Nonlin. Processes Geophys., 21, 869–885, 2014
www.nonlin-processes-geophys.net/21/869/2014/
doi:10.5194/npg-21-869-2014

Nonlinear Processes in Geophysics
Open Access

A non-Gaussian analysis scheme using rank histograms for ensemble data assimilation

S. Metref[1], E. Cosme[1], C. Snyder[2], and P. Brasseur[1]

[1]CNRS – Univ. Grenoble Alpes, LGGE (UMR5183), 38041 Grenoble, France
[2]National Center for Atmospheric Research, Boulder, Colorado, USA

Correspondence to: S. Metref (sammy.metref@legi.grenoble-inp.fr)

Received: 1 August 2013 – Revised: 10 July 2014 – Accepted: 14 July 2014 – Published: 25 August 2014

Abstract. One challenge of geophysical data assimilation is to address the issue of non-Gaussianities in the distributions of the physical variables ensuing, in many cases, from nonlinear dynamical models. Non-Gaussian ensemble analysis methods fall into two categories, those remapping the ensemble particles by approximating the best linear unbiased estimate, for example, the ensemble Kalman filter (EnKF), and those resampling the particles by directly applying Bayes' rule, like particle filters. In this article, it is suggested that the most common remapping methods can only handle weakly non-Gaussian distributions, while the others suffer from sampling issues. In between those two categories, a new remapping method directly applying Bayes' rule, the multivariate rank histogram filter (MRHF), is introduced as an extension of the rank histogram filter (RHF) first introduced by Anderson (2010). Its performance is evaluated and compared with several data assimilation methods, on different levels of non-Gaussianity with the Lorenz 63 model. The method's behavior is then illustrated on a simple density estimation problem using ensemble simulations from a coupled physical–biogeochemical model of the North Atlantic ocean. The MRHF performs well with low-dimensional systems in strongly non-Gaussian regimes.

1 Introduction

The principal goal of data assimilation is to estimate the state of a dynamical system, based on prior information and a time series of observations, while calculating probabilistic measures corresponding to the accuracy of this estimation. Kalman filter theory (Kalman, 1960) became a reference in data assimilation as it provides the optimal solution to the linear and Gaussian filtering problem (Cohn, 1997). The ensemble Kalman filter (EnKF, Evensen, 1994) and closely related methods (Lermusiaux, 1999; Pham, 2001; Whitaker and Hamill, 2002, to cite only a few) are different implementations of the Kalman filter relying on ensembles. For the analysis step, they transform the prior ensemble into a posterior ensemble, using a function that is optimal (an optimal map, Cotter and Reich, 2013) under the assumption of Gaussianity of the prior ensemble and the observation errors. Those methods are applicable to high-dimensional systems in meteorology (Whitaker et al., 2008; Buehner et al., 2010) and oceanography (Lermusiaux, 2006; Sakov et al., 2012). This success is – in part – due to the fact that the dynamics of these systems are weakly nonlinear, that is, do not strongly deviate from a linear evolution within the space and timescales characterizing the density of available observations. Briefly, a weak nonlinearity transforms a Gaussian distribution into a weakly non-Gaussian distribution, with which the EnKF still performs well. Many recipes have been developed to enforce the good behavior of the EnKF with such systems, including localization techniques (Sakov and Bertino, 2010; Greybush et al., 2011), sampling strategies (Pham, 2001; Anderson, 2012), and observational targeting (Bishop et al., 2001); see Bocquet et al. (2010) for more details and other examples. Nevertheless, EnKF-based methods remain sensitive to the violation of the Gaussian assumption (Lawson and Hansen, 2004; Lei et al., 2010) and may lead to unwanted phenomena such as inaccurate estimations, failure to respect nonlinear physical balances, or more dramatically to instability of the filter.

Along with the developments of the EnKF, there is a growing need for non-Gaussian ensemble data assimilation methods. Data assimilation is no longer a tool solely

for meteorology and oceanography. Other disciplinary fields with stronger nonlinearities and much sparser data networks (i.e., geomagnetism; Fournier et al., 2010) increasingly depend upon data assimilation. Even in the traditional fields of application, models' nonlinearity tends to increase along with their complexity. Even with linear models, non-Gaussian observation error densities make the assimilation problems non-Gaussian. Intrinsically, non-Gaussian variables are common in the atmosphere and the ocean, such as humidity (Dee and Da Silva, 2003) and concentrations of sea ice or phytoplankton (Brankart et al., 2012).

The ensemble data assimilation methods can be sorted in two categories: those that transform the prior ensemble particles using a deterministic map (transform methods), and those that sample the posterior probability density (sampling methods). They can also be classified as parametric, non-parametric, or semi-parametric, depending on the assumptions on the shape of the probability densities they use; the EnKF falls in the parametric (with the Gaussian assumption) transform methods. The EnKF with Gaussian anamorphosis, further described in Sect. 2 of the present paper, transforms variables to make their densities Gaussian before applying the EnKF analysis. It can be considered as a semi-parametric transform method, since it is not a fully non-parametric method able to deal with any kind of probability density, as illustrated in Sect. 2. The truncated-Gaussian EnKF as described by Lauvernet et al. (2009) is of the parametric, sampling type. Reich (2013) introduces a sequential method of the non-parametric, transform category. The particle filter (Gordon et al., 1993; van Leeuwen, 2009) is the most popular method of the non-parametric, sampling type, but is well known to be particularly subject to the curse of dimensionality, which makes it difficult to use with high-dimensional systems (Snyder et al., 2008). Finding solutions to make the particle filter applicable to high-dimensional systems is a very active topic of research (Nakano et al., 2007; van Leeuwen, 2010; Morzfeld et al., 2012; Snyder, 2012). A few of them actually rely on some hybridization with the EnKF (Bocquet et al., 2010; Lei et al., 2010; Hoteit et al., 2012).

Ensemble methods of the non-parametric, transform category have rarely been explored, although they could be less sensitive to the curse of dimensionality than the sampling methods, due to the transformation step that helps enforce a better fit of particles to the observations. In this respect, the approach proposed by Reich (2013) would deserve further examination, in particular with high-dimensional systems. Another method that could be somewhat classified as a partly non-parametric transform method is the rank histogram filter (RHF, Anderson, 2010). The RHF is a hybrid between the EnKF and a fully non-Gaussian approach, named that way because it is based on a statistical processing similar to the rank histograms (Anderson, 1996; Hamill, 2001) used for ensemble forecast evaluation. The RHF corrects observed variables by representing their prior densities and the observation likelihoods as piecewise continuous functions in order to directly apply Bayes' rule. This theoretically solves the generalized problem for a single observed variable. However, the other variables are still corrected using a linear regression onto the corrections of observed variables, as in the EnKF. We believe these advantages justify a more detailed exploration of the RHF philosophy. The main objective of this paper is to present an extension of the rank histogram approach of Anderson (2010) to unobserved variables yielding a fully non-parametric transform scheme for ensemble data assimilation, in the spirit of the method of Reich (2013). Throughout the paper, this multivariate RHF is referred to as MRHF.

The article is outlined as follows. In Sect. 2, we present some considerations on the joint non-Gaussianity of two variables, and how ensemble analysis schemes perform with such densities. Emphasis is given to the EnKF and to the RHF of Anderson (2010). Section 3 develops the extension of the RHF to unobserved variables called MRHF along with an approximation of the latter. Numerical experiments are presented in Sect. 4, where the MRHF and its approximation are evaluated with the highly nonlinear and non-Gaussian Lorenz 63 model, and compared in different setups (corresponding to different levels of nonlinearities) to the EnKF, to the RHF and to a particle filter. In Sect. 5, the new schemes are finally illustrated with a density estimation problem based on a realistic ensemble from a coupled marine biogeochemical model. Even though this last experiment is not a data assimilation problem, the results give an insight into the behavior of the method. A discussion and a conclusion are given in the last section.

2 Gaussian and non-Gaussian analysis in ensemble filtering

The increasing popularity of ensemble filters is largely due to the relative simplicity of their implementation. They basically alternate propagation steps and analysis steps. During a propagation step, each particle of the ensemble is advanced in time using the dynamical system model, possibly including some parameterization of the model error. An analysis step occurs after a propagation step, when an observation $\boldsymbol{Y}^{\mathrm{m}}$ is available. The observation $\boldsymbol{Y}^{\mathrm{m}}$ is a realization of the (random) measurement vector $\boldsymbol{Y}^{\mathrm{o}} = h(\boldsymbol{X}) + \epsilon$, where h is a forward observation operator, $\boldsymbol{X}$ the state vector to be estimated, and ϵ the observation error. The analysis step conflates the prior ensemble $\{\boldsymbol{X}_i^f\}_{i=1,\ldots,N_{\mathrm{e}}}$, composed of N_{e} particles resulting from the previous forecast, and the available observation $\boldsymbol{Y}^{\mathrm{m}}$, to provide a posterior (analysis) ensemble $\{\boldsymbol{X}_i^{\mathrm{a}}\}_{i=1,\ldots,N_{\mathrm{e}}}$. Observation errors are often assumed temporally and spatially uncorrelated so that each one can be independently assimilated (Houtekamer and Mitchell, 2001; Evensen, 2003). If the spatial correlations cannot be neglected, a linear transformation of the observation vector is theoretically possible (and exact in the linear and Gaussian

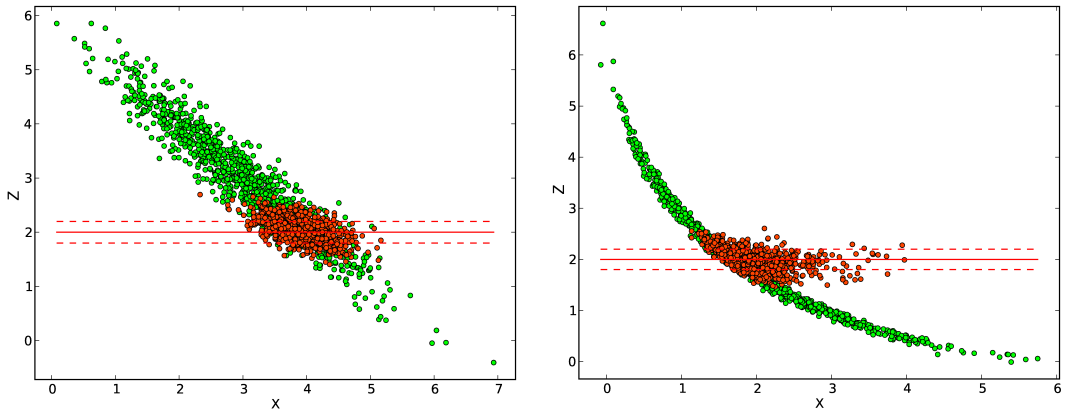

Figure 1. Illustration of EnKF analyses with joint Gaussian (left panel) and weakly non-Gaussian (right panel) prior distributions. The prior ensemble is represented by the green dots and the posterior ensemble is represented by the red dots. The variable corresponding to the z axis is observed with a value shown by the red solid line. Its uncertainty (identical in all the illustrations of Sect. 2) is assumed Gaussian with standard deviations symbolized with the red dashed lines.

context), in which the observation error covariance matrix is diagonal (Anderson, 2003).

A popular implementation of the ensemble analysis is the EnKF with serial processing of observations (Houtekamer and Mitchell, 2001). Following the description given by Anderson (2003), the ensemble update by each observation is performed in two steps, namely, the update of the observed variable followed by the update of unobserved variables. If the observation is not a direct observation of a state variable, then the state vector $\boldsymbol{X}$ can be augmented with the observed function of state variables, $\boldsymbol{Y}^{\mathrm{o}}$, to introduce a directly observed variable. In what follows, only the case of the direct observation of a variable is considered. With two scalar variables, $\boldsymbol{X} = (x, z)^T$, z being subject to a direct measurement z^{o} with realization z^{m}, decomposing the EnKF analysis equations shows that the correction of the observed variable z for ensemble particle i is

$$\delta z_i = \frac{\mathrm{Var}(z)}{\mathrm{Var}(z) + \mathrm{Var}(\epsilon)} \left(z^{\mathrm{m}} - z_i - \epsilon_i\right), \tag{1}$$

where ϵ_i is a perturbation that takes the observation error ϵ into account. The correction for the unobserved variable x, for particle i, is

$$\delta x_i = \frac{\mathrm{Cov}(x, z)}{\mathrm{Var}(z)} \delta z_i. \tag{2}$$

It is clear from the latter equation that the correction of any unobserved variable is a function of the linear correlation between the variable and the observed variable. Such formulation must be questioned when the linear correlation is not a relevant measure of the statistical relationship between these variables, as may occur when the statistics are non-Gaussian. This issue is illustrated in Fig. 1: in the left panel, a prior ensemble of a bivariate Gaussian state (x, z) is depicted (green dots); the second variable z is observed. The corrections for the unobserved variable, based on a relevant linear correlation with the observed variable, leads to an analysis ensemble (red dots) fitting the bivariate Gaussian probability density function (pdf) that would be produced by implementing Bayes' theorem. This analysis ensemble is consistent with both the physics, introduced through the prior information, and the observation. In the right panel, the two (non-jointly Gaussian) variables exhibit a nonlinear statistical relationship, that cannot be fully captured by a linear regression. Consequently, even if the corrections of the observed variable are somewhat correct, those of the unobserved variable can be erroneous. This results in a rather poor analysis ensemble, where particles appear in unexpected parts of the phase space, in violation of the inter-variable relationship, as described by the prior ensemble.

This problem is not new; it falls under the general designation of non-Gaussian data assimilation. Solutions exist, among the "resampling" methods in particular, but their effective application in high dimensions is either impossible or requires further development; see Bocquet et al. (2010) for a review. Some non-Gaussian schemes derive from refinements to the EnKF. A promising one, mostly studied in oceanography, is the Gaussian anamorphosis (Bertino et al., 2003; Simon and Bertino, 2009; Béal et al., 2010; Brankart et al., 2012). Anamorphosis consists in transforming the initial physical variables to make them fit Gaussian distributions. The standard EnKF analysis can be applied to these transformed variables. Then, the physical analysis variables are recovered by the inverse transformation. The transformation can be either analytical or numerical (Bocquet et al., 2010). The following illustration is performed with the numerical transformation described by Brankart et al. (2012). The left panel of Fig. 2 shows the same non-Gaussian prior ensemble as the right panel of Fig. 1 (green dots), along with the analysis ensemble obtained using Gaussian

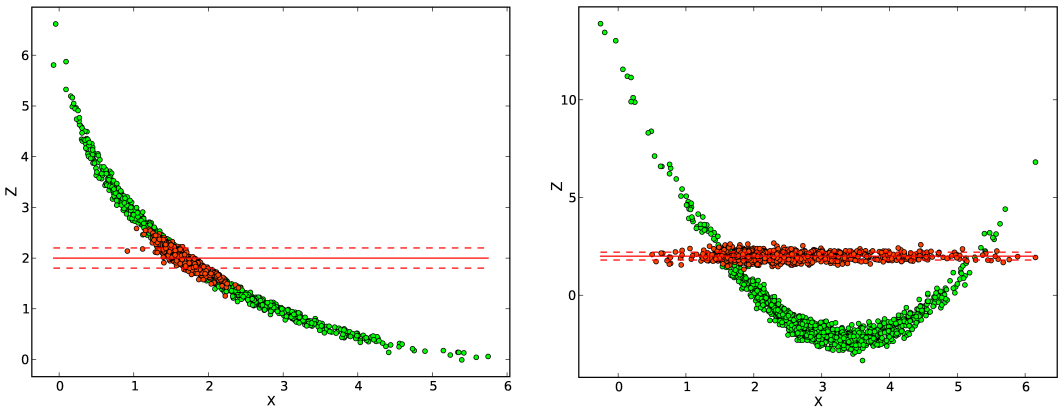

Figure 2. Same as Fig. 1 but for the EnKF with Gaussian anamorphosis and for joint weakly non-Gaussian (left panel) and strongly non-Gaussian (right panel) prior distributions. The weakly non-Gaussian prior is the same as in Fig. 1 (right panel).

anamorphosis (red dots). Anamorphosis clearly improves the EnKF. But anamorphosis presents several limitations, one of which is that it is based on a one-to-one correspondence between the prior and the target (Gaussian) distributions involved in the transformation. If this is not verified, Gaussian anamorphosis fails. Such failure is depicted on the right panel of Fig. 2, where the prior ensemble follows a strongly non-Gaussian density law which exhibits bimodality under conditioning by z. The EnKF with Gaussian anamorphosis (or without; not shown) provides a very poor analysis ensemble.

Fully non-Gaussian ensemble analysis schemes, that is, schemes derived without any assumption on the shape of the prior ensemble density, implement Bayes' rule to solve the analysis step:

$$p\left(\boldsymbol{X}|\boldsymbol{Y}^{\mathrm{o}}\right) \propto p(\boldsymbol{X})\, p\left(\boldsymbol{Y}^{\mathrm{o}}|\boldsymbol{X}\right), \tag{3}$$

where $p(\boldsymbol{X})$ is the prior probability density for the state vector $\boldsymbol{X}$ to estimate, $p(\boldsymbol{Y}^{\mathrm{o}}|\boldsymbol{X})$ the observation likelihood (identical to the observation density for Gaussian observation errors), and $p(\boldsymbol{X}|\boldsymbol{Y}^{\mathrm{o}})$ the posterior density, that is, the density of the state given the observations. A detailed Bayesian description of data assimilation is provided by Wikle and Berliner (2007) for instance. The fully non-Gaussian ensemble data assimilation problem is usually solved by resampling methods of the particle filter type. The particle filter (Gordon et al., 1993; Doucet et al., 2001) is subject to very active developments for its future application with high-dimensional geophysical problems (Nakano et al., 2007; van Leeuwen, 2009, 2010; Morzfeld et al., 2012). The key point in implementing a particle filter is to beat the curse of dimensionality, and that will probably not be solved shortly for large applications (Snyder et al., 2008). One major reason, we believe, is that particle filters have yet to implement localization, in which any given observation affects the update only in a spatially local region near the observation location and as is common in the EnKF (Houtekamer and Mitchell, 1998; Hamill et al., 2001).

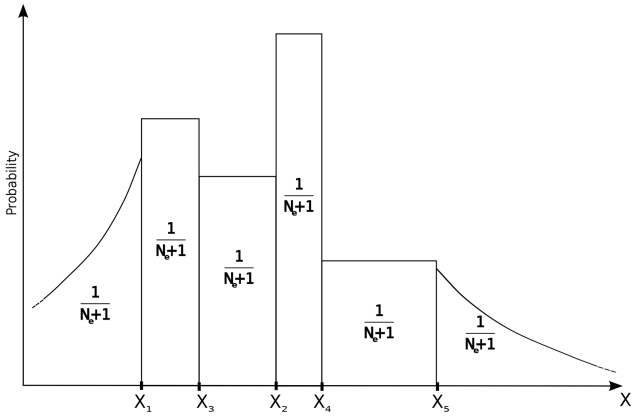

Figure 3. Reconstruction of a density from an ensemble using the rank histogram approach.

Here, we explore a new non-Gaussian ensemble analysis scheme of the "transform" type, in which localization can be implemented. We start from the the rank histogram filter (RHF), a partially non-Gaussian transform scheme, that has been proposed by Anderson (2010). The RHF processes observations serially. For the direct observation z^{o} of variable z, the continuous prior density for z is represented as a rank histogram (by analogy with the rank histogram diagnostic used to evaluate ensemble predictions, as introduced in geophysics by Anderson (1996), and later discussed in Hamill (2001), and Candille and Talagrand (2005)). The histogram is composed of $N_{\mathrm{e}} - 1$ bounded regions partitioned by the sorted ensemble particles (the order statistics of the problem) and two unbounded regions on the tails. In each inner region, a density value is assigned so that the region contains a probability mass of $\frac{1}{N_{\mathrm{e}}+1}$ (Fig. 3). The two outer regions are covered by tails of probability mass $\frac{1}{N_{\mathrm{e}}+1}$ as well; their shape may be chosen freely, and this may actually be a key element for the success of the RHF (Anderson, 2010). In particular,

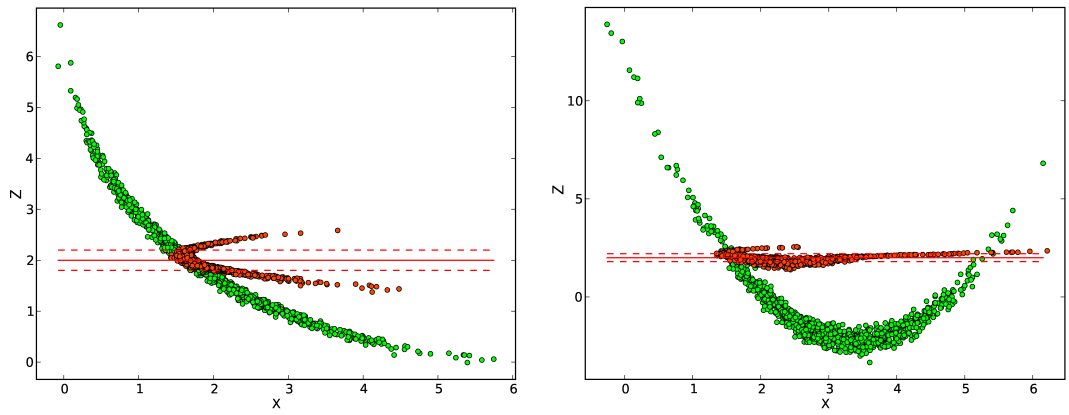

Figure 4. Same as Fig. 2 but for the RHF.

very long tails can help to correct biases and to make the filter more resilient to divergence. More precisely, the prior density of z is written:

$$p(z) = \frac{1}{N_\mathrm{e}+1} \sum_{i=1}^{N_\mathrm{e}-1} \frac{1_{[z_i, z_{i+1}[}(z)}{(z_{i+1}-z_i)} + T(z), \tag{4}$$

with $1_{[z_i, z_{i+1}[}(z)$ the indicator function on the interval $[z_i, z_{i+1}[$ (yielding 1 if z belongs to this interval, 0 otherwise) and $T(z)$ also a combination of indicator functions representing the tails term applied to the two outer regions. The likelihood $p(z^\mathrm{o}|z)$ is known analytically from the observation error density. It is discretized on the same grid as $p(z)$, and the two functions are multiplied pointwise to provide a constant piecewise expression (after normalization) of the posterior density $p(z|z^\mathrm{o})$. The analysis ensemble is finally obtained using a (deterministic) procedure of inversion of the cumulative distribution function. Corrections on z particles are calculated by the difference between the posterior and the prior values of z. These corrections are used to compute the corrections for the unobserved variables with Eq. (2), that is, applying a linear regression. This latter step, inherited from the EnKF, is perhaps the main weakness of the RHF, as illustrated in Fig. 4: the RHF analysis performs rather poorly in both weakly and strongly non-Gaussian cases addressed in the previous illustrations. However, considering the many positive aspects of this scheme (non Gaussian, robust, deterministic, possible to localize), it seems worth trying to correct this weakness and extend the rank histogram approach to unobserved variables.

3 Multivariate rank histogram filter

3.1 Principle

We wish to generalize the RHF to the general Bayesian framework. The RHF first addresses the analysis of the observed variable, then deals with the others. It is thus an implementation of the *sequential realization method* to sample a multivariate probability density, for example, as presented by Tarantola (2005). This method leans on the Knothe–Rosenblalt rearrangement, a decomposition of the joint probability density into a product of marginal and conditional univariate densities. With three scalar variables $\boldsymbol{X} = (x, y, z)$, this decomposition is

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

A sample from the joint density is obtained by deterministically sampling $p(z)$ first, then $p(x|z)$ (using the result for $p(z)$), and $p(y|x, z)$ (using the previous two results). The purpose of data assimilation here is to condition the joint density to an observation z^o of z. Following the usual Markovian memoryless assumption for the observation process, which implies $p(z^\mathrm{o}|x, y, z) = p(z^\mathrm{o}|z)$, and using the decomposition (Eq. 5) it is straightforward to find that

$$p\left(x, y, z|z^\mathrm{o}\right) = p(z|z^\mathrm{o})\, p(x|z)\, p(y|x, z). \tag{6}$$

Here again, the deterministic sampling of both densities conditioned on z are based on the previously sampled densities; hence, the sampling of variables x and y depends on the observation z^o. To obtain the first factor on the right-hand side, the EnKF uses Eq. (1); the RHF implements Bayes' rule for z: $p(z|z^\mathrm{o}) \propto p(z)\, p(z^\mathrm{o}|z)$. But for both methods, the second and third terms on the right-hand side are computed using Eq. (2), which comes from a Gaussian, Kalman filtering perspective. We propose below a new non-Gaussian approach to sample scalar particles from these conditional densities to implement Eq. (6) with non-parametric densities. This scheme is deterministic, in the sense that no random number need be generated during the analysis process. The analyses are then reproducible and the method is of the "transform" type.

3.2 Implementation of the MRHF analysis

Let $\{z_i^\mathrm{a}\}_{i=1,\ldots,N_\mathrm{e}}$ be the posterior ensemble of the observed variable z, that is, a sample of $p(z|z^\mathrm{o})$. Consider the first

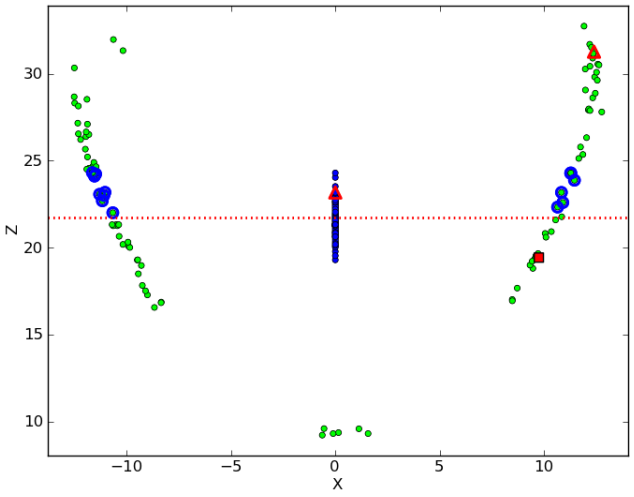

Figure 5. Illustration of the MRHF analysis step for the first unobserved variable: green dots represent the prior ensemble; blue dots vertically aligned at $X = 0$ represent the posterior z ensemble. The red dotted line is the observation of z, the red square is the true state (not used for the analysis). The red empty triangles show the particle being processed and its corresponding z analysis value. Blue circles show the selected particles to form the posterior density $p_i^{\mathrm{a}}(x)$; see text for details.

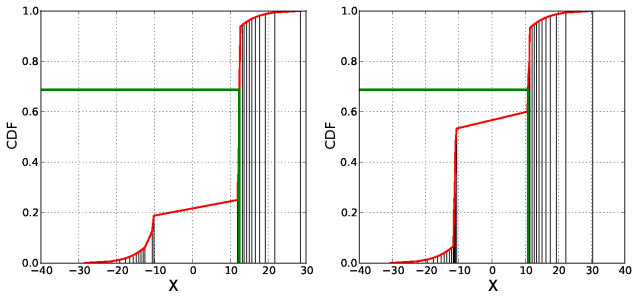

Figure 6. Illustration of the MRHF analysis step for the first unobserved variable (one particle): red lines represent the cumulative distribution functions (cdf) of the prior density $p_i^{\mathrm{b}}(x)$ (left panel) and the posterior density $p_i^{\mathrm{a}}(x)$ (right panel). The vertical black lines show the selected particles used to build these densities. To compute the x analysis value for the prior x particle near 11 on the left panel, the green line must be followed: the cdf for the prior x is computed with the prior cdf (Result is near 0.68); from this cdf value, the x analysis value is obtained on the right panel; see text for details.

unobserved variable, x in Eq. (6). The MRHF analysis determines x_i^{a}, the x analysis value for particle i, by deterministically sampling the conditional density $p_i^{\mathrm{a}}(x) \equiv p(x|z = z_i^{\mathrm{a}})$. This density must first be formed. Some steps of the procedure are illustrated in Fig. 5. The green dots represent the prior ensemble in the X–Z plane; the blue dots at $X = 0$ represent the z analysis ensemble $\{z_i^{\mathrm{a}}\}_{i=1,\ldots,N_{\mathrm{e}}}$. The red line is the observation realization. The following process is repeated for $i = 1, \ldots, N_{\mathrm{e}}$. For a given i, a subset of particles is selected in the prior ensemble (green dots with blue circles), whose z values lie in the neighborhood of z_i^{a} (blue dot with red triangle) along the z direction. The selection process is discussed later. Applying the rank histogram approach to the x values of the selected particles, a one-dimensional density is then formed to represent $p_i^{\mathrm{a}}(x)$.

To follow Eq. (6), one approach would be to draw a random realization from this density to provide x_i^{a}. This, however, is far from optimal from the physical viewpoint, because it can generate large corrections resulting in physical instabilities and imbalances, as previously observed by Anderson (2003) in the EnKF context. In Fig. 5 for instance, the prior particle (green dot with red triangle) is in the right-hand side mode of the distribution. Since the observation does not enable one to know in which mode the truth (red dot) actually is, it makes sense to try to keep this particle in its mode of origin, thus minimizing its modification.

Instead of a random draw in $p_i^{\mathrm{a}}(x)$ which could arbitrarily move the particle to the left-hand side mode, the following steps are proposed:

- with a similar selection and a rank histogram process, form the density of x conditioned to the *background* value of z: $p_i^{\mathrm{b}}(x) \equiv p(x|z = z_i^{\mathrm{b}})$;
- compute the cumulative distribution functions $C_i^{\mathrm{b}}(x)$ and $C_i^{\mathrm{a}}(x)$ from $p_i^{\mathrm{b}}(x)$ and $p_i^{\mathrm{a}}(x)$, respectively;
- compute the position of the prior particle in the prior density: $c_i = C_i^{\mathrm{b}}(x_i^{\mathrm{b}})$;
- preserve the rank of the particle in the posterior density by taking $x_i^{\mathrm{a}} = C_i^{\mathrm{a}(-1)}(c_i)$ as analysis value for x and particle i. This is illustrated in Fig. 6.

Although this method does not always prevent a particle shifting from one mode to another, two neighboring particles (i.e., close to each other) in the prior ensemble are likely to remain neighbors in the posterior ensemble. In a multimodal case for instance, two particles in the same mode are more likely to remain in the same mode after analysis. Figure 7 illustrates the successful behavior of the MRHF analysis in the strongly non-Gaussian case introduced in Sect. 2 (Figs. 2 and 4, right panels).

Once the z and x analysis values are computed for each particle, the analysis values for the third variable y can be computed. The process is strictly similar to the one described above, but for the variable y, and with an additional $x = x_i^{\mathrm{a}}$ term in the conditional statement. In practice, this reduces to selecting particles from the prior distribution in the neighborhood of $(x_i^{\mathrm{a}}, z_i^{\mathrm{a}})$ in the two-dimensional plane (x, z), to form the density $p(y|x, z, z^{\mathrm{o}})$. The other steps remain unchanged.

As a remark, one may notice that this analysis method brings some similarities with the heuristic method presented in Anderson (2003) in the EnKF context. His idea is to compute the covariance term in Eq. (2) using a subset

of neighboring particles. The MRHF goes one step further by considering the nonlinear relationship between those particles.

3.3 Selection of particles and mean-field approximation

We now come back to the selection of particles, and start with the first unobserved variable x. To represent the target density $p_i^{\mathrm{a}}(x)$ accurately, the selected particles must have a z value close to z_i^{a} used in the conditional statement. At the same time, there may be few or no particles whose z values differ from z_i^{a} by less than a specified, small amount, since N_{e} is finite. Thus, there is a trade-off between selecting particles that are very close to z_i^{a} and selecting a sufficient number of particles to represent $p_i^{\mathrm{a}}(x)$.

In the numerical experiments with the Lorenz 63 system, presented in Sect. 4, a maximal distance ($d_{\max}$) is prescribed, along with a minimal and a maximal number of particles. Specifications of these parameters are detailed in Sect. 3.4. No attempt has been made in this work to make the scheme independent of the variables, for example by normalizing all variable by their prior variance.

For the second unobserved variable y, the difference is computed as a distance in the two-dimensional plane (x, z). The maximal distance to consider as a threshold must be prescribed accordingly. As the algorithm proceeds to additional unobserved variables, the curse of dimensionality becomes apparent: each time a new unobserved variable is analyzed, a dimension is added and the volume of states within the maximal distance decreases so fast that each region defined by one particle and a finite radius around it has negligible probability to hold another particle.

As a schematic illustration, assume that $z_i^{\mathrm{a}} = 30$ in Fig. 5, and $d_{\max} = 1$. There exist prior particles with $29 < z < 31$. Thus, the analysis for x can be conducted accurately. Assume now that the x analysis provides $x_i^{\mathrm{a}} = -10$. Prior particles in the two-dimensional neighborhood of $(x_i^{\mathrm{a}}, z_i^{\mathrm{a}})$ (for example, in a circle centered on this point with radius $\sqrt{2}$) are sparse or nonexistent. More distant particles must therefore be included and the accuracy of the analysis for y may be poor.

This obstacle leads us to introduce an approximation, termed the mean-field approximation by Cotter and Reich (2013), which consists in dropping the unobserved variables in the conditional statements in Eq. (6), and thus computing the posterior density as

$$p\left(x, y, z|z^{\mathrm{o}}\right) \simeq p\left(z|z^{\mathrm{o}}\right) p(x|z)\, p(y|z). \tag{7}$$

This amounts to processing each unobserved variable in the same way as the first one of the series. The approximation (Eq. 7) limits the scheme's ability to handle complex, jointly multimodal densities, as will be illustrated with the Lorenz (1963) model in Sect. 4. An important advantage of the approximation is that it makes the analyses of the different unobserved variables independent. Thus, they can be parallelized on a computer.

3.4 MRHF parameters and possible tuning

Several of the MRHF parameters are related to the computation of rank histograms that is used to update both the observed and unobserved variables. Building a pdf with the rank histogram approach implies a division by the distance between two consecutive particles (Anderson, 2010). To avoid possible computational overflow, it is important to set a minimum spacing ϵ_{RHF} between two consecutive particles. This is done by moving each particle at a distance of ϵ_{RHF} from its closest neighbor when necessary. The particles are processed sequentially from the mean toward the tails of the distribution. For the following experiments with the Lorenz 63 system, a wide range of ϵ_{RHF} values are tested, from 10^{-6} to 10^{-2}, and the values that provide the smallest errors are retained. The main and expected conclusion is that experiments with the RHF and MRHF with large ensemble sizes are sensitive to ϵ_{RHF} within this range, because large ϵ_{RHF} tend to excessively diffuse peaked probability densities when those are built by a large number of particles close to each other.

As stressed by Anderson (2010), several choices are possible for the shape of the tails. With the Lorenz 63 system, sensitivity tests have shown that the MRHF with small ensembles and frequent observations ($\Delta t = 10$) is sensitive to the shape of the tails, and performs better with Gaussian tails. For larger ensembles, results are similar with different shapes of tails. Constant tails are specified because it is a bit cheaper computationally. Constant tails extend to prescribed values slightly beyond the model phase space boundaries. Tails are introduced only for the observed variables, because their probability densities are multiplied by the observation densities before resampling. No tail is introduced for unobserved variables. For the Lorenz 63 model, the minimal and maximal values are set to $[-20, -30, 0]$ and $[20, 30, 50]$, respectively.

An additional parameter controls the scheme's behavior when the prior distribution has multiple, well-separated modes. Figure 6 (left panel) shows the cumulative distribution function of a probability density made of two disjoint modes. Between the two modes, this function increases, although it should remain constant because the modes are disjoint. This is due to the rank histogram approach to build the probability density, and emphasized by the limited number of particles in the ensemble. In the analysis step, unrealistic particles may then appear in the region between the two modes. In the Lorenz 63 experiments, the probability density has been set to zero when below a threshold of $1/6 \times 1/(N_{\mathrm{e}} + 1)$.

As discussed in Sect. 3.3, the particle selection depends on three additional parameters. The first is the maximal distance $d_{\max}$. The specification of $d_{\max}$ should account for magnitude and the variability of the variables, the dimension of the space in which the difference is defined, and the ensemble size. With the Lorenz 63 system, we take $d_{\max} = d\sqrt{n}$, where n is the dimension of the space ($n = 1$ for the first unobserved

variable; $n = 2$ for the second unobserved variable) and d is prescribed according to the ensemble size: 1 for small ensembles ($N_e = 8$, 16, 32), 0.1 for medium ensembles ($N_e = 64$, 128), and 0.01 for the larger ensembles ($N_e = 256$, 512). To ensure a sufficient but not too large number of selected particles, it is wise to fix a minimum and a maximum number of particles. In the following experiments, those are set to 5 and 15, respectively.

Finally, like many ensemble methods, the MRHF suffers from sampling errors in the description of the densities. With the EnKF, this is usually corrected with covariance inflation. This does not make real sense with the MRHF, since the analysis does not rely on covariances. After the analysis, the particles are slightly perturbed with a white Gaussian noise, as it is often done with particle filters to avoid collapse toward one single particle. For the Lorenz 63 experiments that follow, a few values of variance have been tested in the range 0.01–0.05 for this noise, and the experiments yielding the smallest errors have been retained.

3.5 Localization

Localization consists in reducing the corrections to some variables, as computed during the analysis step, according to their distance from the observation. Beyond a certain distance, all corrections are set to 0. With the MRHF, localization is straightforward because it is a transform method. Similar to the EnKF with serial processing of observations, the analysis corrections are multiplied by coefficients that are functions of the distance to the observation. Therefore the correction may be restricted to elements of the state vector that are spatially close to the location of the observed variable, and changes to the state vector can then be gracefully tapered to zero as distance from the observation location increases. In the present work, no localization has been applied in the experiments since it is not needed on those assimilation problems. Localization will be studied in future works.

3.6 Connections with other methods

The MRHF is a non-parametric transform ensemble data assimilation method. We have presented it as a generalization of the RHF, but it also has connections with other non-parametric transform methods discussed in the Introduction, such as the method introduced by Reich (2013). This method derives from the theory of optimal transportation (or optimal mapping), whose application to sequential data assimilation has been suggested by El Moselhy and Marzouk (2012), Cotter and Reich (2013), and Reich (2013). Instead of trying to compute an approximation of the posterior pdf, the crucial idea of this theory is to find a "transfer map" f such that $f(\boldsymbol{X})$ is distributed according to the posterior pdf when $\boldsymbol{X}$ is distributed according to the prior pdf. Once such a transfer map is identified, a posterior sample can be generated from the prior sample. The transfer map is a mathematical expression of the consistency of the posterior sample depending on both the prior sample and the observation. However a transfer map is not unique (Villani, 2009). To find one, one may require the map to satisfy some additional optimality condition. Cotter and Reich (2013) and Reich (2013) propose to find the map that minimizes the expected squared distance between $\boldsymbol{X}$ and $f(\boldsymbol{X})$, so as to make the smallest possible changes to go from the prior to the posterior sample. As described in Sect. 3.1, the MRHF uses a transfer map. This map is particularly simple, since only one-dimensional probability densities are involved: we choose the map that preserves the position of the particle during its transfer from the prior to the posterior density. This also makes the smallest possible changes to go from the prior to the posterior sample.

4 Numerical experiments with the Lorenz 63 model

The Lorenz 63 model (L63) is a well-known system of three ordinary differential equations based on a simplification of atmospheric cellular convection (Lorenz, 1963). It is also a widely used test case for developments in data assimilation. The state variables are denoted X, Y, Z. The usual configuration of the model is adopted here: the parameters are set to $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$; the integration time step is $\delta t = 0.01$. In the following experiments a simulation of reference is performed and considered as the true trajectory to recover.

4.1 Fully observed state vector

4.1.1 Experimental setup and diagnostics

In this subsection, the full state (X, Y, Z) is observed. The observations are created by adding to the true trajectory independent perturbations drawn from a white Gaussian noise with standard deviation $\sigma_o = 2$ as in Harlim and Hunt (2007) and Bocquet (2011). Three different experiments are conducted for different observation time intervals: $\Delta t = 0.10$, $\Delta t = 0.25$, and $\Delta t = 0.50$. These observation time intervals are expected to provide mild, medium, and strong nonlinear test cases (Bocquet, 2011).

Each experiment is run over 10^5 assimilation cycles. To avoid any spin-up issues, a burn-in period of 1000 analysis cycles is used. Five filters are compared: the stochastic EnKF, the RHF, a particle filter, the MRHF and the MRHF with mean-field approximation (see Sect. 3). The EnKF and the RHF are tested with a large set of inflation factors and the best in terms of root mean square error are retained for comparisons. The particle filter is implemented in its sequential importance resampling (SIR particle filter) version (Gordon et al., 1993). Resampling is performed using the universal resampling method described by Whitley (1994). After resampling, the particles are perturbed with a white Gaussian noise with variance selected in the [0.01, 0.05] interval to provide the smallest errors.

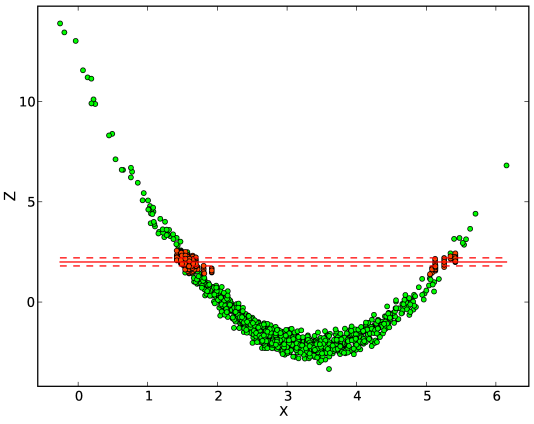

Figure 7. Same as Fig. 4, right panel, but for the MRHF.

The filters are tested for different ensemble sizes: $N_e = [8, 16, 32, 64, 128, 256, 512]$. The filters are first evaluated by the time-averaged value of the root mean square error (RMSE) between the analysis and the simulation of reference. We also evaluate the filters' approximation of the full posterior distribution using the Kullback–Leibler (KL) divergence, or relative entropy (Kullback, 1959), which measures a distance between two probability densities P and Q according to the formula:

$$\mathrm{KL}(P, Q) = \int \log \frac{P}{Q} \mathrm{d}P. \quad (8)$$

Density P refers to the density described by the ensemble after an analysis step. Ideally, the reference density Q should be the analytical solution to the problem. Since this is not available to us, we take the SIR particle filter solution with 2048 particles as a reference. It proves to be the best of all in terms of RMSE, as it will be shown in the results section. Also, by construction, the SIR particle filter provides a physically balanced solution, assuming that the noise added to each particle during the resampling step is prescribed small enough not to affect this balance significantly. For a given assimilation method, a small KL divergence guarantees that the solution is physically balanced. For a perfect computation of the KL divergence, the joint probability densities should be used. To limit the problematic effects of subsampling, especially when the ensemble sizes are small, we stick to the marginal densities of the first Lorenz variable, X. Using Y or Z provides very similar results. The marginal densities of X are recovered from the updated ensembles of each filter by using rank histograms.

4.1.2 Results

When $\Delta t = 0.10$, Fig. 8, upper panel, shows that the EnKF outperforms all the other methods for $N_e \leq 128$. This is because the system is fully and accurately observed, and frequently enough to make the analysis problem close to Gaussian. However, with $N_e \geq 256$, $\Delta t = 0.10$, the fully nonlinear methods perform slightly better. In a rather similar setup, but with an ensemble transform Kalman filter instead of a stochastic EnKF, Bocquet (2011) also concludes that the Kalman filter, perfectly designed for such problems, is extremely hard to beat. Nonetheless, the RHF and MRHFs behave rather well, even if not as well as the EnKF. In a medium nonlinear case ($\Delta t = 0.25$, central panel of Fig. 8) and for $N_e \geq 32$, both the MRHF in its full formulation and the mean-field approximated MRHF produce a smaller RMSE than the EnKF and the RHF. The SIR particle filter needs 128 particles to perform as well as the MRHFs. Finally, in a case of strong nonlinearities (Fig. 8 bottom-panel), the MRHFs outperforms the EnKF and the RHF for any ensemble size. The SIR particle filter needs in this case more than 256 particles in order to achieve similar performance. In all cases, the MRHF and the MRHF with mean-field approximation behave very similarly, the latter being even slightly better most of the time. This suggests that the mean-field approximation has a small negative impact, and that the dimensionality issue that may affect the MRHF, as described in Sect. 3.3, is already present in a three-variable system.

Figure 9 shows the counterpart of Fig. 8 for the KL divergence for the X-variable. The KL divergences are computed with respect to a SIR particle filter solution with 2048 particles. It is remarkable that other, independent 2048-SIR particle filter solutions do not provide null KL divergences. This is because the random perturbations introduced after resampling are different in the test and reference experiments. In all observation scenarios, this KL divergence approaches 1; this can then be considered as the target score for the other methods. The MRHFs perform very well, even in the mildly nonlinear case, with large ensembles. As nonlinearities grow stronger, they perform increasingly well in comparison with the others. In particular, the ensemble size required by the SIR particle filter to reach the performance of the MRHFs increases dramatically. In the strongly nonlinear case (bottom panel), the MRHFs perform better than the EnKF and RHF for any ensemble size and are only outperformed by the SIR particle filter for very large ensemble sizes ($N_e \geq 256$). In any case, the SIR particle filter performs better than the others for large ensembles.

4.2 Bimodal case – Z observed

4.2.1 Experimental setup

The L63 attractor is characterized by two lobes centered on points of attraction and connected to each other at their bottom (where the minimal values of Z are encountered). Figure 10 displays horizontal slices through the L63 attractor represented in its phase space. The two lobes are easily identified in the region $Z > 24$ (bottom row), exhibiting two or four distinct modes. In a data assimilation framework without any prior information other than the whole attractor itself,

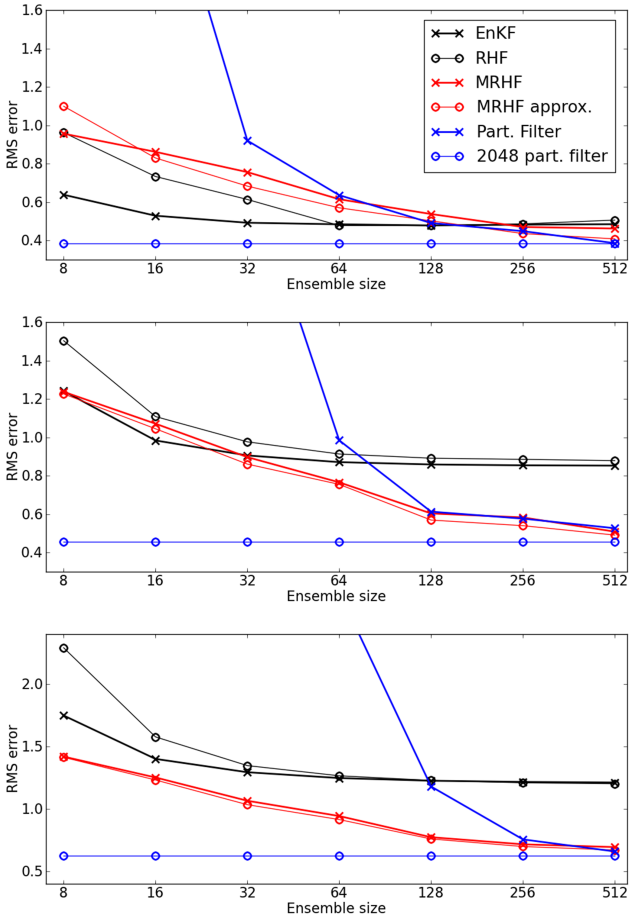

Figure 8. Time-averaged analysis root mean square error (RMSE) for the EnKF (thick black line with crosses), the RHF (thin black line with open circles), the SIR particle filter (thick blue line with crosses), the full MRHF (thick red line with crosses), and the mean-field approximated MRHF (thin red line with open circles); for experiments on the fully observed Lorenz 63 with observation time intervals $\Delta t = 0.10$ (upper panel), $\Delta t = 0.25$ (center panel), and $\Delta t = 0.50$ (bottom panel). The thin blue line with open circles represents the time-averaged analysis RMSE for the SIR particle filter with 2048 particles, which can be considered as a target score.

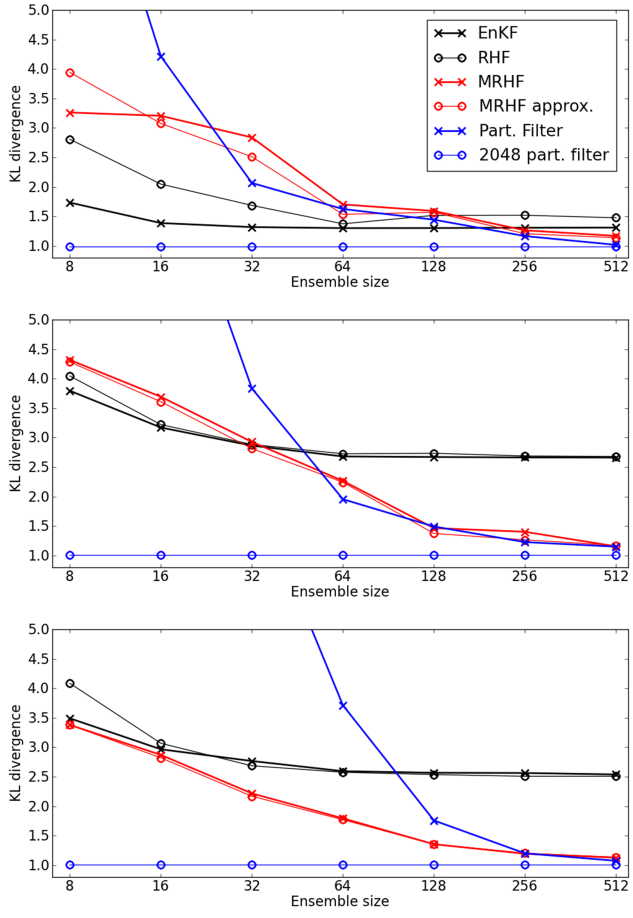

Figure 9. Same as Fig. 8 but for the mean Kullback–Leibler divergence on the X variable (results are similar on the Y and Z variables).

a single observation of Z does not enable us to determine the mode where the truth actually is. The dynamics often help to determine whether it is in an ascending branch or a descending branch of the attractor so we expect bimodal posteriors. It is thus a strongly non-Gaussian data assimilation problem.

In the following experiment, observations of Z are extracted from a "true" trajectory, perturbed with a white Gaussian noise of variance 1, and assimilated every 40 time steps ($\Delta t = 0.40$). The assimilation is conducted over 10^5 analysis cycles after a burn-in period of 1000 time steps.

The evaluation of the MRHF performance is strictly similar to the previous experiments, except that the reference solution to compute the KL divergence comes from the SIR particle filter with 4096 particles, instead of 2048. As it is argued in Sect. 4.2.2 below, the RMSE is not a meaningful diagnostic in this case, making it difficult to verify that the 4096-SIR particle filter provides an accurate solution. However, the objective of that test relies on the fact that an appropriate data assimilation method should be able to maintain the representation of the bimodality in this particular case. The SIR particle filter, given a substantial number of particles, might generate overly dispersive ensembles but does maintain the bimodality (Fig. 12). It has been checked that it is true for the whole integration period. Hence, a small KL divergence between the methods and the 4096-SIR particle filter will confirm that the bimodality is respected.

4.2.2 Results

The time-averaged RMSE is a classical performance diagnostic in data assimilation. However, when the solution is bimodal, the RMSE does not tell much. Here, the RMSE vary between 3 and 7, whatever the method and the ensemble size are, and they do not decrease with increasing ensemble sizes.

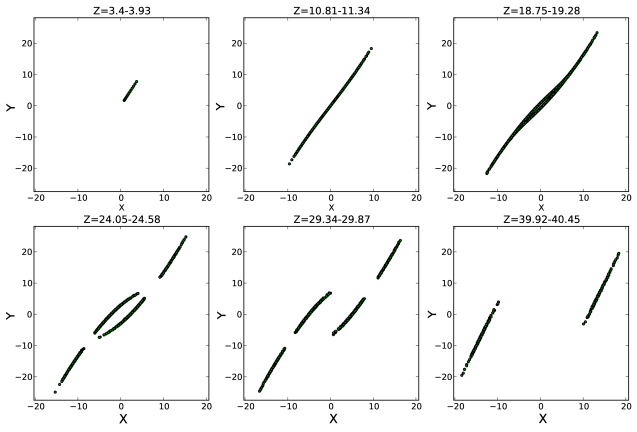

Figure 10. Horizontal slices (in X–Y planes, Z intervals indicated on tops) through L63 attractor. In the two bottom left graphs, the two modes in the center are parts of the descending branches of the lobes (Z decreases with time), while the modes in the corners are the ascending branches.

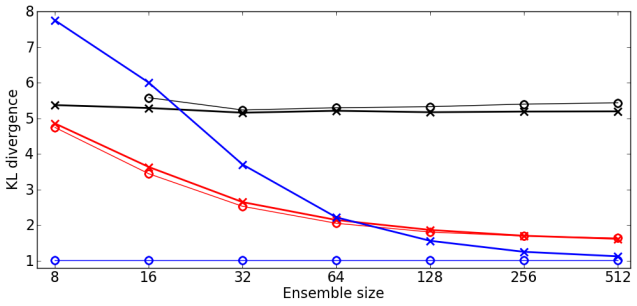

Figure 11. Time-averaged Kullback–Leibler divergence on the X variable for the EnKF (thick black line with crosses), the RHF (thin black line with open circles), the SIR particle filter (thick blue line with crosses), the full MRHF (thick red line with crosses), and the mean-field approximated MRHF (thin red line with open circles); for the experiment on Lorenz 63 when only the third variable (Z) is observed every 40 time steps during 10^5 analysis cycles. The thin blue line with open circles represents the time-averaged Kullback–Leibler divergence for the SIR particle filter with 2048 particles, which can be considered as a target score.

Since they provide no relevant information, the RMSE are not discussed further.

The KL divergence is a much more appropriate diagnostic in this case, especially using the marginal densities of X, since bimodality is expected in the X direction. Figure 11 displays the KL divergences as a function of ensemble size, for the 5 methods considered in the previous setup. The RHF with 8 particles proved to be unstable with any inflation factor. The fully non-parametric methods (SIR particle filter, MRHFs) yield much better scores than the others (EnKF and RHF). The KL divergence of the EnKF and RHF does not depend on the ensemble size, confirming that these methods are not designed to deal with such multimodal problems. The SIR particle filter needs at least $N_e = 128$ particles to obtain a smaller KL divergence than the MRHFs.

To illustrate the differences in the behaviors of the various methods, Fig. 12 depicts scatter plots of the analysis ensembles at the 389th cycle, given by the reference SIR particle filter (with 4096 particles), and the other five methods with 256 particles. The scatter plots are drawn in the X–Y plane, and the true state is shown by the red squares and red dashed lines. The 389th cycle has been chosen arbitrarily for this illustration. Similar behavior of the filters is observed throughout the experiment.

While the correct posterior distribution is bimodal, the EnKF and the RHF tend to create particles between the two modes. This explains their large KL divergences with respect to the reference solution. The SIR particle filter and the MRHF, on the other hand, provide visually correct, bimodal solutions in this case (but they can be subject to mode leakage too, at other cycles). The MRHF with mean-field approximation is again broadly similar to the full MRHF. Because the correction to Y is independent of the correction to X, however, a few particles can switch from one mode to the other along the X (resp. Y) direction without switching along the Y (resp. X) direction. These particles appear in unrealistic regions of the model phase space (regions that cannot be visited by the attractor, near $X = -10$, $Y = 10$ or $X = 10$, $Y = -10$). This illustrates the limitation of the MRHF with mean-field approximation to deal with bimodal distribution. Mode leakage of this kind is significantly reduced by the deterministic resampling method used in the MRHF, in comparison with a stochastic method (result not shown). This is because the MRHF method, described in Sect. 3.2, preserves the relative rank of particles at the analysis step. Also, the Lorenz 63 system does not seem affected by a few outliers, and the KL divergence scores of the MRHF with mean-field approximation are good.

5 An illustration of density estimation

We next consider the analysis step for each scheme in a more complex model, with no forecast loop. The interest of this illustration is to observe their behavior in a first realistic context.

5.1 The marine biogeochemical context

The interactions between ocean dynamics and biogeochemistry are complex. The variations of the mixed layer depth (MLD) strongly influences the nutrient supply, hence the phytoplankton production in the euphotic layer (Dutkiewicz et al., 2001). MLD variations are themselves controlled, at least for a large part, by variations in the wind forcing. With the growing interest in understanding ocean biogeochemical cycles and thanks to the increasing amount of dedicated satellite missions (SeaWiFS, MERIS, MODIS), the

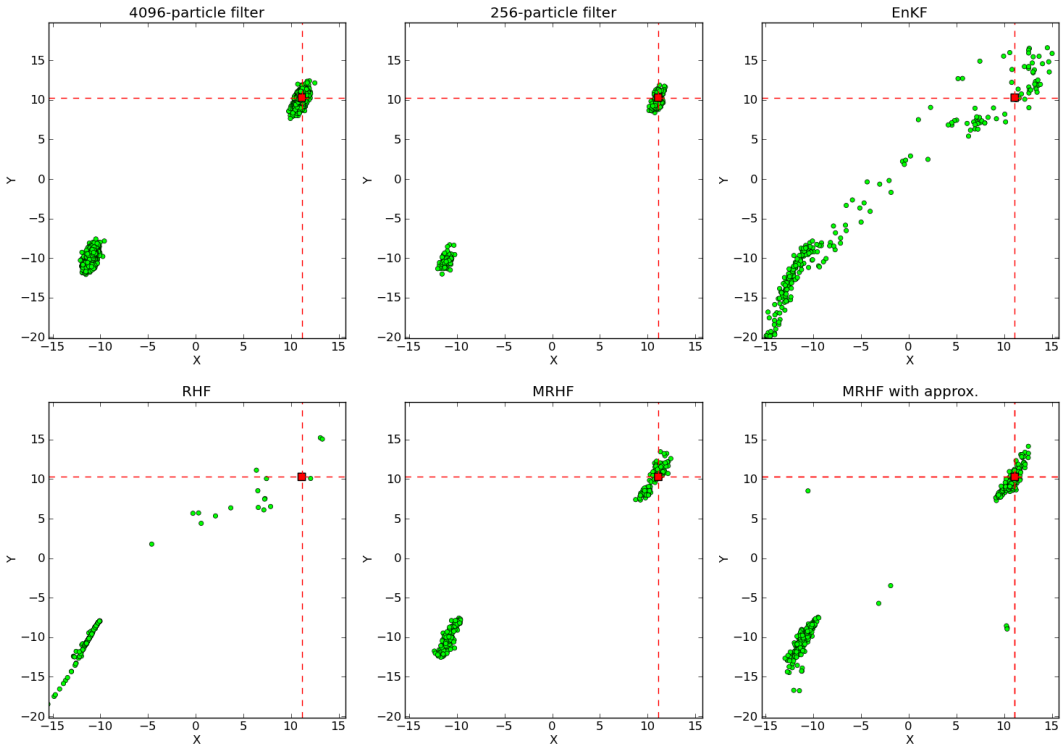

Figure 12. Scatter plots of the analysis ensembles with 256 particles (except upper-left) at the 389th analysis cycle, in the X–Y plane of Lorenz 63 model phase space: 4096-SIR particle filter (considered as the reference solution) (upper left panel); Particle filter (upper center panel); EnKF (upper right panel); RHF (bottom left panel); MRHF (bottom center panel); Mean-field approximation MRHF (bottom right panel). The red dashed lines indicate the truth from which the observation of Z is produced.

assimilation of ocean color data, a proxy of chlorophyll concentration, has taken off in the last few years (Gregg et al., 2009). A better estimation of the dynamical variables, including wind forcing, appears as an interesting by-product. Because of the nonlinear relationships of biogeochemical variables between each other and with dynamical variables, ocean biogeochemical data assimilation is a fundamentally non-Gaussian problem. This is well demonstrated by Béal et al. (2010). They use a three-dimensional coupled physical–biogeochemical model of the North Atlantic with a 1/4° horizontal resolution. A 200-particle ensemble simulation is run, perturbing the wind forcing, during the 1998 North Atlantic spring bloom. The nonlinear model response, expressed in the non-Gaussianity of the bivariate distributions of the ensemble variables, is clearly demonstrated.

5.2 The density estimation experiment

The data used for the illustration below comes from the ensemble simulation of Béal et al. (2010). We focus on the 1 day forecast at the Gulf Stream station (47° W, 40° N).

The upper-left graphs of Figs. 13, 14, and 15 respectively show the forecast ensemble in the chlorophyll (CHL)–mixed layer depth (MLD) plane, the chlorophyll (CHL)–biogeochemical detritus (DET) plane and the biogeochemical detritus (DET)–mixed layer depth (MLD) plane. Each green dot represents an ensemble particle and the blue dot shows the values from a reference model run. A chlorophyll observation is created by perturbing the value from this reference with a Gaussian white noise with variance $\sigma_o = 0.001$.

Each of the assimilation schemes from the previous section is applied, at a fixed time step and a fixed grid point, to the seven-variable control vector. The control vector is composed of seven prognostic variables of the model: one dynamical variable (MLD) and six biological variables (chlorophyll, detritus, dissolved organic matter, NH_4, NO_3 and phytoplankton). The 200-particle forecast ensemble is used as a prior distribution to test various analysis schemes using the chlorophyll observation. The other panels of Figs. 13, 14, and 15 display the estimated ensembles, with red triangles, on, respectively, the chlorophyll (CHL)–mixed layer depth (MLD) plane, the chlorophyll (CHL)–biogeochemical detritus (DET) plane and the biogeochemical detritus (DET)–mixed layer depth (MLD) plane, obtained using the EnKF (upper right graphs), the RHF of Anderson (center left graphs), the full MRHF (center right graphs), the mean-field approximation MRHF (bottom-left graphs), and the SIR particle filter (bottom right graphs) analysis steps. The chlorophyll observation is represented by the blue full line;

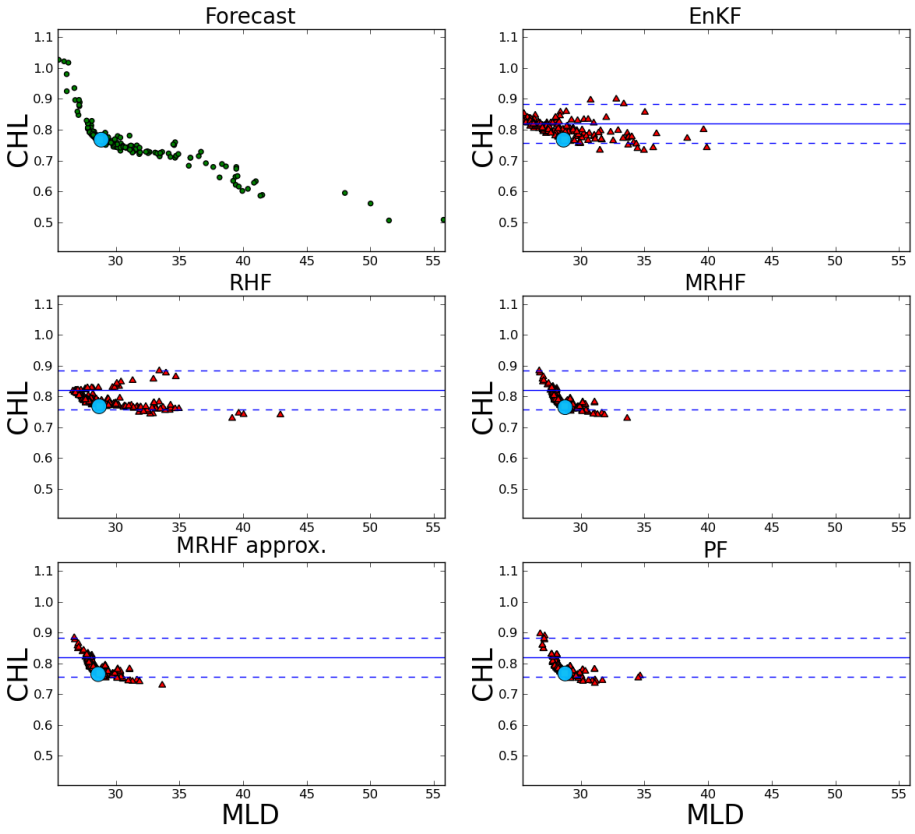

Figure 13. Prior forecast ensemble (upper left panel) and posterior analysis ensembles (red triangles), on the chlorophyll (CHL)–mixed layer depth (MLD) plane, produced by the EnKF (upper right panel), the RHF (center left panel), the full MRHF (center right panel), the mean-field approximation MRHF (bottom-left), and the SIR particle filter (bottom right panel). Only chlorophyll is observed with the value $\mathrm{CHL_o}$ indicated by the blue line on each panel. The two dashed blue lines represent the interval $[\mathrm{CHL_o} - 2\sigma_o, \mathrm{CHL_o} + 2\sigma_o]$, where $\sigma_o^2 = 0.001$ is the observation error standard deviation. The blue dot represents the true state, from which the chlorophyll value has been extracted and perturbed to form the observation $\mathrm{CHL_o}$.

the blue dashed lines are at a distance of $2 \times \sigma_o$ from the blue full line.

A data assimilation method is obviously expected to move the prior particles close to the observations. However, a major requirement for those methods, especially in the non-Gaussian context, is also to maintain the information on the model attractor contained in the shape of the prior ensemble dispersion. In Figs. 13 and 14, the prior ensemble clearly says that the observed variable (CHL) and the unobserved variables (i.e., MLD and DET) have an obvious statistical connection, but this connection is not linear. Thus, the methods using a linear regression in the physical space to correct the unobserved variable, that is, EnKF and RHF, fail to maintain the non-Gaussian shape of the prior density. The SIR particle filter analysis step is a very good approximated implementation of Bayes' theorem (modulo sampling errors). Hence, with a sufficiently large ensemble, it provides a posterior ensemble consistent with the observation and the prior ensemble. In comparison the full MRHF analysis step also manages to produce a posterior distribution consistent with both the observation and the prior ensemble. Nevertheless, it clearly appears in Fig. 15 that the curse of dimensionality striking during the particle selection process (as discussed in Sect. 3.3) degrades the correction on the second unobserved variable. The mean-field approximation discussed in Sect. 3.3, appears to overcome this issue and produces a posterior distribution very similar to the SIR particle filter.

Figure 15 displays the posterior ensembles in the mixed layer depth (MLD)–biogeochemical detritus (DET) plane. Those graphs allow us to observe the bivariate densities between two unobserved state variables. It is known that the EnKF and the RHF appropriately maintain the covariances between all pairs of variables in a linear and Gaussian context. However, in a non-Gaussian case such as this one, Fig. 15 shows that both filters do not provide an appropriate ensemble update (in the sense of Bayes' rule). Meanwhile, the MRHF and the SIR particle filter, provide an estimated ensemble taking into accout the relationship between all pairs of variables by respecting the information contained in the prior distribution.

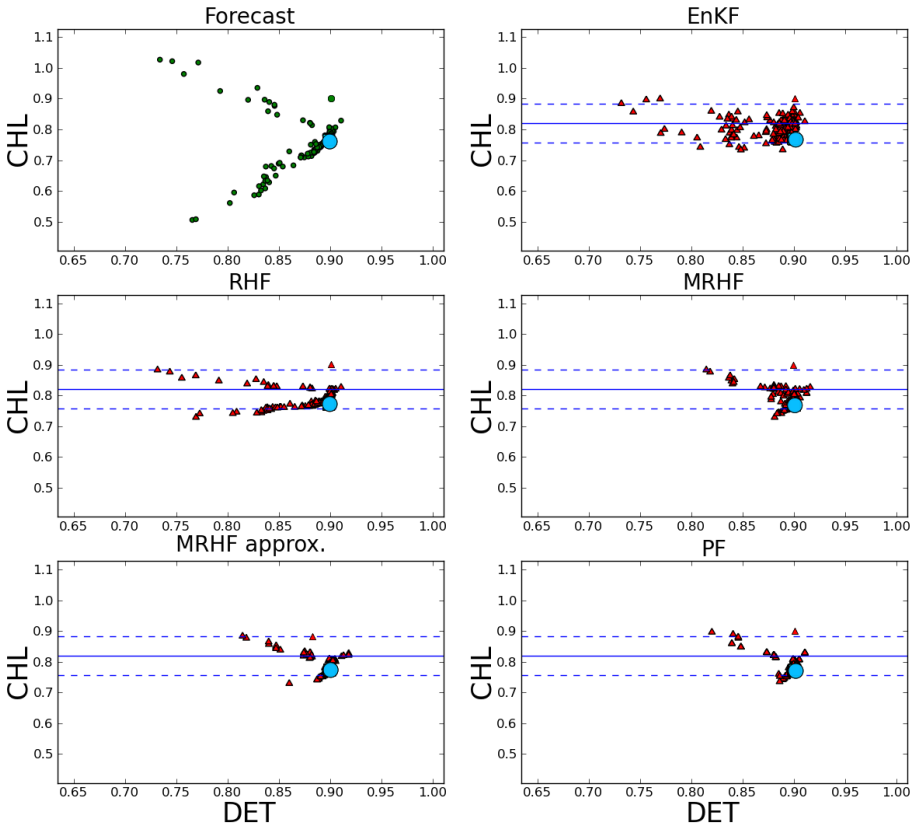

Figure 14. Same as Fig. 13 but on the chlorophyll (CHL)–biogeochemical detritus (DET) plane.

6 Discussion and conclusions

This paper has introduced the multivariate rank histogram filter (MRHF), a fully non-Gaussian analysis scheme for ensemble data assimilation. This filter is an extension of the rank histogram filter introduced by Anderson (2010). In order to set the MRHF in the wider context of non-Gaussian analysis methods, the behavior of some of these methods has been examined in idealized, bivariate frameworks where the variables were jointly Gaussian, weakly non-Gaussian, or strongly non-Gaussian (Sect. 2). The MRHF clearly falls into the category of methods able to deal with strong non-Gaussianity, along with the particle filters. An approximated version of the MRHF, based on the mean-field approximation (Cotter and Reich, 2013), is also proposed.

Numerical experiments with the Lorenz 63 model, in a data assimilation problem with fully observed state vector but for different observation time intervals corresponding to different levels of nonlinearity (Sect. 4.1), confirm that the MRHF does not perform better than the EnKF in quasi-linear context with small ensemble sizes. Nevertheless, in this context, the MRHF performs slightly better with larger ensembles ($N_e \geq 256$). When nonlinearity is stronger, the MRHF considerably reduces the root mean square error and the Kullback–Leibler divergence in comparison with other methods when given a sufficiently large number of particles ($N_e \geq 64$). The MRHF with mean-field approximation exhibits very similar performance. Experiments in the most nonlinear regime, characterized by the bimodality of the state density, confirm the ability of the MRHF to handle strong non-Gaussianities (Sect. 4.2). Finally, an experiment with prior data from a coupled physical–biogeochemical model (Sect. 5) illustrates the behavior of the MRHF analysis (in its full and approximated forms) when facing strongly non-Gaussian densities in a more explicitly geophysical context. This illustration is not a full data assimilation problem but the posterior densities produced by the MRHF analysis show that Bayes' theorem is correctly approximated in this formulation.

Aside from documenting the performance of the MRHF, the experiments in this paper show the importance of matching the assimilation method to the level of non-Gaussianity in the problem at hand. Even though a general method such as the MRHF should perform well in any situation, the fact is that the EnKF (or other linear methods) are perfectly adequate in many applications and often much less expensive computationally. An advantage of the MRHF in this respect is that it is easily hybridized with other serial transform methods, such as the EnKF, because such schemes process observations serially and update observed and unobserved

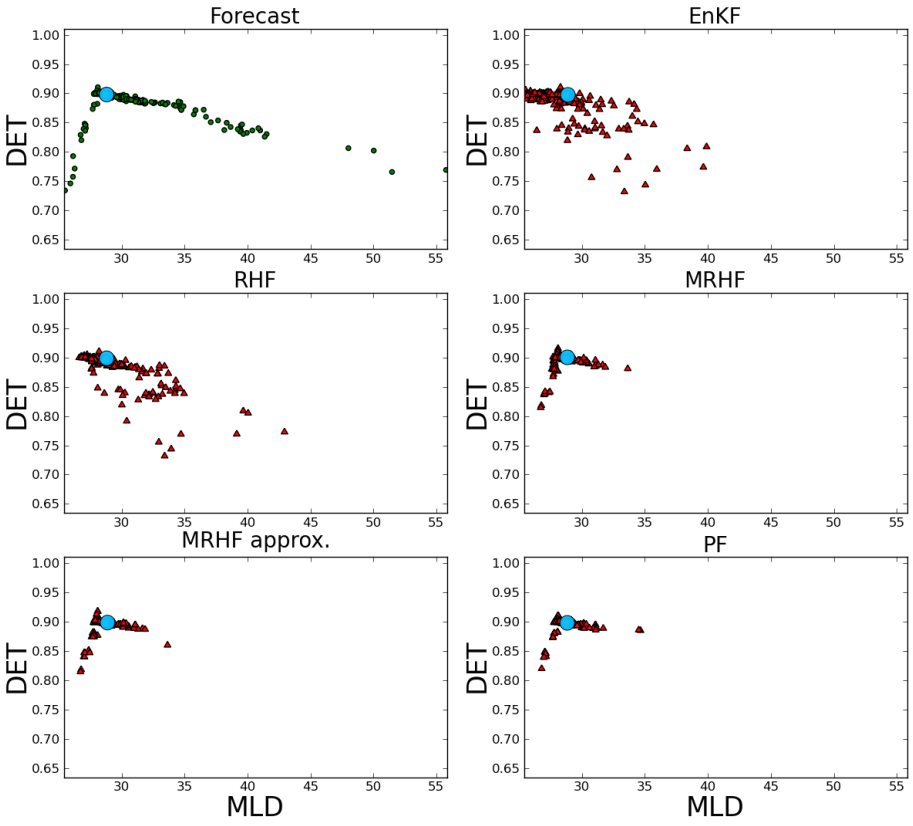

Figure 15. Same as Fig. 13 but on the biogeochemical detritus (DET)–mixed layer depth (MLD) plane.

variables serially. One can then consider, for example, updating wind components with the EnKF but the more non-Gaussian humidity with the MRHF.

Relative to Kalman filters, the MRHF's deterministic scheme has assets to preserve physical balances during the analysis. Relative to particle filters, the MRHF also has advantages in addressing the curse of dimensionality. As explained in Sect. 3.5, spatial localization of the update step comes naturally with the MRHF, although it has not been tested here. Also, effects of the curse of dimensionality on the MRHF can be greatly reduced using the mean-field approximation, as illustrated in Sect. 3.3. In fact, use of this approximation is likely necessary for successful implementation of the MRHF in many realistic problems. Finally, it is relatively easy to reduce any sensitivity of the MRHF to biases in observed variables, and to decrease the associated tendency for filter divergence, because of the freedom to choose the tails of the prior density for these variables (Sect. 2). Then, contrary to the particle filters (at least in their bootstrap formulation, Gordon et al., 1993), a large number of observations helps in avoiding divergence.

According to our experiments, the MRHF is much more expensive than the EnKF in terms of computation. The MRHF analysis proves to be approximately 50, 100, and 200 times more expensive than the EnKF for typical ensemble sizes of 64, 128, and 256, respectively. The MRHF might then be best suited for rather specific problems with strong non-Gaussianity and few observations or as part of hybrid schemes where the MRHF is used only for certain variables. Nevertheless, spatial localization and a more efficient implementation of the mean-field approximation have the potential to greatly reduce computation cost and expand the range of problems for which the MRHF is feasible. Both aspects need to be investigated before the application of the MRHF to realistic problems and these are the next steps of this work.

Acknowledgements. This work was supported by the Région Rhônes-Alpes and NCAR. It is also a contribution to the SANGOMA project supported by European Comissions Seventh Framework Programme FP7/2007–2013, grant agreement no. 283580, and to the CNRS/INSU/LEFE program. The authors are grateful to Sebastian Reich and two other anonymous reviewers for their relevant and constructive comments.

Edited by: T. Gneiting
Reviewed by: S. Reich and two anonymous referees

The publication of this article is financed by CNRS-INSU.

References

Anderson, J. L.: A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations, J. Climate, 9, 1518–1530, 1996.

Anderson, J. L.: A local least squares framework for ensemble filtering, Mon. Weather Rev., 131, 634–642, 2003.

Anderson, J. L.: A non-Gaussian ensemble filter update for data assimilation, Mon. Weather Rev., 138, 4186–4198, 2010.

Anderson, J. L.: Localization and Sampling Error Correction in Ensemble Kalman Filter Data Assimilation, Mon. Weather Rev., 140, 2359–2371, 2012.

Béal, D., Brasseur, P., Brankart, J.-M., Ourmières, Y., and Verron, J.: Characterization of mixing errors in a coupled physical biogeochemical model of the North Atlantic: implications for nonlinear estimation using Gaussian anamorphosis, Ocean Sci., 6, 247–262, doi:10.5194/os-6-247-2010, 2010.

Bertino, L., Evensen, G., and Wackernagel, H.: Sequential Data Assimilation Techniques in Oceanography, Internat. Stat. Rev., 71, 223–241, 2003.

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

Bocquet, M.: Ensemble Kalman filtering without the intrinsic need for inflation, Nonlin. Processes Geophys., 18, 735–750, doi:10.5194/npg-18-735-2011, 2011.

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

Brankart, J.-M., Testut, C.-E., Béal, D., Doron, M., Fontana, C., Meinvielle, M., Brasseur, P., and Verron, J.: Towards an improved description of ocean uncertainties: effect of local anamorphic transformations on spatial correlations, Ocean Sci., 8, 121–142, doi:10.5194/os-8-121-2012, 2012.

Buehner, M., Houtekamer, P. L., Charette, C., Mitchell, H. L., and He, B.: Intercomparison of Variational Data Assimilation and the Ensemble Kalman Filter for Global Deterministic NWP, Part I: Description and Single-Observation Experiments, Mon. Weather Rev., 138, 1550–1566, 2010.

Candille, G. and Talagrand, O.: Evaluation of probabilistic prediction systems for a scalar variable, Q. J. Roy. Meteorol. Soc., 131, 2131–2150, 2005.

Cohn, S. E.: An introduction to estimation theory, J. Meteorol. Soc. Jpn., 75, 257–288, 1997.

Cotter, C. J. and Reich, S.: Ensemble filter techniques for intermittent data assimilation, in Large Scale Inverse Problems, Radon Ser. Comput. Appl. Math., 13, 91–134, 2013.

Dee, D. and Da Silva, A. M.: The choice of variable for atmospheric moisture analysis, Mon. Weather Rev., 131, 155–171, 2003.

Doucet, D., de Freitas, N., and Gordon, N.: An introduction to sequential Monte Carlo methods, in: Sequential Monte Carlo Methods in Practice, edited by: Doucet, D., de Freitas, N., and Gordon, N., Statistics for Engineering and Information Science, Springer-Verlag, New York, 2001.

Dutkiewicz, S., Follows, M., Marshall, J., and Gregg, W. W.: Interannual variability of phytoplankton abundances in the North Atlantic, Deep-Sea Res., Pt. II, 48, 2323–2344, 2001.

El Moselhy, T. A. and Marzouk, Y. M.: Bayesian inference with optimal maps, J. Comput. Phys., 231, 7815–7850, 2012.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res., 99, 10143–10162, 1994.

Evensen, G.: The Ensemble Kalman Filter: theoretical formulation and practical implementation, Ocean Dynam., 53, 343–367, 2003.

Fournier, A., Hulot, G., Jault, D., Kuang, W., Tangborn, A., Gillet, N., Canet, E., Aubert, J., and Lhuillier, F.: An Introduction to Data Assimilation and Predictability in Geomagnetism, Space Sci. Rev., 155, 247–291, 2010.

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

Gregg, W. W., Friedrichs, M. A., Robinson, A. R., Rose, K. A., Schlitzer, R., Thompson, K. R., and Doney, S. C.: Skill assessment in ocean biological data assimilation, J. Mar. Syst., 76, 16–33, 2009.

Greybush, S. J., Kalnay, E., Miyoshi, T., Ide, K., and Hunt, B. R.: Balance and Ensemble Kalman Filter Localization Techniques, Mon. Weather Rev., 139, 511–522, 2011.

Hamill, T. M.: Interpretation of Rank Histograms for Verifying Ensemble Forecasts, Mon. Weather Rev., 129, 550–560, 2001.

Hamill, T. M., Whitaker, J. S., and Snyder, C.: Distance-Dependent Filtering of Background Error Covariance Estimates in an Ensemble Kalman Filter, Mon. Weather Rev., 129, 2776–2790, 2001.

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

Hoteit, I., Luo, X., and Pham, D.-T.: Particle Kalman Filtering: A Nonlinear Bayesian Framework for Ensemble Kalman Filters, Mon. Weather Rev., 140, 528–542, 2012.

Houtekamer, P. L. and Mitchell, H. L.: Data Assimilation Using an Ensemble Kalman Filter Techniqu, Mon. Weather Rev., 126, 796–811, 1998.

Houtekamer, P. L. and Mitchell, H. L.: A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation, Mon. Weather Rev., 129, 123–137, 2001.

Kalman, R.: A new approach to linear filtering and prediction problems, Trans. Am. Soc. Mech. Mech. Eng. J. Basic Eng., 82, 35–45, 1960.

Kullback, S.: Information Theory and Statistics, Wiley, New York, 1959.

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

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

Lei, J., Bickel, P., and Snyder, C.: Comparison of Ensemble Kalman Filters under Non-Gaussianity, Mon. Weather Rev., 138, 1293–1306, 2010.

Lermusiaux, P. F. J.: Data assimilation via error subspace statistical estimation, Part II: Middle Atlantic Bight shelfbreak front simulations and ESSE validation, Mon. Weather Rev., 127, 1408–1432, 1999.

Lermusiaux, P. F. J.: Uncertainty estimation and prediction for interdisciplinary ocean dynamics, J. Comp. Phys., 217, 176–199, 2006.

Lorenz, E.: Deterministic Nonperiodic Flow, J. Atmos. Sci., 20, 130–141, 1963.

Morzfeld, M., Tub, X., Atkinsa, E., and Chorin, A.: A random map implementation of implicit filters, J. Comp. Phys., 231, 2049–2066, 2012.

Nakano, S., Ueno, G., and Higuchi, T.: Merging particle filter for sequential data assimilation, Nonlin. Processes Geophys., 14, 395–408, doi:10.5194/npg-14-395-2007, 2007.

Pham, D. T.: Stochastic Methods for Sequential Data assimilation in Strongly nonlinear systems, Mon. Weather Rev., 129, 1194–1207, 2001.

Reich, S.: A nonparametric ensemble transform method for Bayesian inference, SIAM J. Sci. Comput., 35, A2013–A2024, 2013.

Sakov, P. and Bertino, L.: Relation between two common localisation methods for the EnKF, Comput. Geosci., 15, 225–237, 2010.

Sakov, P., Counillon, F., Bertino, L., Lisæter, K. A., Oke, P. R., and Korablev, A.: TOPAZ4: an ocean-sea ice data assimilation system for the North Atlantic and Arctic, Ocean Sci., 8, 633–656, doi:10.5194/os-8-633-2012, 2012.

Simon, E. and Bertino, L.: Application of the Gaussian anamorphosis 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, doi:10.5194/os-5-495-2009, 2009.

Snyder, C.: Particle filters, the optimal proposal and high-dimensional systems, Proc. Seminar on Data Assimilation for Atmosphere and Ocean, ECMWF, Reading, Berkshire, 161–170, 2012.

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

Tarantola, A.: Inverse problem theory and methods for model parameter estimation, SIAM, Philadelphia, USA, 2005.

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

van Leeuwen, P.: Nonlinear data assimilation in geosciences: an extremely efficient particle filter, Q. J. Roy. Meteorol. Soc., 136, 1991–1999, 2010.

Villani, C.: Optimal transportation: Old and new, Springer-Verlag, Berlin, Heidelberg, 2009.

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

Whitaker, J. S., Hamill, T. M., Wei, X., Song, Y., and Toth, Z.: Ensemble data assimilation with the NCEP Global Forecast System, Mon. Weather Rev., 136, 463–482, 2008.

Whitley, D.: A genetic algorithm tutorial, Statist. Comput., 4, 65–85, 1994.

Wikle, C. and Berliner, L.: A Bayesian tutorial for data assimilation, Physica D, 230, 1–16, 2007.