Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2022-3-RC1, 2022

Comment on gmd-2022-3

Anonymous Referee #1

Referee comment on "Hybrid ensemble-variational data assimilation in ABC-DA within a tropical framework" by Joshua Chun Kwang Lee et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-3-RC1, 2022

Referee Comment
===============

Title: Hybrid ensemble-variational data assimilation in ABC-DA within a tropical framework
Author(s): Joshua Chun Kwang Lee et al.
MS No.: gmd-2022-3
MS type: Development and technical paper

General Comments

The paper provides the description of the implementation of a specific data assimilation algorithm (Hybrid En-Var) in an existing toy-model (ABC-model covering an x-z slice of the atmosphere) and exemplarily shows the performance of the DA-system in a tropical setup of the model.

The paper focusses on the description of the algorithms which are not new but when applied to this specific model setup may have potential to investigate specific issues of the formulation, setup, tuning and characteristics of data assimilation systems. Studies of simplified models can provide considerable insight into the performance of data assimilation algorithms under specific conditions and thus I think it is worth describing the system in a scientific paper.

The description of the implementation is outlined very clearly and detailed and is reproducible. Indeed I think that some passages may be shortened without loss of information. Maybe the authors could try to strengthen the text in this sense.

The performance of the system is outlined by example of a simulation encountering tropical convection. This is certainly a very specific application and in order to allow the reader to grasp the situation it would be nice to give more information on the case and the setup:

The evaluation of the system by means of the above example points at some deficiencies of the setup (for instance the choice of the climatological B_c). This is not a detriment of this paper as the system described here basically should be a tool to gain insight into the possibilities tuning of the setup. Also here some more information should be provided on the specific setup of the B_c and B_e matrices, maybe some more (cross-)correlations implied by them, also by the revised B_c matrix which was mentioned (derived after the spin up), or by contrasting the (time depending) B_e correlations with the actual physical fields at that time. Basically the choices made for B_c and B_e variances would deserve more discussion.
More details will be proposed in the specific comments below.

I think the paper would gain from strengthening the description of the algorithm itself but providing more insight into the details of validation test case.

Scientific significance:

Good, the described data assimilation algorithms algorithms are not new but there implementation in this specific toy-model setup has potential to study details of the data assimilation setup and application to specific situations.

Scientific quality:

Fair, The methods are well described. The description experimental setup would improve with some more details given. The setup of the illustrative test case could gain from further informations and discussion of the tuning parameters involved.

Scientific reproducibility:

Good for the description of the implementation of the algorithms (reproducible).
The testcase cannot be assessed that well.

Presentation quality:

Good

Specific Comments

- 3.2.3 Inter-variable and spatial localisation

The authors state that it is possible to achieve inter-variable localisation within this setup. I think it should be mentioned that the strength of ensemble systems is to provide reasonable time-dependent inter-variable correlations and that therefore one should have good reasons to apply inter-variable localisation in praxis.

The author state that L_horiz has been found to be not positive definite if the length scale is too large (localisation functions exceeds the cycling domain). The fix applied by the authors (setting negative eigenvalues of U^alfa seems problematic to me as the shape of the resulting L at the origin is not smooth). There are better ways to handle this problems:
1) The original article of Gaspari and Cohn shows how positive definite correlation functions can be designed on the sphere. This also works on a cycled domain.
2) another option is to specify U^alfa as a Gaspari Cohn function with half the length scale as L. Then U^a*U^aT will again approximate a Gaussian with the required length scale.

- 3.3 Generation of ABC analysis ensemble

When first introducing the EBV method the authors just mention that "the method .. is uninformed about the observational method". I think this issue should be discussed a little bit further, maybe when the ensemble spread is compared to the rmse error. Only if observation density and observation error are properly accounted for in the ensemble generation process (as done in some other ensemble generation processes mentioned in this section) it can be expected that ensemble spread and rmse matches.

The same discussion applies to the estimation of the climatological matrix B_c. Also here as far as I see only possible balances are taken into account, but not the actual variance which actually depends on the data assimilation setup.

- 3.3.1 Ensemble bread vectors

I thing the normalisation factor E_tot deserves more discussion. This is basically a tuning factor which fixes the ensemble spread. Deriving it from the mean energy norm of the ensemble of differences of independent realisations of the state vector (eq. 7b) would represent the climatological variance, not the uncertainty of the analysis.

- 4.1 implied background error covariances.

I think the covariances shown in Figure 3 would deserve some more discussion. It is mentioned that balances and multi-variate relationships will be explored in a separate study but some more information would be helpful here.
The (time dependent) B_e covariances could be contrasted with physical fields at the respective time. Can the vertically alternating patterns in the B_c correlations be explained ?

- 4.2 Details of observing system simulation experiments

Figure 4 shows that the contribution of J_e to the total cost function is very small, even only 20% in case of only B_e used (and 80% contribution of J_o). Doesn't this indicate some insufficient tuning of the variances? It means that the contribution of the background in this experiment is quite limited.

- 4.3 Sensitivity to weighting of B_c and B_e

Figure 5 shows a variation of the RMSE on a time scale of 8h, figure 6 of the assimilated values itself, can that be explained ?

If the RMSE values of analysis errors are compared to the nominal observational errors the former appear to be very large. It would be illustrative to show the distribution of observations to better understand the performance of the data assimilation procedure. What are the forecast (background) errors.

It is stated that B_c was re-calibrated using other training data after the spin-up process. Wouldn't it be appropriate to show the covariances for this matrix in figures 3, as they are actually used in the assimilation experiment ?

Technical Corrections

- 3.2.3 Inter-variable and spatial localisation

It should be stated how exactly the length scale h for the localisation function is defined, there a several options: There the Gaspari?Cohn function goes to zero, based on the second derivation at the origin (as defined by Daley, ...

- 3.3 Generation of ABC analysis ensemble

The EBV-method is first mentioned in Section 3.3 but the synonym is defined not before section 3.3.1

- Figure 6:

It is hard to see the (gray) ensemble trajectories. The figures could be stretched in the vertical to better resolve this.