Reply to “Comment on ‘Ensemble Kalman filter with the unscented transform’”

X. Luo\textsuperscript{a,c,*}, I.M. Moroz\textsuperscript{b}, I. Hoteit\textsuperscript{d}

\textsuperscript{a} King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
\textsuperscript{b} Mathematical Institute, 24-29 St Giles’, Oxford, OX1 3LB, UK
\textsuperscript{c} Oxford-Man Institute of Quantitative Finance, Eagle House, Walton Well Road, Oxford, OX2 6ED, UK

\textbf{ARTICLE INFO}

Article history:
Received 1 December 2009
Received in revised form
20 April 2010
Accepted 26 April 2010
Available online 12 May 2010
Communicated by T. Sauer

Keywords:
Data assimilation
Ensemble Kalman filter
Ensemble transform Kalman filter
Unscented transform
Ensemble unscented Kalman filter

\textbf{ABSTRACT}

This is a reply to the comment of Dr. Sakov on the work “Ensemble Kalman filter with the unscented transform” of Luo and Moroz (2009) [2].

© 2010 Elsevier B.V. All rights reserved.

1. Discussion

Dr. Sakov cited an earlier work [1], which is similar to [2], but whose conclusion contradicts that in [2]. We note that these conclusions shall be interpreted based on how the filter is implemented. In [1], the authors considered the combination of the ensemble transform Kalman filter (ETKF) with the positive–negative pair (PNP) scheme. In order to keep the size of the analysis ensemble constant at each assimilation cycle, the authors in [1] chose to discard half of the members before propagating them forward (otherwise the ensemble size will double at each cycle). The advantage in doing this is that the computational cost can be very cheap. However, as a potential problem, the sample covariance of the remaining ensemble members might not be a good approximation to the one before halving the size, which may thus deteriorate the performance of the filter.

Dr. Sakov also compared the numerical results of [2] with the existing ones in the literature, for example, [3–5]. In terms of absolute rms error, Dr. Sakov commented that the results in [2] are in the range of 0.73–0.99 [6], while those in [3–5] can be as low as 0.17–0.21. This difference may depend on how the comparison is made. For example, in order to obtain a better performance, in [3] the authors used a bank of ETKFs, while in [4] the ensemble size was up to 91 for the 40-dimensional Lorenz–Emanuel model (L40 after [6]). In contrast, in [2] the number of sigma points is at most 13, since the upper bound $l_u = 6$. If in [4] one chose the ensemble size to be 13 for the L40 system, then it can be seen that the rms errors of the various ETKFs are all above 1 (cf. Fig. 3 of [4]), higher than the range of 0.73–0.99 achieved by the EnUKF in [2].

In [2], we have focused on the scenario where the ensemble size (or the rank of the sample covariance) in the filter is typically (much) lower than the dimension of the system in assimilation. Nevertheless, it should be interesting to compare the performances of the EnUKF and the EnKF (with the ETKF as the representative) beyond the aforementioned scenario, where the ensemble size of a filter could be close to, or even higher than, the dimension of the system in assimilation. To this end, in what follows we would like to conduct one more experiment, supplementary to those in [2].

In the experiment, we examine the performances of the EnUKF and the ETKF under various covariance factors (but without covariance filtering to avoid complicating our discussion). The settings in the experiment are as follows. The truths and observations are generated in the same way as that in [2, Section 4]. For the EnUKF, we let its intrinsic parameters to be $\beta = 2, \lambda = -2,$ the threshold $h_1 = 1000$. We let the lower bound $l_l$ equal to the upper bound $l_u$ (hence the truncation number $I = l_l = l_u$), and take their values in the set $\{5, 10, 15, 20, 30, 40\}$, which implies that the corresponding ensemble size $n = 2I + 1$ takes values in the set $\{11, 21, 31, 41, 61, 81\}$. In addition, we let the inflation factor $\delta$ take values in the set $\{0.05 : 0.05 : 0.5\}, \{1 : 0.5 : 6\}$. 

\textsuperscript{*} Corresponding author at: King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.

E-mail address: xiaodong.luo@kaust.edu.sa (X. Luo).

0167-2789/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
relative rmse of the ETKF is 0.049, which is obtained at $\delta = 0.05$ with $n = 61$, while the lowest relative rmse of the EnUKF is 0.075, achieved at $\delta = 0.05$ with $n = 41$.

The above result reflects the difference between the EnUKF and the ETKF in their philosophies of statistics estimations. As shown in the Appendix of [2], the unscented transform (UT) used in the EnUKF aims to produce the estimations which capture the first few terms in the Taylor expansions of the true mean and covariance, but with an even mismatch higher order terms. In contrast, the Monte Carlo approximation adopted in the ETKF produces the estimations that converge asymptotically to the true statistics. However, given only a small ensemble size, the resulting large sampling errors may lead to large biases and spurious modes. For this reason, the relative superiority between the EnUKF and the ETKF may invert in the small and large sample scenarios.

### 2. Conclusion

Through the supplementary experiment, we showed that, in the small sample scenario, the EnUKF can perform (slightly) better than the ETKF with the same ensemble size, while as the ensemble size increases, the ETKF outperforms the EnUKF instead. Our explanation of this phenomenon is the following: The estimations of the ETKF are based on the Monte Carlo approximations, which converge asymptotically to the true values as the ensemble size increases. However, in the small sample scenario, the Monte Carlo approximations may exhibit large biases and spurious modes due to the relatively large sampling errors, which are partially avoided in the estimations based on the UT. However, one possible problem of the UT is that its estimations may not converge to the true statistics as the ensemble size increases, since its ensemble members are deterministically chosen, rather than randomly drawn as the samples in the Monte Carlo approximations.

### Appendix. Implementation of the covariance filtering technique

A comparison of the results presented in Figs. 1 and 2 (when $n = 11$) in this reply and those in Figs. 2 and 3 of [2] confirms the benefit of covariance filtering. By conducting covariance filtering, one increases the rank of a sample covariance matrix. So in effect, it increases the effective ensemble size. Other benefits (and shortcomings) of covariance filtering are also discussed in, for example, [7,8].

Covariance filtering essentially involves computing the Schur product between a sample covariance and a taper matrix. Mathematically, the Schur product, $C \equiv A \odot B$, of two matrices $A$ and $B$ with the same dimensions, is defined as the matrix with the same dimension as $A$ and $B$, whose components $C_{ij} = A_{ij}B_{ij}$, where $A_{ij}$, $B_{ij}$ and $C_{ij}$ are the components on the $i$th row and the $j$th column of the matrices $A$, $B$ and $C$, respectively. In our discussion, we suppose that $A$ is the covariance matrix, and $B$ is the taper matrix used to reduce the spuriously large correlations in $A$.

The construction of $B$ can be done in the following way:

$$B_{ij} = \rho(d_{ij}),$$

where $d_{ij}$ is a metric measuring the difference between the locations $i$ and $j$, and $\rho$ is a correlation function. Several examples of correlation functions were discussed in [9]. In our experiments, we followed [10] and chose function $\rho$ in the following form:

$$\rho(z) = \begin{cases} 
\frac{1}{4}z^5 + \frac{1}{2}z^3 + \frac{5}{8}z^2 - \frac{5}{3}z + 1, & \text{if } 0 \leq z \leq 1; \\
\frac{1}{12}z^6 + \frac{1}{2}z^4 + \frac{5}{8}z^3 - \frac{5}{3}z^2 - 5z + 4 - \frac{2}{3}z^{-1}, & \text{if } 1 < z \leq 2; \\
0, & \text{if } z > 2.
\end{cases}$$

For illustration, the shape of the function $\rho$ is plotted in Fig. 3. As one can see there, the function $\rho$ has a “cut-off” effect at $z = 2$, in the sense that the values of the function are set to zero for all $z > 2$.

### Table 1

<table>
<thead>
<tr>
<th>Inflation factor</th>
<th>Relative rmse</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>0.7</td>
<td>0.8</td>
</tr>
<tr>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>0.9</td>
<td>1.0</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Inflation factor</th>
<th>Relative rmse</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>0.7</td>
<td>0.8</td>
</tr>
<tr>
<td>0.8</td>
<td>0.9</td>
</tr>
<tr>
<td>0.9</td>
<td>1.0</td>
</tr>
</tbody>
</table>

### Fig. 1

Relative rmse errors of the EnUKF (with different ensemble sizes) as functions of the inflation factors.

### Fig. 2

Relative rmse errors of the ETKF (with different ensemble sizes) as functions of the inflation factors.
For data assimilation in real world, $d_{ij}$ is normally a function of the distance between the locations $i$ and $j$ in the three-dimensional physical world. For example, see [11]. However, in mathematical analysis, this choice might not always be available. For example, the system states of a mathematical model may not have any physical meaning, so that we cannot observe them in the physical world. On the other hand, for a mathematical model, the (physical) distance between locations $i$ and $j$ may also not be well defined as in the physical world.

For the above reasons, in our experiments we chose to calculate $d_{ij}$ based on the covariance matrix $A$ itself. Suppose that $A$ is a matrix with $m$ rows (and the number of its rows is not less than the number of its columns), so that

$$ A = \left[ r_i^T, \ldots, r_m^T \right]^T, $$

where $r_i$ is the $i$th row of $A$. We define

$$ d_{ij} = \| r_i^T - r_j^T \|_2 / l_c, $$

where $l_c$ is a length scale that is introduced to influence where the “cut-off” effect of the function $\rho$ takes place.¹ But note that here, $l_c$ does not have any physical meaning (i.e., it does not correspond to any quantity in the physical world). Instead, it is interpreted as a threshold of the statistical metric $d_{ij}$, which measures the (statistical) difference between the $i$th and $j$th elements of a random vector $x$. In [12, Section 3.3.3.2], one of us showed through a numerical example that covariance filtering conducted in this way can achieve the same effect as those reported in the literature by using the physical distances between different locations to construct the taper matrix.

In [2], covariance filtering was conducted on the background covariance, the cross covariance, and the projection covariance (cf. Eq. (22) of [2]), since all of them are involved in updating the background to the analysis (cf. Eq. (23) of [2]). No covariance filtering was conducted on the quantities other than the above three covariance matrices.

References


¹ Note that we have assumed that the number of the rows of a matrix (not necessarily square) is not less than the number of its columns. If this is not the case, then it is suggested to choose the column vectors to calculate the distances $d_{ij}$ in Eq. (4) instead. In this way, covariance filtering can be applied to non-square matrices like the cross covariance (when the dimension of the state space is not equal to that of the observation space).