, for a definition and discussion on the quasi-static approach).

© 2016 Royal Meteorological Society

Table 1. Average correct ($\log p_1$) and incorrect ($\log p_0$) CME estimated using Gaussian Hermite Quadrature (GHQ) with degree 32 (see text and Appendix for details) and Monte Carlo with 10^6 particles (MC). The RMSE of the best-fit to the power law, $y(x) = a + bx^c$, for the numerical values of MC with $10^2 \leq N \leq 10^6$ are given in the 5^{th} column. The extrapolated asymptotic value for $N \to \infty$ is given in the 6^{th} column.

		GHQ	MC (N=10^6)	RMSE of best-fit to $y = a + bx^c$	MC$^\infty$
L63	Correct Model $\mathcal{M}_1$	-65.44	-65.44	0.004	-65.44
	Incorrect Model $\mathcal{M}_0$	-78.19	-109.19	2.22	-102.25
L95	Correct Model $\mathcal{M}_1$	–	-574.57	0.411	-574.63
	Incorrect Model $\mathcal{M}_0$	–	-744.68	0.440	-729.25

MC is taken as the reference for the L95 correct case, and all four methods being evaluated, except IS, converge to the target with a similar high level of accuracy. When observations are dense and frequent enough, and when the model forecast is only weakly nonlinear in between two observation times, the errors are nearly Gaussian and the three DA methods track the true signal successfully; as a result, their CME estimates almost coincide.

Estimating the CME for the incorrect models is more intricate and, in fact, MC and GHQ no longer provide identical results for the L63 model (see also Table 1). As explained above, GHQ is used here as the reference target value. In contrast to the true-model case, IS is now the least accurate, EnKF and IEnKS are the best, while En-4D-Var is in between the two.

When comparing to the incorrect-model CME for the L95 model, we see that the En-4D-Var and IEnKS methods provide similarly accurate results, followed by the EnKF, while IS is not able to converge to the target given here by the MC. Together, the results for the true- and the incorrect-model CME in the L95 model suggest that, at least in this weakly nonlinear regime, the accuracy of the DA-based estimates of the CME is connected to the level of sophistication of the DA method, with the two smoothers, En-4D-Var and IEnKS, performing better than the filter (EnKF).

Time series of CME values. Figure 3 shows 1000-time steps long time series of the instantaneous CME values, for the correct and incorrect models, over the time interval 100–1100 (after the spin up) for the L63 model (left panels) and the L95 one (right panels). The parameter values appear in the figure caption, and the evidencing window is $K = 10$ in all experiments.

As expected, the CME values in the incorrect model are algebraically smaller than in the correct one, for all four methods,

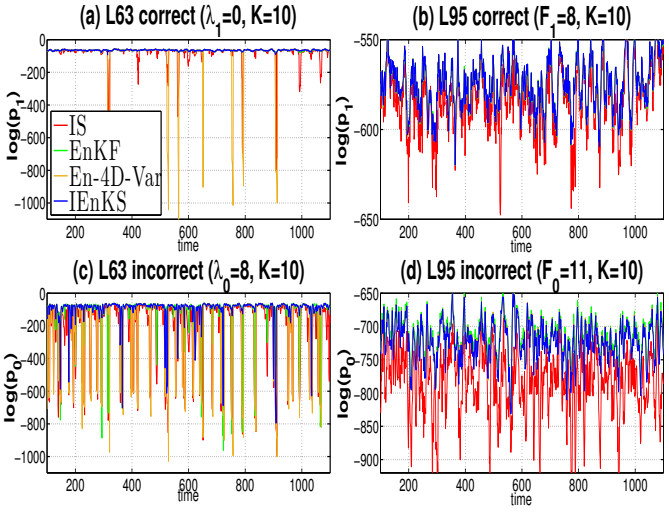

Figure 3. Time series of 1000 instantaneous CME values starting from $t = 100$, for the correct (top panels) and incorrect (bottom panels) configurations of the L63 model (left panels) and L95 model (right panels), respectively; the evidencing window has length $K = 10$ for all four panels. The values of the forcing in the correct vs. the incorrect models are as in Fig. 2; see also panel legends. The CME is computed using the four methods: IS, EnKF, En-4D-Var and IEnKS.

and especially so in the L95 model; see Fig. 3(d). The behavior of the CME time series is, moreover, very different between the L63 and L95 model. In the L63 model, the CME values are predominantly quite small in absolute value, but large on-off bursts are observed, in which the CME jumps to algebraically very low values, *i.e.* to quite large absolute values. These spikes are also observed in the correct model, but they get larger and more frequent in the incorrect one. Moreover, these spikes are mainly found in the IS and En-4D-Var methods, with only some occasional instances in the EnKF, and to an even lesser extent in the IEnKS.

The CME time series of the correct L95 model in panel (b) are almost never as small in absolute value as for the L63 model in panel (a), while three of the four methods to compute CME provide similar results, and IS alone is both negatively biased overall and exhibits strong negative spikes as well. The CME values for the incorrect L95 model in panel (d) are shifted

© 2016 Royal Meteorological Society

downward by about 100 logits with respect to the correct model for the estimates provided by the EnKF, En-4D-Var and IEnKS methods, while the IS departures are even more substantial.

The differences in the numerical results for CME estimation with the L63 and L95 models using DA-based methods can be interpreted by the two models' dynamical and statistical properties. The L63 model has a strange attractor which is bimodal. As shown already by Miller *et al.* (1994) in comparing an EKF with a 4D-Var method, these features can easily lead to situations in which the model solution and the observations are on a different "wing of the butterfly," *i.e.* on a different lobe of the attractor, thus making already the state estimation particularly challenging.

Such situations are even more deleterious when using DA to evaluate model evidence and when the incorrect model is under study. Carrassi *et al.* (2008) and Carrassi and Vannitsem (2010) provide further examples of applying DA to both the L63 and L95 model, and discuss some of the implications of their dynamical features on DA results. The fine, onion-skin structure of the L63 attractor is most likely to be responsible for the large spikes in the CME values obtained by the En-4D-Var method: multiple minima in the cost function arise, and their number increases as the window length K increases, cf. Figs. 6a–c in Miller *et al.* (1994, and discussion therein). This inference is confirmed by some of the numerical experiments below; see Fig. 6 and related discussion.

The use of the quasi-static approximation in the IEnKS, in which one observation vector is sequentially added at each step of the minimization, helps track the global minimum of the cost function and prevents getting trapped in a secondary minimum that would give a strongly biased estimate of the CME. This may explain the relatively smoother profile of the IEnKS-based CME estimates.

Distribution of CME values. The distributions of the CME values are shown in Figs. 4 and 5 for the L63 and L95 models, respectively. The statistics are computed based on 4,000 time steps long experiments, started after the spin-up. The same correct and incorrect forcing values as before were used in panels (a) and (b), respectively, while the corresponding zoomed areas of the pdfs are plotted in panels (c) and (d), respectively. Note that to improve

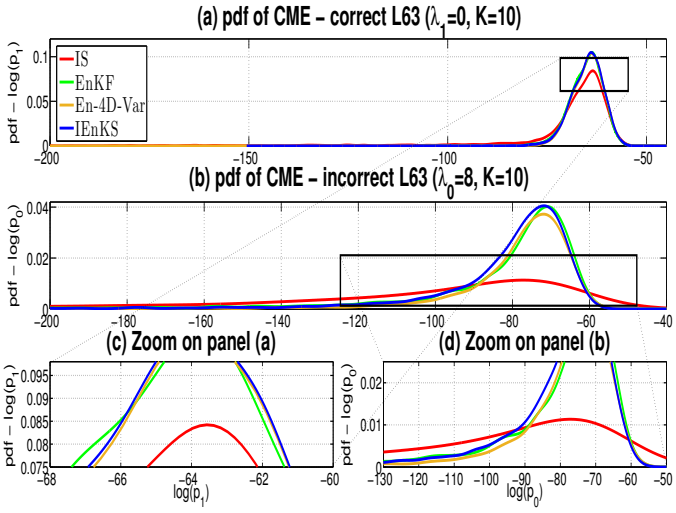

Figure 4. Same as Fig. 2 for the probability density function (pdf) of the CME for the L63 model, and the four methods — IS, EnKF, IEnKS and En-4D-Var — only. (a,b) Full pdf; and (c,d) zoom on the central part of the pdf.

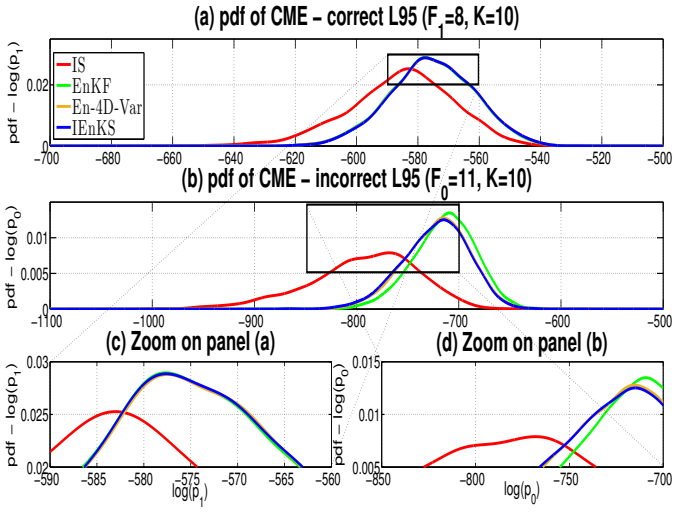

Figure 5. Same as Fig. 4 but for the L95 model.

visualization of the main features of the pdfs we have intentionally limited the displayed range for the L63 to $[-200, -40]$. As a consequence the pdfs in Fig. 4 do not reflect the large, occasional, peaks observed in Fig. 3a,c.

In agreement with Fig. 3, the three Gaussian DA-based CME estimates are quite similar to each other in the factual case and differ from the IS estimate, which shows a negative bias in both models, particularly in L95, cf Fig. 5(a). A closer inspection of the zoomed area for the L63 model in Fig. 4(c) reveals that the EnKF pdf is slightly shifted toward smaller values, whereas the En-4D-Var and IEnKS estimates are still very close to each other. The similarities among the three DA-based estimates are visible when looking at the zoomed area for the L95 model in Fig. 5(c).

Differences among IS, EnKF, En-4D-Var and IEnKS estimates are, however, apparent in the CME estimates with the incorrect

© 2016 Royal Meteorological Society

model. All four distributions are now displaced toward smaller values than in the corresponding correct model cases — since the observations are more likely in the true world — and other discrepancies appear, too. It is thus remarkable that the IS pdfs, for both the L63 and L95 models, show now even stronger departures from the other three distributions. The IS method here does not use any resampling to keep the members on the track of the true signal; this can lead to unrealistic underestimations of the CME, as noted already in Fig. 3.

When looking more attentively at the pdfs for the three DA-based approaches, additional details emerge. In the L63 model, the IEnKS estimates differ from both the EnKF and En-4D-Var, in particular its pdf is slightly larger toward smaller CME values, cf. Fig. 4(d), although the modes of the three distributions are almost indistinguishable, cf. Fig. 4(b). The situation is somewhat different in the L95 model, where the En-4D-Var and the IEnKS pdfs are very close when using the incorrect model, while the EnKF pdf peaks at a slightly larger CME value.

Sensitivity analysis of the results. This section describes the numerical sensitivity analysis of the CME computational methods with respect to the discrepancy between correct and incorrect forcing, $\Delta\lambda$ or ΔF, and the length of the evidencing window, K. We wish to understand how the different methods respond to the effect that these factors have on the CME integral. The comparison among the DA-based methods is guided by the following predictions. First, the CME gets smaller for increasing $\Delta\lambda$ or ΔF. In fact, if $\Delta\lambda$ or ΔF is seen as a measure of the difference between the correct and incorrect models, the marginal likelihood of observations of the latter model will decrease along with the difference between the two. Recall that "correct model" is intended here as the one used to generate the observations so that, by construction, the CME values must be higher in this case. The sensitivity analysis to $\Delta\lambda$ or ΔF aims to assess the extent to which this is the case in the DA-based approximations of the CME and how this compares among the four approaches. Second, the CME gets smaller for increasing K. In fact, the longer the evidencing window the longer the time at disposal for the model to manifest its deviations from the observations, and the more statistically reliable the CME should be. We do not have in this case a guess for the unknown actual growth rate, but we expect

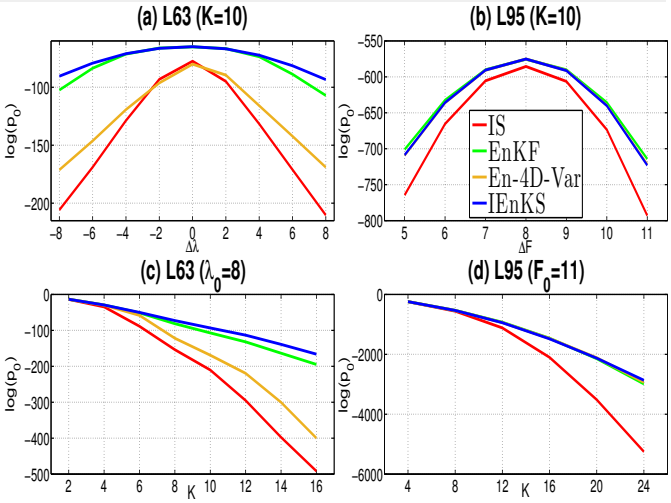

Figure 6. Mean CME as a function (a,b) of the difference between the incorrect and correct model forcing (a) $\Delta\lambda = \lambda_0 - \lambda_1$ and (b) $\Delta F = F_0 - F_1$, and (c,d) of the length of the evidencing window. Results for (a,c) the L63 model and (b,d) the L95 model. In the top panels (a,b) the evidencing window is $K = 10$, while in the bottom panels the forcing parameter equals the incorrect value: (c) $\lambda_0 = 8$ for L63, and (d) $F_0 = 11$ for L95.

that the better the data assimilation is able to keep the incorrect model close to the observations, the bigger the CME values.

The way the four methods respond to the amplitude of the discrepancy between the correct and the incorrect forcing, and to the length K of the evidencing window, is depicted in Fig. 6.

In panels (a) and (b), the evidencing window is kept fixed at $K = 10$, while the difference in the forcing is plotted on the abscissa. It is striking that, roughly speaking, the CME values for all four methods are inversely proportional to the difference, in absolute value, between the two forcings (see Fig. 6(a,b)). Nevertheless, the rate of change of the CME values varies from one method to another, and the differences between methods are noticeably larger in the case of the L63 model in panel (a). Of the four, it is IS that falls off most strongly as the forcing value deviates more from the correct one, in both panels. The IS method is thus the least accurate at a given forcing, but also the most sensitive one to the deviation of the forcing from the correct value.

The En-4D-Var method also exhibits a marked sensitivity to forcing value, although less so than the IS method. Interestingly, the En-4D-Var estimate of CME in the true case, *i.e.* for $\lambda_1 = 0$, is very close to IS, and increasingly closer to EnKF and IEnKS than IS as we move away from the correct forcing value for the L63 model, cf. panel (a). The EnKF and IEnKS estimates peak at a slightly higher value for the true CME and show lesser dependence on the forcing difference between the two models. In the L95 model, cf. panel (b), the three DA-based computational

© 2016 Royal Meteorological Society

techniques — EnKF, En-4D-Var and IEnKS — give a similar response to the difference in forcing, while IS is negatively biased with respect to all three and shows a larger dependency on the forcing difference between the correct and incorrect models.

The impact of the evidencing window length, K, is analyzed in panels (c) and (d), for the L63 model and the L95 one, respectively. Here the incorrect model forcing is kept fixed at $\lambda_0 = 8$ for L63 and at $F_0 = 11$ for L95. The increase of the evidencing window length facilitates the discrimination between the correct and the incorrect models, since the two dynamics have a longer time interval each to manifest their differences, but it also makes the computation of the integral in Eq. (16) more delicate, calling therewith for more sophisticated integration methods.

The L63 model highlights this quandary better, due to the model's greater simplicity but also bimodal character, cf. panel (c). In it, we see how IS changes rapidly by increasing K, but En-4D-Var is also very sensitive. While the correctness of the En-4D-Var estimate cannot be assessed, the large deviation from the other two DA-based methods points to its lower accuracy. Again, the gap between the En-4D-Var estimates and those of EnKF and IEnKS may well be due to the emergence of multiple minima of the cost function as K increases. This is corroborated by the absence of this behavior in the "more unimodal" L95 model, cf. panel (d), where the three DA-based approaches give indistinguishable estimates.

The numerical results so far pertain to the comparison of the four methods of evaluating the CME integral in Eq. (16). One can presume that their accuracy depends strongly on the accuracy of the DA-based scheme on which they rely, leaving IS, which suffers from under-sampling and lack of resampling, as the least accurate option. Nevertheless, trying to achieve an overall ranking of DA schemes may be difficult, since different methods may be optimal for specific tasks and under specific circumstances. Such differences in method performance for model evidence are illustrated here by application to the L63 model vs. the L95 one, with the former exhibiting a bimodal attractor.

6. Applications

We briefly discuss now two possible applications of our proposed schemes; both are straightforward to describe and implement based on the results obtained in section 5.

6.1. *Parameter estimation*

Let us assume now that the true values of the forcing parameters, *i.e.* λ_1 in L63 and F_1 in L95, are unknown, but that we have access to a time series of observations for each one of the two models. Many methods for parameter estimation are available (e.g., Ghil 1997; Kondrashov *et al.* 2008; Kantas *et al.* 2009; Carrassi and Vannitsem 2011) but the present maximum likelihood estimation approach is relatively new. A recent review on parameter estimation for the geosciences can be found in Bocquet (2015).

By maximizing the CME for the parameter of interest, we can use the observations on the model state to evaluate the unknown forcing. In addition, we will apply the standard likelihood ratio approach to derive confidence intervals on the forcing estimates obtained by our CME approach.

The results of this approach to parameter estimation are also available in Figs. 6(a,b). With $K = 10$, the four methods yield an unbiased estimate of the forcing parameter in both the L63 (panel (a)) and the L95 (panel (b)) model, since the CME curves in both panels reach their maxima at the correct parameter value.

The methods differ, however, in the curvature of each of the CME curves at their respective maxima, and thus on the corresponding confidence intervals. More precisely, the less accurate methods yield a higher curvature than the most accurate ones, and thus lower estimates of uncertainty. Indeed, the variance of a maximum likelihood estimator is driven by the inverse of the second derivative of the negative log likelihood, asymptotically (e.g., Millar 2011). Accordingly, less accurate methods appear to still be able to yield unbiased parameter estimates, but to underestimate uncertainty. Furthermore, increasing the length K of the evidencing window in Figs. 6(c,d) does not affect the position of the maxima and so the estimates remain unbiased, but it appears to increase the gap in CME estimates for parameter values that differ from the truth. Thus increasing K reduces the uncertainty estimates, as expected.

© 2016 Royal Meteorological Society

6.2. *Causal attribution of climate-related events*

Providing causal assessments on episodes of extreme weather or unusual climate conditions has become an important topic in the climate sciences over the last decade (Solomon *et al.* 2007; Stocker *et al.* 2013). Its importance arises from the multiple needs for public dissemination, litigation in a legal context, adaptation to climate change or simply improvement of the science associated with these events (Stott *et al.* 2013).

The conventional approach to event attribution so far consists in comparing two probabilities, p_1 and p_0, and in computing the so-called *fraction of attributable risk* or $F_{\rm AR}$ of the event under study, where $F_{\rm AR} = 1 - p_0/p_1$. Here p_1 is the probability of occurrence of the event in a model $\mathcal{M}_1$ representing the observed climatic conditions, which simulates the real world, referred to as *factual*, while p_0 is the probability of occurrence of the event in a second model $\mathcal{M}_0$ that represents this time the alternative world that might have occurred had the forcing of interest been absent, referred to as *counterfactual*. In the conventional approach, p_1 and p_0 are calculated by running an ensemble of simulations of each one of the two models, $\mathcal{M}_1$ and $\mathcal{M}_0$, which is computationally quite costly.

Recently, Hannart *et al.* (2016) have shown that this conventional approach can be improved upon by applying DA to derive the CME of a series of observations of the given event for these two alternative models, $\mathcal{M}_1$ and $\mathcal{M}_0$. In both cases, the CME is estimated for the same sequence of meteorological observations of the event. If the value of the factual CME substantially exceeds that of the counterfactual CME for the event under scrutiny, then it is possible to conclude that the forcing of interest has had a causal influence on the event.

Hannart *et al.* (2016) made use of the EnKF in their DA-based approach to event attribution. Their proposed approach is further investigated here by using the four methods discussed so far to compute model evidence; the effectiveness of these methods is evaluated by calculating the *discriminating power* of each of them. This metric is defined as the ability to discriminate whether the observed sequence of observations is more or less likely to occur in the factual rather than in the counterfactual world, and it is obtained as the ratio between the factual and counterfactual model evidence, p_1/p_0. Figure 7 shows the average of the logarithm $\log(p_1/p_0)$ of the discriminating power for L63 and L95, as a function of the counterfactual model's forcing and of the evidencing window length K.

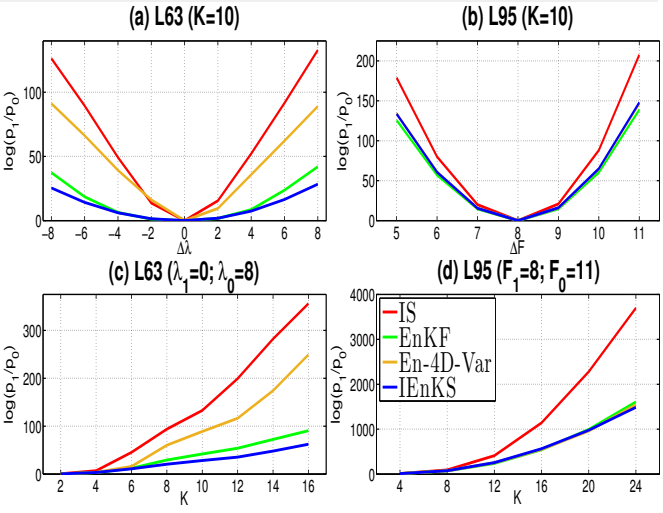

Figure 7. Same as Fig. 6 but for the mean discriminating power.

As expected, the discriminating power increases monotonically as the difference between the factual and the counterfactual forcing increases — *i.e.*, as the cause to whom the event is to be attributed is more salient — as well as with the length K of the evidencing window. Indeed, the larger K is the more time one can observe the difference between the models.

In agreement with our previous results on the CME, Figs. 7(b,d) show that the three DA-based methods provide similar results for the L95 model, while IS systematically overestimates the discriminating power. This means that, if IS is used in an attribution application, it will exaggerate the responsibility of the forcing under scrutiny. The DA-based methods offer a more plausible way to assess the main cause that might have produced the observed sequence.

The situation is slightly less clear for the L63 model, cf. panels (a) and (c), although IS still provides the largest CME estimate for any given K. But the DA-based methods are affected now differently by the increase in the evidencing window, with En-4D-Var displaying the strongest response, as already discussed in relation with Fig. 6.

© 2016 Royal Meteorological Society

7. Concluding remarks

7.1. Summary

This study focused on the problem of quantifying the resulting performance of a state inference by estimating the model evidence, *i.e.* the marginal likelihood $p(\mathbf{y}_{k:})$, which quantifies the accuracy of a model-to-data fit. Model evidence is a natural metric for selection and comparison, and it is relevant to many problems faced by both scientists and practitioners: *e.g.* calibrating model parameters, comparing the skill of several candidate models in representing the observed signal, or evidencing the existence of a causal relationship between an external forcing and an observed response. This latter attribution problem, and in particular the development of methods for attribution of climate related events in real or near-real time (Stott *et al.* 2013; Hannart *et al.* 2016), was one of the original motivations of this study.

Deriving model evidence for high-dimensional models and big datasets – both of which are common in the geosciences – is usually computationally intractable and the issue can only be solved under simplified assumptions (e.g., Hürzeleri and Künsch 2001). In this paper we have shown how this task can be carried out efficiently using data assimilation (DA) techniques specifically designed to deal with large numerical models and dataset subject to partially Gaussian assumptions.

We introduced the original *contextual* formulation of the model evidence (CME), $p(\mathbf{y}_{k:1}|\mathbf{y}_{0:})$ in Eq. (16). As opposed to the standard formulation, the CME is narrowed down to the present systems condition, and the conditional pdf $p(\mathbf{x}_0|\mathbf{y}_{0:})$ is taken as the prior in lieu of the invariant distribution $p(\mathbf{x})$ on the model attractor, as used in the standard case.

This new prior is not only more informative and easier to compute but it is also the outcome, *i.e.* the posterior pdf, of a DA procedure designed for state estimation based on the observations. When a forecast-assimilation cycle is routinely and continuously carried out, as it is in NWP centers, this prior is immediately at hand.

We have then proved that DA can also be used to compute the conditional pdf, $p(\mathbf{y}_{k:1}|\mathbf{x}_0)$, that enters the CME definition, cf. Eq. (16). Together with the DA-based CME prior, using this pdf also allows one to fully accomplish the model evidence estimation task in a consistent and routine way by using DA.

Analytic derivations of the proposed DA-based approach were described for both filtering and smoothing methods. In particular, we presented the Kalman filter (KF), the ensemble Kalman filter (EnKF), the ensemble four-dimensional variational smoother (En-4D-Var) and the iterative ensemble Kalman smoother (IEnKS).

The theoretical formulae for the EnKF, En-4D-Var and IEnKS, along with the importance sampling based on the EnKF, have been compared numerically using two low-order chaotic models, the Lorenz 3-variable (L63) model (Lorenz 1963) and the 40-variable (L95) model of Lorenz and Emanuel (1998). Gauss-Hermite quadrature and a massive Monte Carlo algorithm are used as independent highly accurate reference results for the CME values.

Numerical tests were used to compare the computing methods in terms of the CME time series and their distributions, as well as with respect to their sensitivity to the model forcing and to the length of the evidencing window. In general the accuracy of the estimates increased with the sophistication of the DA method. Thus the IEnKS yielded the best results and the IS the worst ones.

In the comparison of method skill for the two models, the IEnKS also performed better for the L63 model, which possesses a bimodal attractor. In fact the quasi-static approximation employed by the IEnKS helps preventing the appearance of multiple minima in the cost function that hampers the performance of En-4D-Var. In these conditions also the EnKF appears to behave generally better than En-4D-Var, and its skill is the closest to that of the IEnKS. On the other hand, in a weakly nonlinear regime as represented here by the L95 model, the accuracy of the DA-based estimates of the CME appears connected to the level of sophistication of the DA method, with the two smoothers, En-4D-Var and IEnKS, performing better than the filter (EnKF).

We have also considered two applications of estimating the model evidence – namely to parameter estimation and to causal attribution of climate-related events – and have studied the potential of the proposed DA-based approach for these purposes. Results have shown that DA-based model evidence (i) can be efficiently used to estimate unknown model parameters, along with the associated uncertainty; and (ii) it is able to

© 2016 Royal Meteorological Society

discriminate between two competing models and, therewith, to correctly attribute an observed event to one of the two. In both of these applications the IEnKS-based CMEs provide the best performance. Nonetheless the EnKF appears as a good compromise between accuracy and ease of implementation, and its reduced computational cost compared to IEnKS makes it a suitable option for CME computation in a realistic setting.

7.2. *Future directions*

The next step in applying the present DA-based approach to more realistic models and observational scenarios is to consider climate models of intermediate complexity and incomplete and unevenly distributed observations. This application-oriented research activity has to be supported and accompanied by two theoretical lines of investigations, namely the extension of the present results (i) in the presence of model error, and (ii) in conjunction with spatial localization techniques.

Taking into account model error is crucial for obtaining an accurate and robust estimate of the marginal likelihood, and this requirement becomes even more stringent when model evidence is intended for use in model discrimination or selection. Model error can, in fact, masks the difference between two models under-scrutiny, whether one of the two is correct or not. This effect is an issue for both, standard and DA-based, methods, but particularly for the latter as the accuracy of the DA outcome will naturally depend on the model accuracy, as discussed in Hannart *et al.* (2016, Fig. 4d). Methods to incorporate model error in DA procedures for state and parameter estimation have been the subject of an intense, and still active, stream of research (e.g. Dee 1995; Harlim 2013; Raanes *et al.* 2015). The use of these methods and the study of their adaptation to DA-based CME estimation is definitely worth addressing.

The applications of the proposed DA-based approach to large dimensional, spatially extended, dynamical systems necessitate the implementation of localization strategies, commonly used in ensemble-based DA in the geosciences with the aim of compensating for the sampling errors that arise from the use of an insufficiently large ensemble; see Sakov and Bertino (2011) and references therein. The extension of the formulae described here to compute CME in conjunction with localization is a central theme among the authors' current lines of investigation.

Acknowledgments

We acknowledge discussions with the other members of the DADA team (Philippe Naveau, Aurélien Ribes, Juan Ruiz and Manuel Pulido) and with Francis Zwiers and Ted Sheperd on detection and attribution aspects. This research was supported by grant DADA from the Agence Nationale de la Recherche (ANR, France: All authors), and by Grant N00014-16-1-2073 of the Multidisciplinary University Research Initiative (MURI) of the US Office of Naval Research (MG). A. Carrassi has been funded by the Nordic Centre of Excellence EmblA of the Nordic Countries Research Council, NordForsk.

Appendix

Computation of contextual model evidence using Gauss-Hermite quadrature

All quadrature methods approximate an integral as a weighted sum of the integrand's values, evaluated at a finite set of well-specified points called nodes. Gaussian quadrature approximates the integral of an unknown function $f(x)$ over a specified domain $\mathcal{D}$ with a known weighting kernel $\psi(x)$. If the function $f(x)$ is well approximated by a polynomial of order $2m-1$, then a quadrature with m nodes suffices for a good estimate of the integral, according to

$$\int_{\mathcal{D}} \mathrm{d}x f(x)\psi(x) \approx \sum_{i=1}^{m} w_i f(x_i). \tag{A1}$$

The nodes x_i and weights w_i are uniquely determined by the choice of the domain $\mathcal{D}$ and the weighting kernel $\psi(x)$, which in turn determines the type of quadrature. The locations of the nodes $\{x_i : i = 1, \ldots, m\}$ are given by the roots of the polynomial of order m in the sequence of orthonormal polynomials $\{\pi_j(x) : j = 1, \ldots, m\}$, according to the scalar product $(\pi_j|\pi_k) = \int_{\mathcal{D}} \mathrm{d}x\, \pi_j(x)\pi_k(x) = \delta_{jk}$, and the weights are computed once the roots are known.

When the integration domain is the entire real axis and the kernel is given by a Gaussian function, the quadrature method is known as the *Gauss-Hermite* method, since it involves the

© 2016 Royal Meteorological Society

orthogonal Hermite polynomials, and it can be written as

$$\int_{\mathbb{R}} \mathrm{d}x f(x) \exp(-x^2) \simeq \sum_{i=1}^{m} w_i f(x_i)\,, \tag{A2}$$

with x_i being the roots of the Hermite polynomials $\mathcal{H}e_m$ of degree m, and the weights w_i being given by

$$w_i = \frac{2^{m-1} m! \sqrt{\pi}}{m^2 \left\{\mathcal{H}e_{m-1}(x_i)\right\}^2} \qquad i = 1, \ldots, m\,. \tag{A3}$$

Moreover, when the kernel is given by a normal distribution, $\psi = \mathcal{N}(\overline{x}, \sigma^2)$, the Gauss-Hermite approximation becomes

$$\int_{\mathbb{R}} \mathrm{d}x \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\overline{x})^2}{2\sigma^2}\right\} f(x) \simeq \frac{1}{\sqrt{\pi}} \sum_{i=1}^{m} w_i f\left(\sqrt{2}\sigma x_i + \overline{x}\right)\,. \tag{A4}$$

This formula can be obtained from Eq. (A2) after a change of variable and integration by substitution. Gauss-Hermite quadrature is of key importance in many areas of applied science, including statistics and finance, and is described in many textbooks on numerical analysis (e.g., Press *et al.* 1992).

Let us recall the general expression of the CME in Eq. (16):

$$p(\mathbf{y}_{K:1}|\mathbf{y}_{0:}) = \int \mathrm{d}\mathbf{x}_0\, p(\mathbf{y}_{K:1}|\mathbf{x}_0) p(\mathbf{x}_0|\mathbf{y}_{0:}). \tag{A5}$$

In our numerical experiments of section 5, the prior density, *i.e.* the kernel in Eq. (A5), is the multivariate Gaussian posterior obtained from the EnKF applied to the underlying factual trajectory, and it reads

$$p(\mathbf{x}_0|\mathbf{y}_{0:}) = \frac{\exp\left(-\frac{1}{2}\left\|\mathbf{x}_0 - \overline{\mathbf{x}}_0\right\|^2_{\mathbf{P}_0^{\mathrm{f}}}\right)}{\sqrt{(2\pi)^M \left|\mathbf{P}_0^{\mathrm{f}}\right|}}\,. \tag{A6}$$

Here, the forecast error covariance matrix $\mathbf{P}_0^{\mathrm{f}}$ is given in terms of the forecast perturbation matrix, $\mathbf{P}_0^{\mathrm{f}} = \mathbf{X}_0\mathbf{X}_0^{\mathrm{T}}$, as described in section 4.1.

We consider the case $N \geq M+1$, for which $\mathbf{P}_0$ is almost surely non-singular and which applies in the numerical experiments with the L63 model. The choice of a Gaussian kernel allows for the use of Gauss-Hermite quadrature, and the function $f(x)$ in Eq. (A4) is the likelihood $p(\mathbf{y}_{K:1}|\mathbf{x}_0)$. Given Gaussian observational errors with covariance $\mathbf{R}$, $f(x)$ should be replaced with

$$p(\mathbf{y}_{K:1}|\mathbf{x}_0) = \frac{\exp\left(-\frac{1}{2}\sum_{k=1}^{K}\left\|\mathbf{y}_k - \mathcal{H}_k \circ \mathcal{M}_{k:0}(\mathbf{x}_0)\right\|^2_{\mathbf{R}_k}\right)}{\sqrt{(2\pi)^{Kd}\prod_{k=1}^{K}|\mathbf{R}_k|}}. \tag{A7}$$

We now consider the multivariate extension of the Gauss-Hermite quadrature Eq. (A4). In order to make the sampling more efficient, it is convenient to use the principal axes of variability of the covariance matrix $\mathbf{P}_0^{\mathrm{f}}$. To this end, let us decompose the perturbation matrix,

$$\mathbf{X}_0 = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}\,. \tag{A8}$$

Here, $\mathbf{U}$ and $\mathbf{V}$ are the matrices whose columns are the M left and right singular vectors, and $\mathbf{S}$ is the diagonal matrix of the M singular values. We define the operator

$$\mathbf{Z} = \sqrt{2}\mathbf{U}^{\mathrm{T}}\mathbf{S}\mathbf{U}\,. \tag{A9}$$

By proceeding via integration by substitution, as for the univariate case, the multivariate Gauss-Hermite quadrature method can be shown to be given by

$$\int_{\mathbb{R}^M} \mathrm{d}\mathbf{x} \frac{1}{\sqrt{(2\pi)^M\left|\mathbf{P}_0^{\mathrm{f}}\right|}} \exp\left(-\frac{1}{2}\left\|\mathbf{x} - \overline{\mathbf{x}}\right\|^2_{\mathbf{P}_0^{\mathrm{f}}}\right) f(\mathbf{x}) \simeq \sum_{i_1=1, i_2=1, \ldots, i_M=1}^{m} \prod_{j=i_1}^{i_M} w_j f(\overline{\mathbf{x}} + \mathbf{Z}\boldsymbol{\chi}) \tag{A10}$$

with $\boldsymbol{\chi} = (x_{i_1}, x_{i_2}, ..., x_{i_M})$ being the roots of the Hermite polynomials. The final formula to approximate Eq. (A5) is obtained by using Eq. (A7) as $f(\mathbf{x})$.

References

Balmaseda MA, Alves OJ, Arribas A, Awaji T, Behringer DW, Ferry N, Fujii Y, Lee T, Rienecker M, Rosati T, Stammer D. 2009. Ocean initialization for seasonal forecasts. *Oceanography Special Issue* **22**: 154.

Baum L, Petrie T, Soules G, Weiss N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* **41**: 164–171.

Bengtsson L, Ghil M, Källén E (eds). 1981. *Dynamic meteorology: Data assimilation methods*. Springer-Verlag, New York/Heidelberg/Berlin.

Bhend J, Franke J, Folini D, Wild M, Brönnimann S. 2012. An ensemble-based approach to climate reconstructions. *Clim. Past.* **8**: 963–976.

© 2016 Royal Meteorological Society

Blayo E, Bocquet M, Cosme E, Cugliandolo LF (eds). 2015. *Advanced data assimilation for geosciences*, Oxford University Press. Lecture Notes of the Les Houches School of Physics: Special Issue, June 2012.

Bocquet M. 2015. An introduction to inverse modelling and parameter estimation for atmosphere and ocean sciences. In: *Advanced data assimilation for geosciences*, Blayo É, Bocquet M, Cosme E, Cugliandolo LF (eds). Oxford University Press: Les Houches school of physics, pp. 461–493.

Bocquet M. 2016. Localization and the iterative ensemble Kalman smoother. *Q J Roy. Meteor. Soc.* **142**: 1075–1089.

Bocquet M, Pires C, Wu L. 2010. Beyond Gaussian Statistical Modeling in Geophysical Data Assimilation. *Mon. Weather Rev.* **138**: 2997–3023.

Bocquet M, Sakov P. 2013. Joint state and parameter estimation with an iterative ensemble Kalman smoother. *Nonlinear Proc. Geoph.* **20**: 803–818.

Bocquet M, Sakov P. 2014. An iterative ensemble Kalman smoother. *Q J Roy. Meteor. Soc.* **140**: 1521–1535.

Carrassi A, Vannitsem S. 2010. Accounting for model error in variational data assimilation: A deterministic formulation. *Mon. Weather Rev.* **138**: 3369–3386.

Carrassi A, Vannitsem S. 2011. State and parameter estimation with the extended kalman filter: an alternative formulation of the model error dynamics. *Q J Roy. Meteor. Soc.* **137**(655): 435–451.

Carrassi A, Vannitsem S, Nicolis C. 2008. Model error and sequential data assimilation: A deterministic formulation. *Q J Roy. Meteor. Soc.* **134**: 1297–1313.

Carson J, Crucifix M, Preston S, Wilkinson RD. 2015. Bayesian model selection for the glacial-interglacial cycle. *arXiv preprint arXiv:1511.03467* .

Chekroun M, Simonnet E, Ghil M. 2011. Stochastic climate dynamics: random attractors and time-dependent invariant measures. *Physica D* **240**: 1685–1700.

Crisan D, Doucet A. 2002. A survey of convergence results on particle filtering for practitioners. *IEEE T. Signal Proces.* **50**: 736–746.

Dee DP. 1995. On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Weather Rev.* **123**(4): 1128–1145.

Del Moral P. 2004. *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Springer: New York.

Dijkstra H. 2013. *Nonlinear Climate Dynamics*. Cambridge University Press: Cambridge.

Elsheikh A, Hoteit I, Wheeler M. 2014a. Efficient bayesian inference of subsurface flow models using nested sampling and sparse polynomial chaos surrogates. *Comput. Methods Appl. Mech. Engrg.* **269**: 515–537.

Elsheikh A, Wheeler M, Hoteit I. 2014b. Hybrid nested sampling algorithm for bayesian model selection applied to inverse subsurface flow problems. *J. Comput. Phys.* **258**: 319–337.

Evans M, Swartz T. 1995. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Stat. Sci.* **10**: 254–272.

Evensen G. 2009. *Data Assimilation: The Ensemble Kalman Filter*. Springer-Verlag/Berlin/Heildelberg, second edn.

Ghil M. 1997. Advances in sequential estimation for atmospheric and oceanic flows. *J. Meteorol. Soc. Jpn.* **75**: 289–304.

Ghil M, Chekroun MD, Simonnet E. 2008. Climate dynamics and fluid mechanics: Natural variability and related uncertainties. *Physica D* **237**: 2111–2126.

Ghil M, Malanotte-Rizzoli P. 1991. Data assimilation in meteorology and oceanography. *Adv. Geophys.* **33**: 141–266.

Grewal M, Andrews A. 2001. *Kalman filtering : Theory and practice using matlab*. Wiley-Interscience, 2 edn.

Hannart A, Carrassi A, Bocquet M, Ghil M, Naveau P, Pulido M, Ruiz J, Tandeo P. 2016. Dada: data assimilation for the detection and attribution of weather and climate-related events. *Climatic Change* **136**(2): 155–174.

Harlim J. 2013. Model error in data assimilation. *arXiv preprint arXiv:1311.3579* .

Hunt B, Kostelich EJ, Szunyogh I. 2007. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D* **230**: 112–126.

Hürzeleri M, Künsch H. 2001. Approximation and maximising the likelihood for a general state-space model. In: *Sequential Monte Carlo Methods in Practice*, Springer-Verlag: New York.

Jazwinski AH. 1970. *Stochastic processes and filtering theory*. Academic Press, New-York.

Kalman RE. 1960. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering* **82**: 35–45.

Kalnay E. 2002. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press: Cambridge.

Kantas N, Doucet A, Singh S, Maciejowski J. 2009. An overview of sequential Monte Carlo methods for parameter estimation. In: *General State-Space Models*, IFAC System Identification.

Kondrashov D, Sun Cj, Ghil. 2008. Data assimilation for a coupled ocean-atmosphere model. Part II: Parameter estimation. *Mon. Weather Rev.* **136**: 5062–5076.

Liu Q, Pierce D. 1994. A Note on Gauss-Hermite Quadrature. *Biometrika* **81**: 624–629.

Lorenc A. 2013. *Recommended Nomenclature for EnVar Data Assimilation Methods* . Available at:http://www.wcrp-climate.org/WGNE/BlueBook/2013/individual-articles/01_Lorenc_Andrew_EnVar_nomenclature.pdf.

Lorenz E. 1963. Deterministic non-periodic flow. *J. Atmos. Sci.* **20**: 130–141.

Lorenz EN, Emanuel KA. 1998. Optimal sites for supplementary weather observations: simulation with a small model. *J. Atmos. Sci.* **55**: 399–414.

© 2016 Royal Meteorological Society

Millar RB. 2011. *Maximum likelihood estimation and inference: with examples in r, sas and admb*, vol. 111. John Wiley & Sons.

Miller R, Ghil M, Gauthiez F. 1994. Advanced data assimilation in strongly nonlinear dynamical systems. *J. Atmos. Sci.* **51**: 1037–1056.

Palmer T. 1999. A non-linear dynamical perspective on climate prediction. *J. Climate* **12**: 575–591.

Pearl J. 2000. *Causality: Models, reasoning and inference*. Cambridge Univ. Press, Cambridge, UK, and New York, NY, USA.

Pires C, Vautard R, Talagrand O. 1996. On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus A* **48**: 96–121.

Pitt M. 2002. Smooth particle filters for likelihood evaluation and maximisation. *Warwick Economic Research Papers* **651**.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. *Numerical Recipes in Fortran 77*. Cambridge University Press: Cambridge.

Raanes PN, Carrassi A, Bertino L. 2015. Extending the square root method to account for additive forecast noise in ensemble methods. *Mon. Weather Rev.* **143**: 3857–3873.

Reich S, Cotter C. 2015. *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge University Press: Cambridge.

Sakov P, Bertino L. 2011. Relation between two common localisation methods for the enkf. *Computat. Geosci.* **15**(2): 225–237.

Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds). 2007. *Climate Change 2007. The Physical Science Basis: Working Group I Contribution to the Fourth Assessment Report of the IPCC*. Cambridge University Press.

Stocker TF, Qin D, Plattner GK, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex B, Midgley BM (eds). 2013. *IPCC, 2013. Climate Change 2013. The Physical Science Basis: Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Stott P, Allen M, Christidis N, Dole R, Hoerling M, Huntingford C, Pall P, Perlwitz J, Stone D. 2013. Attribution of weather and climate-related events. In: *Climate Science for Serving Society: Research, Modelling and Prediction Priorities. G.R. Asrar and J. W. Hurrell (Eds.)*, Springer.

Talagrand O, Courtier P. 1987. Variational assimilation of meteorological observations with the adjoint vorticity equation, I. Theory. *Q J Roy. Meteor. Soc.* **113**: 1311–1328.

Tandeo P, Pulido M, Lott F. 2015. Offline parameter estimation using EnKF and maximum likelihood error covariance estimates: Application to a subgrid-scale orography parametrization. *Q J Roy. Meteor. Soc.* **141**: 383–395.

Wiener N. 1949. *Extrapolation, interpolation and smoothing of stationary time series, with engineering applications*. M.I.T. Press: Cambridge, MA.

Winiarek V, Bocquet M, Saunier O, Mathieu A. 2012. Estimation of errors in the inverse modeling of accidental release of atmospheric pollutant: Application to the reconstruction of the cesium-137 and iodine-131 source terms from the Fukushima Daiichi power plant. *J. Geophys. Res.* **117**: D05 122.

Winiarek V, Vira J, Bocquet M, Sofiev M, Saunier O. 2011. Towards the operational estimation of a radiological plume using data assimilation after a radiological accidental atmospheric release. *Atmos. Environ.* **45**: 2944–2955.

Zhang Y, Bocquet M, Mallet V, Seigneur C, Baklanov A. 2012. Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects. *Atmos. Environ.* **60**: 656–676.

© 2016 Royal Meteorological Society