Reviewer report for gmd-2023-170 “GHOSH v1.0.0: a novel Gauss-Hermite High-Order Sampling Hybrid filter for computationally efficient data assimilation in geosciences”

Reply on Reviewer comment RC2

This work proposes an ensemble filter (GHOSH) that conducts sampling in such a way that the resulting sample statistics can match the moments of a target distribution up to a specified order. The moment-matching trick is based on Hermite polynomial approximations to the underlying functions, whereas the target distribution is set to be Gaussian. The derived filter is tested in two examples, both indicating that the GHOSH outperforms an existing filter (SEIK) for the experiments conducted in the current work.

The manuscript is clearly written and reasonably organized in general. Below is a list of minor-to-moderate issues spotted in the current manuscript.

We really thank the Reviewer for the positive general comment and for the suggestions provided below. Hereafter, our point-by-point responses to the Reviewer’s comments are provided in bold green.

Firstly, it is worth noticing that, considering the suggestions of the other two Reviewers and having received a positive answer by the Topic Editor on this, we will prepare a revised version of the manuscript divided in two parts.

Spotted issues

1. Page 1 –

ˆ Line 2: Consider replacing “one of” by “among” or something similar,

since “algorithms” is the subject.

We propose to modify the sentence as follows:

“ensemble algorithms are among the most successful data assimilation approaches”.

ˆ Line 20: What does “a higher order of convergence” mean here?

2. Page 2 –

ˆ Line 42: “Montecarlo” → “Monte Carlo”.

We thank the Reviewer for spotting these typos. The manuscript will be corrected with “a higher order of approximation” and changing “Montecarlo” into “Montecarlo”.

ˆ Line 51 – 52: Rephrase the sentence “the second order approximation is more effective the closer the ensemble members are to each other, thus, the larger the ensemble spread the worse will be the approximation error in the mean computation.”

According to this comment and to comments from Reviewer #1 and #3, this sentence will be reformulated as follows:

“At the same time, most of the models used in geoscience applications are based on systems of differential equations that cannot be represented by a second order polynomial and in all of these cases the second order sampling methods provide a non-exact estimation of the mean, affected by an error proportional to the second order approximation error of the model. Furthermore, the second order approximation is more effective the closer the ensemble members are to each other (i.e., small ensemble spread), thus, the higher the uncertainty the worse will be the approximation error in the mean computation. Since the state estimation is often affected by a relatively high uncertainty in data assimilation geoscience applications, this approximation error may be not negligible.”

3. Line 58 – 59, Page 3: The “2r + 1 ensemble members” requirement does not appear exact. For the unscented transform, one can use either principal component analysis (PCA) or truncated singular value decomposition (TSVD) to reduce the number of ensemble members, see

https://doi.org/10.1175/2008JAS2681.1

https://doi.org/10.1016/j.physd.2008.12.003

It may be worth discussing the similarities and differences between the ideas used to control the number of ensemble members in the aforementioned works and the current manuscript.

In the mentioned works, the authors reduce the number of ensemble members, which is still 2r+1, by reducing the subspace dimension r (by a PCA or a TSVD). We thank the Reviewer for suggesting the 2 relevant citations which will be added to the introduction.

4. Line 109 – 117, Page 5: The discussion on the extension of GHOSH to more generic distributions makes sense. A missing part, however, is that the authors did not explain why they confine themselves to Gaussian distribution in the current work, and what could be the challenges for the generalization of GHOSH to more generic distributions

We thank the Reviewer for this suggestion. We propose to add the following sentence in the Discussion:

“Also the statistical moments μ in equation (1) are key hyper-parameters that describe the error pdf moments for orders higher than 2. In our experiments we used the moments of a Gaussian distribution, but it is worth noticing that the GHOSH algorithm does not enforce this choice and other pdfs could be used. However, Kalman filter’s

analysis equations somehow prescribe a Gaussian approximation, thus the GHOSH filter keeps a link to Gaussianity, even if the GHOSH sampling does not. Interestingly other filters, like Hodyss (2011), try to overcome this limitation and could represent good candidates to study the effects of a non-Gaussian GHOSH sampling."

5. Eq. 38, Page 13: The notation w.r.t the Qi component is somewhat confusing. I guess it should be (Qp)−1, but it looks like Qp−1i.

Equation 38 will be improved by adding parentheses as suggested.

6. In the experiments w.r.t the Lorenz96 model, localization does seem used. What is the reason behind this setting?

Localization is a strategy to reduce the degrees of freedom in order to successfully apply an ensemble filter when the number of ensemble members is much smaller than the dimension of the state vector. We did not use localization in the Lorenz96 model, since the number of variables was already comparable with the tested ensemble size (at least in the case of ensemble sizes 31 and 63, while the smallest ensemble size settings were specifically aimed to study the behaviour of the filters in undersampling conditions). We propose to add the following sentence to the Filter setup section:

"No localization has been applied, since the number of variables N is relatively small and comparable with some of the ensemble size settings."

7. Line 434 – 435, Page 19: Why "it implies that the PCA measures the Pearson correlation"?

The statement was misleading. We meant that, after a normalisation (i.e., dividing by the standard deviation), the ensemble covariance matrix and the correlation matrix are the same. And, since Λ is a diagonal scaling matrix containing the variance of the ensemble, the left hand side of equation (7) becomes equivalent to a renormalization of the basis of the ensemble, i.e.,

L^T Λ^(-1) L = (Λ^(-1/2) L)^T (Λ^(-1/2) L),

where (Λ^(-1/2) L) corresponds to the basis L divided by the standard deviation.

Thus, the PCA is applied to the correlation matrix, and the principal components represent the most relevant correlations.

We propose to correct the statement as follows: "it implies that the PCA is computed on the Pearson correlation matrix".

8. Line 490, Page 21: If I've understood correctly, the "best" label corresponds to the configuration that leads to the best DA performance. If so, then in Figure 4 one should use one block to represent it, and I don't see the point to use a single row for the representation.

Often, in realistic applications, the number of ensemble members and the observation frequency are not flexible parameters (the former is usually limited by the computational resources and the latter by data availability). On the other hand, the forgetting factor can be tuned to improve assimilation results. Thus, we believe that it could be informative for some readers to see the comparison between the two filters tuned with their best forgetting factor in a range of ensemble sizes and observation frequencies. Considering also a comment from Reviewer 3, we propose to change line 490 to better clarify that "best" means "best RMSE", i.e.,

"The first line in each colour-map, labelled "best", represents the best result, in terms of lowest RMSE, obtained among the set of tested forgetting factors."

9. Line 661, Page 33: "an higher" → "a higher".

Thank for spotting the typo, that will be corrected.

References

Hodyss, D., 2011: Ensemble State Estimation for Nonlinear Systems Using Polynomial Expansions in the Innovation. Mon. Wea. Rev., 139, 3571–3588, https://doi.org/10.1175/2011MWR3558.1.