Title: Improving Ensemble Data Assimilation through Probit-space Ensemble Size Expansion for Gaussian Copulas (PESE-GC)

Authors: Man-Yau Chan

Recommendation: Minor revision

Summary

This manuscript nicely proposed a method to expand the ensemble size in probit space. Unlike commonly used methods for ensemble size expansion, this PSES-GC ensemble size expansion method can generate virtual members from any multivariate forecast distribution with a Gaussian copula, by using the users' knowledge of the marginal distributions of forecast model variables. This PSES-GC method has been tested with tremendous scenarios, using the L96 model. Results show that PSES-GC can improve the performances of ensemble-based data assimilation methods when the actual ensemble size is small, or the marginal distributions improves the forecast model variable statistics, or the rank histogram filter is used with non-parametric priors. Expanding the ensemble size has advantages to ensemble-based DA in many aspects, while this PSES-GC is an interesting one. The manuscript is well written and pleasant to read. I have several specific comments as below.

1. l28, I'd like to bring up two references here. Sun et al. (2022, 2024) use "analog" ensemble for paleoclimate data assimilation. The paleoclimate data assimilation could be a potential application for the ensemble size expansion, since very limited ensemble members are available.

Sun, H., L. Lei, Z. Liu, L. Ning, and Z.-M. Tan, 2024: A hybrid gain analog offline EnKF for paleoclimate data assimilation. *J. Adv. Model. Earth Syst.,* **16**, e2022MS003414, doi: https://doi.org/10.1029/2022MS003414

Sun, H., L. Lei, Z. Liu, L. Ning, and Z.-M. Tan, 2022: An analog offline EnKF for paleoclimate data assimilation. *J. Adv. Model. Earth Syst.,* **14**, e2021MS002674, doi: https://doi.org/10.1029/2021MS002674

2. l33-34, "However, ensemble modulation assumes that forecast uncertainties have Gaussian statistics." This statement is not right clear to me. Could you explain a little bit why the modulation assumes Gaussian statistics? The modulation just applies localization, not really an ensemble size expansion.

3. l89, it would be nice to have Eq. (6) of Chan et al. (2023) here. Then the reader can immediately see the difference between Eq. (6) of Chan et al. (2023) and the W in this manuscript.

4. Section 2.2 and Figure 3, it is not straightforward to see how the cross-variable relationships are handled? Are the covariances of model state variables naturally conserved by performing PPI for each kind of state variable?

Again at l132, does the adjustment conserve the covariance in model space?

5. l160-165, it is not quite clear the difference between the components A and C. Are they both about p(y|x)?

6. l239, any explanation for the choice of the error variance of 0.25 and 16 for the SQRT and SQUARE observations, given 1.0 for IDEN?

7. l272, the reference for RTPS is missing.

8. Section 4.4, here mechanisms 1-3 are used. It would be better to consistently use components A, B and C as in previous discussions.

9. l323-327, I don't quite understand here. I thought it should be more Gaussian if in probit space.

10. l340-342, I am not convinced here. The RHF can be seen as a non-Gaussian filter.

11. It is nice to virtually increase the ensemble size. But it would be nicer to discuss the computational cost for the DA process, with increased ensemble sizes.