REVIEW

Beyond Gaussian Statistical Modeling in Geophysical Data Assimilation

MARC BOCQUET

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

CARLOS A. PIRES

*Universidade de Lisboa, and Instituto Dom Luiz (IDL), Centro de Geofísica da Universidade
de Lisboa (CGUL), Lisbon, Portugal*

LIN WU

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

ABSTRACT

This review discusses recent advances in geophysical data assimilation beyond Gaussian statistical modeling, in the fields of meteorology, oceanography, as well as atmospheric chemistry. The non-Gaussian features are stressed rather than the nonlinearity of the dynamical models, although both aspects are entangled. Ideas recently proposed to deal with these non-Gaussian issues, in order to improve the state or parameter estimation, are emphasized.

The general Bayesian solution to the estimation problem and the techniques to solve it are first presented, as well as the obstacles that hinder their use in high-dimensional and complex systems. Approximations to the Bayesian solution relying on Gaussian, or on second-order moment closure, have been wholly adopted in geophysical data assimilation (e.g., Kalman filters and quadratic variational solutions). Yet, nonlinear and non-Gaussian effects remain. They essentially originate in the nonlinear models and in the non-Gaussian priors. How these effects are handled within algorithms based on Gaussian assumptions is then described. Statistical tools that can diagnose them and measure deviations from Gaussianity are recalled.

The following advanced techniques that seek to handle the estimation problem beyond Gaussianity are reviewed: maximum entropy filter, Gaussian anamorphosis, non-Gaussian priors, particle filter with an ensemble Kalman filter as a proposal distribution, maximum entropy on the mean, or strictly Bayesian inferences for large linear models, etc. Several ideas are illustrated with recent or original examples that possess some features of high-dimensional systems. Many of the new approaches are well understood only in special cases and have difficulties that remain to be circumvented. Some of the suggested approaches are quite promising, and sometimes already successful for moderately large though specific geophysical applications. Hints are given as to where progress might come from.

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

1. Introduction

a. The context

This is a review of the non-Gaussian aspects of data assimilation, in the context of geophysics. Through references and a few examples, it investigates the difficulties in producing analyses using statistical modeling that goes beyond Gaussian hypotheses or beyond second-order moment closure. The emphasis is on the concepts and promising ideas, rather than on the technicalities or the completeness of the bibliography. However, mathematical details will be given when necessary or appealing. Examples, original in some cases, will be provided using simple models.

Nonlinearity and non-Gaussianity are interlaced topics, and it is difficult to discuss only one facet of the problem, neglecting the other. For instance, nonlinarities of a dynamical model inevitably produce non-Gaussian priors to be used in later analyses. Nevertheless, the focus of this review is more on the non-Gaussian aspects from the statistical modeling viewpoint, and on ways to extend the usual Gaussian analysis of current data assimilation orthodoxy. There are reviews, or relevant reports, that focus more on the nonlinear aspects but less on modeling the non-Gaussian statistics (Miller et al. 1994; Evensen 1997; Verlaan and Heemink 2001; Andersson et al. 2005). The intended scope of the article is broad: meteorology, oceanography, and atmospheric chemistry. Non-Gaussianity may take many forms there, and does not necessarily always come from the dynamics. However, a commonality of these fields is the very large dimension of the state and observation vector spaces. At first glance, this rules out most of the sophisticated probabilistic mathematical methods that are meant to estimate the full state probability density function (pdf), or higher-order moments, or to provide a state estimate without any approximation.

b. Statistical modeling of the estimation problem

A data assimilation system consists of a set of observations, and of a numerical model, that may be static or dynamical, that may be deterministic or stochastic, and that represents the underlying physics. The mathematical modeling of uncertainty for this system implies that one embeds models and observations in a statistical framework and provides uncertainty input about them. From the uncertainty of the components of this data assimilation system, one can ultimately infer, not only an estimate of the true state, but also the uncertainty of that estimate.

This uncertainty could originate from the imprecise initial state of the system. It could also stem from the more or less precise identification of forcings of the dynamical systems, such as emission fields (in atmospheric chemistry), radiative forcing, boundary conditions, and couplings to other models that may be imperfect. The deficiency of the model itself is another source of uncertainty. To account for this type of uncertainty, models could explicitly be made probabilistic. This occurs when some stochastic forcing is implemented to represent subgrid-scale processes in Eulerian models, or when stochastic particles are simulated to represent dispersion in Lagrangian models. The uncertainty could also come from the observations in the form of representativeness or instrumental errors, or indirectly from the models and algorithms used to filter these observations through quality control. Finally, in the case of remote sensing, it could stem from the joint use of a model (a radiative transfer model for instance) and an algorithm that infers data from indirect measurements.

The proper statistical modeling depends on how uncertainty evolves under the full data assimilation system dynamics. In particular, in the context of forecasting, this modeling should properly account for the uncertainty growth–reduction cycle, which is controlled by the forecast–analysis steps of the data assimilation cycle.

Truncating statistics to the first- and second-order moments (bias and error covariance matrix) may be made necessary because of the complexity of the fully Bayesian data assimilation algorithms. It also reduces the computer storage of higher moments, which are gigantic objects in geophysical systems. This truncation may also be justified from the point of view of the evolution of the dynamical model. If, in the vicinity of a trajectory, the model can be replaced by its tangent linear approximation, then initial Gaussian statistics will remain so in this vicinity. Unfortunately, the statistics would diverge from ideal Gaussianity when the model is strongly nonlinear, when the analyses are infrequent, or when the observational data are sparse (Pires et al. 1996; Evensen 1997).

c. Outline of the review of ideas

In this article, these arguments will be developed. At first, the reasons why researchers in geophysical data assimilation have avoided exact Bayesian modeling, even though it may be more natural, will be reviewed (section 2).

Having accepted that a fully Bayesian approach may not be computationally affordable for geophysical or large environmental problems, Gaussian filtering or variational approaches have been developed with success in geophysical data assimilation. Then the reasons why non-Gaussianity is often bound to reemerge in the statistical modeling of well-behaved geophysical problems

are detailed. Next, we will briefly describe the strategies that have been employed, especially in four-dimension variational data assimilation (4D-Var) and the ensemble Kalman filter, to accommodate possible non-Gaussian deviations in the data assimilation system (section 3). However, this is not the main focus of this review, since literature offers excellent discussion papers on these topics (Kalnay et al. 2007). Ways to objectively measure the deviations from Gaussianity will then be discussed (section 4).

In section 5, the following strategies to better control the uncertainty and to make it less non-Gaussian are examined: targeting of observations, specific treatment of highly nonlinear degrees of freedom, localization, and model reduction.

In section 6, we will review a selection of new ideas to make use of non-Gaussian statistical modeling in data assimilation, either perturbatively (about a Gaussian formalism), or nonperturbatively (without direct reference to a Gaussian formalism). Several original examples will serve as illustrations.

To conclude, the future of these ideas and concepts is discussed (section 7).

2. Why not non-Gaussian from the start?

A Gaussian modeling of the uncertainty in geophysical data assimilation problems may not be the most natural approach to begin with. A more natural approach would be to forget about the constraints such as the dimensionality of the state space of geophysical models. In a constraint-free context, rigorous methods have been proposed in applied mathematics to solve analytically or numerically the fully nonlinear estimation problem, either in discrete or continuous time.

a. The estimation problem

The notations of Ide et al. (1999) will be liberally followed to define the model and observation equations. They are

$$\mathbf{x}_k = M_k(\mathbf{x}_{k-1}, \mathbf{w}_{k-1}) \quad \text{and} \quad \mathbf{y}_k = H_k(\mathbf{x}_k, \boldsymbol{v}_k), \quad (1)$$

where $\mathbf{x}_k$ is the state vector in $\mathbb{R}^N$ at time t_k and $\mathbf{w}_k$ and $\boldsymbol{v}_k$ are noises that represent model and observation errors, respectively. They are stochastic in nature and stand for the uncertainty inherent to the system in Eq. (1). In this section, it is assumed for the sake of simplicity that these perturbations are additive, that is, one has instead

$$\mathbf{x}_k = M_k(\mathbf{x}_{k-1}) + \mathbf{w}_k \quad \text{and} \quad \mathbf{y}_k = H_k(\mathbf{x}_k) + \boldsymbol{v}_k, \quad (2)$$

where $M_k(\mathbf{x}_{k-1})$ is the model that links deterministically $\mathbf{x}_{k-1}$ to $\mathbf{x}_k$, $\mathbf{y}_k$ is the vector of observations in $\mathbb{R}^d$ at time t_k,

and $H_k : \mathbb{R}^N \to \mathbb{R}^d$ is the observation operator. Note that the number of observations d may depend on time index k in any realistic context. Here $\mathbf{w}_k$ and $\boldsymbol{v}_k$ are additive noises that represent model and observation errors. The sets of random vectors $\{\mathbf{w}_k\}_{k=1,\ldots,K}$ and $\{\boldsymbol{v}_l\}_{l=1,\ldots,K}$ are taken to be white in both space and time, and they are assumed to be mutually independent.

Let p_W be the pdf of $\mathbf{w}_k$, and let $\boldsymbol{v}_k$ be distributed according to the pdf p_V. Both p_W and p_V may well depend on time t_k, but the dependence is not made explicit in the notation. These laws define the transition kernel, which probabilistically relates state $\mathbf{x}_{k-1}$ to state $\mathbf{x}_k$, as well as the likelihood of the state $\mathbf{x}_k$ with respect to the observations $\mathbf{y}_k$:

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}) = p_W[\mathbf{x}_k - M_k(\mathbf{x}_{k-1})],$$
$$p(\mathbf{y}_k|\mathbf{x}_k) = p_V[\mathbf{y}_k - H_k(\mathbf{x}_k)]. \quad (3)$$

Within this probabilistic framework, one could either be interested in estimating the true state of the system, along with its uncertainty, at the present time, or be interested in estimating the true state for all times.

The smoothing approach is meant to solve the latter problem. Given $\mathbf{X}_k = \{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$, the collection of all state vectors from time t_0 to time t_k, and $\mathbf{Y}_k = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k\}$, the collection of all observation vectors from time t_1 to time t_k, recursive application of Bayes's law and transition rules leads to the conditional pdf (Jazwinski 1970; Lorenc 1986):

$$p(\mathbf{X}_k|\mathbf{Y}_k) \propto p(\mathbf{x}_0)\prod_{l=1}^{k}\{p_V[\mathbf{y}_l - H_l(\mathbf{x}_l)]p_W[\mathbf{x}_l - M_l(\mathbf{x}_{l-1})]\},$$
$$(4)$$

where $p(\mathbf{x}_0)$ is the prior pdf of the initial state vector. The $\ln[p(\mathbf{X}_k|\mathbf{Y}_k)]$ would define an objective function that ranks state trajectories according to their likelihood. This is the usual Bayesian embedding of the variational formalism and in particular of 4D-Var when the statistics are assumed to be Gaussian.

In the sequential approach (filtering problem), the goal is to estimate the final system state (usually present state) and its uncertainty, rather than the full trajectory of the system states. The sequential method decomposes into a forecast step and an analysis step. The forecast step uses the Chapman–Kolmogorov kernel relationship to connect two states related by an integration of the model:

$$p(\mathbf{x}_k|\mathbf{Y}_{k-1}) = \int d\mathbf{x}_{k-1} \, p_W[\mathbf{x}_k - M_k(\mathbf{x}_{k-1})]p(\mathbf{x}_{k-1}|\mathbf{Y}_{k-1}).$$
$$(5)$$

Obviously, it invokes the transition kernel. The analysis step defines the best estimate of the system state and its uncertainty (full pdf) after the assimilation

of new observations $\mathbf{y}_k$. It makes use of Bayes's theorem:

$$p(\mathbf{x}_k|\mathbf{Y}_k) = \frac{p_V[\mathbf{y}_k - H_k(\mathbf{x}_k)]p(\mathbf{x}_k|\mathbf{Y}_{k-1})}{\int d\mathbf{x}_k\, p_V[\mathbf{y}_k - H_k(\mathbf{x}_k)]p(\mathbf{x}_k|\mathbf{Y}_{k-1})}. \quad (6)$$

b. Nonlinear statistical time-continuous estimation

The estimation solution was given for a discrete time problem. There is no fundamental reason why an analog formalism could not be built for continuous time. Let us first consider the case where there is no assimilation of observations. One first considers a general stochastic model. Its equation, to be understood in the Ito sense, is

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\, dt + \mathbf{G}(\mathbf{x}_t, t)\, d\mathbf{w}_t, \quad (7)$$

where $\mathbf{f}$ is the deterministic part of the model. The noise $d\mathbf{w}_t$ drives the uncertainty ($\mathbf{w}_t$ is a standard Wiener process), and is weighted by the deterministic matrix function $\mathbf{G}(\mathbf{x}_t, t)$. It is well known that the pdf of the full system state vector p_t obeys a Fokker–Planck equation [see Gardiner (2004) for an exposition from the physicist's point of view]:

$$\frac{\partial p_t}{\partial t} = \mathcal{L}_{\mathrm{FP}}(p_t), \quad (8)$$

where

$$\mathcal{L}_{\mathrm{FP}}(p_t) \equiv -\sum_i \frac{\partial}{\partial x_i}\{[\mathbf{f}]_i(\mathbf{x}, t)p_t\} + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}([\mathbf{Q}_t]_{ij}p_t), \quad (9)$$

with $\mathbf{Q}_t = \mathbf{G}(\mathbf{x}_t, t)\mathbf{G}(\mathbf{x}_t, t)^{\mathrm{T}}$. The observation equation has also its stochastic formulation:

$$d\mathbf{y}_t = \mathbf{h}(\mathbf{x}_t, t)\, dt + \sqrt{\mathbf{R}_t}\, d\boldsymbol{v}_t, \quad (10)$$

where $\mathbf{h}(\mathbf{x}_t, t)$ is the deterministic observation operator, $\mathbf{R}_t$ is the observation error covariance matrix, and $\boldsymbol{v}_t$ is a standard Wiener process independent from $\mathbf{w}_t$. The evolution of the conditional pdf $p_t^\star$ derived from Eqs. (7) and (10), is governed by the Zakai equation (Zakai 1969) (or alternatively normalized Kushner equation):

$$dp_t^\star = \mathcal{L}_{\mathrm{FP}}(p_t^\star)\, dt + p_t^\star \mathbf{h}_t^{\mathrm{T}}\mathbf{R}_t^{-1}\, d\mathbf{y}_t. \quad (11)$$

Throughout the paper, T designates the matrix and vector transpose operator. From an analytical point of view, regardless of the algorithmic complexity and numerical cost, the full statistical estimation (smoothing or filtering) problem can be solved by exact methods. These ideas have been explored by Miller et al. (1999) from a geophysicist's

perspective. Several examples, from a one-dimensional double-well potential model to a truncated spectral barotropic model, were given as illustrations of strongly nonlinear systems.

From an algorithmic standpoint, by making the knowledge of a dynamical system a probabilistic one, the objects to deal with are not any more in $\mathbb{R}^N$ (estimate of a state vector). Rather, they are functions $p(\mathbf{x})$ of N variables, and possibly time in the smoothing case. This stresses the change of scale in complexity. The mathematics required to account for these problems do exist but their efficiency is questionable when applied to high-dimensional geophysical systems.

c. Particle filters and their curse in high-dimensional systems

When it comes to numerically solving the fully Bayesian filtering (possibly smoothing) problem, the most popular approaches are based on Monte Carlo sampling (Handschin and Mayne 1969; Kitagawa 1987). Sequential Monte Carlo methods to solve these nonlinear filtering equations are called particle filters. A quite complete and very clear review of the potential applications of particle filters to the geophysical estimation problems was offered by van Leeuwen (2009).

In the context of numerical models and forecasting, the most intuitive variant of the particle filters is the bootstrap particle filter (Gordon et al. 1993). As in any Monte Carlo method, the idea is to represent a pdf by particles $\{\mathbf{x}_k^1, \mathbf{x}_k^2, \ldots, \mathbf{x}_k^M\}$, that is, a collection of system vector states $\mathbf{x}_k^i \in \mathbb{R}^N$ at time t_k, which samples the probability density function $p_k(\mathbf{x}_k)$:

$$p_k(\mathbf{x}_k) \simeq \sum_{i=1}^M \omega_{k-1}^i \delta(\mathbf{x}_k - \mathbf{x}_k^i). \quad (12)$$

The weights ω_{k-1}^i are initially equal to $1/M$, but could differ in the following.

When assimilating $\mathbf{y}_k$, the analysis consists in applying the Bayes's formula to the Monte Carlo estimate of the pdf. The weights are then simply altered by the following likelihood:

$$\omega_k^i \propto \omega_{k-1}^i p(\mathbf{y}_k|\mathbf{x}_k^i), \quad (13)$$

and they need to be normalized so that their sum is 1. This analysis step just involves multiplications of weights and likelihoods. In particular, the innovation-statistics matrix is not inverted as would be required by most filtering methods based on first- and second-order moments. This makes the particle filter a simple and beautiful method, but we shall soon recall its curse.

During the forecast step from time t_k to time t_{k+1}, the particles are propagated with the model $\mathbf{x}_{k+1} = M_{k+1}(\mathbf{x}_k) + \mathbf{w}_{k+1}$. In the context of the bootstrap filter, the pdf is then simply updated at time t_{k+1} and satisfies Eq. (12) but at time t_{k+1}. This completes the particle filter cycle.

Unfortunately, for high-dimensional systems, most of the weights vanish and just a few particles remain likely (see e.g., Berliner and Wikle 2007). Therefore, in this case, the particle filter becomes useless for the estimation problem.

The ensemble size required for a proper estimation has been shown to scale exponentially with the system size, or the innovation variance. To prove this, Snyder et al. (2008) studied the statistics of the biggest weight. They demonstrated on a simple Gaussian model that the required size of the ensemble scales like $M \sim \exp(\tau^2/2)$, where τ^2 is the variance of the observation log-likelihood. Under simple independence assumptions, it is expected to scale like the dimension of observation space on one hand, and the dimension of state space on the other.

This behavior is related to the so-called curse of dimensionality (Bellman 1961). Consider one of the particles, a state space vector in $\mathbb{R}^N$. It is meant to be representative of a volume $[-\varepsilon, \varepsilon]^N$ of state space centered on that particle. However, particles similar to that particle are close to it according to the Euclidean metric: they lie in a neighborhood, say within a distance of ε. In high-dimensional systems, a representativeness issue arises because of the shrinking of the hypersphere of radius ε within the hypercube $[-\varepsilon, \varepsilon]^N$. Indeed the volume of the hypersphere relative to that of the hypercube vanishes like $(\pi/4)^{N/2}/\Gamma[(N/2) + 1]$ with the space dimension N. The particle is less and less representative of the cubic volume it is meant to sample. In the context of data assimilation, this implies that the observational prior and the background prior overlap less and less as the state space or observation space dimensions increase. As a consequence, most of the weights of the particles vanish, leading to a poor analysis.

To mitigate the collapse of the weights, a resampling step is often used in the bootstrap filter. Basically the idea is to draw new particles among the old ones according to the probability given by their weight. After this resampling, all the new particles have the same weight. However, it is likely that many of the new particles will be drawn from the same original particle. This will deplete the ensemble. Unless model error is already specified and of stochastic nature, it is necessary to introduce some perturbation (noise) into the forecast step, in order to enrich the ensemble.

However, the resampling does not fundamentally solve the issue since the weights still degenerate, possibly at a lower rate. The collapse of the weights could be avoided using Monte Carlo Markov chain (MCMC) techniques, based for instance on a Metropolis–Hastings selection algorithm (Gilks and Berzuini 2001). Contrary to importance sampling (and importance proposal sampling, which will be detailed later), these approaches have no exponential dependence on the dimensionality [see the illuminating discussion by MacKay (2003), chapter 29]. However, a considerable number of iterations would be required for an MCMC approach to sample a filtering pdf like those met in geophysical applications. That is why it is not clear to the authors whether these techniques will be decisive in a successful particle filter for high-dimensional systems.

d. Illustrations

As an illustration of these concepts, a bootstrap filter is compared to the ensemble Kalman filter (EnKF) of Evensen (1994), on a Lorenz-95 model (Lorenz and Emmanuel 1998). The comparison follows the methodology of Nakano et al. (2007). This EnKF is the original single-ensemble variant, corrected by Burgers et al. (1998). The Lorenz-95 model is implemented with the standard forcing parameter $F = 8$, and $N = 10$ variables, in a perfect model setting. In particular, no stochastic forcing is applied. The absence of a stochastic term does not prevent a reliable assessment of the data assimilation system on long enough runs, thanks to the ergodicity of the dynamical system.

One of every two sites is observed. The time interval between two observations is $\Delta t = 0.05$, which corresponds to a *geophysical time* of 6 h (Lorenz and Emmanuel 1998). The synthetic measurements are perturbed with a normal white noise of root-mean-square $\sigma = 1.5$. The root-mean-square observation error, χ, for the EnKF is also chosen to be 1.5. This places the filter in optimal conditions since the observation error prior statistics coincide with the true observation error statistics. Moreover, in order to reduce the undersampling errors in covariances for small ensemble size, two versions of the EnKF are implemented, with or without localization (Houtekamer and Mitchell 1998). Localization is carried out thanks to a tapering function (Houtekamer and Mitchell 2001; Hamill et al. 2001) of the form given by Eq. (4.10) of Gaspari and Cohn (1999), with an optimally tuned localization length. For the localization length as well as other parameters in the following, optimal values were selected via a sensitivity analysis to minimize the analysis error.

The particle filter to which the EnKF will be compared is the bootstrap filter that was described above. It is tested on the same setup and it uses the same parameters when applicable, as the ensemble Kalman filter. Contrary to the EnKF, no localization is used in the particle filter (this
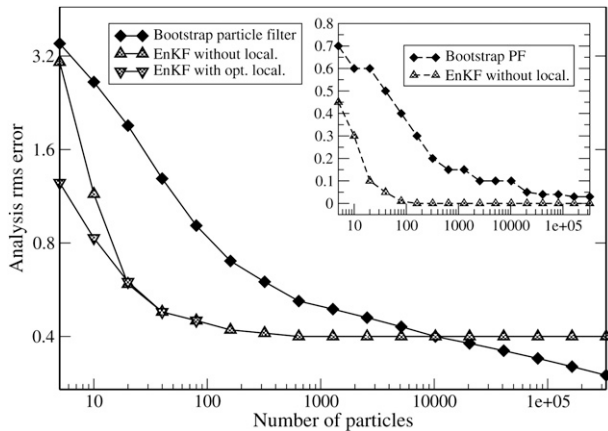
FIG. 1. Comparison of the performance of a bootstrap filter with an ensemble Kalman filter (without localization) on a Lorenz-95 model with $N = 10$ variables, as the ensemble size is increased. Rough estimations of the noise levels required to stabilize the two filters (without localization) are displayed in the insert. For small ensemble size, the performance of an EnKF with localization using an optimally tuned localization length is also displayed. Beyond 20 members, localization hardly makes a difference.

FIG. 2. Empirical statistics of the maximum weights of a Lorenz-95 particle filter for four system sizes: $N = 10, 20, 40,$ and 80, ranging from a well balanced to a degenerate case. Apart from the system size and the ensemble size ($M = 128$), the setup is the same as in Fig. 1.

issue will be adressed later). For the particle filter, the prior observation error standard deviation χ is allowed to differ from the observation perturbation standard deviation $\sigma = 1.5$, since it is optimally tuned. For most of the ensemble size M range, the optimal values are close to $\chi = 2.5$. This implicit error inflation is meant to account indirectly for the sampling error in the representation of the pdf [Eq. (12)]. However, for increasing values of M the optimal χ tends to σ (not shown).

In addition, Gaussian white (in space and time) perturbations are added to all particle state vectors with an amplitude that is optimally tuned for the sake of a fair comparison with the EnKF. Again, the optimal values are obtained thanks to a sensitivity study carried out with varying noise magnitude. Such a noise is also necessary in the EnKF case, at least for small ensemble size, since no inflation was implemented and because a single-ensemble configuration of the EnKF is used (Mitchell and Houtekamer 2009). A (residual) resampling is carried out after each analysis. The two schemes are compared on the analysis root-mean-square error (analysis rms error).

Still, the particle filter requires 10^4 members to match the EnKF performance. Results are shown in Fig. 1. The size of the system $N = 10$ was chosen so that the EnKF/bootstrap filter cross-over could be observed with a reasonable computation load.

The collapse of the particle filter with increasing state space dimension can be illustrated on the same Lorenz-95 model. Four configurations are chosen identical to the one described above, but for four system sizes $N = 10,$
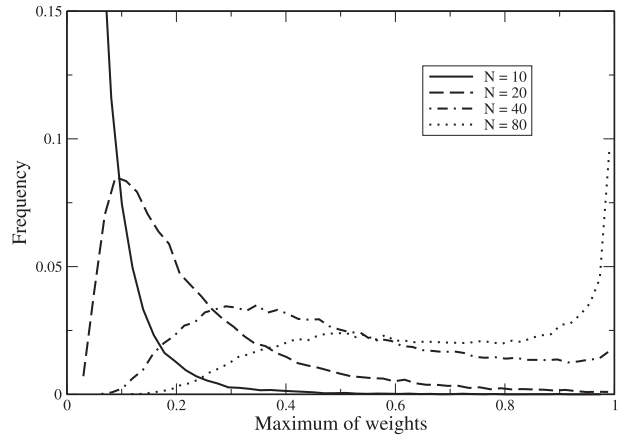
20, 40, and 80. Figure 2 displays the empirical statistics of the maximum weight in the four cases. In the first two cases, the density is rather balanced, with high values of the weight maxima that are not overrepresented. In the last two cases, the weights degenerate. When $N = 80$, the mode is near 1 and the particle filter collapses (it is of no use for estimation).

Therefore, at least with basic (though not naive) algorithms, it is still unreasonable to use particle filters for the state estimation, let alone high-order moments of the errors.

e. Gaussian as a makeshift?

Admittedly, a fully non-Gaussian solution to the estimation problem may still be an intractable problem for large geophysical systems. The first nontrivial approximation to the statistical estimation problem is to truncate the error statistics to second order or to assume these statistics are approximately Gaussian. In this case, the multivariate pdf $p(\mathbf{x})$, with $\mathbf{x} \in \mathbb{R}^N$ can be solely derived from the Gaussian correlations between pairs of variables, which are functions of two variables $p(x_k, x_l)$, with $1 \leq k, l \leq N$. These are equivalent to the full error covariance matrix. Hence, Gaussian estimation still leads to complex objects to deal with, before any reduction. It is computationally tractable only if the full covariance matrix is sampled or reduced.

Gaussian statistics have appealing properties. They are still analytically tractable in multivariate form (i.e., convolution and, integration). Their occurrence and recurrence in physical systems is supported by the central limit theorem. Moreover, Gaussians are the simplest distributions (in the sense of information theory) when only

first- and second-order moments are known, as recalled by Rodgers (2000).

3. Dealing with non-Gaussianity in a Gaussian framework

This section explains the current way of dealing with nonlinearity and non-Gaussianity: reasonable data assimilation should consider non-Gaussian effects as corrections to a Gaussian analysis-based strategy. Variational approaches (4D-Var essentially) and ensemble-based Kalman filters include different approaches to account for model nonlinearities.

Sources of non-Gaussianity can be initially categorized into two families: nonlinearities in models and non-Gaussianity of priors. The latter will be emphasized here since it is the main focus of this review, but also because several recently developed methodologies are available. One has to keep in mind that this classification is arbitrary since nonlinear dynamical models produce inevitably non-Gaussian error statistics, which are often used as background error statistics.

a. Non-Gaussianity from nonlinearities in models

Non-Gaussianity results from nonlinearities in models because, under a nonlinear model transition, a Gaussian pdf becomes non-Gaussian. Nonlinearities in models may come from nonlinearity of Navier–Stokes equations leading to chaos, thresholds of microphysics (cloud, ice, rain), chemistry of atmospheric compounds (including thermodynamics and aerosols size evolution equations), increases in the model resolution that require finer physical schemes such as for precipitation at convective scale, and nonlinearity of observation operator model (remote sensing applications especially: lidar and satellite), etc. Many of these sources of nonlinearities have been discussed by Andersson et al. (2005).

b. Non-Gaussianity from priors

1) PRIOR MODELING OF STATE SPACE OR CONTROL SPACE VARIABLES

Non-Gaussian priors are sometimes more adequate descriptions of the background. This is especially the case for positive variables, with large deviations about their mean. This occurs in many geophysical fields. Humidity in meteorology, species concentrations and emission inventories in atmospheric chemistry, algae population in ocean biogeochemistry, and ice and gas age in paleoglaciology ice cores are just a few examples. In the case of atmospheric chemistry, it is usually considered that the typical errors in emission inventories are of the order of 40%, before any assimilation, which rules

out Gaussian modeling for positive variables. Lognormal distributions are usually used instead.

Pires et al. (2010) take the example of brightness temperature from the High Resolution Infrared Sounder (HIRS) channels to show that non-Gaussianity may also stem from the variability in the specified standard deviations of background errors (a statistical property called heteroscedasticity), in particular when aggregate statistics are used in data assimilation systems.

2) OBSERVATION ERROR PRIOR

Observation error priors may also require non-Gaussian modeling. Otherwise, quality-control (QC) filters are mandatory (Lorenc 1986). Gaussian modeling of observation errors correspond to a least squares penalty. Therefore, the data assimilation system will be forced to comply with outliers, which can be regarded as good when these observations mark correctly an extreme event, or bad when they are actually erroneous observations that nevertheless passed quality filters, or, more likely, that are not compatible with an erroneous model. This may lead to the will to forge a penalty term of a least squares type for a limited range of the error, and a less constraining norm, such as l_1, outside this Gaussian domain. Also, some observations may not pass the QC filter although they are strong indicators of extreme events, as it was the case for the Lothar storm of December 1999 (more information available online at http://4dvarenkf. cima.fcen.uba.ar/Download/Session_3/4DVar_nLnG_Fisher. ppt; C. Tavolato and L. Isaksen 2009, personal communication). A more tolerant filter is affordable if the observation errors are not necessarily Gaussian. This is the so-called Huber norm (Huber 1973), which is differentiable:

$$\mathcal{J}_H(v) = \begin{cases} \frac{1}{2}v^2 & \text{if} \quad |v| < \eta \\ \eta|v| - \frac{1}{2}\eta^2 & \text{if} \quad |v| > \eta \end{cases}. \qquad (14)$$

Another possibility is to complement the l_2 norm by a flat distribution, instead of the l_1 norm, as implemented by Andersson and Järvinen (1999) in the European Centre for Medium-Range Weather Forecasts (ECMWF) forecasting system.

For the sake of clarity, the time index k is dropped here. Rather than the noise additive $y_i = H_i(\mathbf{x}) + v_i$, the observation equation could become $y_i = s_i H_i(\mathbf{x})$, with s_i a strictly positive multiplicative dimensionless factor. In that case, s_i is a relative error.

The vector of relative errors $\mathbf{s}$ may obey a lognormal distribution. Lognormal error statistics are consistent with positive observables, and emerge quite naturally in the modeling of trace constituents and of their emissions,

as mentioned earlier. From experimental or empirical modeling, one may have access to the first moments $\bar{\mathbf{s}} = \mathrm{E}[\mathbf{s}]$ and second-order moments of $\mathbf{s}$:

$$R_{ij} = \mathrm{E}[(s_i - \bar{s}_i)(s_j - \bar{s}_j)^{\mathrm{T}}], \qquad (15)$$

which specifies the distribution of $\mathbf{s}$ completely. As we shall see in section 6, it may be preferable to perform an analysis in a space where random variables are Gaussian. That is why it is convenient to use the Gaussian statistics of the vector $\ln(\mathbf{s})$ (the logarithm applied component-wise to the vector $\mathbf{s}$) instead of $\ln(\mathbf{s}) \sim \mathcal{N}(\mathbf{b}, \widetilde{\mathbf{R}})$. The two sets of first- and second-order moments are related through

$$\widetilde{R}_{ij} = \ln\left(1 + \frac{R_{ij}}{\bar{s}_i \bar{s}_j}\right), \quad b_i = \ln(\bar{s})_i - \frac{1}{2}\widetilde{R}_{ii}. \qquad (16)$$

The term in the cost function related to this likelihood is of the following form:

$$\mathcal{J}_{\mathrm{LN}}(\mathbf{s}) = \frac{1}{2}\sum_{i,j}\{\ln(\mathbf{s}) - \ln[\mathbf{y}/H(\mathbf{x})] - \mathbf{b}\}^{\mathrm{T}}\widetilde{\mathbf{R}}^{-1}\{\ln(\mathbf{s})$$
$$- \ln[\mathbf{y}/H(\mathbf{x})] - \mathbf{b}\}. \qquad (17)$$

If the observation error is unbiased $\mathrm{E}[\mathbf{y} - H(\mathbf{x})] = \mathbf{0}$, then $\bar{s}_i = 1$. Then there is a bias [given by Eq. (16)] that needs to be removed in Gaussian space, which is accounted for in the likelihood Eq. (17). If the median of the observation error is null, then the bias is zero: $\mathbf{b} = \mathbf{0}$.

Dealing with more drastic non-Gaussian constraints, Bocquet (2007) has chosen an l_∞ norm as an error prior penalty:

$$\mathcal{J}_\infty(v) = \begin{cases} 0 & \text{if} \quad v^- \le v \le v^+ \\ +\infty & \text{if} \quad v > v^+ \quad \text{or} \quad v < v^- \end{cases}. \qquad (18)$$

In the context of atmospheric dispersion, this allowed to check rigorously that a transport model was, or was not, compatible with a set of observations, with analyzed errors lying in the predefined interval $[v^-, v^+]$.

Now that possible sources of non-Gaussianity have been recalled, classic solutions to deal with them are reviewed.

c. 4D-Var solutions to deal with nonlinearity

The 4D-Var algorithm was originally proposed to solve the data assimilation problem, described as a constrained optimization problem, using classic descent algorithms. The gradient of the cost function to be minimized can be efficiently computed by optimal control techniques (Le Dimet and Talagrand 1986; Talagrand and Courtier 1987; Courtier and Talagrand 1987).

Suppose that the pdf of the initial state $\mathbf{x}_0$ is given in terms of departure from some known background $\mathbf{x}_b$

[i.e., $p(\mathbf{x}_0) = p_B(\mathbf{x}_0 - \mathbf{x}_b)$] and that the departure vector $\mathbf{x}_0 - \mathbf{x}_b$, the model error $\mathbf{w}_k$, and the observation error $\boldsymbol{v}_k$ are independent Gaussian vectors with zero mean and covariance matrices $\mathbf{B}$, $\mathbf{Q}_k$, and $\mathbf{R}_k$, respectively. Maximizing the conditional pdf $p(\mathbf{X}_K|\mathbf{Y}_K)$ (for a maximum likelihood estimate of $\mathbf{x}_0$) amounts to a minimization of the weakly constrained 4D-Var cost function:

$$\mathcal{J}(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b)$$
$$+ \frac{1}{2}\sum_{k=1}^{K}[H_k(\mathbf{x}_k) - \mathbf{y}_k]^{\mathrm{T}}\mathbf{R}_k^{-1}[H_k(\mathbf{x}_k) - \mathbf{y}_k]$$
$$+ \frac{1}{2}\sum_{k=1}^{K}\mathbf{w}_k^{\mathrm{T}}\mathbf{Q}_k^{-1}\mathbf{w}_k. \qquad (19)$$

4D-Var data assimilation algorithms are used to estimate the initial condition in state of the art operational centers (e.g., for meteorology see Rabier et al. 2000), given the approach's ability to provide flow estimates consistent with the flow evolution and the asynchronous nature of the observations.

When the model and the observation operators are linear or weakly nonlinear (in the case of small time steps of model simulation and of short assimilation time intervals), these Gaussian assumptions may hold reasonably well. However, when the assimilation windows are long enough or the models are strongly nonlinear, the Gaussian assumptions will certainly break down and the conditional pdf could become multimodal. As a result, the maximum likelihood estimates become less informative (Lorenc and Payne 2007). This is equivalent to the existence of multiple minima with the 4D-Var cost function (Gauthier 1992; Miller et al. 1994; Pires et al. 1996), while the deterministic numerical optimization seeks only one relative minimum.

To find the global minimum of $\mathcal{J}$, one might resort to stochastic optimization methods (e.g., simulated annealing; Krüger 1993). Unfortunately, because of the high dimensionality of the flow, such stochastic methods are seldom feasible before a reduction in dimension for practical geophysical applications (Hoteit 2008).

One feasible remedy is to deal with the original nonlinear optimization problem approximately by a succession of inner-loop quadratic optimization problems, in which the model is simplified (at a lower resolution with simpler physics) and linearized (Laroche and Gauthier 1998). The input of one inner-loop iteration is generated by relinearizing the original nonlinear model around the state adjusted by the output of the previous inner-loop iteration. This inner/outer approach may fail when there exist significant inner-loop linearization errors for high-resolution models and longer assimilation windows in a context of perfect models (strong-constraint 4D-Var;

Trémolet 2004). This difficulty can be alleviated by introducing model errors (weak-constraint 4D-Var). Indeed, the propagation of information within the assimilation window with the tangent linear model is shortened as compared to the strong constraint 4D-Var, thanks to model error present at each time step (Andersson et al. 2005; Fisher et al. 2005; Trémolet 2006).

Another approach, the quasi-static variational assimilation, was proposed by Pires et al. (1996) for the assimilation of dense observations. The global minimum was guaranteed by progressively lengthening the assimilation periods, thus always keeping control of the first guesses when using a gradient descent method in the cost function minimization.

d. EnKF solutions to deal with nonlinearity

In the 4D-Var approach, the estimation is performed globally using all the observations falling within the assimilation window. Another approach is to treat the observations sequentially: once a new observation vector $\mathbf{y}_k$ at time t_k is available, an estimate (analysis) $\mathbf{x}_k^a$ can be obtained by optimally combining some a priori state $\mathbf{x}_k^f$ and $\mathbf{y}_k$. A Best Linear Unbiased Estimator (BLUE) analysis reads

$$\mathbf{x}_k^a = \mathbf{x}_k^f + \mathbf{K}_k[\mathbf{y}_k - H_k(\mathbf{x}_k^f)], \tag{20}$$

$$\mathbf{P}_k^a = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)\mathbf{P}_k^f, \tag{21}$$

$$\mathbf{K}_k = \mathbf{P}_k^f \mathbf{H}_k^{\mathrm{T}}(\mathbf{H}_k \mathbf{P}_k^f \mathbf{H}_k^{\mathrm{T}} + \mathbf{R}_k)^{-1}, \tag{22}$$

where $\mathbf{P}_k^f = \mathrm{E}[(\mathbf{x}_k - \mathbf{x}_k^f)(\mathbf{x}_k - \mathbf{x}_k^f)^{\mathrm{T}}]$ is the a priori error covariance matrix, $\mathbf{P}_k^a = \mathrm{E}[(\mathbf{x}_k - \mathbf{x}_k^a)(\mathbf{x}_k - \mathbf{x}_k^a)^{\mathrm{T}}]$ the analysis error covariance matrix, and $\mathbf{K}_k$ is the gain matrix. The form of $\mathbf{K}_k$ requires a linear approximation $\mathbf{H}_k$ of the observation operator H_k around $\mathbf{x}_k^f$. This analysis $\mathbf{x}_k^a$ is optimal (among all possible linear forms) in the sense that the total analysis error variance $\mathrm{Tr}(\mathbf{P}_k^a)$ is minimized. When the a priori and observation errors are assumed to be Gaussian, the analysis $\mathbf{x}_k^a$ in Eq. (20) is also the maximum likelihood estimate of the model state that minimizes the 3D-Var cost function [$K = 1$ in Eq. (19)].

To close the algorithm, the a priori state $\mathbf{x}_{k+1}^f$ and error covariance $\mathbf{P}_{k+1}^f$ at subsequent time t_{k+1} can be chosen to be the forecasts starting from the analyzed state $\mathbf{x}_k^a$ and error covariance $\mathbf{P}_k^a$. When the dynamical model is linear, these forecast and analysis formulas are the Kalman filter equations. For nonlinear models, the extended Kalman filter approximates the evolution of the error covariance using tangent linear approximations of the model equation around $\mathbf{x}_k^a$. The ensemble

Kalman filter (Evensen 1994) uses an ensemble of state vector samples $\{\mathbf{x}_k^i, i = 1, \ldots, M\}$ to approximate the error covariance $\mathbf{P}_k^f$ and $\mathbf{P}_k^a$. This lessens the instability of the covariance evolution equation caused by the truncation errors when linearizing models for strongly nonlinear systems. With a small ensemble [e.g., $O(100)$ members], the EnKF is feasible for large geophysical applications. More algorithmic details can be found in Evensen (2003) and Houtekamer and Mitchell (2005).

Reports about the EnKF can be found in meteorology, oceanography, hydrology, and several other fields (Evensen 2003). The reasons for this popularity are multifold. First, although many environmental systems are nonlinear and high dimensional, there exists low-dimensional subspace (local and global attractors) which represents reasonably well the complete dynamics (Lions et al. 1997; Patil et al. 2001). Thus, the pdf may be represented by a proper ensemble with a limited number of members. Second, it is well known that an ensemble forecast has the advantage against a single control forecast (Leith 1974). Finally, the Gaussian assumption at analysis time may be suitable for many scenarios (e.g., the Gaussian background errors for global numerical weather predictions as in Andersson et al. 2005).

For large systems in meteorology, the most effective EnKF schemes are those that localize the background error covariance (Houtekamer and Mitchell 2001; Hamill et al. 2001) so that spurious correlations at long distance are reduced. In other fields such as oceanography and air quality a related approach using reduced-rank Kalman filters (Cane et al. 1996; Heemink et al. 2001; Pham et al. 1998) has also been tested. Such filters work only in subspaces of the complete error space (Lermusiaux and Robinson 1999; Nerger et al. 2005). The EnKF can also be viewed as a reduced-rank Kalman filter, since the error covariance matrices are approximated by the ensemble statistics in a square root form (Tippett et al. 2003). The analysis in Eq. (20) has two implementations: a deterministic scheme (Whitaker and Hamill 2002) or with perturbations of observations for consistent error statistics (Burgers et al. 1998). They differ from each other in handling non-Gaussianity (Lawson and Hansen 2004).

Improvements to the EnKF are essentially driven by the design of better sampling strategies for the ensemble generation: the second-order exact resampling (Pham 2001), the unscented sampling (Van der Merwe et al. 2000), and the mean-preserving sampling (Sakov and Oke 2008). Increasingly, model deficiencies are simulated using ensemble members generated with different versions of the underlying forecast model (e.g., with different physical parameterizations; Meng and Zhang

2007; Fujita et al. 2007; Houtekamer et al. 2009; or with perturbations of model parameters; Wu et al. 2008).

Another idea is to bridge the gap between variational and sequential approaches, and to improve the EnKF performance (Kalnay et al. 2007) using ideas and techniques developed for 4D-Var. Such attempts are for example: the inner–outer loop to deal with nonlinearities (Kalnay et al. 2007), the variational formulation to treat the non-Gaussian error structure in observation (Zupanski 2005) and background (Harlim and Hunt 2007), and the time interpolation of the background forecasts to the observations so as to produce time-coherent assimilations of all the observations available within the assimilation window (Hunt et al. 2004; Houtekamer and Mitchell 2005). Alternatively, the 3D-Var or 4D-Var can use flow-dependent error covariances computed from the EnKF ensembles (Buehner et al. 2010), which leads to more efficient hybrid algorithms.

4. Measuring non-Gaussianity

In a Gaussian framework, one needs tools to assess the deviation from Gaussianity mainly induced by nonlinearities of the model: objective mathematical measures or statistical tests. These tools will be reviewed in this section, and, moreover, some of them will be used in section 6.

a. Relative entropy

A measure of the discrepancy between two pdfs p and q is given by the Kullback–Leibler divergence or relative entropy (Kullback 1959):

$$\mathcal{K}(p, q) = \int d\mathbf{x}\, p(\mathbf{x}) \ln\left[\frac{p(\mathbf{x})}{q(\mathbf{x})}\right]. \tag{23}$$

Coming from signal and information theory, it quantifies the information gain from q to p. It has (axiomatic) properties that make it very attractive (Cover and Thomas 1991). First of all, it is always nonnegative. It is null if and only if p is equal to q almost everywhere. It is also convex with respect to both p and q. Moreover, it is invariant by any one-to-one reparameterization $\mathbf{x} = \Xi(\boldsymbol{\theta})$. However, it is not a distance in the mathematical sense since it is not symmetric with respect to p and q, nor does it obey the triangle inequality.

The measure has been used in geophysical, high-dimensional applications, in predictability (Kleeman 2002), in the statistical modeling of geophysical dynamical systems (Haven et al. 2005), in inverse modeling (Bocquet 2005b,c), and in the modeling of prior pdfs (Eyink and Kim 2006; Pires et al. 2010).

The Kullback–Leibler divergence can serve as an objective function to measure deviation from Gaussianity.

If p is the full pdf of the uncertainty for the system, and $q \equiv p_G$ is the Gaussian pdf that has the same first- and second-order moments, then $\mathcal{K}(p, p_G)$, often called *negentropy*, is a measure of the non-Gaussianity of p. Obviously if p is Gaussian, the divergence is null and positive otherwise.

The pdf p could be estimated approximately by the use of an ensemble, such as the one used by ensemble-based filters. It is, however, difficult to perform such estimation for high-dimensional systems, especially with a small ensemble. Besides, the presence of a strange attractor complicates the numerical convergence and a proper definition of a continuous limit. Several solutions have been proposed to overcome the difficulty: compute relative entropies of marginals of p or compute expansions of the relative entropy.

The use of the marginals has been explored by Kleeman (2007) with a view to geophysical applications. A lower bound for the divergence $\mathcal{K}(p, p_G)$, with a pdf p depending on N variables can be obtained. For each subset s of $n \leq N$ variables, one integrates out the pdfs p and q on this subset of variables and obtains a divergence $\mathcal{K}(p^s_{N-n}, q^s_{N-n})$. A lower bound is then given by

$$\mathcal{K}(p_N, q_N) \geq \frac{n!(N-n)!}{N!} \sum_s \mathcal{K}(p^s_{N-n}, q^s_{N-n}), \tag{24}$$

where the sum runs over all subsets.

The expansions of the Kullback–Leibler divergence are based on Gram–Charlier or Edgeworth expansion of p/q (e.g., see Barndorff-Nielsen and Cox 1989). These expansions depend on skewness and kurtosis, which is consistent with their use in the diagnosis of non-Gaussianity. They are expressed in terms of cumulants of the distribution. They both represent the same expansion, but the terms are ordered differently. In the Gram–Charlier expansion, the ordering index is the cumulant order, while in Edgeworth, the ordering index is the size M of the ensemble, which samples the distribution. The latter expansion makes the Edgeworth expansion more controlled, though less simple. Using these expansions, the relative entropy can be approximated by

$$\mathcal{K}(p, q) \underset{\text{Gra.}}{\simeq} \frac{1}{12} \sum_{i,j,k} (\kappa_{i,j,k})^2 + \frac{1}{48} \sum_{i,j,k,l} (\kappa_{i,j,k,l})^2, \tag{25}$$

$$\mathcal{K}(p, q) \underset{\text{Edg.}}{\simeq} \frac{1}{12} \sum_{i,j,k} (\kappa_{i,j,k})^2 + O\left(\frac{1}{M^{3/2}}\right), \tag{26}$$

where $\kappa_{i_1, i_2, \ldots, i_n}$ are the *standardized* cumulants of p of order n.

Figures 3 and 4 illustrate the departure from Gaussianity and ways to measure it on a deterministic Lorenz-63 model (Lorenz 1963), where the negentropy can be estimated
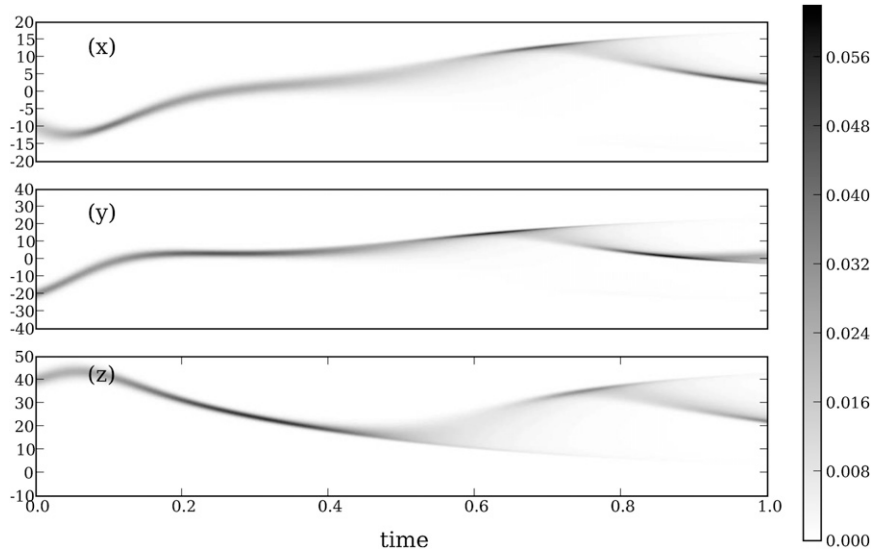
FIG. 3. Evolution of a Gaussian pdf under the Lorenz-63 model equations, with a center of mass initially at $x = -10$, $y = -20$, and $z = 40$. Projections of the full density on the x axis, y axis, and z axis are represented.

numerically. A Gaussian pdf sampled by particles, initially of covariance matrix $\mathbf{P} = \mathrm{diag}(\sigma^2, \sigma^2, \sigma^2)$, with $\sigma = 0.20$, is transformed under the model flow. The full negentropy is estimated via a numerical integration. As explained by Kleeman (2002), relative entropy must be estimated at fixed resolution, possibly fine enough to encapsulate the attractor [a spacing of about $r = \Delta x = \Delta y = \Delta z \simeq 0.1$ has been chosen to discretize the integral Eq. (23)]. Edgeworth expansion, and its estimates based on one- and two-variable marginals are estimated as well. The result of the Edgeworth expansion cannot be directly compared to the full relative entropy estimate of finite resolution. Indeed the ensemble members tend to gather close to, or on, the attractor, which makes their distribution more and more singular with the flow's evolution. For the sake of comparison, the ensemble must therefore be smoothed out by a normal law (of variance 0.5 here) yielding a finite resolution value. Obviously after time $t = 0.5$, the cluster of particles, stretched by the flow, loses its cohesion, and the pdf becomes significantly non-Gaussian. This is confirmed by the indicators based on the negentropy, as well as the Edgeworth expansion.

The deviation from Gaussianity is reported by each one of these indicators, though not with the correct magnitude. Yet, as numerical estimations of the Kullback–Leibler divergence, all these approximations are unsatisfactory.

b. Univariate and multivariate tests of normality

Because such computations cannot easily be generalized to high-dimensional, complex dynamical systems, one could rely on simpler necessary tests of normality.

Hypothesis testing is a well-developed topic in statistics, and many techniques meant to test the Gaussianity of random variables exist.

Among the many tests of normality available in the statistical literature, the skewness and kurtosis, which are directly defined by the cumulants of a distribution, have been used very early. Lawson and Hansen (2004) have used them to assess how differently stochastic and deterministic ensemble-based filters handle non-Gaussianity. There are many other tests such as the Kolmogorov–Smirnov test (Lilliefors 1967), the Anderson–Darling test
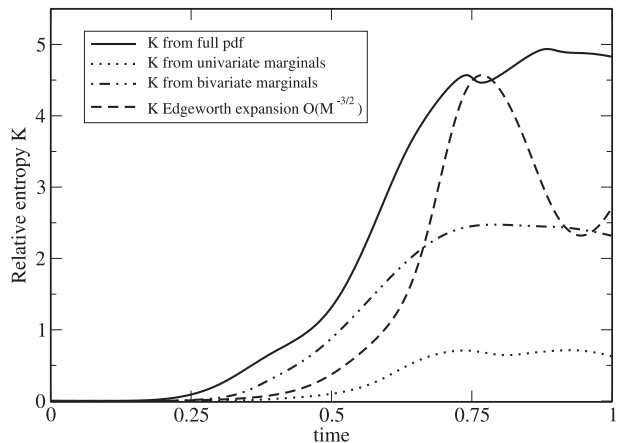

FIG. 4. Estimation of the non-Gaussianity of the pdf evolving under the Lorenz-63 flow depicted in Fig. 3, by the full negentropy, the Edgeworth expansion up to order $3/2$, and the average marginals on 1 and 2 variables. The negentropies are computed at an average resolution of 0.1 units.

(Anderson and Darling 1952), the Shapiro–Wilk test (Shapiro and Wilk 1965), and their many variants.

A few generalizations of these tests to multivariate statistics do exist, but are not meant to handle system sizes as big as those in geophysics. One is therefore compelled to use necessary though insufficient tests. Examples have been given earlier with the use of marginal distributions in the computation of negentropy. One could also rely on combinations of variables in the systems, such as sums (or sums of squares) of individual degrees of freedom, which are supposed to be close to Gaussian. Assuming a Gaussian distribution for the elementary degrees of freedom, it may be possible to compute the distribution of these consolidated random variables. Then univariate null-hypothesis statistical tests, such as those mentioned earlier can be used in turn.

As an example of such a combination of random variables, one could use the Mahalanobis norm of a member state vector [exemplified later in Eq. (27)], and compare it to a χ^2 distribution. Bengtsson et al. (2003) used such a test to obtain a measure of the deviation from normality. They considered a subset of three adjacent variables x_1, x_2, and x_3 in the Lorenz-95 model, so as to reduce the number of degrees of freedom to handle. An ensemble (drawn from an ensemble-based assimilation technique) represents the uncertainty in the system. If $\mathbf{\Sigma}$ is the covariance matrix of this ensemble, defined on the subset, and if $\mathbf{z}_i$ is the deviation of the ith member from the mean (restricted to the subset), then a scalar random variable that combines the three degrees of freedom is

$$d_i = \mathbf{z}_i^{\mathrm{T}} \mathbf{\Sigma}^{-1} \mathbf{z}_i. \tag{27}$$

It should follow a χ^2 distribution with three degrees of freedom. The authors tested this hypothesis using the Kolmogorov–Smirnov test. This way they showed that the forecast produced by an EnKF exhibits significantly non-Gaussian features, at least for long intervals between observation times (cf. $\Delta t = 0.4$, whereas $\Delta t = 0.05$ in this review).

5. Reducing nonlinearity's impact: Divide and conquer

With a denser monitoring network or more frequent observations, the model should remain closer to its tangent linear trajectory between analyses. Therefore, non-Gaussianity will not develop as much as in a system less constrained by observations. Nonetheless, if, with an increasing number of observations, the model resolution is increased as well and subgrid processes are explicitly represented at a finer scale, new sources of nonlinearity and non-Gaussianity might appear as discussed in section 3. In the context of meteorology, the finer the horizontal space scale, the bigger the error growth rate (Lorenz 1969; Tribbia and Baumhefner 2004; Lorenc and Payne 2007). As a consequence how non-Gaussian the errors are may well depend on their scale. Increasing both the space and time resolution and the observation density may lead to a data assimilation system with the appearance of significantly non-Gaussian errors statistics at the convective scale while synoptic-scale errors become smaller and more Gaussian.

In this section, following this paradigm but without going as far as adopting a broad multiscale view on non-Gaussianity, we review some of the ideas put forward to reduce non-Gaussianity and nonlinearity, so that classic data assimilation based on Gaussian hypotheses could become more efficient.

One idea consists in using targeted (also called adaptive) observations in order to obtain a better control. A second one consists in dividing the system between degrees of freedom that are more or less prone to nonlinearities, and hence require more or less accounting for non-Gaussian effects. Another idea consists in representing non-Gaussian features, such as multimodality, via a sum of individual Gaussian components.

a. Better control with adaptive observations

The analysis improvement and the reduction of its computational cost can be obtained by an assimilation that adapts to the properties of the dynamical flow, in particular its instability. For instance, Pires et al. (1996) have shown that the efficient variational assimilation length $\tau_{\mathrm{eff}}(\mathbf{x})$ is proportional to $\lambda^{-1}(\mathbf{x})$, where $\lambda(\mathbf{x})$ is the leading local Lyapunov exponent at $\mathbf{x}$. From Eq. (3.15) of Pires et al. (1996), and relying on their simplifying assumptions [i.e., perfect model, frequent and regular observations within the assimilation window, and $\lambda(\mathbf{x})\Delta t(\mathbf{x}) \ll 1$], the leading analysis error variance $e^2(\mathbf{x})$ is constrained by: $e^2(\mathbf{x}) \leq 2\lambda(\mathbf{x})\Delta t(\mathbf{x})\sigma^2$, where observations of variance σ^2 are obtained each $\Delta t(\mathbf{x})$ (much shorter that the assimilation window). Thus, smaller observation intervals are required for cases that are more unstable.

Adaptive techniques in data assimilation also call for the deployment of targeted observations (TOs), pioneered by the singular vectors approach (Buizza and Montani 1999). Then, Daescu and Navon (2004) use the adjoint sensitivity approach, evaluating the sensitivity function $\|\nabla J\|$, the norm of the gradient of a forecasting error functional J. Local maxima of this sensitivity function in the physical space determine the set $\mathcal{T}$ of targeted observations whose path is sequentially updated taking into account previous $\mathcal{T}$-sets and the set of routine observations. The ensemble transform Kalman

filter (Bishop et al. 2001) predicts the reduction of the eigenvalues of forecast error covariance, with the leading reductions determining the TOs set. Uzunoglu (2007) presents a maximum likelihood Kalman filter where TOs are determined based upon a criterion of Shannon entropy and the condition number of the ensemble subspace covariance matrix.

The number of tracking observations, necessary for the stabilization of a sequential prediction-assimilation system and tracking unstable flow, depends on the number and magnitude of the system's m positive Lyapunov exponents (Carrassi et al. 2008). The efficient monitoring of the m-dimensional unstable space $\mathcal{E}$ is achieved through the blending of a fixed observational network with updated TOs. Efficient analyses with fixed observations are obtained through the assimilation in the unstable subspace (Carrassi et al. 2007) where the analysis increment is confined to the updated unstable subspace $\mathcal{E}$, obtained by the method of breeding of the data assimilation system.

b. Bayesian filtering in reduced-rank system or subsystem

Following the divide and conquer strategy, the use of exact Bayesian techniques, such as particle filters, could be restricted to the significantly non-Gaussian degrees of freedom of the geophysical system. For instance, Lagrangian assimilation of data from oceanographic drifters is highly non-Gaussian since the positions of the drifters need to be controlled too. Spiller et al. (2008) have successfully tested several particle-filtering strategies on such drifters in a flow generated by point vortices. Berliner and Wikle (2007) and Hoteit et al. (2008) explore theoretically the use of particle filters, but on identified low-dimensional manifolds of the dynamics, or on a reduced-order model of a large geophysical system [e.g., through empirical orthogonal functions (EOFs)].

c. Localizing strategies for particle filters

In the context of the fully Bayesian estimation problem, non-Gaussian uncertainty could be reduced by localization of the analyses. Indeed the smaller the area, the smaller the number of degrees of freedom to handle, the less complex (e.g., multimodal) the local pdf of these degrees of freedom should be. As a consequence, the smaller the area, the lower is the necessary number of particles for a given precision estimate. However, contrary to localization in the EnKF, the analyses cannot be simply glued together to get a complete set of updated global particles. That is why localization was not used on the particle filter in the illustrations of section 2. This issue has been largely discussed by van Leeuwen (2009).

d. Gaussian mixtures

In this approach, a background non-Gaussian pdf $p(\mathbf{x})$ is parameterized by Gaussian kernels as (Silverman 1986)

$$p(\mathbf{x}) \simeq \frac{1}{M} \sum_{i=1}^{M} n(\mathbf{x} - \mathbf{x}^i, \mathbf{P}), \qquad (28)$$

where n is the pdf of the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{P})$, of zero mean and covariance matrix $\mathbf{P}$. With Gaussian assumptions on observational errors, it turns out that, by applying Bayes's rule, the posterior pdf is a weighted mixture of Gaussian kernels (Anderson and Anderson 1999), leading to a convenient chaining of the assimilation cycles.

With a view to particle filtering, one could replace each particle of the ensemble by a broader Gaussian kernel. Unfortunately, for high-dimensional systems, this kernel representation also suffers from the curse of dimensionality (Silverman 1986). Recently proposed remedies are essentially filtering in the low-dimensional subspace related to the attractor of the complete system, as mentioned in the previous section. This can be implemented either by a localization and smoothing procedure (Bengtsson et al. 2003). Or it can be carried out thanks to a low rank representation of the error covariance matrix (Hoteit et al. 2008) inherited from the reduced rank Kalman filters. Note that in Bengtsson et al. (2003), the error covariance matrices associated with the kernels are not identical but generated by a Kalman filtering for each sample $\mathbf{x}^i$. Nevertheless, Gaussian mixture models are distinct from the unscented particle filter (Van der Merwe et al. 2000) where the sequential Monte Carlo sampling is performed according to locally linearized importance sampling functions given by the posterior pdf of Kalman filters for each of the particles (see section 6).

6. Bridging the gap between Gaussian and non-Gaussian data assimilation

There have been recent attempts to make use of non-Gaussian ideas in geophysical (or geophysically inspired) data assimilation. They remain quite specific in their application, because of their underlying hypotheses. They are, nevertheless, promising and a discussion on their relevance to geophysical data assimilation is presented. Contrary to section 3 where non-Gaussian errors were described and modeled mathematically, the emphasis here is on producing the analyses that cope with those non-Gaussian errors.

a. Statistical expansion about the climatology

In an ensemble-based filtering system, the estimates of the first- and second-order moments rely on the ensemble itself. Instead of forming a Gaussian pdf as a prior for the analysis using these statistics, one computes the pdf that is *closest* to the climatology and that has the same first- and second-order moments, as shown by Eyink and Kim (2006).

A measure of *distance* between a pdf p and a climatology q is provided by the relative entropy in Eq. (23). Let us assume that the observation operator $\mathbf{H}$ is linear. The minimization of Eq. (23) with respect to p, under the constraints of the first- and second-order statistics:

$$\hat{\mathbf{y}} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{H}\mathbf{x}_i \quad \text{and} \quad \hat{\mathbf{Y}} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{H}\mathbf{x}_i(\mathbf{H}\mathbf{x}_i)^{\mathrm{T}}, \quad (29)$$

where M is the ensemble size, yields the generic exponential solution:

$$p(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}) \propto q(\mathbf{x})\exp\left(\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{H}\mathbf{x} - \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{H}\mathbf{x}\right), \quad (30)$$

where the vector $\boldsymbol{\lambda} \in \mathbb{R}^d$ is the Lagrange parameter conjugated to the mean constraint, and the matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times d}$ is a Lagrange parameter matrix conjugated to the variance constraint. Next, inserting this exponential law in the relative entropy, one is led to the optimization on these dual parameters:

$$\overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\Lambda}} = \operatorname*{argmin}_{\boldsymbol{\lambda}, \boldsymbol{\Lambda}}\left\{\ln[Z(\boldsymbol{\lambda}, \boldsymbol{\Lambda})] - \boldsymbol{\lambda}^{\mathrm{T}}\hat{\mathbf{y}} + \frac{1}{2}\operatorname{Tr}(\hat{\mathbf{Y}}\boldsymbol{\Lambda})\right\}, \quad (31)$$

where $Z(\boldsymbol{\lambda}, \boldsymbol{\Lambda})$ is the partition function

$$Z(\boldsymbol{\lambda}, \boldsymbol{\Lambda}) = \sum_{\mathbf{x}} q(\mathbf{x})\exp\left(\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{H}\mathbf{x} - \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{H}\mathbf{x}\right), \quad (32)$$

with $\sum_{\mathbf{x}}$ an integration symbol that represents a sum when a discrete distribution is considered or an integral when the target space of the distribution is continuous.

Assuming Gaussian errors (observation $\mathbf{y}$ with error covariance matrix $\mathbf{R}$), the pdf is updated using Bayes's rule, within a dual framework. The dual parameters update reads as follows (in a fashion similar to the so-called *information* Kalman filter):

$$\boldsymbol{\lambda}^+ = \boldsymbol{\lambda}^- + \mathbf{R}^{-1}\mathbf{y} \quad \text{and} \quad \boldsymbol{\Lambda}^+ = \boldsymbol{\Lambda}^- + \mathbf{R}^{-1}. \quad (33)$$

This work has also been put forward by van Leeuwen (2009) in his review on particle filters. However, we do not consider the method to be a particle filter, since it involves the truncation of the moments to second order, and the most innovative part is the treatment of the

prior. Though the idea is very appealing, it remains to be proven that an attractor of the dynamics can be described analytically or numerically so that this information can be used in the method. Eyink and Kim (2006) tested their method on the Lorenz-63 model, using a mixture of two Gaussians to describe the two lobes of the attractor. The results show that the method eventually outperforms the EnKF, but in a regime where the filter is very nonlinear. This occurs when the time interval between two analyses reaches about $\Delta t = {}^2\!/_3$, possibly when the climatology starts having a significant impact on the filter trajectories. This might not reflect realistic conditions, since the time interval between two analyses in weather forecasting would rather correspond to $\Delta t = 0.05$.

b. Gaussian anamorphosis

One way to treat non-Gaussianities is to attempt to transform, analytically or numerically, non-Gaussian random variables into Gaussian ones, on which a BLUE-based analysis can appropriately be carried out.

1) ANALYTICAL TRANSFORMATION

If the observables and the state variables are positive, the errors are likely to be of multiplicative nature. Following section 3, a lognormal statistical modeling for these errors may be appropriate. Cohn (1997) has shown how to pass rigorously from the description of the variables with lognormal errors to a space where their statistics are Gaussian. The observation model is

$$\ln(\mathbf{y}) = \widetilde{H}[\ln(\mathbf{x})] + \ln(\mathbf{s}), \quad (34)$$

where $\widetilde{H}$ may be deduced from the original observation model H through $\widetilde{H} = \ln \circ H \circ \exp$. Symbol $\circ$ is the function composition operator. The change of state variable $\tilde{\mathbf{x}} \equiv \ln(\mathbf{x})$, enables the construction of a BLUE estimator in this *Gaussian space*:

$$\tilde{\mathbf{x}}^a = \tilde{\mathbf{x}}^b + \widetilde{\mathbf{K}}[\ln(\mathbf{y}) - \widetilde{H}(\tilde{\mathbf{x}}^b) - \mathbf{b}], \quad (35)$$

$$\widetilde{\mathbf{P}}^a = (\mathbf{I} - \widetilde{\mathbf{K}}\widetilde{\mathbf{H}})\widetilde{\mathbf{B}}, \quad (36)$$

where the bias $\mathbf{b}$ has been defined by Eq. (16) of section 3, and with the usual optimal gain, though in Gaussian space:

$$\widetilde{\mathbf{K}} = \widetilde{\mathbf{B}}\widetilde{\mathbf{H}}^{\mathrm{T}}(\widetilde{\mathbf{H}}\widetilde{\mathbf{B}}\widetilde{\mathbf{H}}^{\mathrm{T}} + \widetilde{\mathbf{R}})^{-1}. \quad (37)$$

Here $\widetilde{\mathbf{H}}$ is the tangent linear operator of $\widetilde{H}$, and $\widetilde{\mathbf{B}}$ and $\tilde{\mathbf{x}}^b$ are the background covariance matrix and first guess, respectively, in *Gaussian space*. One can pull the fields back into the original space and obtain the optimal estimators:

$$x_i^a = \exp\left(\tilde{x}_i^a + \frac{1}{2}\widetilde{R}_{ii}\right), \quad P_{ij}^a = \tilde{x}_i^a \tilde{x}_j^a[\exp(\widetilde{P}_{ij}^a) - 1]. \qquad (38)$$

One weak point of the approach is that the *Gaussian space* observation operator $\widetilde{\mathbf{H}}$ may become nonlinear, in particular if it was linear in the original space. For instance, the dispersion of a tracer is linear. Performing the analysis in Gaussian space would spoil this linearity.

The variational version of this analysis, essentially based on Eq. (17), was examined by Fletcher and Zupanski (2006), including thorough discussions on how to choose a proper estimator and how to precondition the minimization of a cost function such as Eq. (17). Mapping the lognormal errors to a Gaussian space is a particular (analytical) case of a *Gaussian anamorphosis*.

2) NUMERICAL TRANSFORMATIONS

When an analytical transformation to Gaussian space is not possible because the errors do not necessarily follow a lognormal behavior, then numerical methods can be used to achieve a similar goal. This is called a (numerical) Gaussian anamorphosis. This technique is well known in geostatistics (Wackernagel 2003). Its use has been advocated by Bertino et al. (2003) in the context of geophysical data assimilation (see also the next section). The idea of performing the analysis in the *Gaussian space* is the same as that of Cohn (1997), but for a general, albeit numerical transformation.

Consider one scalar random variable $X \sim P_X$, with values in the state space $\mathcal{E}_X$ distributed according to the density $p_X: x \in \mathcal{E}_X \rightarrow p_X(x)$. Its cumulative distribution function (cdf) is $F: x \in \mathcal{E}_X \rightarrow F(x) = P_X(X < x) \in [0, 1]$. Then consider a Gaussian random variable $\Gamma \sim P_\Gamma$, with values in the state space $\mathcal{E}_\Gamma$ distributed according to the density $p_\Gamma: \gamma \in \mathcal{E}_\Gamma \rightarrow p_\Gamma(\gamma)$. Its cdf is $G: \gamma \in \mathcal{E}_\Gamma \rightarrow G(\gamma) = P_\Gamma(\Gamma < \gamma) \in [0, 1]$. Provided F is invertible, then the anamorphosis function is defined by

$$\varphi : \gamma \in \mathcal{E}_\Gamma \rightarrow \varphi(\gamma) = F^{-1} \circ G(\gamma) \in \mathcal{E}_X. \qquad (39)$$

This deterministic map transforms a Gaussian random variable into a non-Gaussian one. The inverse mapping pulls the non-Gaussian variable back to a random Gaussian variable:

$$\varphi^{-1} : X \in \mathcal{E}_X \rightarrow \varphi^{-1}(X) = G^{-1} \circ F(X) \in \mathcal{E}_\Gamma. \quad (40)$$

There is a natural numerical counterpart to this analytical construct, the so-called empirical Gaussian anamorphosis. If one has n samples of X in $\mathcal{E}_X$ that can be ordered: $x_1 < x_2 < \cdots < x_n$ under the simplifying assumption that all x_i are distinct, then the empirical anamorphosis function $\varphi_n: \mathcal{E}_\Gamma \rightarrow \mathcal{E}_X$, is

$$\varphi_n = \frac{1}{n}\sum_{i=1}^{n} x_i I_{]G^{-1}((i-1)/n),G^{-1}(i/n)]}, \qquad (41)$$

where $I_{]a,b]}$ is the support function on interval $]a, b]$ (i.e., a excluded while b is included), equals to 1 on the interval, and 0 everywhere else. However, since φ_n is a stepwise function because of the discrete data, it is not invertible and needs to be smoothed out. One proper and convenient filtering is obtained by a truncated expansion of the empirical Gaussian anamorphosis on a basis of Hermite polynomials. Details can be found in (Wackernagel 2003).

In principle, a Gaussian anamorphosis is needed in both state space and observation space, then analysis equations similar to Eqs. (35) and (36) can be applied in Gaussian space. An inverse Gaussian anamorphosis is then built to pull the analyzed fields back into the original space.

It has recently been implemented on a large ocean and biogeochemical model by Simon and Bertino (2009) with success in a twin experiment. The transformation was applied to a chlorophyll field. Applying this methodology to such a large-scale experiment is not simple. As a first reasonable step, the authors neglected the correlations, and considered some climatological statistical univariate distributions of the non-Gaussian variables, when the anamorphosis is well defined and simple to implement. To take into account the full correlations, one would need to consider multivariate anamorphosis. With multivariate statistics, one would have to rotate the state space to get uncorrelated variables, by principal component analysis or independent component analysis (Hyvärinen and Oja 2000), and then apply Gaussian anamorphosis to each of the marginals. However, a non-Gaussian part of mutual information (in other words, residual correlations) remains in the rotated space (Pires and Perdigão 2007).

3) HUMIDITY TRANSFORM IN METEOROLOGICAL MODELS

Specific humidity q and relative humidity RH fields have intrinsically non-Gaussian distributions due to their finite interval supports. As a consequence, background and observational errors are also non-Gaussian by nature, also exhibiting largely inhomogeneous statistics both in latitude and height as well as presenting cross correlations with different control variables. To optimally apply a BLUE-based analysis, one has to find a proxy control humidity variable $\phi(\Phi)$ of some set Φ of the thermodynamic background variables (e.g., q, pressure p, temperature T), whose conditional background error $\epsilon_b|\Phi$ is at least approximately Gaussian. Hólm et al. (2002) have proposed

several control variables $\phi(\Phi)$ from the distribution of the corresponding forecast differences $\delta\phi$, extracted from observing system simulation experiments (OSSEs) performed with the ECMWF data assimilation system. Since $\delta\phi$ is a difference between background errors, the conditional pdf $p_{\delta\phi}(\delta\phi|\Phi)$ is the convolution of the conditional pdf $p_b(\epsilon_b|\Phi)$ of background errors ϵ_b:

$$p_{\delta_\phi}(\delta\phi|\Phi) = \int d\epsilon_b \, p_b(\epsilon_b|\Phi) p_b(\delta\phi - \epsilon_b|\Phi). \quad (42)$$

Therefore, Gaussian forecast differences $\delta\phi$ lead to a Gaussian ϵ_b of variance var(ϵ_b) = ½var$(\delta\phi)$ and a quadratic background log-likelihood function $J_b(\epsilon_b)$ with a single minimum and thus a simpler procedure of minimization. Given those advantages, one then aims to get at least quasi-Gaussian $\delta\phi$ values. For $\delta\phi$ equal to δq or $\delta \ln(q)$, one obtains approximately exponential distributions, whereas δRH is more closely Gaussian (Hólm et al. 2002). To get collected Gaussian statistics of $\delta\phi$ for all grid points together and Φ states, it is preferable to use the normalized forecast difference:

$$\widetilde{\delta\phi} = \frac{\delta\phi - b(\delta\phi|\Phi)}{\sigma(\delta\phi|\Phi)}, \quad (43)$$

where $b(\delta\phi|\Phi)$ and $\sigma(\delta\phi|\Phi)$ are, respectively, the bias and standard deviation of $\delta\phi$ conditioned on Φ. Thanks to the homoscedasticity (uniform standard deviation) of $\widetilde{\delta\phi}|\Phi$ and the central limit theorem, one achieves an overall Gaussian distribution of $\widetilde{\delta\phi}$, provided that the pdf of $\widetilde{\delta\phi}|\Phi$ is independent of Φ. An alternative to find a Gaussian control is through the *Gaussianization* procedure where one applies the (inverse) anamorphosis defined in the previous section to the forecast differences for all conditions Φ by the transform:

$$f(\delta\phi|\Phi) = G^{-1} \circ F(\delta\phi|\Phi), \quad (44)$$

where F is the cdf of $\delta\phi|\Phi$ and G is the Gaussian cdf. A standard Gaussian homogeneous control increment is readily obtained by the normalization of $f(\delta\phi|\Phi)$ with the corresponding bias and standard deviation, conditioned on Φ (Hólm 2007). This Gaussian anamorphosis applied to different humidity datasets, sometimes called *Hólm transform* in this context, has been shown to have a significant impact in the medium-range ECMWF weather forecasts (Andersson et al. 2007).

4) GAUSSIAN ANALYSES UNDER LINEAR INEQUALITY CONSTRAINTS

Some additional constraints may render a Gaussian data assimilation scheme non-Gaussian. This may happen when the prior forces the control variables to lie in a polytope (to satisfy linear inequalities), or when an observation error prior must account for outliers (as in section 2). If the unconstrained priors are Gaussian, then the constrained priors are truncated Gaussian priors. Remarkably, several Gaussian data assimilation schemes can be extended to the truncated Gaussian case in a mathematically rigorous way, with limited complications.

In variational data assimilation, Lagrangian duality (Borwein and Lewis 2000) can be used to lift these constraints, either on the observational errors (as made explicit in section 2), or in the state background errors (Bocquet 2008), through the use of Lagrange multipliers. The transformation is essentially exact if the cost functions are convex, a requirement that may not be satisfied if the models are nonlinear.

In ensemble-based Kalman filtering, filters can be extended to deal with linear inequalities. Assume the background is a truncated Gaussian, whereas the observation errors are normal. Then the analysis as seen from the Bayes's formula yields the product of a truncated Gaussian by a Gaussian, which is in turn a truncated Gaussian. Besides, the analysis uses the same set of operators (such as the Kalman gain) as in the unconstrained case. This makes the use of such a scheme very practical. The major change comes from the need to sample from a truncated Gaussian, which is not straightforward for high-dimensional problems. The truncated Kalman filter was developed by Lauvernet et al. (2009) in a geophysical context and successfully tested on a one-dimensional mixed layer ocean model.

c. Using non-Gaussian deviations in the priors to improve analysis

Given the sampling statistics of innovations, it is possible to compute the mean and covariance matrix of the innovation vector. This is useful to correct error biases and for tuning the prescribed error covariance matrices (Desroziers et al. 2005). Beyond those statistics, innovation histograms and higher-order moments of the innovations can also be computed, for instance some measures of non-Gaussianity like the skewness s_d and kurtosis k_d. Pires et al. (2010, manuscript submitted to *Physica D*) have computed diagnostics of s_d, k_d for the quality-controlled ECMWF innovations of brightness temperatures of a set of HIRS channels. Their results emphasized the statistically significant non-Gaussianity of the errors in several channels. They estimate a joint non-Gaussian prior pdf for the observations errors ϵ_o and background errors ϵ_b in the observation space, using the maximum entropy on the mean (MEM) method. The method follows the same principle as the one used and exemplified in sections 6a and 6e. The output of the method is a pdf of ϵ_o and ϵ_b, compatible with the

prescribed innovation statistics. Moreover, it is minimally committed in the sense that, from the information theory point of view, it is the simplest pdf (with minimal extra information), that explains the prescribed statistics. This prior modeling can be shown to be beneficial to the subsequent analyses that go beyond the BLUE result.

d. Particle filtering with Gaussian filters as importance proposals

We come back to the ideas of the particle filter. We will see how Gaussian analyses can help to numerically solve the Bayesian estimation problem. The ideas are fairly recent in the geophysical community and the extrapolation to complex systems is speculative.

1) CONCEPT OF IMPORTANCE SAMPLING

Within the smoothing approach, an empirical representation of the pdf of trajectories $\mathbf{X}_k = \{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ conditional on the collection of observations $\mathbf{Y}_k = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k\}$ is considered, up to time t_k. For any time index $k = 1, \ldots, K$:

$$p_k(\mathbf{X}_k|\mathbf{Y}_k) \simeq \sum_{i=1}^{M} \omega_k^i \delta(\mathbf{X}_k - \mathbf{X}_k^i) \quad \text{with} \quad \sum_{i=1}^{M} \omega_k^i = 1. \quad (45)$$

This representation is a combination of M particles' trajectories $\mathbf{X}_k^i = \{\mathbf{x}_0^i, \mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_k^i\}$ and weights ω_k^i attached to each one of them. One has the freedom to draw the particle trajectories from a known *proposal* pdf q_k. These particles can have any distribution, provided the support of q_k includes that of p_k. In order for the particle filter to still solve the Bayesian problem, the weights need to be corrected so that the discrete pdf is still representative of the system pdf:

$$\omega_k^i \propto \frac{p_k(\mathbf{X}_k^i|\mathbf{Y}_k)}{q_k(\mathbf{X}_k^i|\mathbf{Y}_k)}. \quad (46)$$

In the Monte Carlo methods literature, this is known as importance sampling (Doucet et al. 2001).

There is a more practical sequential version of importance sampling. Assuming Markovian dynamics, and conditional dependence of the observation on the current state only, one factorizes $p_k(\mathbf{X}_k|\mathbf{Y}_k)$ according to

$$p_k(\mathbf{X}_k|\mathbf{Y}_k) \propto p_k(\mathbf{y}_k|\mathbf{x}_k)p_k(\mathbf{x}_k|\mathbf{x}_{k-1})p_{k-1}(\mathbf{X}_{k-1}|\mathbf{Y}_{k-1}). \quad (47)$$

If, in addition, one assumes that the proposal distribution is obtained by filtering (not smoothing, i.e., the state pdf depends only on current and past observations), then one relates the importance proposal at successive times according to

$$q_k(\mathbf{X}_k|\mathbf{Y}_k) = q_k(\mathbf{x}_k|\mathbf{X}_{k-1}, \mathbf{Y}_k)q_{k-1}(\mathbf{X}_{k-1}|\mathbf{Y}_{k-1}). \quad (48)$$

Thus, in a sequential context, the recursion formula for the weights reads as follows:

$$\omega_k^i \propto \omega_{k-1}^i \frac{p_k(\mathbf{y}_k|\mathbf{x}_k^i)p_k(\mathbf{x}_k^i|\mathbf{x}_{k-1}^i)}{q_k(\mathbf{x}_k^i|\mathbf{X}_{k-1}^i, \mathbf{Y}_k)} \quad \text{with} \quad \sum_{i=1}^{M} \omega_k^i = 1. \quad (49)$$

The importance proposal pdf of the bootstrap filter is the transition operator of the model: if the proposal is $q_k(\mathbf{x}_k|\mathbf{X}_{k-1}, \mathbf{Y}_k) \equiv p_k(\mathbf{x}_k|\mathbf{x}_{k-1})$ then the bootstrap filter is recovered, since $\omega_k^i \propto \omega_{k-1}^i p_k(\mathbf{y}_k|\mathbf{x}_k^i)$. Another particular choice, the proposal $q_k \equiv p_k(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{y}_k)$, minimizes the variance of the weights. It is called the optimal importance function (Doucet et al. 2000).

In our opinion, the same importance sampling principle can be used to justify two attempts to improve particle filtering from the geophysicist point of view. Xiong et al. (2006) make use of a Gaussian resampling, based on the particles' first- and second-order moments. It was shown to improve the forecast ability of the particle filter, often beating the EnKF, in the case of the Lorenz-63 model. The merging particle filter of Nakano et al. (2007) is, in a similar flavor, a Gaussian resampling (matching of first- and second-order moments) and is used to enrich the sampling of the particle filter. The authors demonstrate a significant improvement with the Lorenz-63 and Lorenz-95 models, but the particle filter still necessitates too many particles as compared to the EnKF, even on these toy models. Although it is not stated in those words, these papers illustrate the use of Gaussian hypotheses on rigorously non-Gaussian estimation, through the use of importance sampling. However, to our knowledge, the necessary correction to the weights for the particle filters, in order to guarantee the proper Bayesian asymptotics, were not computed.

Now we come back to the problem of the collapse of the particle filter. The basic idea that was fostered in the applied mathematics community, and advocated by van Leeuwen (2009) in geophysics, is that in order to avoid too unlikely trajectories, particles should be drawn at time t_{k-1} from a proposal making use of $\mathbf{y}_k$. For high-dimensional applications and complex models, this is certainly not trivial to implement. The following section gives clues, and original numerical examples based on this idea.

2) CURRENT-OBSERVATION-DEPENDENT PROPOSAL WITH GAUSSIAN ANALYSES

Let us give an example of a fully Bayesian filter that makes use of local Gaussian analyses. As an original contribution, let us resort to the optimal importance function $q_k \equiv p_k(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{y}_k)$. Its implementation is only practical (and without approximation) when the observation operator is linear (Doucet et al. 2000). Let us use a data assimilation system of the form in Eq. (2), with

$\mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$ and $\boldsymbol{v}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k)$, and a linear observation operator $\mathbf{H}_k$. Even though the local error assumptions are Gaussian, the filter is meant to account for any nonlinear dynamical model. For each particle, a simple BLUE analysis is carried out which strikes the balance between the uncertainty $\mathbf{R}_k$ of the observation $\mathbf{y}_k$ and the uncertainty of the model $\mathbf{Q}_k$ at time t_k. Therefore, the particles are propagated according to the pdf:

$$q_k(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{y}_k) \equiv p_k(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{y}_k) = n(\mathbf{x}_k - \mathbf{x}_k^a, \boldsymbol{\Sigma}_k), \quad (50)$$

with $\boldsymbol{\Sigma}_k^{-1} = \mathbf{Q}_k^{-1} + \mathbf{H}_k^{\mathrm{T}}\mathbf{R}_k^{-1}\mathbf{H}_k$, $\mathbf{x}_k^a = M_k(\mathbf{x}_{k-1}) + \mathbf{K}_k[\mathbf{y}_k - \mathbf{H}_k M_k(\mathbf{x}_{k-1})]$, and $\mathbf{K}_k = \boldsymbol{\Sigma}_k \mathbf{H}_k^{\mathrm{T}}\mathbf{R}_k^{-1}$. Here $n(\mathbf{x}_k - \mathbf{x}_k^a, \boldsymbol{\Sigma}_k) \propto \exp[-\frac{1}{2}(\mathbf{x}_k - \mathbf{x}_k^a)^{\mathrm{T}}(\boldsymbol{\Sigma}_k)^{-1}(\mathbf{x}_k - \mathbf{x}_k^a)]$ is the pdf of $\mathcal{N}(\mathbf{x}_k^a, \boldsymbol{\Sigma}_k)$, the multivariate Gaussian distribution with mean x_k^a and covariance matrix $\boldsymbol{\Sigma}_k$. The updating recursive law on the weights is simply

$$\omega_k^i \propto \omega_{k-1}^i p_k(\mathbf{y}_k|\mathbf{x}_{k-1}^i), \quad (51)$$

and, on the previous assumptions, can be computed explicitly since it is Gaussian:

$$p_k(\mathbf{y}_k|\mathbf{x}_{k-1}^i) = n(\mathbf{y}_k - \mathbf{H}_k M_k(\mathbf{x}_{k-1}^i), \mathbf{R}_k + \mathbf{H}_k \mathbf{Q}_k \mathbf{H}_k^{\mathrm{T}}). \quad (52)$$

Contrary to the bootstrap filter, only particles with a reasonable likelihood given the observations will be sampled. We have tested this optimal importance sampling particle filter (OISPF) on the Lorenz-95 model, using the exact same setup as in section 2, but with the original $N = 40$ system. Performance tests that compare the merits of the EnKF, the bootstrap filter, and the OISPF are reported in Fig. 5. The OISPF improvement is spectacular for moderate ensemble size, as compared to the bootstrap filter, but insufficient beyond.

In the derivation of the proposal Eq. (50), the uncertainty of the particle position is given by the system noise that is employed to enrich the ensemble. It would be preferable to use the absolute uncertainty of the particle position, which could be estimated from a Gaussian filter. Let $\bar{\mathbf{x}}_k$ and $\mathbf{P}_k$ be the analyzed state and analysis error covariance matrix, respectively, of the Gaussian filter: either extended Kalman filter, unscented Kalman filter, ensemble Kalman filter, or ensemble transform Kalman filter, etc. The proposal would then be of the following form:

$$q_k(\mathbf{x}_k|\mathbf{X}_{k-1}, \mathbf{Y}_k) \equiv n(\mathbf{x}_k - \bar{\mathbf{x}}_k, \mathbf{P}_k). \quad (53)$$

This leads to other variants of particle filter with a guiding BLUE-based proposal.

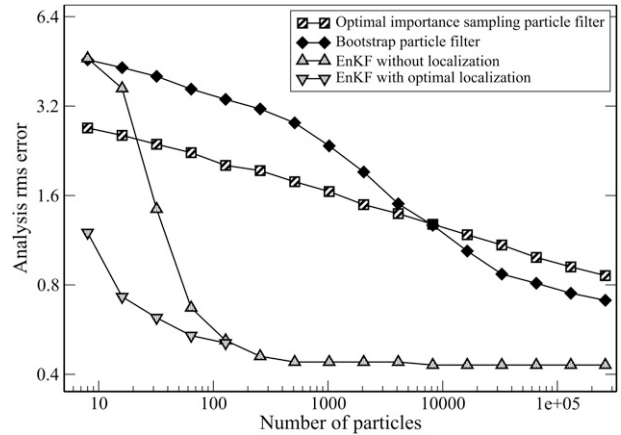Van der Merwe et al. (2000) have built a particle filter with importance sampling given by (several)

FIG. 5. Comparison of the performance of three filters (the EnKF, the bootstrap particle filter, and the OISPF) on a Lorenz-95 model with $N = 40$ variables. For each analysis, the parameters of the filter (essentially model noise) are optimally tuned as reported in Fig. 1. This is necessary, as it is deceiving to penalize one of the filters with inadequate choices. Two versions of the EnKF are tested: with or without localization. The localization length is optimally tuned. Optimally tuned localization only makes a difference for small ensemble sizes (i.e., below 100 members).

Kalman-based filters, which track the system. The idea is to attach a Kalman filter to each particle, and there is therefore a proposal for each of the particles:

$$q_k(\mathbf{x}_k^i|\mathbf{X}_{k-1}^i, \mathbf{Y}_k) \equiv n(\mathbf{x}_k^i - \bar{\mathbf{x}}_k^i, \mathbf{P}_k^i). \quad (54)$$

Since the proposal distribution is not unique, this scheme is a suboptimal importance particle filter. Although this is suitable for signal theory problems and was shown to outperform significantly standard Kalman filtering, this may be much too expensive for geophysical applications. That is why, in the following, the methods to be introduced perform analysis using a single BLUE-based filter.

Let us narrow our choice of proposal to the Gaussian filters that are built on an ensemble. For each particle, one could consider a BLUE analysis based on the observation uncertainty, and on the uncertainty attached to the particle positions, estimated through the empirical covariance matrix $\mathbf{P}_k^f$ as in any ensemble-based method. The proposal of Eq. (49) is

$$q_k(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{y}_k) \equiv n(\mathbf{x}_k - \{M_k(\mathbf{x}_{k-1}) + \mathbf{K}_k[\mathbf{y}_k - \mathbf{H}_k M_k(\mathbf{x}_{k-1})]\}, \boldsymbol{\Sigma}_k), \quad (55)$$

but with $\mathbf{K}_k$ now given by Eq. (22), and $\boldsymbol{\Sigma}_k$ given by

$$\boldsymbol{\Sigma}_k = \mathbf{K}_k \mathbf{R}_k \mathbf{K}_k^{\mathrm{T}} + (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{Q}_k (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)^{\mathrm{T}}. \quad (56)$$

Contrary to the OISPF, there is an implicit approximation in the use of $\mathbf{P}_k^f$ in $\mathbf{K}_k$, as it uses the positions of all particles. Making use of this analysis, Papadakis (2007) proposed to use an ensemble-based Kalman filter as a *channeling* Gaussian filter. That way, he can define a (tentatively) true particle filter (meaning with a proper Bayesian asymptotic limit) with weights attached to each member of the ensemble. After the analysis, the weights are updated thanks to Eq. (55), and the ensemble is possibly resampled. The forecast that follows the analysis is the typical step of an ensemble Kalman filter, using the particles of the filter as ensemble members. The only difference with the EnKF lies in the weights that must be taken into account in the estimation of the mean and error covariance matrix:

$$\bar{\mathbf{x}}_{k+1} = \sum_{i=1}^{M} \omega_k^i \mathbf{x}_{k+1}^i \quad \text{and}$$

$$\mathbf{P}_{k+1}^f = \frac{1}{1 - \sum_{i=1}^{M} (\omega_k^i)^2} \sum_{i=1}^{M} \omega_k^i (\mathbf{x}_{k+1}^i - \bar{\mathbf{x}}_{k+1})(\mathbf{x}_{k+1}^i - \bar{\mathbf{x}}_{k+1})^{\mathrm{T}},$$

$$\tag{57}$$

where $\sum_{i=1}^{M} \omega_k^i = 1$. Although, practically, it has the flavor of an ensemble Kalman filter, it is meant to be a particle filter, with the Bayesian asymptotics, the elegance and the potential pitfall that comes with it. Consequently, this weighted ensemble Kalman filter (WEnKF) may offer a rigorous framework to implement localization in a particle filter via its EnKF proposal.

Two words of caution are in order about the WEnKF. Firstly, the particles are interacting (following the terminology of Del Moral 2004), not only through standard resampling, but also through the estimation of the covariance matrix. Second, as a Gaussian pdf the proposal function has strongly vanishing tails. Therefore, on one hand, the WEnKF should be very efficient in the regime where the EnKF outperforms particle filters as compared to other particle filters. On the other hand, it may be weaker in a regime where simpler particle filters outperform the EnKF. Thus, we believe that the overall interest in the WEnKF is debatable, even though very appealing.

e. Nonperturbative non-Gaussian methods for high-dimensional linear models

There are relevant geophysical cases where the models are approximately linear. But the priors may be intrinsically different from Gaussians. This is the case for tracer transport, radionuclides dispersion, dust, several greenhouse gases including CO_2, etc. In that context (i.e., linear models and non-Gaussian priors), and unlike previous examples, a non-Gaussian analysis can be performed thoroughly without approximation, using nonlinear convex analysis (see Borwein and Lewis 2000). The theory is based on nonquadratic cost-functions that generalize 4D-Var and the Physical space Statistical Analysis System (PSAS) in the specific linear model case (Bocquet 2005b,c, 2007; Krysta and Bocquet 2007; Bocquet 2008).

Let us assume that the system is driven by a forcing field or an initial condition $\mathbf{x} \in \mathbb{R}^N$ with model/observation error $\boldsymbol{v} \in \mathbb{R}^d$:

$$\mathbf{y} = \mathbf{Hx} + \boldsymbol{v}, \tag{58}$$

where $\mathbf{H} \in \mathbb{R}^{d \times N}$ is the Jacobian that combines the observation operator and the model linking the observations to the forcing field or initial condition. The joint prior pdf in control space and in observation error space is $\nu(\mathbf{x}, \boldsymbol{v})$. Finding the posterior pdf $p(\mathbf{x}, \boldsymbol{v})$ is the goal of the analysis in this framework. However, the object $p(\mathbf{x}, \boldsymbol{v})$, might be too complex and of difficult interpretation. That is why an estimator must be chosen that extracts some precise information from the full pdf mode also known as maximum a posteriori (MAP), the mean value estimator, etc.

1) BAYESIAN INFERENCE AND MAXIMUM A POSTERIORI

As an inference rule, one can apply Bayes's theorem using Eq. (58), and straightforwardly derive a variational principle with the following cost function (the primal cost function):

$$\mathcal{L}(\mathbf{x}) = -\ln \nu(\mathbf{x}, \mathbf{y} - \mathbf{Hx}). \tag{59}$$

The MAP estimator of the full pdf is obtained by minimizing this cost function over $\mathbf{x}$. In a Gaussian context, this would give back the usual 4D-Var cost function. The dual cost function

$$\widehat{\mathcal{L}}(\boldsymbol{\lambda}) = [-\ln(\nu)]^*(\mathbf{H}^{\mathrm{T}}\boldsymbol{\lambda}, \boldsymbol{\lambda}) - \mathbf{y}^{\mathrm{T}}\boldsymbol{\lambda}, \tag{60}$$

could be equivalently solved. The $\boldsymbol{\lambda} \in \mathbb{R}^d$ are the Lagrange multipliers that enforce Eq. (58). Here ζ^* is the Legendre-Fenchel conjugate of ζ, and is defined by

$$\zeta^*(\boldsymbol{\alpha}) = \sup_{\mathbf{x}} [\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x} - \zeta(\mathbf{x})]. \tag{61}$$

If ν is Gaussian, one obtains the PSAS formalism (Courtier 1997; Cohn et al. 1998).

However, such a dual equivalence is possible only if $\mathcal{L}$ and $\widehat{\mathcal{L}}$ are both convex, which is an obstacle to a non-Gaussian generalization of the dual formalism (see Auroux 2007).
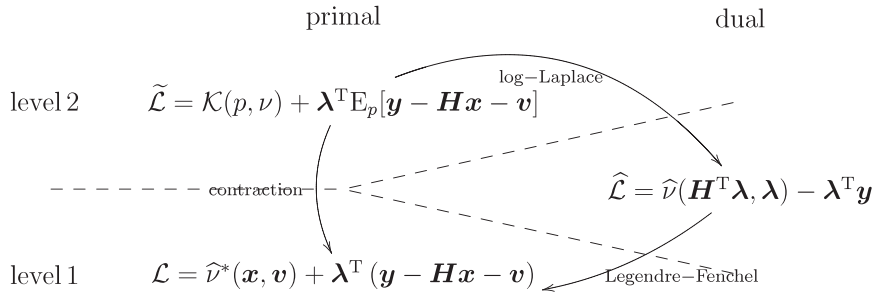
primal   dual

level 2 $\widetilde{\mathcal{L}} = \mathcal{K}(p, \nu) + \boldsymbol{\lambda}^{\mathrm{T}}\mathrm{E}_p[\boldsymbol{y} - \boldsymbol{Hx} - \boldsymbol{v}]$

log−Laplace

                                    $\widehat{\mathcal{L}} = \widehat{\nu}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{\lambda}, \boldsymbol{\lambda}) - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{y}$

− − − − − − contraction − < −

level 1 $\mathcal{L} = \widehat{\nu}^*(\boldsymbol{x}, \boldsymbol{v}) + \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{Hx} - \boldsymbol{v})$

Legendre−Fenchel

FIG. 6. Relations between the primal and dual cost functions.

2) MAXIMUM ENTROPY ON THE MEAN INFERENCE

A second type of non-Gaussian inference is given by the maximum entropy on the mean (MEM) principle, which is also considered to be of Bayesian nature. In a linear model context, it allows one to exploit the machinery of nonlinear convex analysis with efficiency. The inference is again formulated in terms of the prior pdf ν and of the posterior pdf p. The optimal p is the one that minimizes the gain of information (maximizes the entropy), except for what is gained from the observations. Since this gain of information is objectively measured by the Kullback–Leibler divergence, the related cost function (often called level 2 primal cost function) is

$$\widetilde{\mathcal{L}}(p) = \mathcal{K}(p, \nu) + \boldsymbol{\lambda}^{\mathrm{T}}\mathrm{E}_p[\mathbf{y} - \mathbf{Hx} - \boldsymbol{v}], \qquad (62)$$

where $\mathcal{K}$ has been defined in Eq. (23), $\mathrm{E}_p[\cdot] = \Sigma_{\mathbf{x},\boldsymbol{v}}\, p(\mathbf{x}, \boldsymbol{v})\cdot$ is the expectation operator, and $\Sigma_{\mathbf{x},\boldsymbol{v}}$ is a symbol for a sum (discrete variables) or an integral (continuous variables). The constraint that p should satisfy the observation on the mean is enforced through Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^d$. The direct optimization of $\widetilde{\mathcal{L}}(p)$ on p leads to the intermediary:

$$p_{\boldsymbol{\lambda}}(\mathbf{x}, \boldsymbol{v}) \propto \nu(\mathbf{x}, \boldsymbol{v}) \exp[\boldsymbol{\lambda}^{\mathrm{T}}(\mathbf{Hx} + \boldsymbol{v})], \qquad (63)$$

up to the normalization constant of the pdf $p_{\boldsymbol{\lambda}}$. By inserting $p_{\boldsymbol{\lambda}}$ into Eq. (62), one obtains the dual cost function [similarly to Eq. (60)], which depends on $\boldsymbol{\lambda}$:

$$\widehat{\mathcal{L}}(\boldsymbol{\lambda}) = \widehat{\nu}(\mathbf{H}^{\mathrm{T}}\boldsymbol{\lambda}, \boldsymbol{\lambda}) - \mathbf{y}^{\mathrm{T}}\boldsymbol{\lambda}, \qquad (64)$$

where $\widehat{\nu}$ is the log-Laplace transform of ν, and is defined by

$$\widehat{\nu}(\boldsymbol{\alpha}, \boldsymbol{\delta}) = \ln\left[\sum_{\mathbf{x},\boldsymbol{v}} \nu(\mathbf{x}, \boldsymbol{v}) \exp(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x} + \boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{v})\right]. \qquad (65)$$

By evaluating the pdf in Eq. (63) at the the minimum $\overline{\boldsymbol{\lambda}}$ of the cost function Eq. (64), one obtains the posterior pdf $p_{\overline{\boldsymbol{\lambda}}}(\mathbf{x}, \boldsymbol{v})$ from which the most sensible estimators that can be derived are the state mean $\overline{\mathbf{x}}$ and error mean $\overline{\boldsymbol{v}}$.

There is an equivalent cost function formulated in state (or control) space *level-1* primal cost function:

$$\mathcal{L}(\mathbf{x}) = \widehat{\nu}^*(\mathbf{x}, \mathbf{y} - \mathbf{Hx}). \qquad (66)$$

Contrary to the strictly Bayesian and MAP inference, the cost functions in Eqs. (62), (64), and (66) are always convex by construction because of the convexity of the Kullback–Leibler functional on a vector subspace, which was mentioned in section 4. Therefore, these cost functions have always a single minimum, and the primal and dual cost functions are always equivalent (there is no gap in between the minimum of $\mathcal{L}$ and the maximum of $-\widehat{\mathcal{L}}$):

$$\mathcal{L}(\overline{\mathbf{x}}) = -\widehat{\mathcal{L}}(\overline{\boldsymbol{\lambda}}). \qquad (67)$$

This constitutes a non-Gaussian generalization of the duality 4D-Var/PSAS when models are linear. Consistently, when ν is Gaussian, 4D-Var and PSAS cost functions for linear assumptions are recovered. The schematic of the different transformations and equivalences is displayed in Fig. 6. Note that the primal part of the formulation can be generalized to nonlinear models, but the duality correspondence is not valid any more.

An application to the forecast of an accidental plume of pollutant is given in the context of the European tracer experiment (Nodop et al. 1998) in Fig. 7. About 10^3 observations are used for 2×10^4 control variables. The analyses are hence conveniently carried out in the observation space. The plume contours obtained by the MEM method are much finer than with 4D-Var, which is of utmost importance for such a dispersion event, especially in the regions where the concentration field exhibits strong gradients.

The strictly Bayesian solution of the previous section is different from the MEM solution: the exponential pdf in Eq. (63) is generally not the posterior pdf obtained from Bayes's rule. Instead, the MEM method convexifies the objective function that would be obtained from Bayes's rule. Therefore, if the existence of multiple minima matters in the problem, the MEM approach may differ significantly from the strictly Bayesian solution. However,
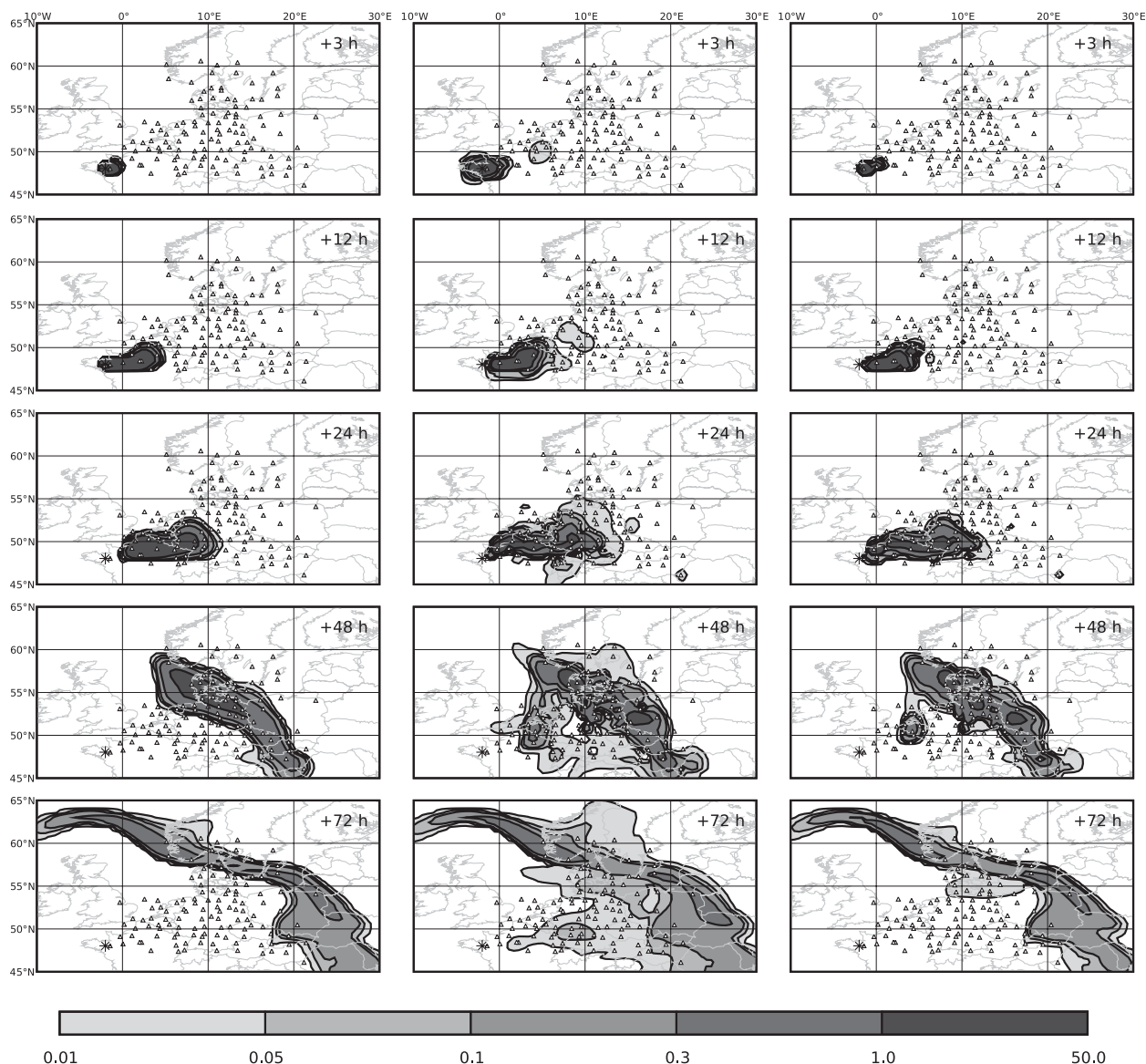
FIG. 7. The asteriks give the true source location, while the triangles indicate the monitoring stations: (top to bottom) time $= +3$ to 72 h. (left) The reference tracer dispersion event European Tracer Experiment, first campaign (ETEX-I) simulated assuming perfect knowledge of the source. Since the source is the main uncertainty beyond the uncertainty in the meteorological fields, this reference simulation can be considered quite close to the truth. (middle) Gaussian-based 4D-Var with sequentially assimilated real observations, not knowing the source. (right) Data assimilation using a non-Gaussian prior and the MEM formalism, not knowing the source. Units are ng m^{-3} of perfluoromethylcyclohexane.

the analysis of multiple minima in geophysical applications is rather speculative. This would likely happen when considering a strongly constrained 4D-Var, with a sufficiently large assimilation window. Rather than facing such a multiple minima optimization problem, it is tempting to convexify anyway the objective function by, for instance, incorporating model error (weakly constrained 4D-Var). The problem of convexification is the arbitrariness of the regularization of penalty functions, which can result in an unlikely solution. However, for pollutant source

reconstruction problems, Bocquet (2008) has shown that the difference between the two approaches is small.

So far, this discussion was relevant for the state estimation problem. For second-order and higher-order moments, the MEM method would not give the correct estimates even in the Gaussian prior case. To circumvent this drawback of the MEM method, Bocquet (2008) showed that correctly defined moments of the MEM inference with a prior ν could be obtained from the strictly Bayesian inference but using the prior $\exp[-(\hat{\nu})^*]$.

Within this extension of the MEM method, a precise correspondence is defined between the two approaches.

3) SECOND-ORDER SENSITIVITY ANALYSIS

When applicable, the MEM method has several technical advantages. One of them is that a second-order sensitivity analysis can be performed analytically. Second-order analysis allows one to compute the sensitivity of an analysis to any parameter (Le Dimet et al. 1997). It usually implies the choice of a sensitivity scalar function to focus on a limited number of relevant degrees of freedom. Building on the properties of the MEM inference, Bocquet (2008) chooses the information gain in the analysis as a sensitivity function. In this context, the information gain $\mathcal{K}_{\mathbf{x}}$ identifies with the part of the cost function attached to the background departure. If the prior pdf splits according to $\nu(\mathbf{x}, \boldsymbol{v}) = \nu_{\mathbf{x}}(\mathbf{x})\nu_{\boldsymbol{v}}(\boldsymbol{v})$, the total cost function splits according to

$$\mathcal{K}_{\mathbf{x},\boldsymbol{v}} = \mathcal{K}_{\mathbf{x}} + \mathcal{K}_{\boldsymbol{v}}. \qquad (68)$$

Here $\mathcal{K}_{\mathbf{x}}$, $\mathcal{K}_{\boldsymbol{v}}$ can be either analytically or numerically computed following for instance in the case of $\mathcal{K}_{\mathbf{x}}$:

$$\mathcal{K}_{\mathbf{x}} = \sum_{\mathbf{x}} p_{\mathbf{x}}(\mathbf{x}) \ln\left[\frac{p_{\mathbf{x}}(\mathbf{x})}{\nu_{\mathbf{x}}(\mathbf{x})}\right] = \underbrace{\overline{\mathbf{x}}^{\mathrm{T}} \mathbf{H}^{\mathrm{T}} \overline{\boldsymbol{\lambda}} - \hat{\nu}_{\mathbf{x}}(\mathbf{H}^{\mathrm{T}}\overline{\boldsymbol{\lambda}})}_{\text{dual form}} = \underbrace{\hat{\nu}_{\mathbf{x}}^{*}(\overline{\mathbf{x}})}_{\text{primal form}} . \qquad (69)$$

Note that in the Gaussian case, one recovers the natural splitting of cost functions into signal and noise degrees of freedom (Rodgers 2000). In this non-Gaussian context, each one of these $\mathcal{K}$ values is relative entropy. The derivative of this splitting with respect to the measurements gives

$$\overline{\boldsymbol{\lambda}} = \partial_{\mathbf{y}}\mathcal{K}_{\mathbf{x}} + \partial_{\mathbf{y}}\mathcal{K}_{\boldsymbol{v}}. \qquad (70)$$

Here $\partial_{\mathbf{y}}\mathcal{K}_{\mathbf{x}}$ is interpreted as the marginal gain in information induced by the variation of the measurements.

This non-Gaussian second-order analysis is illustrated with an original experiment on the inversion of the Chernobyl accident radionuclides source term. The physics is essentially linear so that the methodology applies without approximation. The positivity of the source term requires a non-Gaussian background error prior modeling. Figure 8 illustrates several sensitivities in both Gaussian and non-Gaussian analysis cases. The global marginal gain of information $\partial_{\mathbf{y}}\mathcal{K}_{\mathbf{x},\boldsymbol{v}} = \overline{\boldsymbol{\lambda}}$ depends little on the a priori hypotheses, Gaussian or not. However, the marginal gain of information $\partial_{\mathbf{y}}\mathcal{K}_{\mathbf{x}}$ that goes into the reconstruction of the source (the signal part) significantly depends on the statistical modeling of the prior. For instance, the Scandinavian observations are relatively less significant in the non-Gaussian case than in the Gaussian case. Put differently, the information content of remote observations is stronger in the non-Gaussian case. This may be due to the positivity constraint that rules out the sources with both negative and positive rates that are more compatible with these remote observations.

4) SCORE

Let us assume the prior statistics are Gaussian $\mathbf{x} \sim \mathcal{N}(\mathbf{x}^{b}, \mathbf{B})$ and $\boldsymbol{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$. One indicator that measures the reduction of uncertainty carried out in an analysis on Eq. (58) is given by the following ratio:

$$\rho = 1 - \frac{\|\mathbf{x}^{t} - \overline{\mathbf{x}}\|_{\mathbf{B}^{-1}}^{2} + \|\boldsymbol{v}^{t} - \overline{\boldsymbol{v}}\|_{\mathbf{R}^{-1}}^{2}}{\|\mathbf{x}^{t} - \mathbf{x}^{b}\|_{\mathbf{B}^{-1}}^{2} + \|\boldsymbol{v}^{t}\|_{\mathbf{R}^{-1}}^{2}}$$

$$= \frac{\|\overline{\mathbf{x}} - \mathbf{x}^{b}\|_{\mathbf{B}^{-1}}^{2} + \|\overline{\boldsymbol{v}}\|_{\mathbf{R}^{-1}}^{2}}{\|\mathbf{x}^{t} - \mathbf{x}^{b}\|_{\mathbf{B}^{-1}}^{2} + \|\boldsymbol{v}^{t}\|_{\mathbf{R}^{-1}}^{2}}, \qquad (71)$$

where $\mathbf{x}^{t}$ and $\boldsymbol{v}^{t}$ are the true state and error vector, respectively. The norm $\|\cdot\|_{\mathbf{A}}$, with $\mathbf{A}$ a positive definite matrix, is defined by $\|\mathbf{x}\|_{\mathbf{A}}^{2} = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$. The transformation from the second to the third member is due to the Pythagorean theorem. The analysis $(\overline{\mathbf{x}}, \overline{\boldsymbol{v}})$ in $\mathbb{R}^{N} \oplus \mathbb{R}^{p}$ is the orthogonal projection of $(\mathbf{x}^{b}, \mathbf{0})$ (with respect to the scalar product defined by $\mathbf{B}^{-1} \oplus \mathbf{R}^{-1}$) on the hyperplane of couples $(\mathbf{x}, \boldsymbol{v})$ such that $\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{v}$. This is equivalent to the geometrical interpretation of the analysis in terms of projection using the Mahalanobis norm as a scalar product (Desroziers et al. 2005; Chapnik et al. 2006). This ensures that $0 \leq \rho \leq 1$, with $\rho = 1$ when the reduction of uncertainty is maximum.

Bocquet (2005a) has generalized without approximation this score to non-Gaussian priors provided that the analysis is performed using the MEM principle. In a condensed form, the generalization of Eq. (71) is

$$\rho = \frac{\mathcal{K}(p_{\overline{\mathbf{x}},\overline{\boldsymbol{v}}}, \nu)}{\mathcal{K}(p_{\mathbf{x}_{t},\boldsymbol{v}_{t}}, \nu)}. \qquad (72)$$

Here $p_{\overline{\mathbf{x}},\overline{\boldsymbol{v}}}$ and $p_{\mathbf{x}_{t},\boldsymbol{v}_{t}}$ are the pdfs, belonging to the exponential family Eq. (63), whose state and errors averages are $\overline{\mathbf{x}}$ and $\overline{\boldsymbol{v}}$ in the former case, and $\mathbf{x}_{t}$ and $\boldsymbol{v}_{t}$ in the latter case. The Kullback–Leibler terms, $\mathcal{K}(p_{\overline{\mathbf{x}},\overline{\boldsymbol{v}}}, \nu)$ and $\mathcal{K}(p_{\overline{\mathbf{x}}_{t},\overline{\boldsymbol{v}}_{t}}, \nu)$, are expressed here in their abstract (level 2) form, but they can numerically be estimated using their primal form in Eq. (66). In particular, $\mathcal{K}(p_{\overline{\mathbf{x}},\overline{\boldsymbol{v}}}, \nu)$ often identifies with the minimum of the objective function, and is just a numerical by-product of a data assimilation variational scheme. In the Gaussian case and with independent background and error priors, Eq. (72) simplifies to Eq. (71). This result was useful in the evaluation of a European radionuclides monitoring network whose
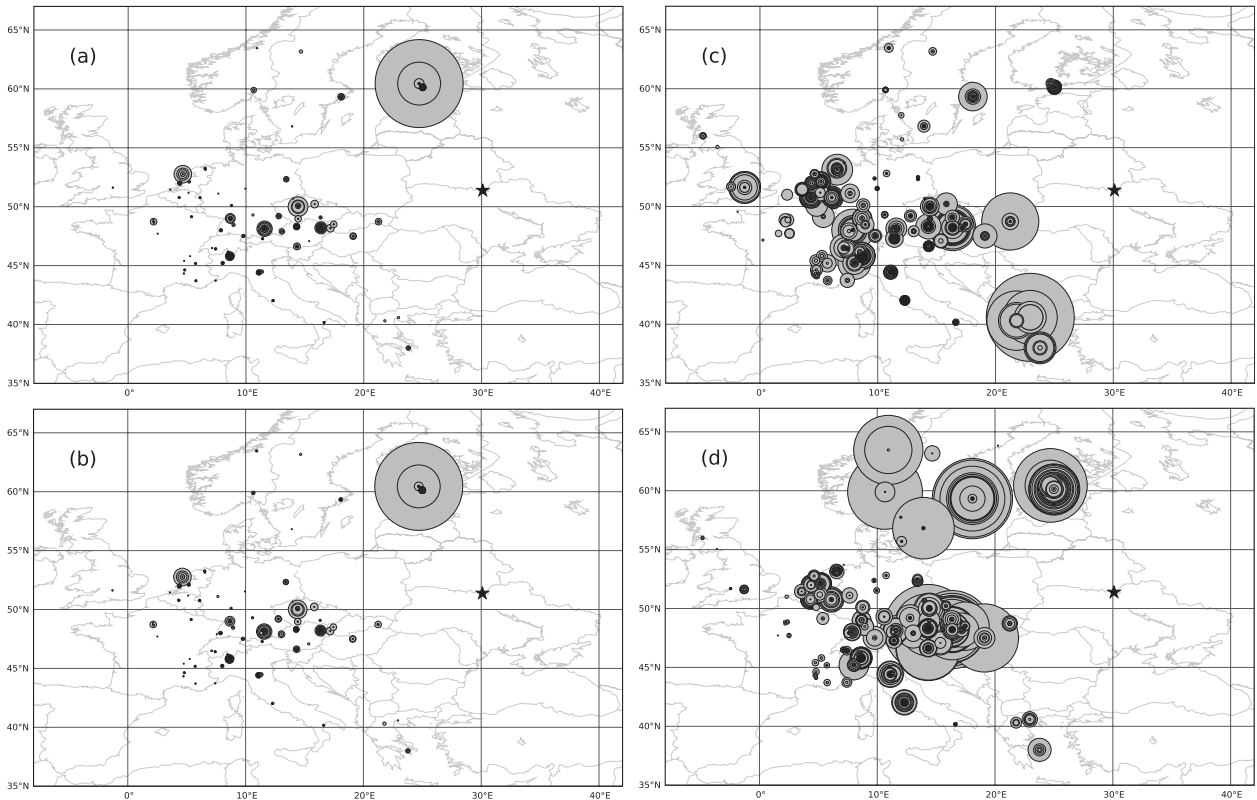
FIG. 8. Chernobyl inversion sensitivity to Gaussian or non-Gaussian analysis, using the setting of Davoine and Bocquet (2007). A disk is drawn for each observation with a radius proportional to the sensitivity. (a),(b) The global marginal sensitivity $\overline{\lambda}$ to the measurements. (c),(d) The marginal gain of information for the signal (source) $\partial_{\mathbf{y}}\mathcal{K}_{\mathbf{x}}$ to the measurements. Both (a) and (c) correspond to a non-Gaussian inversion, while (b) and (d) correspond to a Gaussian inversion.

observations were assimilated in a set of OSSEs (Krysta and Bocquet 2007).

7. Perspectives

The theoretical and numerical Bayesian solutions to the estimation problem have been shown not only to be appealing but also quite natural. Yet, the computational complexity that prevents their use ultimately leads to the success of 4D-Var and the ensemble Kalman filter. However, non-Gaussianity generated by the nonlinearities of model or intrinsic to the priors has not vanished. The ever-increasing computing power and widespread parallel architectures, even on cheap systems, make the use of more sophisticated applied mathematical solutions tempting. Nevertheless, the complexity of the estimation from these solutions does not scale reasonably (e.g., linearly) with the high dimensionality and complexity of geophysical systems, and relying on computing power is insufficient.

In this review, it has been shown that one can use sophisticated tools to diagnose non-Gaussianity in data assimilation with concepts inherited from statistics and information theory. More importantly, several examples of solution were given with promising performances. A few were of perturbative nature using an expansion of the Gaussian analysis system (weakly non-Gaussian prior construction). A few were of nonperturbative nature with little approximation (e.g., maximum entropy filter and Gaussian anamorphosis). Others were of fully nonperturbative nature taking advantage of Gaussian analysis guidance, or some linearity in the models (e.g., the optimal importance function particle filter, the weighted ensemble Kalman filter, and maximum entropy variational inference). These examples all remain specific, either because they rely on an assumption difficult to generalize, or because they were only tested on relatively low-dimensional systems so far. However, they can already be used on highly nonlinear subsystems of larger geophysical systems, such as Lagrangian drifters in a flow, or a submanifold of the dynamics. Alternatively, they can be used on real applications that do possess simplifying features, such as model linearity.

Increasing computing power might not only serve advanced data assimilation techniques, but also allows one to process more observations (denser coverage in space

and time), and finer model resolution. As a consequence, models may become locally more and more linear and more and more Gaussian between analyses. However, this argument has been partly refuted by the nonlinearity of small-scale physics, in conjunction with the fundamentally multiscale nature of geophysical systems. That is why we believe proper handling of non-Gaussianity will remain an important issue.

We think that more general solutions for high-dimensional systems than those exemplified in this review will require the simultaneous reduction of the model's dynamical degrees of freedom, spatial dividing–localization strategies, and an efficient sampling strategy (in connection with model error characterization). Particle filters or variants that are more advanced will then eventually be useful.

The number of possibilities to build up new solutions is tremendous, especially for filtering. It is to be expected that more and more applications will make use of an increasing number of theories mixing sequential and variational approaches, or combining Gaussian analysis and fully Bayesian ones. We expect that comparisons of all these methodologies and contexts will be a (very) difficult task. Theoretical and general guidance will then be needed to sort them all, with both mathematical analysis and the use of high-dimensional geophysical benchmarking models.

REFERENCES

Anderson, J. L., and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.,* **127,** 2741–2758.

Anderson, T. W., and D. A. Darling, 1952: Asymptotic theory of certain ''goodness of fit'' criteria based on stochastic processes. *Ann. Math. Stat.,* **23,** 193–212.

Andersson, E., and H. Järvinen, 1999: Variational quality control. *Quart. J. Roy. Meteor. Soc.,* **125,** 697–722.

——, M. Fisher, E. Hólm, L. Isaksen, G. Radnóti, and Y. Trémolet, 2005: Will the 4D-Var approach be defeated by nonlinearity? Tech. Rep. 479, ECMWF, 28 pp. [Available online at http://www.ecmwf.int/publications/library/ecpublications/_pdf/tm/401-500/tm479.pdf.]

——, and Coauthors, 2007: Analysis and forecast impact of the main humidity observing systems. *Quart. J. Roy. Meteor. Soc.,* **133,** 1473–1485.

Auroux, D., 2007: Generalization of the dual variational data assimilation algorithm to a nonlinear layered quasi-geostrophic ocean model. *Inverse Probl.,* **23,** 2485–2503.

Barndorff-Nielsen, O. E., and D. R. Cox, 1989: *Asymptotic Techniques for Use in Statistics. Meteor. Monogr.,* No. 31, Chapman & Hall, 252 pp.

Bellman, R., 1961: *Adaptive Control Processes: A Guided Tour.* Princeton University Press, 255 pp.

Bengtsson, T., C. Snyder, and D. Nychka, 2003: Toward a nonlinear ensemble filter for high-dimensional systems. *J. Geophys. Res.,* **108,** 8775, doi:10.1029/2002JD002900.

Berliner, M. L., and C. K. Wikle, 2007: Approximate importance sampling Monte Carlo for data assimilation. *Physica D,* **230,** 37–49.

Bertino, L., G. Evensen, and H. Wackernagel, 2003: Sequential data assimilation techniques in oceanography. *Int. Stat. Rev.,* **71,** 223–241.

Bishop, C. H., B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.,* **129,** 420–436.

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

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

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

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

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

Borwein, J. M., and A. S. Lewis, 2000: *Convex Analysis and Nonlinear Optimization: Theory and Examples.* Springer, 273 pp.

Buehner, M., P. L. Houtekamer, C. Charette, H. L. Mitchell, and B. He, 2010: Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part II: One-month experiments with real observations. *Mon. Wea. Rev.,* **138,** 1567–1586.

Buizza, R., and A. Montani, 1999: Targeted observations using singular vectors. *J. Atmos. Sci.,* **56,** 2965–2985.

Burgers, G., P. J. van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.,* **126,** 1719–1724.

Cane, M. A., A. Kaplan, R. N. Miller, B. Y. Tang, E. C. Hackert, and A. J. Busalacchi, 1996: Mapping tropical Pacific sea level: Data assimilation via a reduced state space Kalman filter. *J. Geophys. Res.,* **101** (C10), 22 599–22 617.

Carrassi, A., A. Trevisan, and F. Uboldi, 2007: Adaptive observations and assimilation in the unstable subspace by breeding on the data-assimilation system. *Tellus,* **59A,** 101–113.

——, M. Ghil, A. Trevisan, and F. Uboldi, 2008: Data assimilation as a nonlinear dynamical systems problem: Stability and convergence of the prediction-assimilation system. *Chaos,* **18,** 023112.

Chapnik, B., G. Desroziers, F. Rabier, and O. Talagrand, 2006: Diagnosis and tuning of observational error in a quasi-operational data assimilation setting. *Quart. J. Roy. Meteor. Soc.,* **132,** 543–565.

Cohn, S. E., 1997: An introduction to estimation theory. *J. Meteor. Soc. Japan,* **75,** 257–288.

——, A. da Silva, J. Guo, M. Sienkiewicz, and D. Lamich, 1998: Assessing the effects of data selection with the DAO physical-space statistical analysis system. *Mon. Wea. Rev.,* **126,** 2913–2926.

Courtier, P., 1997: Dual formulation of four-dimensional variational assimilation. *Quart. J. Roy. Meteor. Soc.,* **123,** 2449–2461.

——, and O. Talagrand, 1987: Variational assimilation of meteorological observation with the adjoint vorticity equation. II: Numerical results. *Quart. J. Roy. Meteor. Soc.,* **113,** 1329–1347.

Cover, T. M., and J. A. Thomas, 1991: *Elements of Information Theory.* Wiley Series in Telecommunications, Wiley-Interscience, 542 pp.

Daescu, D. N., and I. M. Navon, 2004: Adaptive observations in the context of 4D-Var data assimilation. *Meteor. Atmos. Sci.,* **85,** 205–226.

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

Del Moral, P., 2004: *Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications.* Springer-Verlag, 566 pp.

Desroziers, G., L. Berre, B. Chapnik, and P. Poli, 2005: Diagnosis of observation, background and analysis-error statistics in observation space. *Quart. J. Roy. Meteor. Soc.,* **131,** 3385–3396.

Doucet, A., S. Godsill, and C. Andrieu, 2000: On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. Comput.,* **10,** 197–208.

——, N. de Freitas, and N. Gordon, Eds., 2001: *Sequential Monte Carlo Methods in Practice.* Springer-Verlag, 612 pp.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.,* **99** (C5), 10 143–10 162.

——, 1997: Advanced data assimilation for strongly nonlinear dynamics. *Mon. Wea. Rev.,* **125,** 1342–1354.

——, 2003: The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dyn.,* **53,** 343–367.

Eyink, G. L., and S. Kim, 2006: A maximum entropy method for particle filtering. *J. Stat. Phys.,* **123,** 1071–1128.

Fisher, M., M. Leutbecher, and G. A. Kelly, 2005: On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation. *Quart. J. Roy. Meteor. Soc.,* **131,** 3235–3246.

Fletcher, S. J., and M. Zupanski, 2006: A data assimilation method for log-normally distributed observational errors. *Quart. J. Roy. Meteor. Soc.,* **132,** 2505–2519.

Fujita, T., D. J. Stensrud, and D. C. Dowell, 2007: Surface data assimilation using an ensemble Kalman filter approach with initial condition and model physics uncertainties. *Mon. Wea. Rev.,* **135,** 1846–1868.

Gardiner, C. W., 2004: *Handbook of Stochastic Methods: For Physics, Chemistry and the Natural Sciences.* 3rd ed. Springer Series in Synergetics, Springer, 415 pp.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.,* **125,** 723–757.

Gauthier, P., 1992: Chaos and quadri-dimensional data assimilation: A study based on the Lorenz model. *Tellus,* **44A,** 2–17.

Gilks, W. R., and C. Berzuini, 2001: Following a moving target-Monte Carlo inference for dynamic Bayesian models. *J. Roy. Stat. Soc. B,* **63,** 127–146.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith, 1993: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F,* **140,** 107–113.

Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.,* **129,** 2776–2790.

Handschin, J., and D. Mayne, 1969: Monte Carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering. *Int. J. Control,* **9,** 547–559.

Harlim, J., and B. R. Hunt, 2007: A non-Gaussian ensemble filter for assimilating infrequent noisy observations. *Tellus,* **59A,** 225–237.

Haven, K., A. Majda, and R. Abramov, 2005: Quantifying predictability through information theory: Small sample estimation in a non-Gaussian framework. *J. Comput. Phys.,* **206,** 334–362.

Heemink, A. W., M. Verlaan, and A. J. Segers, 2001: Variance reduced ensemble Kalman filtering. *Mon. Wea. Rev.,* **129,** 1718–1728.

Hólm, E., 2007: Humidity control variable and supersaturation. *Proc. Workshop on Flow-Dependent Aspects of Data Assimilation,* Reading, United Kingdom, ECMWF, 143–150.

——, E. Andersson, A. Beljaars, P. Lopez, J.-F. Mahfouf, A. J. Simmons, and J.-N. Thépaut, 2002: Assimilation and modelling of the hydrological cycle: ECMWF's status and plans. Tech. Rep. 383, ECMWF, 57 pp. [Available online at http://www.ecmwf.int/publications/library/ecpublications/_pdf/tm/301-400/tm383.pdf.]

Hoteit, I., 2008: A reduced-order simulated annealing approach for four-dimensional variational data assimilation in meteorology and oceanography. *Int. J. Numer. Methods Fluids,* **58,** 1181–1199.

——, D.-T. Pham, G. Triantafyllou, and G. Korres, 2008: A new approximate solution of the optimal nonlinear filter for data assimilation in meteorology and oceanography. *Mon. Wea. Rev.,* **136,** 317–334.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.,* **126,** 796–811.

——, and ——, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.,* **129,** 123–137.

——, and ——, 2005: Ensemble Kalman filtering. *Quart. J. Roy. Meteor. Soc.,* **131,** 3269–3289.

——, ——, and X. Deng, 2009: Model error representation in an operational ensemble Kalman filter. *Mon. Wea. Rev.,* **137,** 2126–2143.

Huber, P. J., 1973: Robust regression: Asymptotics, conjectures, and Monte Carlo. *Ann. Stat.,* **1,** 799–821.

Hunt, B. R., and Coauthors, 2004: Four-dimensional ensemble Kalman filtering. *Tellus,* **56A,** 273–277.

Hyvärinen, A., and E. Oja, 2000: Independent component analysis: Algorithms and applications. *Neural Networks,* **13,** 411–430.

Ide, K., P. Courtier, M. Ghil, and A. Lorenc, 1999: Unified notation for data assimilation: Operational, sequential and variational. *J. Meteor. Soc. Japan,* **75,** 181–189.

Jazwinski, A. H., 1970: *Stochastic Processes and Filtering Theory.* Academic Press, 376 pp.

Kalnay, E., H. Li, T. Miyoshi, S.-C. Yang, and J. Ballabrera, 2007: 4D-Var or ensemble Kalman filter. *Tellus,* **59A,** 758–773.

Kitagawa, G., 1987: Non-Gaussian state-space modeling of nonstationary time series. *J. Amer. Stat. Assoc.,* **82,** 1032–1063.

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

——, 2007: Statitical predictibility in the atmosphere and other dynamical systems. *Physica D,* **230,** 65–71.

Krüger, J., 1993: Simulated annealing: A tool for data assimilation into an almost steady model state. *J. Phys. Oceanogr.,* **23,** 679–688.

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

Kullback, S., 1959: *Information Theory and Statistics.* Wiley, 395 pp.

Laroche, S., and P. Gauthier, 1998: A validation of the incremental formulation of 4D variational data assimilation in a nonlinear barotropic flow. *Tellus,* **50A,** 557–572.

Lauvernet, C., J.-M. Brankart, F. Castruccio, G. Broquet, P. Braseur, and J. Verron, 2009: A truncated Gaussian filter for data assimilation with inequality constraints: Application to the hydrostatic stability condition in ocean models. *Ocean Modell.,* **27,** 1–17.

Lawson, W. G., and J. A. Hansen, 2004: Implications of stochastic and determinisitic filters as ensemble-based data assimilation methods in varying regimes of error growth. *Mon. Wea. Rev.,* **132,** 1966–1981.

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

——, H.-E. Ngodock, B. Luong, and J. Verron, 1997: Sensitivity analysis in variational data assimilation. *J. Meteor. Soc. Japan,* **75,** 245–255.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecast. *Mon. Wea. Rev.,* **102,** 409–418.

Lermusiaux, P. F. J., and A. R. Robinson, 1999: Data assimilation via error subspace statistical estimation. Part I: Theory and schemes. *Mon. Wea. Rev.,* **127,** 1385–1407.

Lilliefors, H. W., 1967: On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *J. Amer. Stat. Assoc.,* **62,** 399–402.

Lions, J.-L., O. P. Manley, R. Temam, and S. Wang, 1997: Physical interpretation of the attractor dimension for the primitive equations of atmospheric circulation. *J. Atmos. Sci.,* **54,** 1137–1143.

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

——, and T. Payne, 2007: 4D-Var and the butterfly effect: Statistical four-dimensional data assimilation for a wide range of scales. *Quart. J. Roy. Meteor. Soc.,* **133,** 607–614.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.

——, 1969: The predictability of a flow which possesses many scales of motion. *Tellus,* **21,** 289–307.

——, and K. E. Emmanuel, 1998: Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atmos. Sci.,* **55,** 399–414.

MacKay, D. J. C., 2003: *Information Theory, Inference and Learning Algorithms.* Cambridge University Press, 550 pp.

Meng, Z., and F. Zhang, 2007: Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part II: Imperfect model experiments. *Mon. Wea. Rev.,* **135,** 1403–1423.

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

——, E. F. Carter, and S. T. Blue, 1999: Data assimilation into nonlinear stochastic models. *Tellus,* **51A,** 167–194.

Mitchell, H. L., and P. L. Houtekamer, 2009: Ensemble Kalman filter configurations and their performance with the logistic map. *Mon. Wea. Rev.,* **137,** 4325–4343.

Nakano, S., G. Ueno, and T. Higuchi, 2007: Merging particle filter for sequential data assimilation. *Nonlinear Processes Geophys.,* **14,** 395–408.

Nerger, L., W. Hiller, and J. Schröter, 2005: A comparison of error subspace Kalman filters. *Tellus,* **57A,** 715–735.

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

Papadakis, N., 2007: Assimilation de données images: Application au suivi de courbes et de champs de vecteurs (Image data assimilation: Application to curve and vector fields tracking). Ph.D. thesis, Université de Rennes I, 242 pp.

Patil, D., R. Hunt, E. Kalnay, J. A. Yorke, and E. Ott, 2001: Local low dimensionality of atmospheric dynamics. *Phys. Rev. Lett.,* **86,** 5878–5881.

Pham, D. T., 2001: Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Mon. Wea. Rev.,* **129,** 1194–1207.

——, J. Verron, and M. Roubaud, 1998: A singular evolutive extended Kalman filter for data assimilation in oceanography. *J. Mar. Syst.,* **16,** 323–340.

Pires, C. A., and R. Perdigão, 2007: Non-Gaussianity and asymmetry of the winter monthly precipitation estimation from NAO. *Mon. Wea. Rev.,* **135,** 430–448.

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

——, O. Talagrand, and M. Bocquet, 2010: Diagnosis and impacts of non-Gaussianity of innovations in data assimilation. *Physica D,* **239,** 1701–1717.

Rabier, F., H. Järvinen, E. Klinker, J.-F. Mahfouf, and A. Simmons, 2000: The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Quart. J. Roy. Meteor. Soc.,* **126,** 1143–1170.

Rodgers, C. D., 2000: *Inverse Methods for Atmospheric Sounding.* Series on Atmospheric, Oceanic and Planetary Physics, Vol. 2, World Scientific, 238 pp.

Sakov, P., and P. R. Oke, 2008: Implications of the form of the ensemble transformation in the ensemble square root filters. *Mon. Wea. Rev.,* **136,** 1042–1053.

Shapiro, S. S., and M. B. Wilk, 1965: An analysis of variance test for normality (complete samples). *Biometrika,* **52,** 591–611.

Silverman, B. W., 1986: *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 175 pp.

Simon, E., and L. Bertino, 2009: Application of the Gaussian ana-
morphosis to assimilation in a 3-D coupled physical-ecosystem
model of the North Atlantic with the EnKF: A twin experiment.
Ocean Sci., **5,** 495–510.

Snyder, C., T. Bengtsson, P. Bickel, and J. L. Anderson, 2008:
Obstacles to high-dimensional particle filtering. *Mon. Wea.
Rev.,* **136,** 4629–4640.

Spiller, E. T., A. Budhiraja, K. Ide, and C. K. R. T. Jones, 2008:
Modified particle filter methods for assimilating Lagrangian
data into a point-vortex model. *Physica D,* **237,** 1498–1506.

Talagrand, O., and P. Courtier, 1987: Variational assimilation of
meteorological observation with the adjoint vorticity equa-
tion. I: Theory. *Quart. J. Roy. Meteor. Soc.,* **113,** 1311–1328.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and
J. S. Whitaker, 2003: Ensemble square root filters. *Mon. Wea.
Rev.,* **131,** 1485–1490.

Trémolet, Y., 2004: Diagnostics of linear and incremental approxi-
mations in 4D-Var. *Quart. J. Roy. Meteor. Soc.,* **130,** 2233–2251.
——, 2006: Accounting for an imperfect model in 4D-Var. *Quart. J.
Roy. Meteor. Soc.,* **132,** 2483–2504.

Tribbia, J. J., and D. P. Baumhefner, 2004: Scale interactions and
atmospheric predictability: An updated perspective. *Mon.
Wea. Rev.,* **132,** 703–713.

Uzunoglu, B., 2007: Adaptive observations in ensemble assimila-
tion. *Comput. Methods Appl. Mech. Eng.,* **196,** 4207–4221.

Van der Merwe, R., A. Doucet, N. de Freitas, and E. Wan, 2000: The
unscented particle filter. Tech. Rep. CUED/F-INFENG/TR
380, Engineering Department, Cambridge University, 45 pp.

van Leeuwen, P. J., 2009: Particle filtering in geophysical systems.
Mon. Wea. Rev., **137,** 4089–4114.

Verlaan, M., and A. W. Heemink, 2001: Nonlinearity in data as-
similation applications: A practical method for analysis. *Mon.
Wea. Rev.,* **129,** 1578–1589.

Wackernagel, H., 2003: *Multivariate Geostatistics.* 3rd ed. Springer,
403 pp.

Whitaker, J. S., and T. M. Hamill, 2002: Ensemble data assimilation
without perturbed observations. *Mon. Wea. Rev.,* **130,** 1913–1924.

Wu, L., V. Mallet, M. Bocquet, and B. Sportisse, 2008: A comparison
study of data assimilation algorithms for ozone forecasts.
J. Geophys. Res., **113,** D20310, doi:10.1029/2008JD009991.

Xiong, X., I. M. Navon, and B. Uzunoglu, 2006: A note on the particle
filter with posterior Gaussian resampling. *Tellus,* **58A,** 456–460.

Zakai, M., 1969: On the optimal filtering of diffusion processes.
Z. Wahrsch., **11,** 230–243.

Zupanski, M., 2005: Maximum likelihood ensemble filter: Theo-
retical aspects. *Mon. Wea. Rev.,* **133,** 1710–1726.