Nonlinear Processes in Geophysics Discussions

Nonlin. Processes Geophys. Discuss., referee comment RC1
https://doi.org/10.5194/npg-2021-28-RC1, 2021

Comment on npg-2021-28

Anonymous Referee #1

Referee comment on "Ensemble Riemannian Data Assimilation: Towards Large-Scale Dynamical Systems" by Sagar Kumar Tamang et al., Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2021-28-RC1, 2021

General comments:

The authors extend their previous research from low-dimensional to a relatively higher-dimensional problem. They include Lorenz-96 model and a two-layer quasi-geostrophic model as examples of medium -dimensional problem. The analysis is inferred from (optimized) joint distribution that couples possibly non-Gaussian probability distributions (PD) of the background PD with observation PD, which, according to authors, enables formal reduction of systematic biases. The system is compared to particle filter and stochastic Kalman filter, with results suggesting the mean squared analysis error can be reduced by 20%-30%. The manuscript is well written, and the figures are adequate.

Specific comments:

(1) Title: Including "high-dimensional" in title is presumptuous in my view. There is a long way between the dimensions considered here and in the realistic NWP system. The actual state vector dimensions are 40 for Lorenz-96 and 4,290 for QG model, while a realistic NWP model deals with a state dimension of the order of 10^8. This is at least 5 orders of magnitude larger than QG model. Technically they may be correct by saying "*towards* high-dimensional" but going from Lorenz-63 with state dimension 3 to Lorenz-96 with state dimension 40 is hardly an important step towards realistic state dimension 100,000,000. I suggest using words "moderate" or "intermediate" instead of "high-dimensional, or completely changing the title to reflect better the experiments conducted in the manuscript.

(2) Multivariate aspect: All problems considered in this (and previous) manuscript are univariate. This clearly reduces the complexity of the problem and does not address the possibility of extending this methodology to more realistic situation. Unless the authors want to include multivariate experiments and results, they should clearly state these systems are univariate and elaborate on potential difficulties of applying the system to multivariate problems.

(3) Non-Gaussian probability distribution of errors: Although authors imply throughout the manuscript that the presented methodology is suitable for non-Gaussian errors, they use only Gaussian probability distribution in the Lorenz-96 and QG-model experiments (e.g., p.14, L.324-326; p.5, L.142; p.16, L.354-355, etc). I believe that including experiments

with skewed non-Gaussian probability distributions, such as Lognormal, Gamma, or Beta, would strengthen the authors' arguments and improve the presentation of the new method. Skewed distributions would automatically address the systematic bias and non-Gaussian errors. If the authors would like to postpone such experiments for the future, that is okay, but they need to clearly state this as a limitation of the current experimental setup.

(4) I feel that possible benefits of bias correction in SEnKF have not been sufficiently explored to make a more realistic comparison between the experiments. While the authors acknowledge this, I think implementing some basic bias correction methodology (e.g., moving average etc) in SEnKF would make the strength of EnRDA method clearer.

(5) p.1, L. 11: Although in NWP practice DA is mostly about optimizing the initial conditions, I do not agree that DA "science" is only about initial conditions: it is about estimating the probability Density Function (PDF) or its discrete equivalent (e.g., probability mass function). I am not requesting this, only suggesting.

(6) p.1, L.14-15: Not sure that Is DA formulation is a "penalization of second-order statistics"? It is a penalization of the cost function, which defines weighted (Euclidian) distances, as authors mention. However not sure why/how a penalization of (Euclidian) distances is a "penalization of second-order statistics" as it could be implied from text. Please explain.

(7) p.9, L.220: the model error term omega-t has zero bias (p.8, L.213). Why is there a systematic error in prediction? Please explain.

(8) p.13, L. 318; p.14, L.325: "heteroscedastic" implies that variance (second-order feature) depends on the point. However, systematic bias is a first order feature. Please explain.