“Simplified Kalman smoother and ensemble Kalman smoother for improving reanalyses”

By Bo Dong, Ross Bannister, Yumeng Chen, Alison Fowler, and Keith Haines

Referee comment

The manuscript introduces a simplified smoother algorithm which can be used to post-processes the analysis state in the filter-based reanalysis datasets. Based on the authors results, this algorithm could fix the time-discontinue problems in the filter-based reanalysis, and further reduce their errors. The algorithm can be easily applied to existing filter-based datasets, and therefore are of interest to the ocean and earth system reanalysis community. However, there are still some problems and deficiencies that the authors should clarify and fix before the manuscript is accepted and published. I suggest the authors further improve the languages and grammar since I feel like some of the statements cannot be understood easily. I suggest a Minor revision.

General Comments:

1. Introduction:

The readers likely prefer to see how much more computational, storage, memory requirements of KS and EnKF than KF and EnKF.This is a major reason for simplifying the original EnKS in this study. I see that there are already studies using KS (L50-L55), probably in small models. Please clarify how many times the original KS algorithms than the KF roughly in L55, so we can see that it is not possible to use KS or EnKS algorithms in large operational forecasting systems.

2. Methods

1) For the purpose of introducing a new algorithm or method, in equation (1)-(5), the authors should clarify Why they use the coefficients in the forms of $\gamma^0\gamma^1\gamma^2\gamma^3$? May I use other forms of decaying coefficients like in ln forms and Gaussian Forms? And how do I decide the parameter γ?

From my expectations, the errors of nonlinear systems grows exponentially, and their growth rate depends on singular values locally or Lyapunov exponent globally. Therefore, it is more nature to use a ln forms of the decaying coefficients. In this sense, based on the predictability of the nonlinear systems and their variables, we can use equation (5) to roughly decide the γ for different state variables.

2) In L100, please clarify: do I need to run the assimilation system, especially the forward model again? Or the algorithm (1)-(4) just do an weighted average on the analysis fields and their errors without running the forward model again (e.g. give S0-S1 back to the forward model and run the model again to produce the model state between the analysis time t-t+1)?

Specific comments:

L35: please clarify the word “analysis”. I think these studies use data assimilation to produce the initial condition for predictions. But there are systems that produce reanalysis datasets now, for instance, the TOPAZ system by Sakav et al, 2012. Therefore, please clarify that they are used to optimize the initial conditions.

L 35, “For example ……”, incomplete sentence, please consider rewriting the sentence.

L105, What do you mean by saying “which is not given an explicit maximum cutoff”? “given as an explicit maximum cutoff”? or which is similar as an maximum cutoff? Please clarify.

L 110,”in which a tangent linear model is used when the model is nonlinear”, this statement is confusing. I think tangent linear model is only used to propagate the model errors in the Kalman gain matrix, right?
L115, “The nonlinear operators have to be replaced by their tangent linear approximations in the forecast and analysis”.
Do you mean that in $\mathrm{y} - \mathrm{H(x)}$, H should also be replaced by their tangent linear approximation?
The model matrix *M* and its tangent linear form M are not explained explicitly.

L120 the meaning of T in equation (7) is not explained.

L 130, “We note that index ℓ could be defined as future analysis timesteps rather than **model timesteps** if data are only introduced at regular analysis intervals”
This statement is confusing. I believe that the analysis steps must be in model timesteps right? Please clarify that l means analysis timesteps, I think this is enough.

L140, “cross time error covariance”, do you have any cross variable error covariance in KS or ExtKS? I think in you simplified algorithms with time-lag, the parameter γ doesn’t include cross-variable information. Please clarify in the comparison between KS and you simplified algorithm.

L150, “, with the spatial covariances being determined by the KF equations, but the temporal covariances (from times $k + \ell$ to k) being approximated by a simple decay”
How about the cross-variable covariances?

L175 “error standard deviation of 2”
Is that mean “add Gaussian errors with standard deviations of 2”? please clarify.

L180, “Dong et al. (2021) used 3DVar for assimilation into L63 and they used a fixed background error covariance from a climatological L63 run”
This statement makes no sense.

Figure 1,2,4, please consider include the dashed lines in the legends. Also, please clarify that KS/KF means the extended KS/KF in the legends. Please added what the x-axis means? Time units.

L205 “RMSE time series in Fig. 1 where the red line declines sharply where data are available”
Do you mean the extended KF results? I don’t think it is a red line.

Figure 3, please clarify that the x axis means time steps rather than time units!
L 235, “the TLM is not always reliable for a system as non-linear as the L63 model”
Until now, no error models away work for the nonlinear system. But they work at limited time steps or conditions. Therefore, I suggest the authors to be more precise “ the TLM can represent the model error propagation within limited time”, maybe.
L 235 “This improves the quality of the forecast error covariance matrix”.
I don’t agree with this statements. If you have any citations for this statement, please add it. For L63 model, I guess TLM should be more accurate than small ensembles (e.g. ~10s) within short assimilation window (e.g., 0.4 time units).
L 240 “While ensemble filter methods are starting to be adopted for larger environmental models, the cost to store, update and apply posterior ensemble covariances still makes ensemble smoother methods generally infeasible”
What is the logic of this statement? The application of ensemble smoother is not related to ensemble filters. Please rewrite the statement.
L290 “the effective temporal smoothing window timescales are generally short reflecting atmospheric timescales.”
What do you mean “reflecting atmospheric timescales”? I think the predictability of atmosphere system should be ~2 weeks. The reasons for using a 24-hour reanalysis window is mostly related to the usefulness of the adjoint model, since the parameterizations processes the accuracy of the adjoint model significantly. Please added citations here.
L 315 “This shielding of longer lag influences if shorter lag data are available is missing in the simple smoother as presented and could cause the application of the smoother to give poor results when very frequent observations are available.”
Not sure what the authors mean, please rewrite and clarify.

L320, “In particular localisation is often required to remove unrealistic error covariances arising from limited ensemble sizes eg. Petrie and Dance (2010), and when extended to ensemble smoothing that localization may need to vary with lag for the cross time error covariances eg. Desroziers et al. (2016).”
The citation should be (eg. Petrie and Dance, 2010), (eg. Desroziers et al., 2016.), please confirm.
After reading the method, I feel like the parameters work the same way as a decay time localization. I think the more challenging things in ensemble smoother are localization spatio-temporal together. In your simplified algorithm, you use the localized analysis from the filters, and assume time-decay γ. So. I am not sure whether there are some challenges when applying to a larger model. As the authors mentioned, the more flexible and challenging things should be deciding different γ with different variables and potentially different locations and time.

L345 This paragraph is not clear, please consider rewrite it in a clear way.

For instance, please clarify, you use extended and ensemble Kalman smoothers.

What is the meaning of "improved RMSE results"?

What is "from the full smoothing"? is that means smoothers?

"when the truth comparison data is not independent," is that means the assimilated data? The statements can be expressed in a more neaty ways.

technical corrections

L85, should be In Dong et al. (2021), a simple smoother m

L 185 please consider deleting "The reasons for this appear later."

L190 "Across the 100 member ensembles"

L240 "are starting to be adopted for larger environmental model"----have been adopted

L245 (Bishop et al., 2001, ETKF)--- (ETKF, Bishop et al., 2001)

L285 "20 time units of the Ensemble runs"-- ensemble

L295, "However there are still further challenges to applying smoothing in real large systems"---- smoother algorithms

L300 "when the same increment gets repeatedly assimilated by the filter" is this means repeatedly produced by the filters?

L305-L310, "Another option not explored here," ---"because L63 is too simple, Another option not explored here"

L310" eg after deployment of new observing platforms."—"for instance"

"and in doing so" --- and therefore.

L325 "decay away"---decay with time.

L330 "for both operational work and for reanalysis systems"---"for both operational and reanalysis systems"

L 330"it seems sensible" is that means "it makes sense"?

"At the same time" , there should be "," behand it. Please consider remove "therefore". and "also".

L 335 "treat as an exponentially parameters"? "rather than to model it directly"? please consider rewrite it .

"post processed provided the increments (changes) in the error covariances between the"----"post processed, provided that the increments (changes) in the error covariances between the"

"and analysis for each filter assimilation window are also stored"---please consider remove "also".

"error variance". what is error variance? Please clarify

L 340 "Lorentz 1963" ---"Lorenz 1963".

L350 "We also demonstrate the smoothing of the uncertainty estimates in both systems" Please clarify what "both systems" means?

L 370 "the increments (change from filter forecast to analysis)"—"the analysis increments" should be enough.