Paper entitled: "Applying prior correlations for ensemble-based spatial localization"

Reviewer 1

This manuscript applies a newly developed spatial localization method (YK18) for the ensemble-based data assimilation using Lorenz (1996) model. This correlation cutoff method is developed in their previous work as a variable localization strategy for coupled systems. They claimed that it can be further utilized as a spatial localization method. They performed twin experiments with Lorenz 1996 model, and compared the YK18 method with the conventional spatial localization method. Overall, the work is useful and the manuscript is well written. However, some details of the method should be better explained. And I still have questions about the foundation of the method. The authors should answer my questions and make some revisions before it could be accepted for publication. Please see my questions and comments below.

Reply:

We sincerely thank the reviewer for the questions that help us to improve the clarity of our manuscripts. The reviewer posted comments related to the feasibility and the foundation of our method. A major concern from the reviewer is the extensive use of YK18 from the L96 model to sophisticated models, which is an important topic that deserves further discussion. We would like to note that it is common to use the L96 model as a first step to examine the feasibility of new approaches, such as another amazing localization work done by Anderson (2007). The applications to large models, including the multivariate effects and practical strategies, will be explored in our future studies.

In this revised manuscript, we have incorporated the reviewer's comments and updated our experiments with an offline run that has the same ensemble size as the experiment. The new results proved that the YK18 can be derived from past data or offline runs with limited ensembles, by which the reviewer's main concerns are likely to be resolved. The detailed answers for each question are attached below.

If our answer and interpretation are not clear enough to the reviewer, we are pleased to have a meeting with the reviewer to discuss it in more detail.

1.   In section 2.3, the authors imply that "The prior square error correlations are collected from a preceding offline run." According to section 3.2 (line 185), I realize the offline run is a DA experiment using a relatively large ensemble (50 in this case) without localization for a long period. Of course it works with toy models such as the Lorenz model in this work and that in YK18. However, for more complicated models (such as GCMs), it is meaningless to perform DA

experiment without localization, and it is impossible to use an ensemble large enough to get rid of localization. That would weaken the argument about the usefulness of the method.

Reply:

We appreciate and agree with your comments. In this revised manuscript, we have updated all experiments with YK18 using only 10 ensembles and localization for the offline runs. We have updated our results in Sections 4.2 and 4.3. The results proved that it is possible to use limited ensembles or historical ensemble forecasts for YK18 with a proper tunning of Eq (6). Thus, it is practical to apply YK18 on large models.

2.   Equation (6) implies that the temporal mean of the squared correlation over all analysis steps is computed to serve as "prior error correlation" to estimate the localization function. I have some questions about that:

2.1 Why do you compute the temporal mean over all analysis steps, instead of all steps including forecast and analysis?

Reply:

There are two main reasons why we choose the ensemble background (forecast) at each analysis step for estimating the error correlation:

- **The availability of data**

In practice, the model states (i.e., forecasts and analysis) and observations are mainly available at the regular analysis times (i.e., 00Z, 06Z, 12Z, 18Z) on a DA routine basis. Therefore, it is reasonable to assume that the data we can use are only available at each analysis time.

- **The potential issue of the error correlation in the analysis**

We avoid computing the error correlations from the analysis because the relationship between model variables and the observation in the analysis could differ from the relations in the physical system. Also, the assumptions (i.e., linear regression, localization) used in DA could induce artificial sampling errors in the analysis (Anderson 2007), thereby affecting the error correlation estimates. For example, the use of the traditional localization could introduce an unbalance of the analysis state (i.e., ageostrophic wind) (Kepret 2009; Greybush et al., 2011) if no post-processing (i.e., IAU) proceeds for the analysis. Usually, the imbalance can be eliminated by the model itself after a few integration steps. Therefore, obtaining the prior error correlations from the forecasts at the end of the DA window (= the background at each DA time) and not including the analysis would be an appropriate configuration.

Considering the analyses are from the data assimilation without localization, does that indicate the ensemble size is large enough such that the true correlation can be recovered without localization? This is still impossible for large models.

Yes, we initially applied large ensembles for the offline run without localization to better capture the true correlations. As we mentioned earlier, we have updated all the experiments with offline runs with limited ensembles in the revised manuscript.

2.2 You use the temporal mean of the squared correlation. So, does the period of the offline run have an impact on the correlations?

Once the temporal mean of the squared correlation is converged to the climatology, the result from different periods of offline runs would be the same. The required period of the offline run (i.e., number of samples) depends on the configurations (i.e., ensemble size) and model complexity.

In this revised manuscript, we added a new Figure 2 (b) discussing the required periods of runs for different models and ensemble sizes.
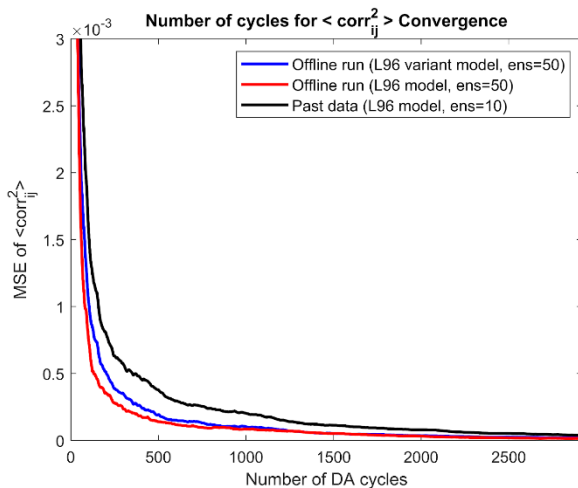
2.3 Is the assimilation process necessary in the offline run?

Yes, the DA process is necessary to constrain the system, preventing the state from jumping to another phase or having totally diverged ensembles, especially for a very chaotic system such as the atmosphere. Besides, the offline run was designed to mimic the DA cycle adopted by operational models, where the analysis cycle is performed every 6 hours.

3. I wonder, is it possible to compute the correlation using an EnOI-like idea? i.e., running a single model and computing the correlation using members at different time steps. This seems much more

practical for real applications. You have already used the temporal mean in the current method anyway.

<u>Reply:</u>

Thank you for this interesting suggestion. After careful research, we found that the feasibility of EnOI-like method depends on the characteristics of the dynamic system. Below we will interpret this argument in detail.

First, assume we have two methods to obtain the prior error correlation:

- **YK18**: correlation derived from ensembles at each DA time, then do the temporal mean over the running period.
- **EnOI-like**: correlation derived from sampling the single model states (ex. ens-mean analysis/forecast) at different time steps along the running period.

Assuming observations are located at the model grid (so, h(x)=x), then the correlation formulations for YK18 and EnOI-like are:

$$Corr_{ij}(YK18) = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\boxed{\sum_{k=1}^{K}\left[x_{k_i}(t) - \overline{x_i(t)}\right]\left[x_{k_j}(t) - \overline{x_j(t)}\right]}}{\sqrt{\sum_{k=1}^{K}\left[x_{k_i}(t) - \overline{x_i(t)}\right]^2}\sqrt{\sum_{k=1}^{K}\left[x_{k_j}(t) - \overline{x_j(t)}\right]^2}}\right)$$

$$Corr_{ij}(EnOI - like) = \frac{\boxed{\sum_{t=1}^{T}[\overline{x_i}(t) - \overline{C_i}][\overline{x_j}(t) - \overline{C_j}]}}{\sqrt{\sum_{t=1}^{T}[\overline{x_i}(t) - \overline{C_i}]^2}\sqrt{\sum_{t=1}^{T}[\overline{x_j}(t) - \overline{C_j}]^2}}$$

Here, we follow the same symbol as Eqs (5) and (6) in our paper, C_i is the temporal mean of the selected samples of $\bar{x}$ (equals to the climate mean of the samples).

One can notice that their covariance parts (i.e., part A) actually reflect different physical meanings: YK18 is the temporal-mean <u>uncertainty</u>, representing the background error in EnKFs, and the EnOI-like is the <u>anomaly to the climatology</u>. The figure below is a schematic of these two differences:
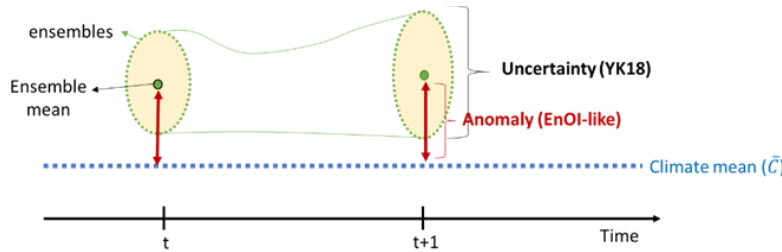

For a dynamic system like the ocean that is overall stable (less chaotic) and its variability is generally proportional to the anomaly, it is possible to use the EnOI-like method to represent the error statistics of the system. However, for a very chaotic system such as the atmosphere, the errors could grow very quickly while its anomaly remains small (like the schematic figure below). For

such a case, <u>using an EnOI-like method may lose the signals of short-term perturbations and thus underestimate the error correlations of the system.</u>
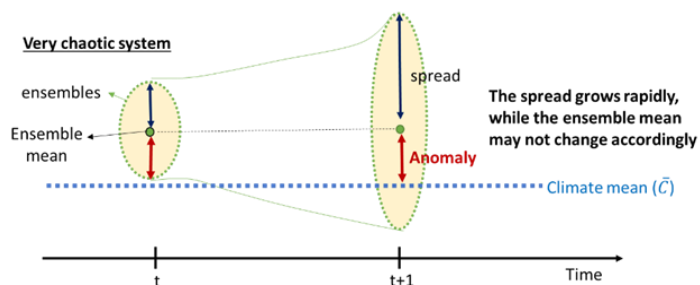
To examine the above hypothesis, we conducted a 2-year experiment on the L96 model with YK18 and EnOI-like methods. The figure below is the prior error correlation estimated by the EnOI-like method (green line) and YK18 with offline runs (red) and past runs (blue). The offline run with a large ensemble (red) can be seen as the "true" correlation. **<u>The result confirmed our expectations that the EnOI-like method may underestimate the error correlations for a very chaotic system such as the L96 model.</u>**
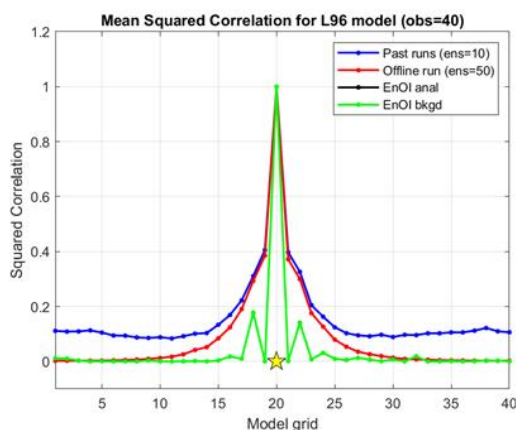
4.   *About Figure 2a. I cannot see any connection between error correlation and observation size from equations (5) and (6). But the connection between error correlation and ensemble size is very clear from equation (5). Do you use this figure to explain the parameters in the offline run?*

<u>Reply:</u>

Yes, the aim of Figure 2 is to show how different parameters (observation and ensemble size) used in the offline run could affect the prior error correlation estimations.

As we stated in [Line 212 - 213], Figure 2(a) confirms that "<u>the prior correlation estimation (Eq (6)) is not very sensitive to the observation size changes, as long as the analysis of the offline run is well constrained</u> ". This conclusion is very important for the YK18 application with new instruments. Assuming we have ensembles from past runs that only assimilated 20 observations. Then, we want to increase the observations to 40 for the new experiment. From Figure (2a), we

know that it is possible to directly obtain the prior error correlation from the 20 obs past runs (because eq(5) and (6) can be computed directly from past ensembles as long as we know the observation location j). This characteristic, in other words, provides evidence for YK18 to use past data to estimate the localization function for newly added observations, which is a significant feature for the applicability of YK18 in modern DA.

We have modified sentences in Section 4.1 to highlight the importance of the result of Figure 2 (a) in this revised manuscript.

5.   The comparison results in figure 5 and figure 6 are not impressive. Though the authors declaim that YK18 can accelerate the spin-up. However, the parameters in GDL and YK18 may be not optimal, so the conclusions are not very persuasive.

Reply:

We appreciate your comments and would like to make further clarifications here.

First, our conclusions are drawn from the optimal configurations for all methods. As we mentioned in the manuscript [Line 190], all the parameters used in GDL and YK18 were optimally tuned for a minimum mean analysis RMSE. So, the parameters we applied in the paper were optimal for GDL and YK18. We conducted experiments with careful assumptions and configurations, so the results we presented are scientifically meaningful and compelling.

In this revised manuscript, we have added Tables listing the parameters used for each experiment. Also, we have improved the clarity of the statement of our configurations in Section 3.2.

Detailed comments

- *Line 175 "GDL: Distance-dependent localization introduced in Section 2.3." I think it is section 2.2.*

Reply:

Thank you for pointing this mistake out. We have corrected it in the revised manuscript.

- *For equations (1) - (3), there are linear observation operator H, for equation (5), it is a potentially nonlinear operator h(x)*

Reply:

Thank you for the comment about the symbol. The capital H in Eqs (1)-(3) represents the matrix with elements of linear operator h. As we specified in [line 115], h(x) is an element-wised interpolation:

"$hj(xk(t))$ is the linear interpolation to the background state $xk(t)$ from the analysis grid to the *jth* observation location".

- *Line 99 "Equation (4) is a smooth and static Gaussian-like function that offers the same localization effect as the GC99 when applied to LETKF." It is inaccurate, because GC99 uses a compact-support function, and it cutoff at some distance, but Eq. (4) does not cutoff.*

<u>Reply:</u>

Thank you for the question. Please note that we already covered the compact-support part in the same paragraph [line 105] in our manuscript:

"When the compact support is presented with the localization function, the observations located beyond a certain distance (in this study is 3.65 times L) from the analysis grid would be discarded by assuming $\rho ij=0$. "

- *R1: Line 91 and line 102, whether the R localization multiplies the elements of R inverse or R itself? Please clarify that.*

<u>Reply:</u>

Thank you for pointing this out. It should be R itself. We have corrected this part for [Line 91] in the revised manuscript.