Unscented/Ensemble Transform-based Variational Filter

Ming Lei[d,e], Christophe Baehr[e,f]

[a]*School of Electronic, Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai*
[b]*Météo-France, National Meteorological Research Centre/CNRS, Toulouse*
[c]*Institut Mathématiques de Toulouse, Université de Toulouse III - Paul Sabatier*

Abstract

To deal with high dimensional non-linear filtering problems, an hybrid scheme called Unscented/Ensemble transform Variational Filter (UEVF) is introduced. This is the combination of an Unscented Transform (UT), an Ensemble Transform (ET) and a rank-reduction method to compute the background covariance error matrices as well as a variational minimization to conduct the mean correction. The proposed UEVF is more efficient than the Unscented Kalman Filter (UKF) to estimate the ensemble mean and covariance by the blending of a variational optimization instead a Kalman linear correction as well as the ET-like covariance estimation into the update step. Moreover, in order to tackle the high dimension dynamics, truncated singular value decomposition is applied to provide a size reduction of sigma-points set with an adaptive fashion. For performance verifications, we present two numerical experiments with different dynamics. The first system is the chaotic and high dimensional Lorenz-95 model. We show the performance of different filters including the UEVF as the increasing of dimensionality or noise level. The second simulation is a model based on the 2D shallow water equation. The same tests are provided on this hydrodynamical system. All the numerical experiments confirm that the UEVF outperforms the widely applied Kalman-like filters explicitly.

Keywords: Variational Filter, Ensemble Kalman Filter(EnKF), Unscented Transform, Ensemble Unscented Kalman Filter(EnUKF), Ensemble Transform, Rank Reduction Method

Email addresses: minglei.sa@gmail.com (Ming Lei), christophe.baehr@meteo.fr (Christophe Baehr)

Unscented/Ensemble Transform-based Variational Filter

Ming Lei[d,e], Christophe Baehr[e,f]

[d]*School of Electronic, Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai*
[e]*Météo-France, National Meteorological Research Centre/CNRS, Toulouse*
[f]*Institut Mathématiques de Toulouse, Université de Toulouse III - Paul Sabatier*

1. Introduction

To determine the initial conditions of a numerical weather forecast (NWF) is one of the major challenge of computational methods [1, 2]. The assimilation is not only necessary to provide an initial state to the NWF, but also to filter observational noises and to learn the model errors. The techniques have evolved with the computational capabilities, since the optimal interpolator to the 3DVAR [3] and recently the 4DVAR [4]. In the classical variational filters for meteorological problems, the covariance error matrix is learned statistically offline. At the same time, the engineering community provides numerous versions of the classical Kalman Filter (KF) adapted to large type of problem and all, taking advantages of the different systems with Gaussian noises, and learning their covariance error matrices online.

Regardless of this progress, during the 90's the particle filter was developed to solve directly the filtering Bayes formulae ([5] or later [6]). G. Evensen [7] first combine the KF conception and Monte-Carlo methods with his Ensemble Kalman Filter (EnKF) which has encountered a real echo in the geophysical sciences [8]. Then F. LeGland [9] proves that the EnKF is not a consistent estimator of the filtering problem. Papadakis et al [10] shows that to regularize the estimator and have a convergence to the filtering process, the EnKF has to be weighted.

The filtering problem is the estimation of a conditional probability following the Bayes rule. Thinking about a probability density function (PDF)

Email addresses: minglei.sa@gmail.com (Ming Lei), christophe.baehr@meteo.fr (Christophe Baehr)

transportation, J. Uhlmann and S. Julier proposed [11] the Unscented Transform (UT) which is supposed to rotate and shape the PDF. Some UT variants of the common filters have been derived (see [12] for the UT Kalman Filter (UKF) and [13] for the UT Particle Filter (UPF)).

In meteorological or geophysical sciences, the greatest challenge is the dimensionality of the system. But the dimensionality is not the degrees of freedom. High dimensional physical systems have correlated dimensions. The dynamics organizes itself along few directions. Taking up the idea of B. Moore [14] for engineering, B. Farrell and P. Ioannou suggest [15] to use the rank-reduced KF for some linearized geophysical problems (see also [16]). In order to decrease the computational cost, independently some authors have developed the square root decomposition and some derived KF [17]. Combining with the UT, Van der Merwe have suggested an unscented KF with a square root decomposition [18].

The ensemblist community has been in keeping with these ideas and settle some variants of the EnKF using a rank reduction or square root decomposition. X. Luo and I. Moroz have proposed [19] a full mixing of these techniques with the Ensemble Unscented Kalman Filter (EnUKF). The EnUKF is then able to treat some high dimensional systems with a variable sample size, which depends of the possible rank reduction and a pre-computed sample covariance. Although there exists debate on the combination method [20, 21], but the numerical results are convincing.

Our work takes place at this sensitive point. For all the KF-like filters the estimate is assumed to be a linear regression with the observation (see section 2.2.1 for details). This is the core of KF but this is hard to accept for general situations especially for a dynamics with strong non-linearities and non-Gaussian perturbations. In certain sense it becomes a bottleneck for some complex and large dimension applications. Therefore we develop an assimilation method with an update step that based on variational minimization instead of the Kalman regression. Naturally the linear regression is replaced by quadratic terms of observation. First an UT is applied for the sample generation and the statistical mean estimation. Then an Ensemble Transform (ET) is used to empirically compute the error covariance. To deal with the high dimensionality, we incorporate a modified rank reduction into the UT scheme in order to maintain a size-reduced ensemble generation. Therefore the new scheme is called the Unscented/Ensemble transform-based Variational Filter (UEVF). This filter is able to deal with the nonlinear and chaotic systems with high dimensionality as the numerical experiments show

in section 5.

It is clear that the suggested method, outlined in section 3, is at the junction of different technologies: The UT approach in order to give an evolution to the error covariance matrices, the ensemble transform to implement the covariance update, the variational minimization to get a state update for a nonlinear dynamics, the rank reduction to achieve a significant dimension reduction with an adaptive fashion.

This paper is divided into 6 sections. In section 2, we introduce the background methodologies, including three update schemes. In section 3, we derive the UEVF and introduce the variational minimization update, and the sigma-points generation with a size-reduced dynamics. In section 4, three implementation issues about the UEVF are discussed. First we present the rank-reduced covariance approximation, then the conjugate-gradient method to minimize the cost-function, and finally the computation of the background error matrix. The section 5 concerns the numerical experiments. There are two examples, the chaotic high dimensional Lorenz-95 model and a simulation based on the 2D shallow water equation. Finally the conclusion and a discussion are provided in section 6.

2. Background methods

We assume that we have an n-dimensional discrete dynamical system

$$\mathbf{x}_{k+1} = \mathcal{M}_{k,k+1}(\mathbf{x}_k, \mathbf{u}_k) \tag{1}$$

and to observe the state $\mathbf{x}_k$ we could (partly) obtain a measurement sequence $\mathbf{y}_k$ using the observer $\mathcal{H}_k : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$, i.e.,

$$\mathbf{y}_k = \mathcal{H}_k(\mathbf{x}_k, \mathbf{v}_k)\,, \quad k = 1, 2, \cdots \tag{2}$$

where $k \in [0, T]$ is the discrete-time index and T the total time step, $\mathbf{x}_k \in \mathbb{R}^n$ and $\mathbf{y}_k \in \mathbb{R}^m$ are the state and the noisy observation at the time k, $\mathcal{M}_{k,k+1}$ is the nonlinear transition operator with $\mathcal{M}_{k,k+1} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$. $\mathbf{u}_k \in \mathbb{R}^n$ is the dynamical noise with a zero mean and non null covariance $\mathbf{Q}_k$, which includes both the ordinary process noise and mismatch errors of mathematical model. We assume that $\mathbf{u}_k$ is independent of the observation noise $\mathbf{v}_k \in \mathbb{R}^m$ with a zero mean and known covariance $\mathbf{R}_k$. $\mathbf{u}_k$ is also supposed to be independent of the initial state $\mathbf{x}_0$.

The problem of data assimilation is equivalent to estimate the probability law $\eta_k = \text{Law}(\mathbf{x}_k|\mathbf{y}_{1:k})$. Two probability laws are used to describe it completely: the first is the predictor law using the Markov transition given by the model, $\eta_k^b = \text{Law}(\mathbf{x}_k^b|\mathbf{x}_{k-1}^a)$, the second is the update law $\eta_k^a = \text{Law}(\mathbf{x}_k^a|\mathbf{x}_k^b, \mathbf{y}_{1:k})$. Finding these two laws solves the assimilation or the filtering problem.

Actually the update process has not an unique representation as a Markov transition. The transformation η_k^b may be represented by different transportation processes that are equivalent in mean. This can be see with the Feynman-Kac formulae [22].

2.1. Unscented transform

The scheme of unscented transform (UT) is designed to solve the following estimation problem[19, 23, 24]: at time $k-1$, we have a Gaussian random variable $\mathbf{x}_{k-1} \in \mathbb{R}^n$ with mean $\mathbf{x}_{k-1}^a$ and covariance $\mathbf{P}_{xx,k-1}^a$, and a Gaussian perturbations $\mathbf{u}_{k-1}$ with zero-mean and covariance $\mathbf{Q}_{k-1}$, which are assumed to be independent of each other. Without loss of generality, we can apply a nonlinear transition $\mathcal{M}$ on the $\mathbf{x}_{k-1}$ to obtain a new random variable $\mathbf{x}_k = \mathcal{M}(\mathbf{x}_{k-1})$ [1]. The interesting thing is to estimate the mean and covariance of the transformed random variable $\mathbf{x}_k$.

The UT generates a set of $2L+1$ states $\{\mathcal{X}_{k-1,i}^a\}_{i=0}^{2L}$, which is called *sigma-points* and defined by

$$\{\mathcal{X}_{k-1,i}^a\}_{i=0}^{2L} = \left\{\mathbf{x}_{k-1}^a, \mathbf{x}_{k-1}^a \pm (\sqrt{(L+\lambda)\mathbf{P}_{xx,k-1}^a})_i, i = 1, \cdots, L\right\}, \quad (3)$$

where $(\sqrt{(L+\lambda)\mathbf{P}_{xx,k-1}^a})_i$ denotes the i-th column of the square root matrix $\sqrt{(L+\lambda)\mathbf{P}_{xx,k-1}^a}$. λ is a constant used for scaling adjusting. L is the dimension of augmented state $[\mathbf{x}_k^T, \mathbf{u}_k^T]^T$.

Associated to the sigma-points, a set of weights $\{W_{k-1,i}\}_{i=0}^{2L}$ is allocated

[1] A more general scenario considers the system $\mathbf{y} = \mathbf{f}(\mathbf{x}, \mathbf{u})$, where $\mathbf{x}$ represents system states and $\mathbf{u}$ the perturbations. They are assumed to be independent, and follow Gaussian distributions. However by introducing the augmented state $\mathbf{z} = [\mathbf{x}^T, \mathbf{u}^T]^T$, the general form reduces to ours, $\mathbf{y} = \mathbf{f}(\mathbf{z})$.

by

$$
\begin{aligned}
W_{k-1,0} &= \frac{\lambda}{L+\lambda}\,, \\
W_{k-1,i} &= \frac{1}{2(L+\lambda)},\ i = 1, \cdots, 2L\,,
\end{aligned}
\tag{4}
$$

It can be proved that, the weighted sample mean $\bar{\mathbf{x}}_{k-1}$ and sample covariance $\bar{\mathbf{P}}_{xx,k-1}$ of the finite set $\{\mathcal{X}_i\}_{i=0}^{2L}$ perfectly match the mean $\mathbf{x}_{k-1}^a$ and the covariance $\mathbf{P}_{xx,k-1}^a$ of $\mathbf{x}_{k-1}$ respectively,

$$
\bar{\mathbf{x}}_{k-1} = \sum_{i=0}^{2L} W_{k-1,i} \mathcal{X}_{k-1,i}^a = \mathbf{x}_{k-1}^a\,, \tag{5a}
$$

$$
\bar{\mathbf{P}}_{xx,k-1} = \sum_{i=0}^{2L} W_{k-1,i} (\mathcal{X}_{k-1,i}^a - \mathbf{x}_{k-1}^a)(\mathcal{X}_{k-1,i}^a - \mathbf{x}_{k-1}^a)^T = \mathbf{P}_{xx,k-1}^a\,. \tag{5b}
$$

The above identities are independent of the choice of parameters α, λ and L [23] and this feature will be employed to implement a strategy of PDF re-approximation in construction of UEVF in section 3.

Once the sigma-points are evolved from analysis set $\{\mathcal{X}_{k-1,i}^a\}_{i=0}^{2L}$ to the background $\{\mathcal{X}_{k,i}^b\}_{i=0}^{2L}$ by $\mathcal{M}_{k-1,k}$, the pairs of $\{\mathcal{X}_{k,i}^b, W_{k-1,i}\}_{i=0}^{2L}$ are used to estimate the background mean and covariance via

$$
\mathbf{x}_k^b = \sum_{i=0}^{2L} W_{k-1,i}\, \mathcal{X}_{k,i}^b\,, \tag{6a}
$$

$$
\begin{aligned}
\mathbf{P}_{xx,k}^b = &\sum_{i=0}^{2L} W_{k-1,i} (\mathcal{X}_{k,i}^b - \mathbf{x}_k^b)(\mathcal{X}_{k,i}^b - \mathbf{x}_k^b)^T + \\
&\beta\, (\mathcal{X}_{k,0}^b - \mathbf{x}_k^b)(\mathcal{X}_{k,0}^b - \mathbf{x}_k^b)^T + \mathbf{Q}_k\,,
\end{aligned}
\tag{6b}
$$

where $\mathcal{X}_{k,i}^b = \mathcal{M}_{k-1,k}(\mathcal{X}_{k-1,i}^a)$. The parameter β in Eq.(6b) is used to compensate the high-order errors introduced by the weighted sample approximation. An optimal choice of $\beta = 2$ is suggested when the state follows a Gaussian distribution [25].

Unlike the EnKF [26, 27] there the random ensemble members being used, the UT employs a deterministic sampling scheme [24], as shown, with the performance superior to the EnKF [19] in some situations.

2.2. Statistical moment update methods

We divide the estimation procedure into two steps: the statistical propagation step and update step. This corresponds to the approximation of the two probability laws, η_k^b and η_k^a, discussed previously in Eqs.(1)(2).

For nonlinear problems, different approaches are explored to retrieve the prediction law η_k^b, e.g., EKF, EnKF and UKF as well as their variants [25]. To our knowledge there is no attempt to improve the update law η_k^b out of the Kalman technologies. The linear regression scheme is preserved in above Kalman-like filters to deal with nonlinear systems. However the linear regression criterion cannot be accepted for highly nonlinear and complex systems.

We will review the concept of three update schemes, incorporating the spirit of the UT reviewed in section 2.1. Then we propose a modified version of the UKF with the variational correction in section 3.

2.2.1. LUMV-based moment update

The standard KF is a linear optimal estimator and suits for linear/Gaussian dynamics. The KF-like filters such as the EKF, the UKF and the EnKF as well as their variants, were proposed to deal with nonlinear problems. But they all use the same linear regression formulae for the update step to renew the background statistics.

Let be two matrices of $A \in \mathbb{R}^n$ and $B \in \mathbb{R}^{n\times m}$ and the analysis $\mathbf{x}^a$ of true state $\mathbf{x} \in \mathbb{R}^n$ represented as a linear function of an observation $\mathbf{y} \in \mathbb{R}^m$ with $m \leq n$, i.e.,

$$\mathbf{x}^a = A + B\mathbf{y}\,, \tag{7}$$

$\mathbf{x}^a$ is a *linear* analysis of true state $\mathbf{x}$; it realizes a *linear minimum variance*(LMV) estimate and minimizes the mean square error. This is a *linear unbiased minimum variance*(LUMV) analysis if the LMV estimate $\mathbf{x}^a$ is unbiased.

Now, we assume that both of the state $\mathbf{x}$ and observation $\mathbf{y}$ are random variables with unknown distribution and are conditioned by a random variable $\mathbf{z}$. We denote $\mathbf{x}^b = \mathbb{E}[\mathbf{x}|\mathbf{z}]$, $\mathbf{y}^b = \mathbb{E}[\mathbf{y}|\mathbf{z}]$, the estimate error covariances $\mathbf{P}^b_{xx} = \mathbb{E}[(\mathbf{x}-\mathbf{x}^b)(\mathbf{x}-\mathbf{x}^b)^T|\mathbf{z}]$, $\mathbf{P}^b_{xy} = \mathbb{E}[(\mathbf{x}-\mathbf{x}^b)(\mathbf{y}-\mathbf{y}^b)^T|\mathbf{z}]$ and $\mathbf{P}^b_{yy} = \mathbb{E}[(\mathbf{y}-\mathbf{y}^b)(\mathbf{y}-\mathbf{y}^b)^T|\mathbf{z}]$, here $\mathbf{P}^b_{yy}$ is non-singular. By the LUMV criterion, we can determine the coefficients $B = \mathbf{P}^b_{xy}(\mathbf{P}^b_{yy})^{-1}$, $A = \mathbf{x}^b - B\,\mathbf{y}^b$.

Then reordering the terms we get

$$\begin{aligned}\mathbf{x}_k^a &= \left[\mathbf{x}_k^b - \mathbf{P}_{xy,k}^b(\mathbf{P}_{yy,k}^b)^{-1}\mathbf{y}_k^b\right] + \left[\mathbf{P}_{xy,k}^b(\mathbf{P}_{yy,k}^b)^{-1}\right]\mathbf{y}_k \\ &= \mathbf{x}_k^b + \mathbf{P}_{xy,k}^b(\mathbf{P}_{yy,k}^b)^{-1}\left[\mathbf{y}_k - \mathbf{y}_k^b\right], \qquad (8a)\\ \mathbf{P}_{xx,k}^a &= \mathbf{P}_{xx,k}^b - \mathbf{P}_{xy,k}^b(\mathbf{P}_{yy,k}^b)^{-1}(\mathbf{P}_{xy,k}^b)^T, \qquad (8b)\end{aligned}$$

which is exactly the update formulae in KF.

2.2.2. ET-based linear mean update

There are some suboptimal KF as the ensemble transform KF (ETKF) [28]: it provides new framework for observation assimilation and forecast covariance estimation.

In particular, if we suppose that there are N members in a ensemble assimilation cycle, by using a numerical decomposition algorithm, a $n \times N$ square root matrix $\mathbf{S}_{x,k}^b$ can be obtained from the background error covariance such that $\mathbf{P}_{xx,k}^b = \mathbf{S}_{x,k}^b(\mathbf{S}_{x,k}^b)^T$. Similarly let be $\mathbf{S}_{x,k}^a$ a $n \times N$ square root matrix of analysis error covariance $\mathbf{P}_{xx,k}^a$, then $\mathbf{S}_{x,k}^a$ can be updated from $\mathbf{S}_{x,k}^b$ by multiplied a $N \times N$ matrix $\mathbf{T}_k$ [28], such that

$$\mathbf{S}_{x,k}^a = \mathbf{S}_{x,k}^b\mathbf{T}_k\,, \qquad (9)$$

where the transform matrix $\mathbf{T}_k = \mathbf{V}_k(\mathbf{D}_k + \mathbf{I}^{N\times N})^{-1/2}$ with $\mathbf{V}_k$ the eigenvector matrix of $(\mathbf{H}_k^x\mathbf{S}_{x,k}^b)^T\mathbf{R}_k^{-1}(\mathbf{H}_k^x\mathbf{S}_{x,k}^b)$. $\mathbf{H}_k^x$ is the linearization of $\mathcal{H}_k$ with respect to $\mathbf{x}_k$, i.e., $\mathbf{H}_k^x = \partial\mathcal{H}_k/\partial\,\mathbf{x}_k$. $\mathbf{T}_k$ is linked with a singular value decomposition as

$$(\mathbf{H}_k^x\mathbf{S}_{x,k}^b)^T\mathbf{R}_k^{-1}(\mathbf{H}_k^x\mathbf{S}_{x,k}^b) = \mathbf{V}_k\mathbf{D}_k\mathbf{V}_k^T\,, \qquad (10)$$

where $\mathbf{D}_k$ is a diagonal matrix containing the eigenvalues of $(\mathbf{H}_k^x\mathbf{S}_{x,k}^b)^T\mathbf{R}_k^{-1}(\mathbf{H}_k^x\mathbf{S}_{x,k}^b)$. The analysis error covariance $\mathbf{P}_{xx,k}^a$ is given by

$$\mathbf{P}_{xx,k}^a = \mathbf{S}_{x,k}^a(\mathbf{S}_{x,k}^a)^T = \mathbf{S}_{x,k}^b\mathbf{T}_k\mathbf{T}_k^T(\mathbf{S}_{x,k}^b)^T\,. \qquad (11)$$

Once $\mathbf{x}_k^a$ and $\mathbf{S}_{x,k}^a$ are determined, the analysis ensemble $\{\mathbf{x}_{k,i}^a\}_{i=1}^N$ for the next assimilation cycle can be computed by

$$\mathbf{x}_{k,i}^a = \mathbf{x}_k^a + \sqrt{N-1}(\mathbf{S}_{x,k}^a)_i\,, i = 1, \cdots, N\,, \qquad (12)$$

where $(\mathbf{S}_{x,k}^a)_i$ is the i-th column of the square root matrix $\mathbf{S}_{x,k}^a$.

2.2.3. Variational-based nonlinear mean update

For the mean update involved in a nonlinear assimilation, one attractive way is to apply a Variational Filtering (VF) [29, 30]. One assumes that the background and observation errors are independent and the total PDF can be expressed by $P = P_b P_o = \exp(\ln P_b + \ln P_o)$. The maximum of the total PDF is equivalently the minimization of the cost-function $J = -\ln P_b - \ln P_o$. Then we define $\varepsilon_k^b = \mathbf{x} - \mathbf{x}_k^b$ and the linearized error $\varepsilon_k^o = \mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k^b) - \mathbf{H}_k^x(\mathbf{x} - \mathbf{x}_k^b)$. $\mathbf{B}_k$ and $\mathbf{R}_k$ denote the matrices of background and observation error covariance. The expression of J at time k, $J_k(\mathbf{x}) = J_k^b(\mathbf{x}) + J_k^o(\mathbf{x})$, is

$$J_k^b(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_k^b)^T \mathbf{B}_k^{-1}(\mathbf{x} - \mathbf{x}_k^b)\,, \tag{13a}$$

$$J_k^o(\mathbf{x}) = \frac{1}{2}\left(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k^b) - \mathbf{H}_k^x(\mathbf{x} - \mathbf{x}_k^b)\right)^T \mathbf{R}_k^{-1} \times \left(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k^b) - \mathbf{H}_k^x(\mathbf{x} - \mathbf{x}_k^b)\right)\,, \tag{13b}$$

$$\mathbf{H}_k^x = \left.\frac{\partial \mathcal{H}_k\left(\mathbf{x}_k, \mathbf{v}_k\right)}{\partial\, \mathbf{x}_k}\right|_{\substack{\mathbf{x}_k = \mathbf{x}_k^b \\ \mathbf{v}_k = \mathbf{0}}}\,, \tag{13c}$$

The background error matrix $\mathbf{B}_k$ is a function of $\mathcal{M}_{k-1,k}$. A further discussion about its implementation will be seen in section 4.3.

Finally, we minimize at each time step the $J_k(\mathbf{x})$ from a point $\mathbf{x} \neq \mathbf{x}_k^a$ with a constraint of its gradient $J_k'(\mathbf{x}) = \partial J_k(\mathbf{x})/\partial\, \mathbf{x}$ being nonnegative, such that $\mathbf{x}_k^a = \arg\min_{\mathbf{x}} J_k(\mathbf{x})$ with $J_k'(\mathbf{x}) \geq 0$, where $J_k'(\mathbf{x})$ is derived as

$$J_k'(\mathbf{x}) = \mathbf{B}_k^{-1}(\mathbf{x} - \mathbf{x}_k^b) - (\mathbf{H}_k^x)^T \mathbf{R}_k^{-1}\left(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k^b) - \mathbf{H}_k^x(\mathbf{x} - \mathbf{x}_k^b)\right). \tag{14}$$

To minimize the cost-function as a quadratic form, there are a number of efficient algorithms [38, 39]. A further discussion about the cost-function minimization can be seen in section 4.2.

3. Unscented/Ensemble transform Variational Filter (UEVF)

We propose the UEVF to meet the following requirements:

- Accuracies of the mean estimation and sample size: The EnKF estimation sometimes introduces spurious modes even if the ensemble mean and covariance are correct [19]. On the contrary for UT mean estimation, the sigma-points are chosen to match the true mean and covariance. The sample size is fixed, i.e., twice the degrees-of-freedom plus

one. Actually the sample size determined with trial and errors in the EnKF is not usable especially for high dimension problems. The adjustable parameters in the UT provide a capability to counterbalance the mismatch in system model or the observation perturbations.

- Accuracies of the mean update: The KF-like estimations, such as the UKF [12], perform a correction with respect to the LUMV criterion. The variational mean update reviewed in section 2.2.3 implements a minimization with nonlinear quadratic terms. Theoretically in terms of update accuracies, the later is better than the former. This motivates us to replace the KF-like update by a variational minimization update. Nevertheless the background error covariance $\mathbf{P}^b_{xx,k}$ has also to be updated, to deal with the online covariance renew, we adopt the ensemble transform like designed in the ETKF in Eq.(11).

Now we will present our new method. It incorporates the scheme of unscented mean estimation with the variational minimization update and an ensemble covariance computation. On the other hand, to implement our method for systems with large degrees-of-freedom, we adopt the technique of rank reduction as presented in geophysical literature [19, 31, 32].

We express $\mathbf{P}^b_{xx,k}$ with a square root matrix $\mathbf{S}^b_{x,k}$ and a residue $\mathbf{Q}_k$, and have $\mathbf{P}^b_{xx,k} = \mathbf{S}^b_{x,k}(\mathbf{S}^b_{x,k})^T + \mathbf{Q}_k$.

For the initial step, the state $\mathbf{x}^a_0$ and its error covariance $\mathbf{P}^a_{xx,0}$ are given. We get $\mathbf{S}^a_{x,0}$ from the square root decomposition of $\mathbf{P}^a_{xx,0} = \mathbf{S}^a_{x,0}(\mathbf{S}^a_{x,0})^T$.

We assume that the sigma-points have been computed during the initialization step. Details will be seen at the end of the time loop.

Then the iterative filtering procedure begins, and for the kth time step we have:

- Mean estimation step :
  Propagate forward the analyzed sigma-points with the model $\mathcal{M}_{k-1,k}$ and generate the background sigma-points, $\{\mathcal{X}^b_{k,i} = \mathcal{M}_{k-1,k}(\mathcal{X}^a_{k-1,i})\}_{i=0}^{2\ell_{k-1}}$. The associated weights $\{W_{k-1,i}\}_{i=0}^{2\ell_{k-1}}$ are inherited from the previous

cycle $k-1$. Then we get

$$\mathbf{x}_k^b = \sum_{i=0}^{2\ell_{k-1}} W_{k-1,i}\, \mathcal{X}_{k,i}^b\,, \tag{15a}$$

$$\mathbf{S}_{x,k}^b = \Big[\sqrt{W_{k-1,0}^{\beta}}(\mathcal{X}_{k,0}^b - \mathbf{x}_k^b), \sqrt{W_{k-1,1}}(\mathcal{X}_{k,1}^b - \mathbf{x}_k^b), \cdots, \sqrt{W_{k-1,2\ell_{k-1}}}(\mathcal{X}_{k,2\ell_{k-1}}^b - \mathbf{x}_k^b)\Big]\,, \tag{15b}$$

$$\mathbf{P}_{xx,k}^b = \mathbf{S}_{x,k}^b(\mathbf{S}_{x,k}^b)^T + \mathbf{Q}_k\,, \tag{15c}$$

where $W_{k-1,0}^{\beta} = W_{k-1,0} + \beta$.

- Mean update step :

$$\mathbf{x}_k^a = \arg\min_{\mathbf{x}} J_k(\mathbf{x})\,, \text{subject to } J_k'(\mathbf{x}) \geq \mathbf{0}\,, \tag{16a}$$

$$J_k(\mathbf{x}) = \frac{1}{2}(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k^b, \mathbf{0}) - \mathbf{H}_k^x(\mathbf{x} - \mathbf{x}_k^b))^T\mathbf{R}_k^{-1}\times (\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k^b, \mathbf{0}) - \mathbf{H}_k^x(\mathbf{x} - \mathbf{x}_k^b)) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k^b)^T(\mathbf{P}_{xx,k}^b)^{-1}(\mathbf{x} - \mathbf{x}_k^b)\,, \tag{16b}$$

$$J_k'(\mathbf{x}) = (\mathbf{P}_{xx,k}^b)^{-1}(\mathbf{x} - \mathbf{x}_k^b) - (\mathbf{H}_k^x)^T\mathbf{R}_k^{-1}\times (\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k^b, \mathbf{0}) - \mathbf{H}_k^x(\mathbf{x} - \mathbf{x}_k^b))\,, \tag{16c}$$

$$\widehat{\mathbf{V}}_k\widehat{\mathbf{D}}_k\widehat{\mathbf{V}}_k^T \approx \mathbf{V}_k\mathbf{D}_k\mathbf{V}_k^T = (\mathbf{H}_k^x\mathbf{S}_{x,k}^b)^T\mathbf{R}_k^{-1}(\mathbf{H}_k^x\mathbf{S}_{x,k}^b)\,, \tag{16d}$$

$$\mathbf{T}_k = \widehat{\mathbf{V}}_k(\widehat{\mathbf{D}}_k + \mathbf{I}^{\ell_k\times\ell_k})^{-1/2}\,, \tag{16e}$$

$$\mathbf{S}_{x,k}^a = \mathbf{S}_{x,k}^b\mathbf{T}_k\,, \tag{16f}$$

where a singular value decomposition is applied to the covariance matrix $(\mathbf{H}_k^x\mathbf{S}_{x,k}^b)^T\mathbf{R}_k^{-1}(\mathbf{H}_k^x\mathbf{S}_{x,k}^b)$ (Eq.(16d)). In $\mathbf{D}_k$ and $\mathbf{V}_k$, the eigenvalues $\{\sigma_{k,i}^2\}_{i=1}^L$ and the eigenvectors $\{\mathbf{e}_{k,i}\}_{i=1}^L$ are sorted in descending order. $\widehat{\mathbf{D}}_k = \mathrm{diag}(\sigma_{k,1}^2, \cdots, \sigma_{k,\ell_k}^2)$ is a $\ell_k \times \ell_k$ diagonal matrix formed with the first ℓ_k-th bigger eigenvalues and is the rank-diminished version of $\mathbf{D}_k$. Associated with, $\widehat{\mathbf{V}}_k = [\mathbf{e}_{k,1}, \cdots, \mathbf{e}_{k,\ell_k}]$ is a $L \times \ell_k$ eigenvectors matrix. The approximation by truncation will be explained later in section 4.1.

- sigma-point generation step :
  Generate the sigma-points $\{\mathcal{X}_{k,i}^a\}_{i=0}^{2\ell_k} = \{\mathbf{x}_k^a, \mathbf{x}_k^a \pm (\sqrt{\ell_k + \lambda}\ \mathbf{S}_{x,k}^a)_i, i =$

$1, \cdots, \ell_k\}$, where the symbol $(\cdot)_i$ denotes the i-th column of the square root matrix $\sqrt{\ell_k + \lambda}\ \mathbf{S}^a_{x,k}$. To the sigma-points, the associated weights given by $\{W_{k,0} = \lambda/(\ell_k + \lambda), W_{k,i} = 0.5/(\ell_k + \lambda), i = 1, \cdots, 2\ell_k\}$. The sampling size ℓ_k is determined to satisfy $\ell_k < L$, where L is the dimension of augmented state $[(\mathbf{x}^a_k)^T, \mathbf{u}^T_k]^T$.

4. Implementation of the UEVF

4.1. Construction of the sigma-points based on a truncated Covariance

For high dimensional problems, the sampling size of the ensemble methods is a difficulty. To tackle the problem, similarly to the works suggested by literatures [31, 32, 33, 34, 19], we use the technique of truncated singular value decomposition (TSVD) to generate perturbations of the ensemble forecast. A rank-reduction of the covariance matrix is adopted to approximate the original one using the TSVD. This approximation is a compromise between the computational cost and the accuracy [19].

A singular value decomposition is applied to $\mathbf{P}^a_{xx,k} = \mathbf{V}_k \mathbf{D}_k (\mathbf{V}_k)^T$, where $\mathbf{D}_k = \mathrm{diag}(\sigma^2_{k,1}, \cdots, \sigma^2_{k,L})$ is the eigenvalues $\sigma^2_{k,i}$'s of $\mathbf{P}^a_{xx,k}$ sorted in descending order. L is the dimension of $[\mathbf{x}^T_k, \mathbf{u}^T_k]^T$ and $\mathbf{V}_k = [\mathbf{e}_{k,1}, \cdots, \mathbf{e}_{k,L}]$ is the matrix of eigenvectors. Then we get

$$\mathbf{P}^a_{xx,k} = \sum_{i=1}^{L} \sigma^2_{k,i} \mathbf{e}_{k,i} \mathbf{e}^T_{k,i} \approx \sum_{i=1}^{\ell_k} \sigma^2_{k,i} \mathbf{e}_{k,i} \mathbf{e}^T_{k,i} = \widehat{\mathbf{V}}_k \widehat{\mathbf{D}}_k (\widehat{\mathbf{V}}_k)^T\,, \quad \ell_k < L\,, \tag{17}$$

where ℓ_k is called the truncation size of the sigma-points. Therefore the $\widehat{\mathbf{V}}_k$ and $\widehat{\mathbf{D}}_k$ are with a reduced dimension, $L \times \ell_k$ and $\ell_k \times \ell_k$, respectively.

With too small values of ℓ_k, we lost some important structures of $\mathbf{P}^a_{xx,k}$ and too big values lead to prohibited computational costs. ℓ_k is an integer and can be determined by an efficient scheme [19] below:

$$\begin{aligned} &\sigma^2_{k,i} > \mathrm{trace}(\mathbf{P}^a_{xx,k})/\gamma_k\,, \quad i = 1, \cdots, \ell_k\,, \\ &\sigma^2_{k,i} \le \mathrm{trace}(\mathbf{P}^a_{xx,k})/\gamma_k\,, \quad i > \ell_k + 1\,, \end{aligned} \tag{18}$$

where γ_k is an adjustable threshold with a lower bound ℓ_l and an upper bound ℓ_u specified in order to prevent ℓ_k to be too large or too small. We adjust the threshold γ_k to keep $\ell_l \le \ell_k \le \ell_u$.

To determine γ_k the scheme is also given by [19]: at the initial step we specify a threshold γ_1, if γ_1 is a proper value such that ℓ_1 satisfies $\ell_l \le \ell_1 \le \ell_u$,

then in next iteration, we put $\gamma_2 = \gamma_1$. If γ_1 is too small, $\ell_1 < \ell_l$, then γ_1 increases. If γ_1 is too large, $\ell_1 > \ell_u$, then replace γ_1. This procedure will be continued until ℓ_1 falls into the specified range. After this adjustment, let $\ell_2 = \ell_1$ at the next iteration. Then adjust it to let ℓ_2 fall into the specified range. By this way, the truncation size ℓ_k varies in time.

When the truncation size ℓ_k is determined, the $L \times L$ square root matrix $\mathbf{S}^a_{x,k}$ is replaced by the $L \times \ell_k$ matrix $\{(\widehat{\mathbf{S}^a_{x,k}})_i \triangleq \sigma_{k,i}\mathbf{e}_{k,i}, i = 1, \cdots, \ell_k\}$. Following the Eq.(3), the (deterministic) sigma-point truncated set is given by

$$\{\mathcal{X}^a_{k-1,i}\}_{i=0}^{2\ell_k} = \left\{\mathbf{x}^a_{k-1}, \mathbf{x}^a_{k-1} \pm (\ell_k + \lambda)^{1/2}(\widehat{\mathbf{S}^a_{x,k}})_i \,, i = 1, \cdots, \ell_k\right\} . \tag{19}$$

Its associated weights set is :

$$\begin{aligned} W_{k-1,0} &= \frac{\lambda}{\ell_k + \lambda} \,, \\ W_{k-1,i} &= \frac{1}{2(\ell_k + \lambda)} \,, \quad i = 1, \cdots, 2\ell_k \,. \end{aligned} \tag{20}$$

Consequently the truncated version of the sigma-points, given by Eq.(19) and (20), balances the requirements in terms of computational cost and accuracy.

4.2. Minimization of the cost-function

The minimization of cost function in Eq.(16) plays a key role in the UEVF implementation. For meteorological models, the number n of state variables easily exceeds 10^8. If the degrees of freedom $N < n$, the calculation of the full background matrices $(\mathbf{x} - \mathbf{x}^b_k)^T(\mathbf{P}^b_{xx,k})^{-1}(\mathbf{x} - \mathbf{x}^b_k)$ (Eq.(16b)) has order of $\mathcal{O}(N^2)$ complexity. For a typical NWF, N^2 is generally about 10^{16} [37]. Therefore a direct solution is not feasible for operational applications.

Different minimization algorithms are available [35], such as the steepest descent method, Newton and quasi-Newton methods, etc. However conjugate-gradient algorithms outperforms them in storage requirement and convergent rate [36, 37, 38, 39]. The method becomes often the only implementable choice for large-scale nonlinear minimization considering the computational efficiency and the accuracy as main criteria [37] and is used in the Variational Filters (VF) involved in operational data assimilation systems.

In order to apply the conjugate-gradient method efficiently, we rewrite the background term via the relation $\varepsilon^b_k = \mathbf{U}_k\mathbf{z}_k$, where $\varepsilon^b_k = \mathbf{x} - \mathbf{x}^b_k$ is the analysis

increment. The transform matrix $\mathbf{U}_k$ is well designed to dimensionalize the variational problem. The condition number of $\mathbf{U}_k$ is small and the product $\mathbf{U}_k\mathbf{U}_k^T$ match the full background error covariance $\mathbf{P}_{xx,k}^b$, i.e., $\mathbf{U}_k\mathbf{U}_k^T \approx \mathbf{P}_{xx,k}^b$. In terms of analysis increments, the Eq.(16b) and (16c) can be rewritten

$$J_k(\mathbf{z}_k) = \frac{1}{2}\mathbf{z}_k\mathbf{z}_k^T + \frac{1}{2}(\mathbf{Y}_k - \mathbf{H}_k^x\mathbf{U}_k\mathbf{z}_k)^T\mathbf{R}_k^{-1}(\mathbf{Y}_k - \mathbf{H}_k^x\mathbf{U}_k\mathbf{z}_k)\,, \tag{21a}$$

$$J_k'(\mathbf{z}_k) = \mathbf{z}_k - \mathbf{U}_k^T(\mathbf{H}_k^x)^T\mathbf{R}_k^{-1}(\mathbf{Y}_k - \mathbf{H}_k^x\mathbf{U}_k\mathbf{z}_k)\,, \tag{21b}$$

where $\mathbf{Y}_k \triangleq \mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k^b, \mathbf{0})$ is innovation vector and $\mathbf{H}_k^x$ the Jacobian of the $\mathcal{H}_k$ defined in Eq.(13c). The second derivative of cost function, $J_k''(\mathbf{z}_k) \triangleq \partial J_k'(\mathbf{z}_k)/\partial \mathbf{z}_k$, can also be obtained by

$$J_k''(\mathbf{z}_k) = \mathbf{I} + \mathbf{U}_k^T(\mathbf{H}_k^x)^T\mathbf{R}_k^{-1}\mathbf{H}_k^x\mathbf{U}_k\,, \tag{22}$$

where $\mathbf{I}$ denotes a unit matrix with an appropriate dimension. For the conjugate-gradient method we denote by $\mathcal{G}_k$ the residual

$$\mathcal{G}_k = J_k'(\mathbf{z}_k)\,. \tag{23}$$

We use the convention $\beta_{-1} = \mathbf{d}_{-1} = 0$, and we give initialization values to $\mathbf{U}_0$, $\mathbf{z}_0$ and $\mathbf{H}_0^x$. The conjugate-gradient minimization [40, 37] for the time step k is the procedure:

- Compute the gradient $J_k'(\mathbf{z}_k)$ and the second derivative $J_k''(\mathbf{z}_k)$:

$$\mathcal{G}_k^I \triangleq J_k'(\mathbf{z}_k) = \mathbf{z}_k - \mathbf{U}_k^T(\mathbf{H}_k^x)^T\mathbf{R}_k^{-1}(\mathbf{Y}_k - \mathbf{H}_k^x\mathbf{U}_k\mathbf{z}_k)\,, \tag{24a}$$

$$\mathcal{P}_k \triangleq J_k''(\mathbf{z}_k) = \mathbf{I} + \mathbf{U}_k^T(\mathbf{H}_k^x)^T\mathbf{R}_k^{-1}\mathbf{H}_k^x\mathbf{U}_k\,. \tag{24b}$$

- Compute the descent direction $\mathbf{d}_k$, the step size α_k and update the $\mathbf{z}_k$:

$$\mathbf{d}_k = -\mathcal{G}_k^I + \beta_{k-1}\mathbf{d}_{k-1}\,, \tag{25a}$$

$$\alpha_k = \frac{(\mathcal{G}_k^I)^T\mathcal{G}_k^I}{\mathbf{d}_k^T\mathcal{P}_k\mathbf{d}_k} = \frac{\|\mathcal{G}_k\|^2}{\mathbf{d}_k^T\mathcal{P}_k\mathbf{d}_k}\,, \tag{25b}$$

$$\mathbf{z}_k = \mathbf{z}_k + \alpha_k\mathbf{d}_k\,. \tag{25c}$$

  where $\|\cdot\|$ denotes Euclidean norm.

- Compute a new gradient $J'_k(\mathbf{z}_k)$ and update β_k:

$$\mathcal{G}_k^{II} \triangleq J'_k(\mathbf{z}_k) = \mathbf{z}_k - \mathbf{U}_k^T(\mathbf{H}_k^x)^T\mathbf{R}_k^{-1}(\mathbf{Y}_k - \mathbf{H}_k^x\mathbf{U}_k\mathbf{z}_k)\,, \tag{26a}$$

$$\beta_k = \frac{\|\mathcal{G}_k^{II}\|^2}{\|\mathcal{G}_k^{I}\|^2}\,. \tag{26b}$$

- A convergence criterion for stopping the iterations is tested: even $\|\mathcal{G}_k^{II}\|$ is less than a threshold ξ_{eps}, even the iteration number is larger than a specified number. Finally we get the analyzed state $\mathbf{x}_k^a = \mathbf{U}_k\mathbf{z}_k + \mathbf{x}_k^b$.

The conjugate-gradient method may be seen as the solution of $J'_k(\mathbf{z}_k) = 0$ in the m-th iteration [42], where m is the number of distinct eigenvalues of $\mathbf{I} - \mathbf{U}_k^T(\mathbf{H}_k^x)^T\mathbf{R}_k^{-1}\mathbf{H}_k^x\mathbf{U}_k$. Moreover when these eigenvalues are clustered into groups of approximately equal number, the method will converge even faster.

4.3. Computation of the background error matrix $\mathbf{B}_k$

From the Eq.(13), the quadratic cost function is completely characterized by the background error matrix $\mathbf{B}_k$ and observation error covariance $\mathbf{R}_k$. If $\mathbf{R}_k$ is nearly fixed and mostly a constant, the matrix $\mathbf{B}_k$ plays a key role. Different strategies for the determination of the guess covariance error matrix $\mathbf{B}_k$ were explored (see [26, 27, 43, 44] for details):

- In the VF framework, $\mathbf{B}_k$ is seen as a function of nonlinear evolution $\mathcal{M}_{k-1,k}$ and supposed to be known and time-variant. It can be computed with the empirical statistics of a one-step-ahead prediction of an ensemble. In the UEVF we simply choose to put $\mathbf{B}_k \approx \mathbf{P}_{xx,k}^b$ in Eq.(16b).

- In 3DVar (Pointwise Variational Filter) or 4DVar (Trajectorial Variational Filter), $\mathbf{B}_k$ is assumed to be constant in time and learned statistically offline.

- In other situations, $\mathbf{B}_k$ is a time-variant statistic and can empirically computed as a conditional background covariance online. Particularly, if $\mathbf{B}_k$ comes from a random ensemble members $\{\mathcal{X}_{k,i}^b\}$ with size N, one may show that $\mathbf{B}_k^N \approx \overline{\mathbf{P}_{xx,k}^b} = \frac{1}{N-1}\sum_{k=1}^N(\mathcal{X}_{k,i}^b - \overline{\mathbf{x}_k^b})(\mathcal{X}_{k,i}^b - \overline{\mathbf{x}_k^b})^T$, where $\overline{\mathbf{x}_k^b} = \sum_{i=1}^N \mathcal{X}_{k,i}^b/N$. We call EnVar a Variational Filter using the empirical $\mathbf{B}_k^N$ matrix. This scheme is simple and direct, we will use it for the numerical comparisons.

In next section, we propose two numerical simulations that based on the Lorenz-95 model and the 2D shallow water equation. We want to test the ability of our filter to estimate the reference signal.

5. Simulation and discussion

This section examine the performance of the UEVF using numerical simulations. Two other filters are used to compare the performances: the EnUKF and the EnVar. The EnUKF uses deterministic members to perform an UT estimation, see [19, 45] for details. The EnVar have an estimation with random members to reduce the collapse effects. We add to the EnVar a spherical simplex centering scheme following [19, 46] for a convenient comparison.

To estimate the filters errors, we choose the dimension-averaged relative root mean square (RMS) error given by

$$E_k = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{M}\sum_{j=1}^{M}\left(\frac{\mathbf{x}_{i,k}^{a,j} - \mathbf{x}_{i,k}^{\text{true},j}}{\mathbf{x}_{i,k}^{\text{true},j}}\right)^2\right]^{1/2}, \quad k = 1, \cdots, T, \tag{27}$$

where n is the dimension of state. A time $k < T$, $\mathbf{x}_{i,k}^{\text{true},j}$ and $\mathbf{x}_{i,k}^{a,j}$ are the i-th component of the truth and the analyzed state for the j-th Monte Carlo simulation. M is the total number of Monte Carlo runs.

The positive semi-definiteness of the root square matrices in Eq.(15) have to be guaranteed. The problem comes from the varying sampling size ℓ_k in Eq. (17). Indeed the weight $W_{k-1,0}$ in Eq.(20) may be negative if $\lambda < 0$ or $\ell_k + \lambda < 0$. Consequently the effective weight $W_{k-1,0}^{\beta}$ in Eq.(15b) is defined by $W_{k-1,0} + \beta$ where $\beta \geq 0$ is a constant. Then the parameters λ and β verify $W_{k-1,0} + \beta \geq 0$ and $\ell_k + \lambda > 0$. It means that $\lambda \geq -\beta\ell_k/(1+\beta)$. Since ℓ_k is bounded within the interval $[\ell_l, \ell_u]$, we choose $\lambda \geq -\beta\ell_l/(1+\beta)$.

Three additional techniques are implemented both to the EnUKF and the UEVF in order to improve the performance of these filters. Nevertheless they are not necessary for good working order of the algorithms. We put them to be consistent with the experiments of Luo and Munoz using the EnUKF [19]:

To apply the spherical simplex centering scheme to the ensemble transform in the UEVF and the EnUKF, we follow the algorithm in [25] to build a centering matrix $\mathbf{U}$, where $\mathbf{U}$ is given by Eq. (C15) in [47]. It is time-invariant and does not involve the dynamical model and observation operator.

The second technique is the covariance inflation. Many investigations point out the systematic underestimation of the covariance of analysis errors in the EnKF. The inflation of these covariance matrices is the solution suggested by some authors [45, 48]. We choose to multiply the perturbations to the mean $\mathbf{x}_k^a$ of the analysis by a coefficient $1+\delta$. It is equivalent to increase the covariance matrix by a constant $(1+\delta)^2$. We set the covariance inflation factor $\delta = 0.5$ in the experiments.

At least we use a covariance filter method [49, 50]. It is based on the Schur-product in order to reduce the effect of sample error of the covariance matrix. Following [19], the scale length ℓ_c of the covariance filter is seen as an optimum within a certain range. It minimizes the relative RMS error and we follow the suggestion of the authors to set $\ell_c = 200$.

5.1. The Lorenz-95 model

First to analyze our results, we propose to use the Lorenz-95 model. This is a chaotic dynamical system introduced by E.N. Lorenz ([51] or [52]). It describes a simplified propagation of an atmospheric wave along a meridian circle. The circle is divided into n intervals at time k and the simplified model is given by

$$\frac{\partial x_{i,k}}{\partial t_k} = (x_{i+1,k} - x_{i-2,k})x_{i-1,k} - x_{i,k} + F\,, \tag{28}$$

where $i = 1, \cdots, n$ is the dimension index. The cyclic boundary conditions, $x_{0,k} = x_{n,k}$ and $x_{-1,k} = x_{n-1,k}$ as well as $x_{n+1,k} = x_{1,k}$, are adopted to determine the state components x_i. The constant F is set to $F = 8$ (for $F > 4.4$ the system is chaotic with positive Lyapunov exponents). The solutions of the system are obtained by a numerical integration with a fourth-order Runge-Kutta method. The time-step $\Delta t = 0.05$ unit corresponds to a 6-hours physical time mesh [52]. Moreover we add a dynamical Gaussian perturbation $\mathbf{u}_k$ with a zero-mean and a covariance $\mathbf{Q}_k$.

The observer $\mathcal{H}_k$ in Eq.(2) is simply chosen as a time-invariant identity matrix with an additional Gaussian noise. Therefore $\mathbf{y}_k = \mathbf{x}_k + \mathbf{v}_k$ where $\mathbf{x}_k = [x_{1,k}, \cdots, x_{n,k}]^T$ is the state vector and $\mathbf{v}_k$ denotes a n-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{R}_k)$.

The UEVF is nearly the same as the EnUKF except for the correction step. In particular for the UEVF, the mean update formulae in Eq.(16) can

be reduced to

$$\mathbf{x}_k^a = \arg\min_{\mathbf{x}} J_k(\mathbf{x}), \quad \text{with } J_k'(\mathbf{x}) \geq \mathbf{0}, \tag{29a}$$

$$J_k(\mathbf{x}) = \frac{1}{2}(\mathbf{y}_k - \mathbf{x})^T \mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{x}) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k^b)^T (\mathbf{P}_{xx,k}^b)^{-1}(\mathbf{x} - \mathbf{x}_k^b), \tag{29b}$$

$$J_k'(\mathbf{x}) = -\mathbf{R}_k^{-1}(\mathbf{y}_k - \mathbf{x}) + (\mathbf{P}_{xx,k}^b)^{-1}(\mathbf{x} - \mathbf{x}_k^b). \tag{29c}$$

Therefore the cost function $J_k(\mathbf{x})$ is a linear combination of the two quadratics terms $(\mathbf{y}_k - \mathbf{x})^T(\mathbf{y}_k - \mathbf{x})$ and $(\mathbf{x} - \mathbf{x}_k^b)^T(\mathbf{x} - \mathbf{x}_k^b)$. At the opposite the EnUKF update mean given in Eq.(8a) is directly a linear combination between observation $\mathbf{y}_k$ and background $\mathbf{x}_k^b$:

$$\mathbf{x}_k^a = \left(\mathbf{P}_{xy,k}^b (\mathbf{P}_{yy,k}^b)^{-1}\right) \mathbf{y}_k + \left(\mathbf{I} - \mathbf{P}_{xy,k}^b (\mathbf{P}_{yy,k}^b)^{-1}\right) \mathbf{x}_k^b, \tag{30}$$

where $\mathbf{I}$ denotes an unit matrix. The performance improvement achieved by the UEVF compared to the EnUKF comes from the difference of the previous update formulae in Eq.(29-30).

In this experiment, we set other parameters: $\lambda = -2, \beta = 2$, the initial threshold $\gamma_1 = 1000$. For a state dimension n, we set the initial ensemble size $\ell_1 < n/2$.

To explore the effect of different levels of dynamic and observation noises, as well as the dimension n on the filtering performance, we consider five cases as follows:

Case 1: Low level of noise

In this case, we set the state dimension $n = 40$, the initial sample size $\ell_1 = 15$, the lower bound $\ell_l = 10$ and the upper bound $\ell_u = 20$, $\mathbf{Q}_k = 5. \times \mathrm{Id}(40, 40)$, $\mathbf{R}_k = 0.1 \times \mathrm{Id}(40, 40)$. We use 50 Monte Carlo simulations. The relative RMS error, and the evolution in time of the sampling size ℓ_k correspond to the figures 1(a) and 1(b) respectively. Averaged computation time : EnUKF is 8.4448s, UEVF is 8.6661s, and EnVar is 2.562s.

It seems (Fig.1(a)) that the UEVF has the best accuracy and the EnVar behaves as an approximation of the UEVF. However the EnVar is unstable and has many biases with surprising amplitudes. The EnUKF has a constant gap with the UEVF. Fig.1(b) shows that, with the same initial sample size ℓ_1, the evolution of ℓ_k is different: EnUKF quickly increases to the upper bound ℓ_u and keeps the level to the end, while UEVF tends to decline with a constant adjustment. A more interesting point is about the ensemble size.

A larger value does not necessarily guarantee a smaller RMS error. This is relevant with the well-known behaviour of the ensemble. They reach a saturation score as the number of elements increase [53]. For the computation time, the UEVF is almost equal to EnUKF and both are more than three time of EnVar.

Case 2: High level of noise

Now we use the same dimension $n = 40$ and initial sample size $\ell_1 = 15$, but we increase the noise levels to $\mathbf{Q}_k = 10. \times \mathrm{Id}(40, 40)$ and $\mathbf{R}_k = 0.5 \times \mathrm{Id}(40, 40)$. The relative RMS error and evolution of ℓ_k are computed with 50 Monte Carlo runs (Figs. 2(a) and 2(b)). Averaged computation time : EnUKF is 8.9178s, UEVF is 9.2549s, and EnVar is 2.7443s.

The relative RMS UEVF errors is still oscillating with a small amplitude (Fig.2(a)). EnUKF series appears more flat but keeps a large gap with the UEVF. The RMS error of the EnVar is most of the time far bigger than the UEVF and the EnUKF. The sample size of the UEVF tends to increase quickly (Fig.2(b)) and reaches the upper bound. The computation times of the three methods are slightly increased compared with case 1.

Case 3: High state dimension and low noise level

We increase the dimension to $n = 200$, and we specify a lower level of noise $\mathbf{Q}_k = 5. \times \mathrm{Id}(200, 200)$, $\mathbf{R}_k = 0.1 \times \mathrm{Id}(200, 200)$ and initial sample size $\ell_1 = 15$. We perform 50 Monte Carlo simulations. The relative RMS error and the evolution of ensemble size ℓ_k, are plotted in Figs. 3(a) and 3(b). Averaged time elapsing: EnUKF is 144.1208s, UEVF is 150.3785s, and EnVar is 14.4913s.

As the dimension have been increased from 40 to 200 with a low noise level, the EnUKF becomes sensible to the dimension effects (Fig.3(a)) and its RMS error tends to a larger amplitude and maintains high values. Both the UEVF and the EnVar do not seem to be impacted by the increased dimension and still keep lower error amplitude. The dimension size affects the computational time: for both the UEVF and the EnUKF, the average time is more than 17 times that the time used in case 1. There is no surprise to this increase of time: the matrix algebra and optimization are time-consuming mathematical operations in high dimension. This can be also observed for the EnVar but only with a factor about 6.

Case 4: High state dimension and noise level

Then we do not change the initial ensemble size $\ell_1 = 15$, but we use a dimension $n = 200$ and higher noises levels $\mathbf{Q}_k = 10. \times \mathrm{Id}(200, 200)$ and $\mathbf{R}_k = 0.5 \times \mathrm{Id}(200, 200)$. The results with 50 Monte Carlo runs are given in Figs. 4(a) and 4(b) respectively. Averaged computation time: EnUKF is 152.8557s, UEVF is 156.8607s, and EnVar is 15.4724s.

The RMS errors produced by both the EnUKF and the EnVar have large gap compared with UEVF (Fig.4(a)). The UEVF is the only filter not affected by the coupled of strong perturbations and high dimensions. Talking about the sample size, the two methods behave with the fashion (Fig.4(b)). To compare the time consumption, the average time is roughly multiply by 1.05 comparing with the case 3. It should be remarked that the EnVar occasionally reports numerical problems (ill-conditioned minimizations) with a divergent filtering process. When more Monte-Carlo runs are performed, the relative RMS error curve of EnUKF tends to diverge.

Case 5: Higher state dimension

In this last case there are no change in the initial ensemble size and noise levels, but we use a 4-fold state dimension $n = 800$. Thus we have again $\mathbf{Q}_k = 10. \times \mathrm{Id}(800, 800)$ and $\mathbf{R}_k = 0.5 \times \mathrm{Id}(800, 800)$. The lower and upper bound of sample size are changed to $[\ell_l, \ell_u] = [10, 50]$. The results with only one run are given in Figs.5(a) and 5(b). Averaged computational time: EnUKF is 4285.2601s, UEVF is 4361.1175s, and EnVar is 580.8696s.

The error curve of the UEVF (Fig.5(a)) shows a remarkable stability and maintains lower errors compared with the other filters. The EnVar error curve has again large oscillations and stays largely above the UEVF errors. For the ensemble size, both the UEVF and the EnUKF quickly reach the top bound and stay in the level to the end. The other effect of high dimensions is on the computational times. The computation time increase due to the sigma-points methods is roughly 17 times the cost in the case 4 and 36 times the computation cost of the EnVar. During the minimization procedure in the EnVar, a number of error reports appears and indicates the appearance of ill-conditioned minimization problem and lost of computational accuracy.

To finish with the numerical tests, we compare the averaged trajectories produced by the three methods (Fig.6) and the reality according to

parameters used in case 5. The averaged trajectories are calculated by $\overline{\mathbf{x}_k^a} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{k,i}^a$, where $\mathbf{x}_{k,i}^a$ denotes the i-th component of analysis $\mathbf{x}_k^a$. We present figure 6 one of the component of the state vector. The behaviour of the 3 filters is then clear with an EnVar divergence from time to time.

In term of the relative RMS error and the evolution of the ensemble size or the computation time via 5 cases, we have showed that the UEVF exhibits a better accuracy in state estimate than the EnUKF and the EnVar. However the UEVF's time cost is far larger than the EnVar's and roughly equals to the EnUKF. The similar conclusion held when facing the super-large-dimension situation.

5.2. 2D-Shallow Water simulation

The shallow water equations (SWE) are a set of hyperbolic PDE's. It used to model the perturbations propagation of the water height (or other incompressible fluids). The equations are derived from depth-integrating Navier-Stokes equations considering the horizontal length scale much greater than the vertical length scale. In the case of no frictional forces, the SWE can be written as :

$$\frac{\partial Q}{\partial t} + \frac{\partial F(Q)}{\partial x} + \frac{\partial H(Q)}{\partial y} = 0\,, \tag{31a}$$

$$Q = [\mathbf{h}, \mathbf{ha}, \mathbf{hb}]^T\,, \tag{31b}$$

$$F(Q) = [\mathbf{ha}, \mathbf{ha^2} + g\mathbf{b}^2/2, \mathbf{hab}]^T\,, \tag{31c}$$

$$H(Q) = [\mathbf{hb}, \mathbf{hab}, \mathbf{hb}^2 + g\mathbf{h}^2/2]^T\,, \tag{31d}$$

where the time t and the two space coordinates x and y are independent. The dependent variables are the fluid height $\mathbf{h}$ and the 2D fluid velocities $\mathbf{a}$ and $\mathbf{b}$. The gravitational constant set to $g = 9.8$. We add a dynamical noise $\mathbf{u}$ assumed to be a 2D centered Gaussian process with covariance $\mathbf{Q}$.

Similarly to the tests with the Lorenz-95 model, we choose a time invariant observer $\mathcal{H}_k$ with $\mathbf{y}_k = \mathcal{H}_k(Q_k, \mathbf{v}_k) = \mathbf{h}_k + \mathbf{v}_k$ where the $\mathbf{h}_k$ is a function of space coordinates (x, y), $\mathbf{v}_k$ denotes a 2D Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{R}_k)$.

For this experiment, we confine our simulation domain to a square region with the size of 40×40 unit. We choose a Dirichlet boundary conditions with a reflection. By hypotheses $\mathbf{a} = \mathbf{0}$ on the vertical sides of the square and $\mathbf{b} = \mathbf{0}$ on the horizontal sides.

We adopt a second order Lax-Wendroff scheme [54, 55] to compute the numerical solution. At the initial step we set $\mathbf{a}_1 = 0$ and $\mathbf{b}_1 = 0$ for all

the square domain. Then we choose randomly a position within the domain, and we add there a 2-dimensional Gaussian shaped peak $\mathbf{h}_1$ (central height of 2 units). This Gaussian spike simulates a local disturbance like a droplet hitting the surface. Then the generated wave propagates forward and backward in our domain. The observations are obtained by adding Gaussian noise $\mathbf{v}_k$ to the state vector. In the following simulations, we choose the EnVar to compare with the proposed UEVF. For the two filters we start with the same initial condition $\mathbf{x}_1^a$, and use the same observation sequence $\{\mathbf{y}_0 \cdots \mathbf{y}_k\}$.

Concerning the UEVF, we set the parameters β, λ, the threshold γ_1, the covariance inflation factor δ to the same values as in the experiment 5.1. The covariance filtering scale is $\ell_c = 250$, the sample size bound are $[\ell_l, \ell_u] = [5, 25]$, the time step is $\Delta t = 0.1$s, the space mesh is $\Delta x = \Delta y = 1.0$ units, the initial sampling size is $\ell_1 = 15$.

We consider different scenarios in terms of depth $\mathbf{h}_k(x, y)$ with different noise levels, i.e., $\mathbf{Q}_k = \mathbf{R}_k = 0.003 \times \mathrm{Id}(40, 40)$ and $\mathbf{Q}_k = \mathbf{R}_k = 0.03 \times \mathrm{Id}(40, 40)$. These different cases may explore the effect of noises on the performance of the UEVF and the EnVar. The corresponding depth $\mathbf{h}_k(x, y)$ at a same time step and the relative RMS error, are plotted in Fig.7 and Fig.8.

Different noise levels leads to different behaviour of estimates. The UEVF estimates perceptibly outperform the EnVar. During the first experiment, the noise is relatively weak and the shape of the estimates is similar to the reality. The UEVF filtered field is sharper than the EnVar with a cleaner background. This is obvious for $k = 20$ (Fig.7(a)) or $k = 40$ (Fig.7(b)). Moreover when the time increases, the EnVar filtered signal is more fuzzy and only a little improvement over the background noise is achieved.

Then, the noise level is increased by 10-fold the previous one (Fig.8). The estimate of the UEVF still preserves the shape of the reality (Fig.8(a)) with a tiny background noise. On the other hand the EnVar estimates show a worse shape. The deformation and shrinkage are more than the previous experiment. After 20 steps propagation ($k = 40$) the situation is similar. The EnVar has no more improvement, while UEVF still maintains a level of confidence.

In order to quantify the effect of different noise on the accuracy, to compute the relative RMS error (Fig.9), we adopt the relative 2-norm based RMS error [19] to quantify the matrix-posed sequence $\mathbf{h}_k(x, y)$, which is defined by $E_k = \| \mathbf{x}_k^a - \mathbf{x}_k^{\text{true}} \|_2 \,/\, \| \mathbf{x}_k^{\text{true}} \|_2$, where $\| \cdot \|_2$ denotes L_2 norm. One can see by comparing Fig.9(a) and Fig.9(b) that whatever the level of noise, the EnVar

tends to diverge with big rate. As the noise level increases, the RMS errors for both methods become larger. The gap between the two curves becomes wider. This indicates that the EnVar is more sensible to the noise level. In the case of high noise level, although the RMS error of the UEVF undergoes to an increase signal-to-noise ratio, the height field is still well-reconstructed (Fig.8(a)). Finally whatever the experiment, the UEVF seems to be more consistent with the nonlinear filtering estimation while the EnVar seems to be noise dependent and its results could be divergent.

6. Conclusions

In this work, an efficient estimator based on the concept of the variational optimization and the unscented transform as well as the ensemble transform, therefore called the Unscented/Ensemble transform-based Variational Filter (UEVF), is developed.

Roughly speaking the improvements achieved by the UEVF holds in two points: to attempt to break the linear assumption of the KF and to learn online the background error covariance matrix. Indeed in practice considering the complex dynamics characterized by strong nonlinearity and high dimensional, the relation between estimate and observation is not necessary linear and the linear assumption may introduce strong errors. It could wreck all efforts of performance enhancement. Therefore, we suggest in UEVF that, the scheme of KF-like mean update could be replaced by a variational minimization. Then the estimate is a quadratic function of the observation and the background state. Moreover, the background error matrix $\mathbf{B}_k$ in the variational minimization is replaced by a rank-reduced error covariance, which is designed by a deterministic statistics and updated from a set of size-truncated ensembles.

The generation of ensembles in the UEVF inherits from the Unscented Transform (UT), as that in EnUKF [19]. In the UT for the purpose to generate a symmetric ensemble associated to the weights and to propagate it forward, the weights are centered and spread out in positive-negative directions. It helps to localize the distribution of ensemble and benefits to the convergence of filtering.

In the UEVF, mean update is implemented via a variational analysis. It reduces the problem of statistical moment estimation to a quadratic optimization. This scheme extends the linear update of the KF to a quadratic

minimization and therefore it is not surprising to find an unique numerical solution which enhances the performance with a considerable quantity.

With the comparison of nonlinear statistics estimation between the EnKF and the UKF in appendix, we show that the UT adopted by the UEVF has a best accuracy, at least up to second order of Taylor series expansion for an arbitrary distribution of random variable. The accuracy of estimation can be increased to third order when a symmetric distribution is given by the random variable. In addition, the filter parameters (λ, β, the ensemble size ℓ_k, the threshold γ_k, lower and upper boundaries ℓ_l and ℓ_u) provide extra flexibilities to guarantee better accuracies with respect to the filtering problem and the dynamics of the model.

With numerical experiments, we have demonstrated the improvement of estimation accuracies, the filtering convergence as well as a correct behavior in front of the high dimensional effects via two kinds of dynamical systems: the Lorenz-95 model and the 2D perturbed shallow water equation. We have compared the proposed UEVF with the EnUKF and the EnVar in terms of the relative RMS error and time consumption by a number of Monte Carlo runs, the estimation accuracy and the computation cost confirm the outperforming and efficiency of the UEVF.

Acknowledgments

This research was partially supported by the ANR PREVASSEMBLE project.

Appendix: Comparison the accuracies of mean estimation between the EnKF and the UKF

We compare the mean estimation accuracy of the EnKF and the UKF is performed as below.

The EnKF uses an ensemble of state models to represent the empirically statistic errors [26, 44]. Accordingly we assume that at the end $k-1$ time step, the of N members are updated and form the analysis ensemble $\{\mathcal{X}^a_{k-1,i}\}_{i=1}^N$. The analyzed members are propagated forward by the nonlinear dynamics in Eq.(1) and generate the background members, $\{\mathcal{X}^b_{k,i} : \mathcal{X}^b_{k,i} = \mathcal{M}_{k-1,k}(\mathcal{X}^a_{k-1,i}, \mathbf{0})\}_{i=1}^N$.

For the convenience of the discussion, we only focus on the observation function $\mathcal{H}_k$ in Eq.(2). We assume that $\mathcal{H}_k$ can be expanded in a Taylor

series with finite differential order. The time index in subscript is dropped in this appendix. So based on the background members $\mathcal{X}^b \triangleq \{\mathcal{X}_i^b\}_{i=1}^N$, the ensemble mean via Taylor expansion can be written as

$$\begin{aligned}
\mathbf{y}^b &= \mathbb{E}[\mathcal{H}(\mathcal{X}^b, \mathbf{v})] = \mathbb{E}[\mathcal{H}(\overline{\mathbf{x}^b} + \widetilde{\mathcal{X}^b}, \mathbf{v})] \\
&= \mathcal{H}(\overline{\mathbf{x}^b}, \mathbf{0}) + \frac{1}{2!}\nabla^T\left[\frac{1}{N}\sum_{i=1}^N \widetilde{\mathcal{X}_i^b}(\widetilde{\mathcal{X}_i^b})^T\right]\nabla\mathcal{H} \\
&\quad + \frac{1}{3!}\nabla^T\left[\frac{1}{N}\sum_{i=1}^N \widetilde{\mathcal{X}_i^b}(\widetilde{\mathcal{X}_i^b})^T(\nabla\widetilde{\mathcal{X}_i^b})^T\right]\nabla\mathcal{H} + \cdots \\
&\approx \mathcal{H}(\overline{\mathbf{x}^b}, \mathbf{0}) + \frac{N-1}{N\times 2!}\nabla^T\overline{\mathbf{P}_{xx}^b}\nabla\mathcal{H} + \frac{\sum_{i=1}^N \mathbf{D}^3_{\widetilde{\mathcal{X}_i^b}}\mathcal{H}}{N\times 3!} + \frac{\sum_{i=1}^N \mathbf{D}^4_{\widetilde{\mathcal{X}_i^b}}\mathcal{H}}{N\times 4!} + \cdots,
\end{aligned} \tag{32}$$

where $\{\widetilde{\mathcal{X}_i^b} = \mathcal{X}_i^b - \overline{\mathbf{x}^b}\}_{i=1}^N$ denotes the background errors, the statistical mean is $\overline{\mathbf{x}^b} = \sum_{i=1}^N \mathcal{X}_i^b/N$, the statistical covariance $\overline{\mathbf{P}_{xx}^b} = \sum_{i=1}^N \widetilde{\mathcal{X}_i^b}(\widetilde{\mathcal{X}_i^b})^T/(N-1)$. The theoretical mean $\mathbf{x}^b$ and covariance $\mathbf{P}_{xx}^b$ are substituted by their empirical approximation, i.e., $\overline{\mathbf{x}^b} \approx \mathbf{x}^b$, $\overline{\mathbf{P}_{xx}^b} \approx \mathbf{P}_{xx}^b$. The symbol $\mathbf{D}_{\widetilde{\mathcal{X}_i^b}} \triangleq (\widetilde{\mathcal{X}_i^b})^T\nabla$ is used to simplify expression. Similarly, the ensemble covariance $\mathbf{P}_{yy}^b$ is given by the expansion

$$\begin{aligned}
\mathbf{P}_{yy}^b &= \mathbb{E}[(\mathbf{y}-\mathbf{y}^b)(\mathbf{y}-\mathbf{y}^b)^T] \\
&\approx \nabla^T\overline{\mathbf{P}_{xx}^b}\nabla\mathcal{H} + \frac{\sum_{i=1}^N(\mathbf{D}_{\widetilde{\mathcal{X}_i^b}}\mathcal{H})(\mathbf{D}^2_{\widetilde{\mathcal{X}_i^b}}\mathcal{H})^T}{(N-1)\times 2!} + \frac{1}{N-1}\left[\frac{\sum_{i=1}^N \mathbf{D}_{\widetilde{\mathcal{X}_i^b}}\mathcal{H}(\mathbf{D}^3_{\widetilde{\mathcal{X}_i^b}}\mathcal{H})^T}{3!}\right. \\
&\quad \left. + \frac{\sum_{i=1}^N(\mathbf{D}^2_{\widetilde{\mathcal{X}_i^b}}\mathcal{H})(\mathbf{D}^2_{\widetilde{\mathcal{X}_i^b}}\mathcal{H})^T}{2!\times 2!} + \frac{\sum_{i=1}^N(\mathbf{D}^3_{\widetilde{\mathcal{X}_i^b}}\mathcal{H})(\mathbf{D}_{\widetilde{\mathcal{X}_i^b}}\mathcal{H})^T}{3!}\right] \\
&\quad - \frac{N-1}{N}\frac{(\nabla^T\overline{\mathbf{P}_{xx}^b}\nabla\mathcal{H})(\nabla^T\overline{\mathbf{P}_{xx}^b}\nabla\mathcal{H})^T}{2!\times 2!} + \cdots.
\end{aligned} \tag{33}$$

We can see from the above expression with the higher terms (more than second order in RHS of (32) and (33)) that spurious modes may appear. They vanish as the ensemble size N tends to infinity. Moreover if the size N is finite, the term $\nabla^T\overline{\mathbf{P}_{xx}^b}\nabla\mathcal{H}$ in (32) and the term $(\nabla^T\overline{\mathbf{P}_{xx}^b}\nabla\mathcal{H})(\nabla^T\overline{\mathbf{P}_{xx}^b}\nabla\mathcal{H})^T$ in (33) are always uncentered.

In order to analyze the accuracies of the unscented transform estimation in the UKF, similarly we apply the Taylor expansion to the observation function $\mathcal{H}$. Then for the sigma-points $\mathcal{X}^b \triangleq \{\mathcal{X}_i^b\}_{i=0}^{2L}$ generated from the background mean $\mathbf{x}^b$ and its covariance $\mathbf{P}_{xx}^b$, the unscented mean can be written as

$$\begin{aligned}\mathbf{y}^b &= \sum_{i=0}^{2L} W_i \mathcal{H}(\mathcal{X}_i^b, \mathbf{0}) \\ &= \mathbb{E}[\mathcal{H}(\mathcal{X}^b, \mathbf{v})] = \mathbb{E}[\mathcal{H}(\mathbf{x}^b + \widetilde{\mathcal{X}^b}, \mathbf{v})] \\ &= \mathcal{H}(\mathbf{x}^b, \mathbf{0}) + \frac{\nabla^T \mathbf{P}_{xx}^b \nabla \mathcal{H}}{2!} + \frac{1}{2(L+\lambda)} \sum_{i=1}^{2L} \Big(\mathbf{D}^4_{\widetilde{\mathcal{X}_i^b}} \mathcal{H} + \mathbf{D}^6_{\widetilde{\mathcal{X}_i^b}} \mathcal{H} + \cdots \Big) \ , \end{aligned} \tag{34}$$

where $\widetilde{\mathcal{X}^b} \triangleq \{\widetilde{\mathcal{X}_i^b} = \mathcal{X}_i^b - \mathbf{x}^b\}_{i=1}^N$ stands for the background error. One can see that all the odd terms will vanish because of the symmetry of the sigma-points. The unscented covariance may be expanded in the series of

$$\begin{aligned}\mathbf{P}_{yy}^b &= \sum_{i=0}^{2L} W_i [\mathcal{H}(\mathcal{X}_i^b, \mathbf{0}) - \mathbf{y}^b][\mathcal{H}(\mathcal{X}_i^b, \mathbf{0}) - \mathbf{y}^b]^T \\ &= \mathbb{E}[(\mathbf{y} - \mathbf{y}^b)(\mathbf{y} - \mathbf{y}^b)^T] \\ &= \nabla^T \mathbf{P}_{xx}^b \nabla \mathcal{H} + \frac{1}{2(L+\lambda)} \Bigg[\frac{\sum_{i=1}^{2L} (\mathbf{D}_{\widetilde{\mathcal{X}_i^b}} \mathcal{H})(\mathbf{D}^3_{\widetilde{\mathcal{X}_i^b}} \mathbf{f})^T}{3!} \\ &\quad + \frac{\sum_{i=1}^{2L} (\mathbf{D}^2_{\widetilde{\mathcal{X}_i^b}} \mathcal{H})(\mathbf{D}^2_{\widetilde{\mathcal{X}_i^b}} \mathcal{H})^T}{2! \times 2!} + \frac{\sum_{i=1}^{2L} (\mathbf{D}^3_{\widetilde{\mathcal{X}_i^b}} \mathcal{H})(\mathbf{D}_{\widetilde{\mathcal{X}_i^b}} \mathcal{H})^T}{3!} \Bigg] \\ &\quad - \frac{(\nabla^T \mathbf{P}_{xx}^b \nabla \mathcal{H})(\nabla^T \mathbf{P}_{xx}^b \nabla \mathcal{H})^T}{2! \times 2!} + \cdots \ , \end{aligned} \tag{35}$$

The Eq.(34) shows that the first three-order terms are exact. The approximation begins with the fourth order. In Eq.(35) there is no spurious mode attributable to the sample size due to the higher terms than the third order. For the accuracies of nonlinear estimation, explicitly the unscented transform is better than the pure ensemble algorithm. Thus for the UEVF suggested in Section 3, we have chosen the unscented transform to shape the covariance.

References

[1] Daley R., Atmospheric Data Analysis. Cambridge Atmospheric and Space Science Series, 1991, Cambridge University Press.

[2] Lorenc A., Analysis methods for numerical weather prediction. Quart. J. Roy. Meteor. Soc., 112, 1986: 1177-1194

[3] Courtier P., Andersson E., Heckley W., Pailleux J., Vasiljevic, D. and co-authors, The ECMWF implementation of three-dimensional variational assimilation (3D-VAR). Part 1: formulation. Quart. J. Roy. Meteorol. Soc. 124, 1998: 1783-1807

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

[5] Del Moral P., Rigal G. , and Salut G. , Filtrage non-linéaire non-gaussien appliqué au recalage de plates-formes inertielles, Rapport LAAS No.92207, 1991

[6] Gordon N. J., Salmond D. J. and Smith A. F. M. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. IEEE Proceedings on Radar and Signal Processing 140 (2), 1993: 107-113. doi:10.1049/ip-f-2.1993.0015

[7] Evensen G., The Ensemble Kalman Filter: theoretical formulation and practical implementation, Ocean Dynamics, 53(4), 2003: 343-367

[8] Houtekamer P. L., Mitchell H.L., A sequential ensemble Kalman filter for atmospheric data assimilation. Mon. Wea. Rev., 129 (1), 2001, 123-137

[9] Le Gland F., Monbet V., Tran V. D., Large Sample Asymptotics for the Ensemble Kalman Filter. Research Report, inria-00409060,2009, URL = http://hal.inria.fr/inria-00409060/PDF/RR-7014.pdf.

[10] Papadakis N., Mémin E., Cuzol A. and Gengembre N., Data assimilation with the weighted ensemble Kalman filter, Tellus A, 62, 2010, 673-697. doi: 10.1111/j.1600-0870.2010.00461.x

[11] Uhlmann J. K., Julier S. J., and Durrant-Whyte H. F., A new approach for the nonlinear transformation of means and covariances in linear filters. IEEE Trans. Automatic Control, 1996

[12] Julier, S.J., and Uhlmann, J.K., IDAK Ind., Unscented filtering and nonlinear estimation, Jefferson City, MO, USA, Proceedings of the IEEE, 92(3), 2004, 401-422

[13] Van der Merwe R. , Doucet A. , de Freitas N. and Wan E., the Unscented Particle Filter, Advances in Neural Information Processing Systems (NIPS13), ,2000, MIT Press

[14] Moore B. C., Principal component analysis in linear systems: controllability, observability, and model reduction. IEEE Trans. Autom. Control, 26 (1), 1981, 17-32

[15] Farrell B.F. and Ioannou P.J , State estimation using a reduced order Kalman Filter, J. Atmos Sci, 58, 2001, 3666-3680

[16] Buehner, M., and Malanotte-Rizzoli P., Reduced-rank Kalman filters applied to an idealized model of the wind-driven ocean circulation, J. Geophys. Res., 108, 2003, 3192-3207, doi:10.1029/2001JC000873

[17] Sayed A. H. and Kailath T., A State-Space Approach to Adaptive RLS Filtering, IEEE Sig. Proc. Mag., 1994, 11 (3)

[18] Van der Merwe R. , Wan E. A., The square-root unscented Kalman filter for state and parameter-estimation, IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings, 6, 2001, 3461-3464

[19] Luo X., Moroz I.M., Ensemble Kalman filter with the unscented transform Physica D: Nonlinear Phenomena, 238 (5), 2009, 549-562

[20] Sakov P., Comment on 'Ensemble Kalman filter with the unscented transform', Physica D: Nonlinear Phenomena, 238 (22), 2009, 2227-2228

[21] Luo X., Moroz I.M., Hoteit I. , Reply to 'Comment on 'Ensemble Kalman filter with the unscented transform", Physica D: Nonlinear Phenomena, 239 (17), 2010, 1662-1664.

[22] Del Moral, P., Feynman-Kac Formulae Genealogical and Interacting Particle Systems with Applications. Series: Probability and Applications, 2004, Springer, New York

[23] Julier S.J., Uhlmann J.K., A New Extension of the Kalman Filter to Nonlinear Systems, Proceedings of the 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls, 1997, 182-193

[24] Julier S.J., Uhlmann J.K., Durrant-Whyte H., A new method for the nonlinear transformation of means and covariances in filters and estimators, IEEE Transactions on Automatic Control 45 (2000) 477-482

[25] Julier S. J. , Uhlmann J. K., Unscented filtering and nonlinear estimation, Proc. IEEE 92, 2004, 401-422

[26] Evensen G., The Ensemble Kalman Filter: Theoretical Formulation and Practical Implementation, Ocean Dynamics, 53 (4), 2003, 343-367

[27] Evensen G., The ensemble Kalman filter for combined state and parameter estimation IEEE Control Systems Magazine, 29 (3), 2009, 83-104

[28] Bishop C. H., Etherton B. J., Majumdar S. J., Adaptive sampling with ensemble transform kalman filter. Part I: Theoretical Aspects, Mon. Wea. Rev. 129 (2001) 420-436

[29] Hamill T. M., and Snyder C., A Hybrid Ensemble Kalman Filter-3D Variational Analysis Scheme. Mon. Wea. Rev., 128, 2000, 2905-2919

[30] Rihan F. A., Collier C. G., Ballard S. P. and Swarbrick S. J., Assimilation of Doppler radial winds into a 3D-Var system: Errors and impact of radial velocities on the variational analysis and model forecasts, Quart. J. Roy. Meteor. Soc., 134, 2008, 1701-1716

[31] Hansen P. C., The truncated svd as a method for regularization, BIT, 27 (4), 1987, 534-553

[32] Turner M. R. J., Walker J. P., and Oke P. R., Ensemble member generation for sequential data assimilation, Remote Sensing of Environment, 112 (4), 2008, 1421-1433

[33] Ehrendorfer M. and Tribbia J. J. , Optimal prediction of forecast error covariances through singular vectors, J. Atmos. Sci., 54, 1997, 286-313

[34] Uzunoglu B. , Fletcher S. J., Zupanski M., and Navon I. M., Adaptive ensemble reduction and inflation, Quart. J. Roy. Meteor. Soc., 133, 2005, 1281-1294

[35] Wright M.H., Practical Optimization. Academic Press, 1981

[36] Steihaug T., The Conjugate Gradient Method and Trust Regions in Large Scale Optimization, SIAM Journal on Numerical Analysis, 20 (3), 1983, 626-637

[37] Navon, I. M. and Legler, D. M., Conjugate-Gradient Methods for Large-Scale Minimization in Meteorology, Mon. Wea. Rev., 115 (8), 1987, 1479-1502

[38] Gilbert J.C., Nocedal J., Global Convergence Properties of Conjugate Gradient Methods for Optimization, SIAM J. Optim, 2 (1), 1992, 21-42.

[39] Dai Y. H. and Yuan Y. , A Nonlinear Conjugate Gradient Method with a Strong Global Convergence Property, SIAM J. Optim, 10 (1), 1999, 177-182

[40] Hestenes M. R. and Stieffel E., Methods of conjugate-gradients for solving linear systems, J. Res. Natl. Bur. Stand., 48 (6), 1952, 409-436

[41] Fletcher R. and C. M. Reeves, Function minimization by conjugate-gradients, The Computer Journal, 7 (2), 1964, 149-153

[42] Hestenes, M.R., Conjugate directions methods in optimization, Applications of Mathematics, Vol.12, 1980, Springer-Verlag

[43] Sakov P., Evensen G. and Bertino L., Asynchronous data assimilation with the EnKF, Tellus A., 62 (1), 2010, 24-29

[44] Evensen G., Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res, 99 (C5), 1994, 10143-10162

[45] Whitaker J. S. , and Hamill T. M. , Ensemble data assimilation without perturbed observations, Mon. Wea. Rev., 130, 2002, 1913-1924

[46] Livings D. M. , Dance S. L., Nichols N. K., Unbiased ensemble square root filters, Mon. Wea. Rev. , 131 (7), 2001, 1485-1490

[47] Wang X. , Bishop C. H., Julier S. J., Which Is Better, an Ensemble of Positive-Negative Pairs or a Centered Spherical Simplex Ensemble ?, Mon. Wea. Rev. 132, 2004, 1590-1605

[48] Anderson J. L., An ensemble adjustment kalman filter for data assimilation, Mon. Wea. Rev. 129, 2001, 2884-2903

[49] Hamill T. M., Whitaker J. S., and Snyder C. , Distance-dependent filtering of background error covariance estimates in an ensemble kalman filter, Mon. Wea. Rev. 129 (11), 2001, 2776-2790

[50] Houtekamer P. L. , and Mitchell H. L., Data assimilation using an ensemble kalman filter technique, Mon. Wea. Rev., 126, 1998, 796-811

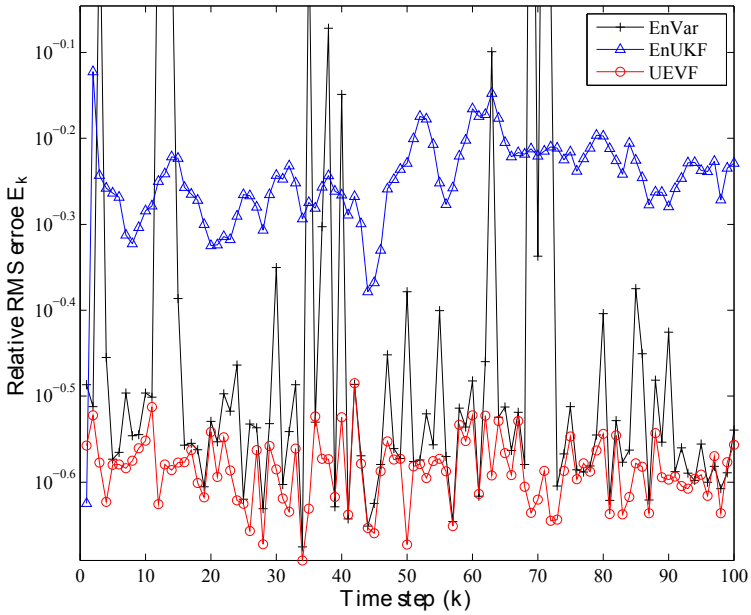
[51] Lorenz E. N., Predictability: A problem partly solved, In Seminar on Predictability, V1, ECMWF - Reading, 1995

[52] Lorenz E. N., and Emanuel K. A., Optimal sites for supplementary weather observations: Simulation with a small model, J. Atmos. Sci., 55, 1998, 399-414

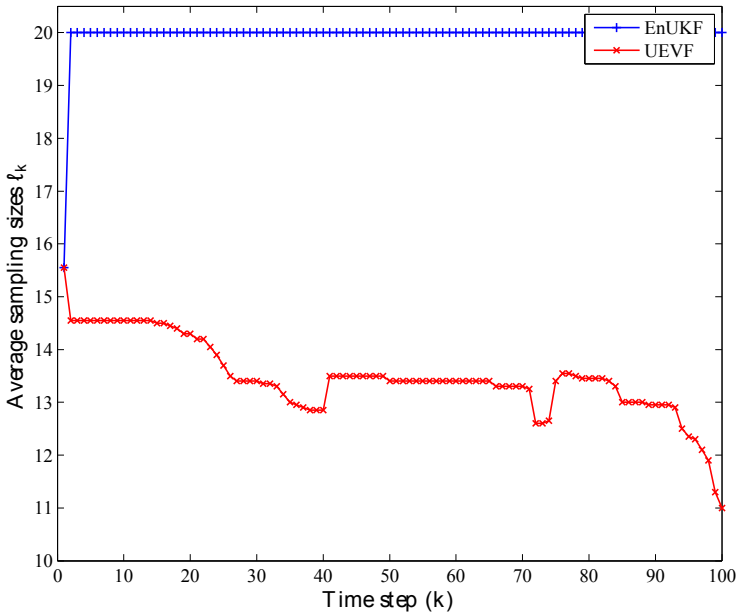
[53] Weigel P. , Liniger A. , and Appenzeller C. , The Discrete Brier and Ranked Probability Skill Scores, Mon. Wea. Rev., 135 (1), 2007, 118-124

[54] Lax P.D , and Wendroff B. , Systems of conservation laws, Commun. Pure Appl Math, 13, 1960, 217-237

[55] Thompson M.J., An Introduction to Astrophysical Fluid Dynamics, Imperial College Press, London, 2006

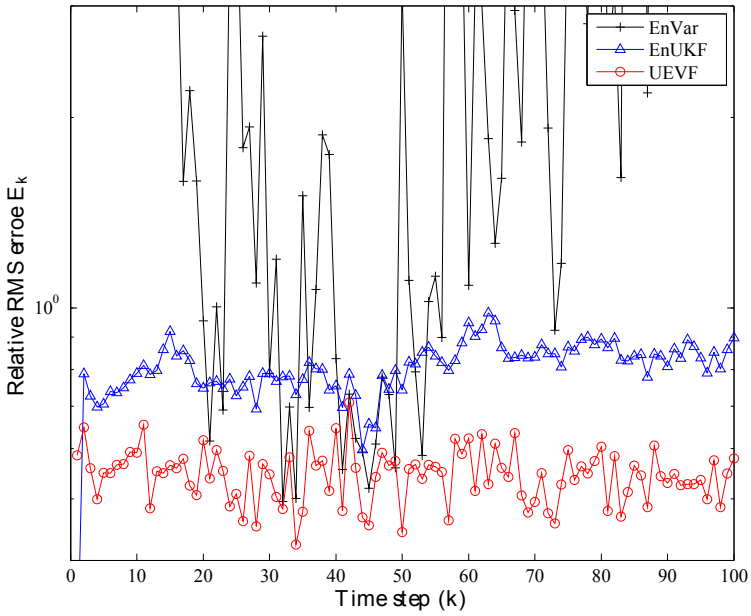


(a) Comparison between the relative RMS errors E_k of the UEVF, the EnUKF and the EnVar

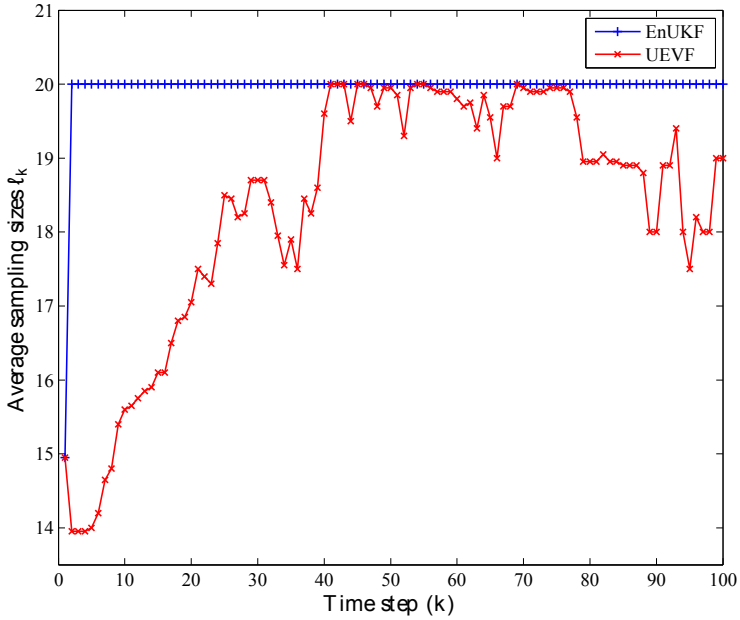


(b) Comparison between the averaged sampling size ℓ_k of the UEVF and the EnUKF

Figure 1: Effects of the low level of state and observation noise covariance $\mathbf{Q}_k = 5. \times \mathrm{Id}(40, 40)$ and $\mathbf{R}_k = 0.1 \times \mathrm{Id}(40, 40)$ on the estimate accuracy and the evolution of sampling size ℓ_k. State dimension $n = 40$ and 50 Monte Carlo simulations of the Lorenz-95 model.

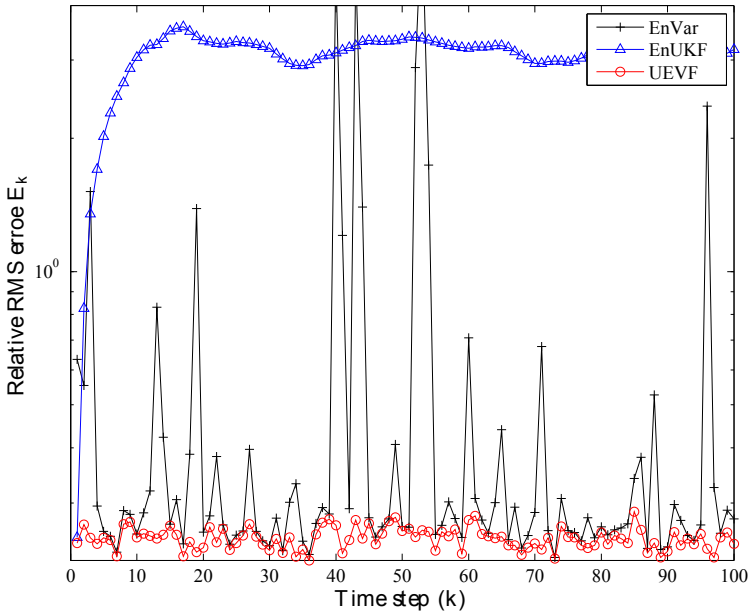


(a) Comparison between the relative RMS errors E_k of the UEVF, the EnUKF and the EnVar

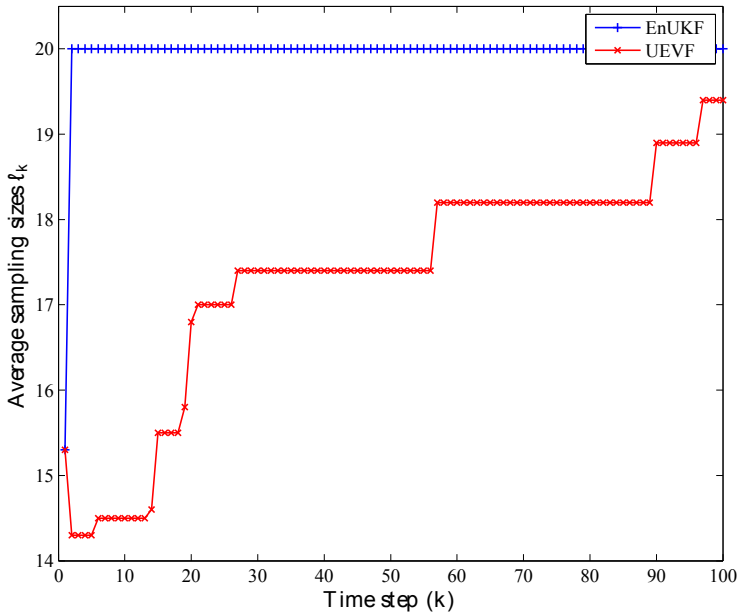


(b) Comparison between the averaged sampling size ℓ_k of the UEVF and the EnUKF

Figure 2: Effects of the high level of noise with covariance $\mathbf{Q}_k = 10. \times \mathrm{Id}(40, 40)$ and $\mathbf{R}_k = 0.5 \times \mathrm{Id}(40, 40)$ on the estimate accuracy and the evolution of sampling size ℓ_k. State dimension $n = 40$ and 50 Monte Carlo simulations of the Lorenz-95 model.

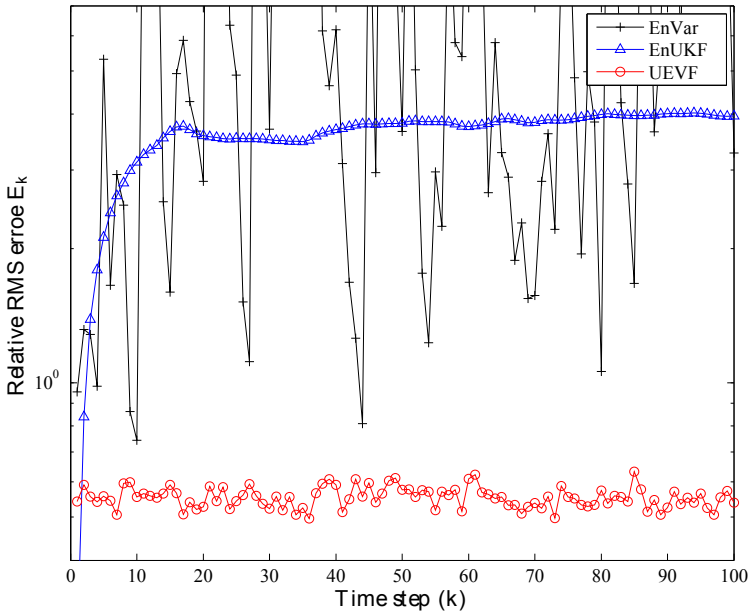


(a) Comparison between the relative RMS errors E_k of the UEVF, the EnUKF and the EnVar

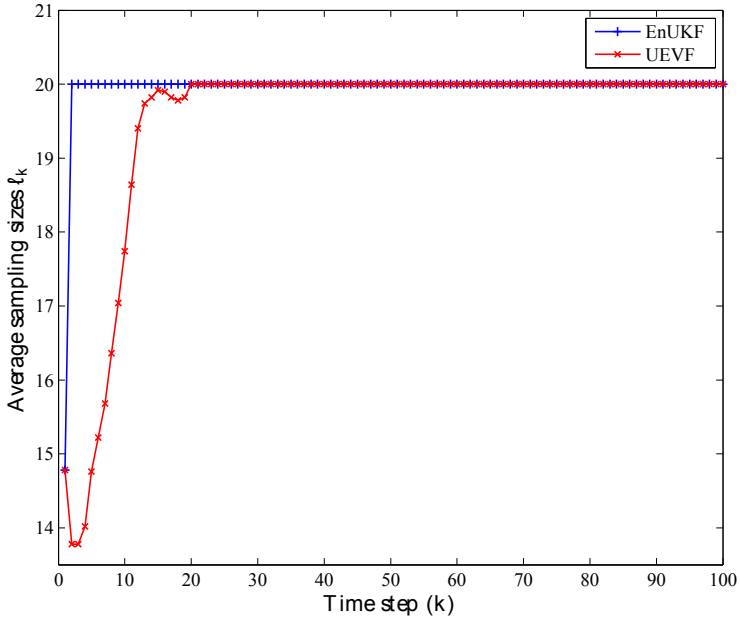


(b) Comparison between the averaged sampling size ℓ_k of the UEVF and the EnUKF

Figure 3: Effects of the dimension $n = 200$ with ordinary covariance $\mathbf{Q}_k = 5. \times \mathrm{Id}(200, 200)$ and $\mathbf{R}_k = 0.1 \times \mathrm{Id}(200, 200)$ on the estimate accuracy and the evolution of sampling size ℓ_k. 50 Monte Carlo simulations of the Lorenz-95 model.

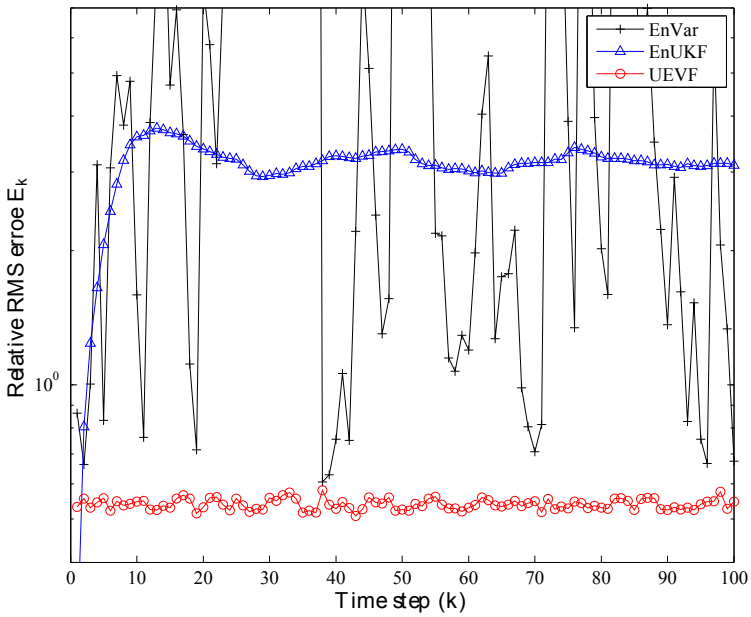


(a) Comparison between the relative RMS error E_k of the UEVF, the EnUKF and the EnVar

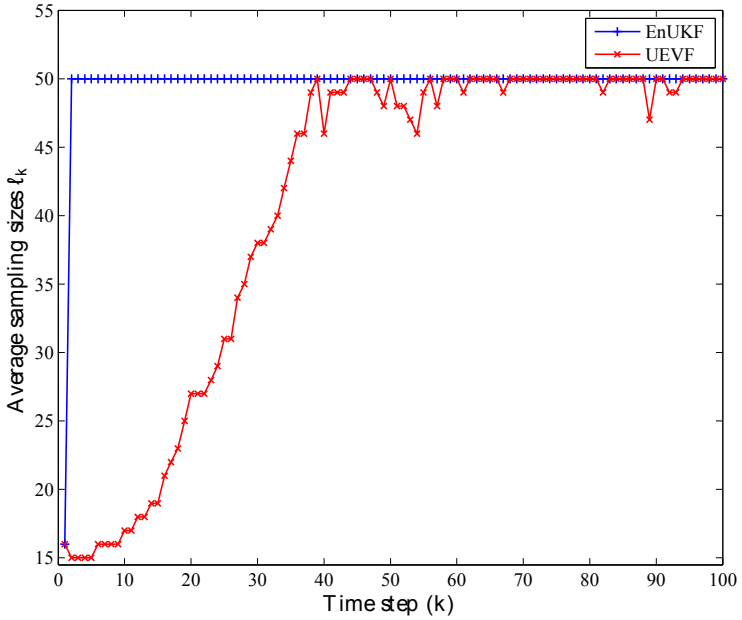


(b) Comparison between the averaged sampling size ℓ_k of the UEVF and the EnUKF

Figure 4: Effects of the high dimension $n = 200$ with ordinary covariance $\mathbf{Q}_k = 10. \times \mathrm{Id}(200, 200)$ and $\mathbf{R}_k = 0.5 \times \mathrm{Id}(200, 200)$ on the estimate accuracy and the evolution of sampling size ℓ_k. 50 Monte Carlo simulations of the Lorenz-95 model.



(a) Comparison between the relative RMS error E_k of the UEVF, the EnUKF and the EnVar



(b) Comparison between the averaged sampling size ℓ_k of the UEVF and the EnUKF

Figure 5: Effects of the high state dimension $n = 800$ with ordinary covariance $\mathbf{Q}_k = 10. \times \mathrm{Id}(800, 800)$ and $\mathbf{R}_k = 0.5 \times \mathrm{Id}(800, 800)$ on the estimate accuracy and the evolution of sampling size ℓ_k. 50 Monte Carlo simulations of the Lorenz-95 model.

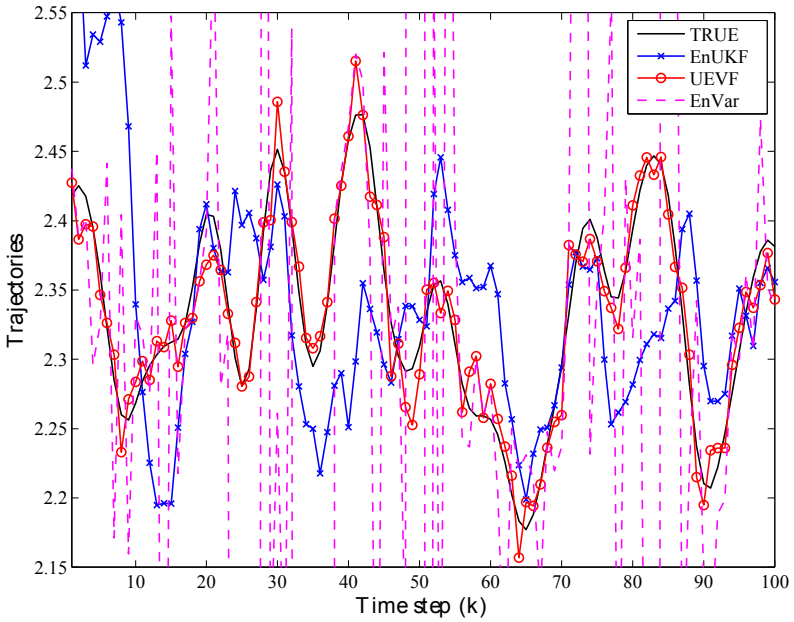
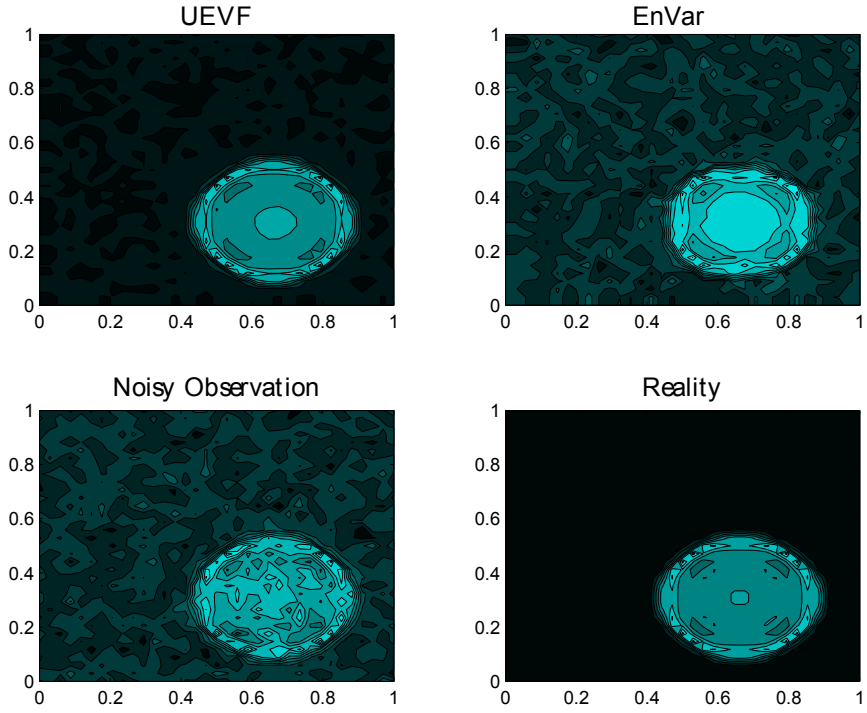
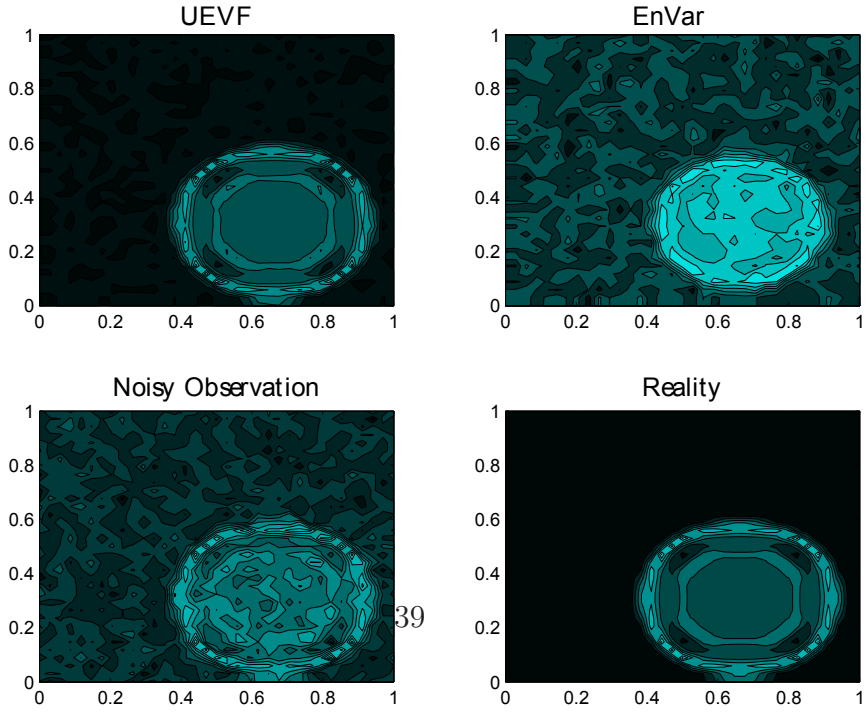


Figure 6: Comparison between the averaged trajectories of the UEVF, the EnUKF, the EnVar and the reality for one component of the state vector of the Lorenz-95 model

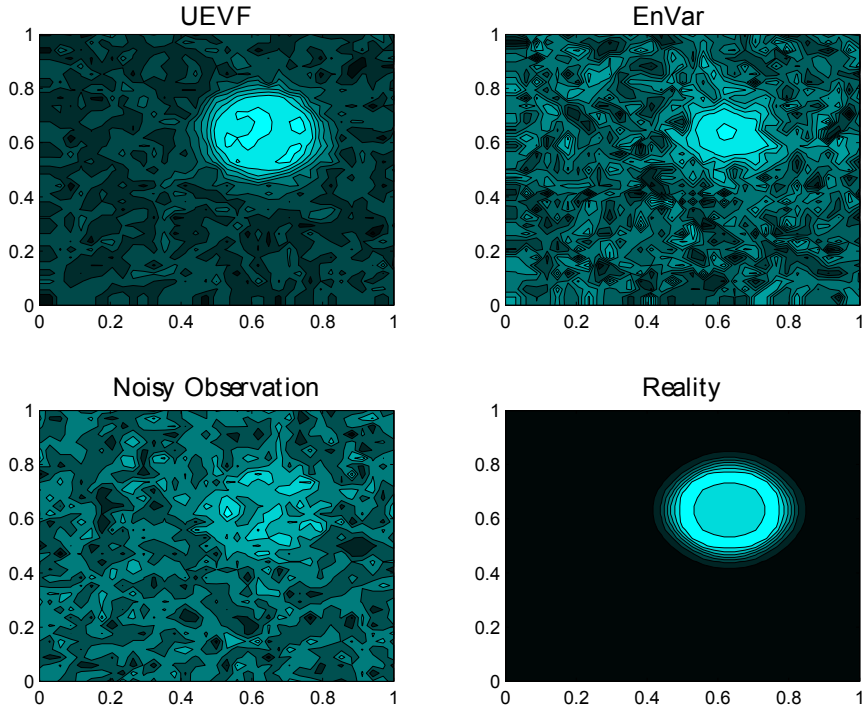


(a) The depths $\mathbf{h}_k(x, y)$ estimate by the UEVF and the EnVar compared to the noisy observation and the reality, at time $k = 20$

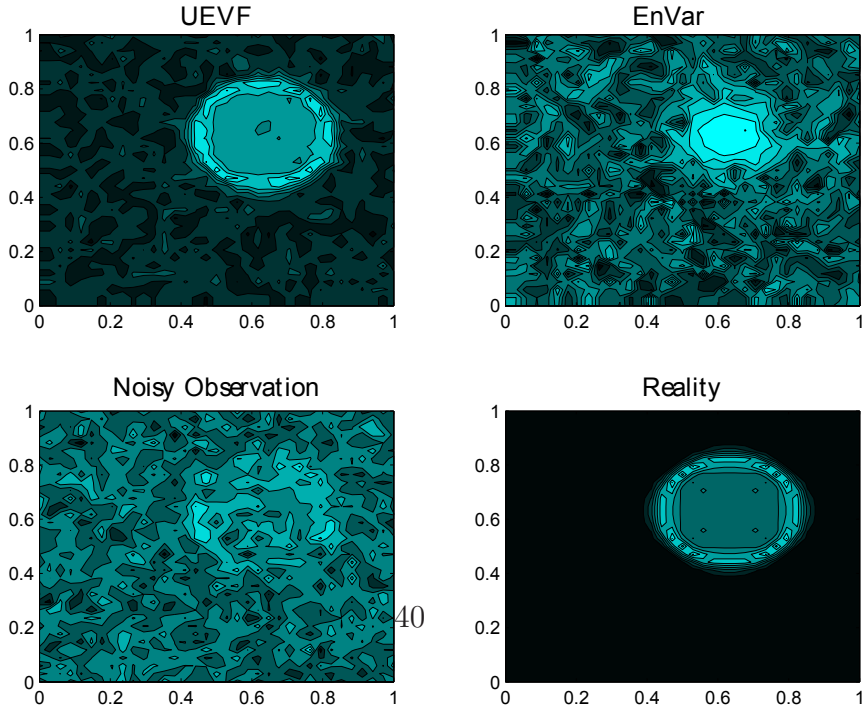


(b) The depths $\mathbf{h}_k(x, y)$ estimate by the UEVF and the EnVar compared to the noisy observation and the reality, at instant $k = 40$

Figure 7: Effects of dynamical and observational noises with $\mathbf{Q}_k = \mathbf{R}_k = 0.003 \times \mathrm{Id}(40, 40)$ on the estimates, for two time step $k = 20, 40$ of the shallow water simulation.

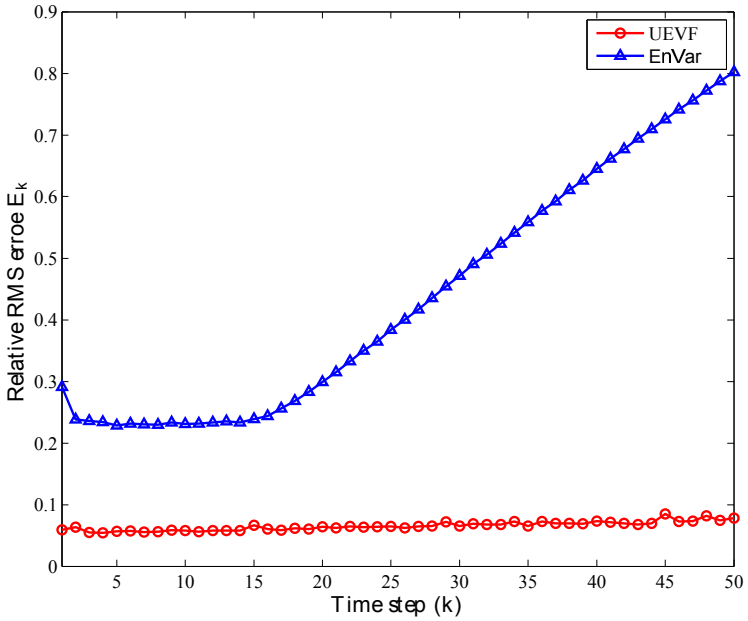


(a) The depths $\mathbf{h}_k(x, y)$ estimate by the UEVF and the EnVar compared to the noisy observation and the reality, at time $k = 20$

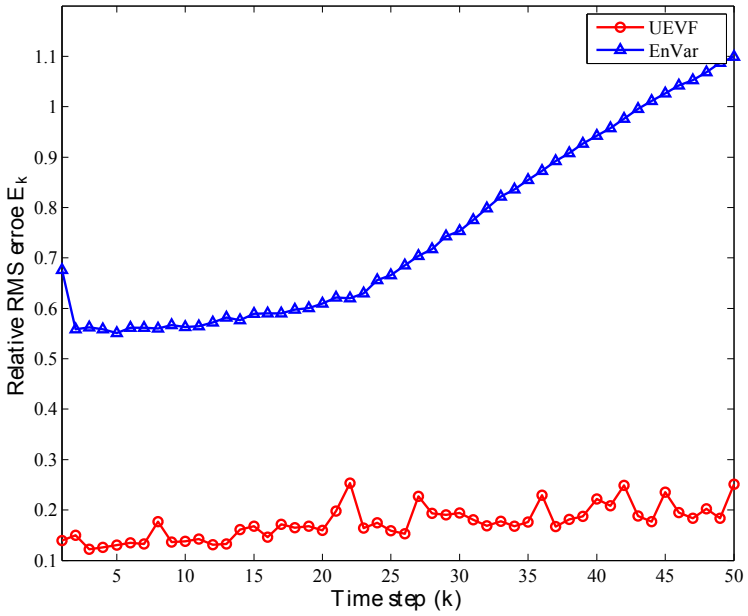


(b) The depths $\mathbf{h}_k(x, y)$ estimate by the UEVF and the EnVar compared to the noisy observation and the reality, at time $k = 40$

Figure 8: Effects of dynamical and observational noises with $\mathbf{Q}_k = \mathbf{R}_k = 0.03 \times \mathrm{Id}(40, 40)$ on the estimates, for two time step $k = 10, 40$ of the shallow water simulation.



(a) Comparison between the relative RMS errors E_k of the UEVF and the EnVar in the case of $\mathbf{Q}_k = \mathbf{R}_k = 0.003 \times \mathrm{Id}(40, 40)$



(b) Comparison between the relative RMS errors E_k of the UEVF and the EnVar in the case of $\mathbf{Q}_k = \mathbf{R}_k = 0.03 \times \mathrm{Id}(40, 40)$

Figure 9: Effects of the different level of state and observation noises on the accuracies in the estimation of water heights in the shallow water simulation.