Second review of "A novel Gauss-Hermite High-Order Sampling Hybrid filter for computationally efficient data assimilation in geosciences: PythonDA v1.2.1" by S. Spada et al.

Reply on Reviewer 2 comment

Here, I only review the first part of the now split manuscript. Splitting the submission into two parts helped to avoid overloading and missing focus. It should also give space for some better analysis of the properties of the proposed GHOSH filter. The authors revised the manuscript taking into account many of my comments. Unfortunately, instead of improving the methodological descriptions themselves the authors rather added examples, e.g. in Sec. 2.1 and Appendix C. With this the authors missed the chance to significantly improve the text because essential aspects are not clearly explained, but only illustrated afterwards. The authors also include long explanations in the response to the reviewers, but only shorter changes into the manuscript. Obviously, this approach does not help much to improve the manuscript. Despite these weaknesses in the presentation, the proposed method looks interesting and promising. The main requirement for revision is to improve the explanation of the method and the assessment of the experiments by taking the review comments seriously into account. At the moment, the manuscript seems to be written in the approach of mathematicians giving first the abstract equations and potentially later explanations. In addition, the text seems to using a jargon which is unusual for Earth science of modeling journals (e.g. using 'coordinate' for entries of a matrix). This approach is difficult to follow for typical readers of GMD (see my comment later on Equation 1). Given that the authors of the manuscript have different background and not all come from mathematics, I'm convinced that they would be able to improve the manuscript to a level so that the equations building the basis of the high-order sampling and the new filter and their motivation are easy to understand. To this end, I recommend another revision to enhance the explanations to make the manuscript accessible for typical readers of GMD.

We thank the reviewer for the general comment and for all the specific comments that we carefully addressed in the following. Hereafter, our point-by-point responses to the Reviewer's comments are provided in italic. Lines and equation numbers refers to the last reviewed version of the manuscript.

As in my first review, I like to comment on the suitability of the manuscript for GMD:
The manuscript is an algorithmic work introducing a new filter variant for data assimilation. The implementation of the filter with a toy model is only used to assess its behavior (which is the usual practice when developing new assimilation methods). As such, the manuscript is not 'about the model/software' but 'about the algorithm'. To this end, I am surprised about the Editor's statement 'Your manuscript presents a new algorithm implementation' in the 'MS records' on November 11, 2023. This seems to be a misunderstanding, because there is no 'new algorithm implementation' presented. In fact, the implementation or software is never described in the manuscript.
The authors responded to my comment on the suitability of the manuscript for GMD mentioning that the description of manuscript types (https://www.geoscientific-model-development.net/about/manuscript_types.html) mentions data assimilation for 'development and technical papers'. However, the actual 'Aim and scope' of GMD is not defined on that web page, but on

https://www.geoscientific-model-development.net/about/aims_and_scope.html. The description of manuscript types has obviously to be understood within the frame defined by this defined 'Aim and Scope' of GMD. In fact, the data assimilation papers in GMD that I found in a search for 'assimilation' use the type 'development and technical paper'.
While there have been several papers been published in GMD which are concerned with data assimilation all of them focus on software aspects, but I'm now aware about any algorithmic paper on DA published in GMD. This, this manuscript would be the first paper focusing on the introduction of the new algorithm. This would, at least, be unusual.
Related to the fact that the manuscript is not about the software, the title is misleading. Including 'PythonDA v1.2.1' implies that the manuscript is about this software, which is not the case. Thus, as a reviewer, I would need to recommend 'reject' with this title, because the manuscript fully misses to introduce the software. This consideration is also relevant because the Gitlab repository claims "This is a Python Data Assimilation tool. It contains classes for DA experiments, twin experiments, DA filters and metrics." Thus, the software has the ambition to be a general data assimilation software and not just an implementation of the new data assimilation algorithm. The current title claims that this would be published, which is invalid without proper description of the software.
Overall, since according to the Editor's assessment the manuscript fits into the scope of GMD also after the first review round, I recommend to classify this manuscript as 'development and technical paper' and to remove 'Python DA v1.2.1' from the title. For me this is essential to avoid the misleading impression that the paper introduces a new software or model. If the manuscript cannot be published without the 'name of the software' and a version number, this again shows to me that the scope of GMD does not fit for this manuscript.

1. We thank the reviewer for helping to find the right frame for this manuscript. Even if the manuscript type falls in a gray area, the editor agreed that 'development and technical paper' could be the best fitting choice, changing the title in "A novel Gauss-Hermite High-Order Sampling Hybrid ensemble filter for computationally efficient data assimilation in geosciences – Part 1: application to Lorenz-96 in PythonDA (v1.2.1)".

In the revision the authors showed little attempts to improve the descriptions of the introduced equations im Section 2 and the Appendix, but rather added examples that illustrate the meaning of equations. For example, equation (1) and its description is unchanged and can still hardly be understood, but the authors now added the example of some moments that can be computed. (This is not what I meant by my initial review comment on adding lines - it was recommending to add lines to Eq. 1 to make the big set of indices better understandable, while still being a general system of equations). Thus, the authors demonstrate little skill in improving their explanations and still give readers a hard time when they try to understand the text. The text is written like one already knows the resulting method, but is not really suited to introduce it. Actually, with equation (1) the authors introduce a set of equations with many indices and symbols. However, u and v, \mu, but also the subscripts j_1, j_\xi, and \xi are all unknown at this point. There are some 'real numbers' (line 112) indiced by different subscripts in some system of nonlinear equations. It is also mentioned that the "system represents the matching between the statistical moments of the ensemble and the normalized uncertainty pdf." (line 116), without explaining where the statistical moments are. Only later in line 125 the reader learns that \mu are statistical

moments. Before this, the authors added the many lines with examples. Here, the right hand side show fixed numbers, but the authors missed to explain why a zero mean of the ensemble or an identity matrix for the ensemble covariance matrix is expected. In fact, the authors miss to provide the information that the system is used to determine ensemble perturbations, so there the relation to an ensemble is just missing. Only toward the end of the sub-section, the readers learn in line 137 that the v's are used to define the ensemble weights and the the u's will define some matrix \Omega_h, but the meaning of this matrix is still unclear at the end of Section 2.1 and it is only clarified when equation (6) is finally introduced in Sec. 2.2. I see many ways to explain the motivation and purpose of Eq. (1) in a much clearer way and recommend that the authors rethink and revise their descriptions. It would be good to first explain what they want to achieve and write up in form of equations (it is simple: one wants to use a system of equations computing the statistical moments to determine ensemble perturbations and related ensemble weights. To be able to solve the system one has to define the moments that should be fulfill, e.g. a Gaussian with mean zero and the identity matrix of covariance matrix). Then one would need to introduce the used symbols and indices. The authors should also clarify the relation between 'approximation order' and 'moments' (see also my later comment on lines 32/33), which both are mentioned around equation (1). If this is clearly explained, the aim is not that difficult any more (I think, I mainly understood it from reading Appendix A, but also from having read the manuscript twice (the original and the revised one). In any case, since an Appendix should not be required to be read, it's important to improve the main text.) The explanation should also clarify why this particular system of equations is used. I commented on the initial manuscript on line 130 that it needs to be explained why the generated distribution 'must be uncorrelated, normalized and have 0 mean'. In the response the authors stated that they now explain this in line 109-138. However, lines 115/116 still plainly state "The system represents the matching between the statistical moments of the ensemble and the normalized uncertainty pdf" without giving any reason why this is the case. Here, the authors again miss the chance to first explain what they want to achieve before they show how they achieve it.

2. We thank the reviewer for the very useful comments that helped us to improve the clarity of the presentation. We rethought Section 2.1 entirely and implemented many changes to text in order to explain ideas and symbols before equations. The changes includes (but are not limited to): (i) improved motivation for the moment-matching system in equation (2), (ii) explaining that the uncertainty pdf must be normalized and uncorrelated because the algorithm takes mean and covariance as input and rescales all the moments accordingly, (iii) description of the role of omega matrix before explaining how to build it, (iv) clarification of the relation between order and moments.

The approach to add examples to an otherwise difficult to understand explanation is then brought to an extreme in Appendix C. Here the authors added 10 pages of examples for the sample values and means for different cases. This is enormous and completely violates the rule that a scientific paper has to be 'as short as possible'. Obviously, one could try to make Appendix C at least visually shorter by concatenating more information in single lines (e.g. lines 871-881 include a lot of redundant information and could be compressed into 3 lines (for \mu_i, \mu_ij, \mu_ii), but nonetheless the authors should try hard to make this shorter.

3. Appendix C is motivated by the fact that also other reviewers pointed out the confusion that can arise between improving higher order moments and using them to improve the mean (which is the main objective of the method described in the manuscript). Thus, Appendix C has been introduced to help clarify this non-obvious approach, and it has a pedantic and didactical structure, starting from dimension 1, then passing to 2, 3, eventually showing a simple example in higher dimensions. We feel it is not strictly necessary for every reader, but probably very helpful for a number of readers, keeping in mind that, being an Appendix, it can be completely skipped. For these reasons, we would prefer to keep the present Appendix C structure.

Other changes that are aimed to clarify comments don not really seem to help. For example in lines 32/33 it is now explained "with the term 'order' we refer to the polynomial order of approximation of the filter or of its sampling method." Here it is still unclear what a 'polynomial order of approximation of the filter' is. Usually, the different ensemble Kalman filter variants assume a random sample and use this in their analysis step. There is no polynomial computed or approximated. As such it is unclear where this 'polynomial order' arises from. This likewise holds for the model, which is mentioned as 'polynomial of order h' (lines 34/35). The model is usually assumed to be an arbitrary model, and not approximated by a polynomial. As such, a polynomial order of a model is not considered. Here, the background of the authors might be different than that of the likely readers of GMD. To this end, these aspects need to be better explained.

4. We thank the reviewer for pointing out this possible source of misunderstanding. Our method does not require that the model is explicitly approximated with a polynomial, instead it relates the error that one would make approximating the model with a polynomial (known as “best polynomial approximation error”) with the error made in the ensemble forecast. We changed the text between lines 31-44 in order to: (i) clarify the definition of order of approximation, (ii) rephrase the sentence about polynomial approximation of the model so that it is clear that no explicit computation of the polynomial is required or expected.

All reviewers showed irritations why a Gaussian needs higher moments given it is fully described by the first two moments. This aspect was actually shortly explained now in one of the responses to the reviewers, but the authors missed to explain this in the manuscript. I recommend to clarify that a Gaussian is fully defined by it is first two moments, but that the further moments are relevant, too. I understand, that a second-order exact sampling could lead to a skewed ensemble. Thus, while the mean and covariance would be correct, the skewness would not be zero. Here, a higher order sampling helps and this seems to be what the proposed sampling method achieves. This also links to the example in Figure 1, which also lead to irritations: The ensemble is symmetric in 2D and 1D, but a valid sampling in 3D does not need to have this property. It would really help the clarity of the manuscript would be clearly explained to avoid such irritations (and I have the impression that this is not yet sufficient in the revised manuscript).

5. We included in the discussion the explanation about the Gaussian case. Furthermore, we completely reworked the text referred to Fig. 1 to improve clarity and we added the motivation why the high-order sampling is beneficial for Gaussian uncertainties.

The authors explain in the introduction that if the model is a polynomial of order 2, the mean would be correctly calculated when a second-order sampling is used. Now, the Lorenz-96 model equations are second-order polynomials, but still the higher-order sampling in the new filter is better than second-order sampling and filtering. This looks like a contradiction. Please clarify this case.

6. The model is a function m_t that takes the state of the system x_t at time t and gives the state of the system x_{t+1} at time t+1, i.e., m_t(x_t)=x_{t+1}.
In the Lorenz case, this is realized by setting initial conditions x(t)=x_t, and solving a first order differential equation involving a second order polynomial. Then m_t(x_t) is defined as x(t+1), being x the solution of the differential equation. Thus, in this case, m_t represents the relation between the initial condition at t and the solution evaluated at t+1, which is not a second order polynomial.

Unfortunately, the authors mostly defend their choices for the experiments with the Lorenz-96 model. To this end they only extended the range of tested forgetting factors and added a small set of longer experiments. For me this did not resolve my comments on the initial submission.
In addition to my previous comments, I like to point to some other issue which became clearer also from performing more tests with the provide software:

7. Thanks to the previous review comments, the set of experiments has been extended to include longer experiments and additional forgetting factors. The manuscript has been enriched by this addition, however the inclusion of further tests is out of the scope of the present manuscripts, which already discusses the main results of the novel GHOSH algorithm.

- The main difference between the SEIK and GHOSH filters seems to be their stability. For the observation frequency 0.15 the SEIK filter shows in longer runs (I used 200 time steps) large variations of the quality of the estimate. Here, the filter seems to either converge to RMSEs very similar to those of the GHOSH filter, or it diverges. Actually, this behavior cannot really be quantified by averaging RMSEs. Please explain this behavior more clearly.

8. We agree completely with the reviewer. When an assimilation experiment is successful, the filter stays more or less near to the truth. When an assimilation experiment fails, it usually deviates significantly.
It is worth noting that each RMSE shown in the manuscript is computed over all the tests, and not averaging the RMSEs of the single tests. Computing the aggregated RMSE in this way helps to keep the original meaning of the RMSE score (i.e., quantifying how often and how far the trajectory deviates from the truth) even if many experiments are summarized with a single number. As such, we think that our aggregated RMSE is a good metric to evaluate the reported behavior.

Thanks to the reviewer's comment, we clarified in a footnote the aggregation formula. Furthermore, at line 412, we motivated the use of RMSE: "Representing both how often and how far the filters deviate from the truth, RMSE is a common proxy for a filter skill".

- As I commented on before, the experiments with 20 time units are too short. Now the authors argue with physical arguments, e.g. from an ocean application. However, it is not at all clear whether an ocean or atmosphere application would need 100 days for convergence, and whether the new filter would be significantly faster in these applications because we don't know in which regime a real assimilation application is compared to the highly idealized Lorenz-96 system. However, having done a small number of longer experiments, I have the impression that apart from better convergence of the SEIK, one would likely obtain results that are not very much different. To this end, I recommend to clarify the motivation for the short experiments at the beginning of Section 5.1. Currently, the argumentation about the length is in the discussion section, which is too late.

9. We agree with the reviewer about the importance of stating the motivation for the different experiment lengths in the experimental design section. At line 453 we included the following sentences:
"The relatively short time series in the first experiment phase are motivated by the aim to focus on comparing DA convergence after the assimilation and not on long-term convergence. On the other hand, the 150-time-units have been carried out to verify the robustness of the filter on longer time scales".

- The experiments use global filters and observations at each second grid point. The results show that the RMSEs of observed and unobserved grid points are nearly identical. I would expect this because the decorrelation length scale of the Lorenz-96 model is larger than zero, so that neighboring grid points are correlated. To this end, the plots showing the RMSE for non-observed grid points do not seem to provide essential information and could be safely removed.

10. We agree that evaluating uncorrelated variables is important for checking possible degradation on non-assimilated variables. However, variables that are non-assimilated but correlated are, in our opinion, interesting as well to test data assimilation skills. Indeed, the evaluation of the novel method effects on non-assimilated variables shows the capability of transferring information gathered with observations to the whole state of the system, exploiting correlations. If the model is linear, transferring information is an easy task, but in the nonlinear case, it is much harder for the filter to integrate the information from one variable to another. For that reason, we believe that showing SEIK and GHOSH skills also for non-assimilated variables can be informative for the reader. At line 414, we included the sentences "Separately assessing filter skills on non-assimilated variables provides indications on the capability of transferring information gathered with observations to the whole state of the system, exploiting correlations" in the "Experimental setup" section.

- The statement on maximum differences between the RMSE of the SEIK and GHOSH filters refers to cases where both filters use the same inflation but none is optimally tuned (e.g. the

statement on 'up to 70% reduction' in line 14 of the abstract). I don't see how this provides a reasonable statement. Of course one tunes the inflation and since we know that both filters need a different inflation, we immediately know that using the same inflation for both is not a good choice. To this end, I recommend to focus the comparison on the optimal cases in the first row of the figures. The ~50% error reduction by the GHOSH with optimally tuned filters is impressive enough. The dependence of the RMSE on the varying forgetting factor would then mainly serve to show that the filters differ in this respect.

11. We agree with the reviewer that focusing on the best tuned forgetting factor is the most reasonable choice and we changed the abstract accordingly.

- I tested the GHOSH without the eigenvalue decomposition and related projection (mainly for timing reasons, see below). Actually, in this case the GHOSH becomes similarly unstable as a SEIK filter. This indicates that the projection/transformation with the eigenvectors is essential. It would be good if the authors could clearly explain the purpose of this projection. Thus, can you explain what is really achieved with this projection? (This is likely an addition to section 2.2)

12. The eigenvalue decomposition is a crucial part of the algorithm, since it identifies the principal components of the uncertainty subspace, which is a key point for correctly projecting weights and members, aligning the direction with higher order with the directions of bigger uncertainty. We agree with the reviewer that this point should be made more evident in Section 2.2. In the new version of the manuscript, the motivation for the eigenvalue decomposition is now presented at the beginning of the sampling section (2.2) and then recalled after the relevant equations.

- The code allows to run the GHOSH with different orders, but the authors only show results with order 5. With this they miss the chance to answer the question which order is actually required.
For this, it would be very useful to add experiments with orders 3 and 4 (I'm wondering whether GHOSH with order 2 simply reduces to the SEIK filter). Actually, I did some experiments with order 3 (after correcting a small bug in the code) and in this case the GHOSH seems to be similarly unstable as the SEIK filter (thus constraining the skewness is not sufficient). Unfortunately, order 4 is not implemented, so that one cannot test whether order 5 is necessary or whether order 4, thus constraining the kurtosis, would be sufficient. I recommend to add such experiments (given the splitting of the manuscript, there should be space for this).

13. We thank the reviewer for pointing out this interesting observation. For informational purposes only, we made some preliminary tests in this direction before starting to work on this manuscript, and the outcome was that the improvement was already present with order 3, but less impressive than with order 5. However, we think that such analysis would be more related to understanding the Lorenz96 model-specific nonlinearities than to explain something about the GHOSH filter itself. Thus, it will possibly be included in a future work, leaving this one focused on the presentation of the high-order sampling and GHOSH filter.

Detailed comments:

Abstract:
While the abstract was improved, there are still some issues:
- Line 1: please rephrase 'model knowledge' (a model does not 'know' something)
- Lines 4-5: Please rephrase 'Among other options to improve the capability of generating an effective ensemble ... represent a novelty. Indeed,'. This misses the relevant information that the manuscript does introduce "a sampling method featuring a higher polynomial order of approximation" and does not have the conciseness we would expect in an abstract.
- Line 7: remove 'version 1.0.0'
- Lines 11-12: The sentence ' The comparison between the GHOSH and the SEIK filter has been done by varying a number of data assimilation settings.' doesn't contain relevant information and can be removed.
- The statement 'up to a 70% reduction' is misleading as I commented before. Since the filters are tuned in any realistic case and also in the Lorenz-96 model, it is only valid to compare the optimally tuned filters.

14. We thank the reviewer for the suggested changes that improve the readability and clarity of the abstract. We implemented all the suggestions.

Lines 44-46: Please clarify the sentence "A potential strategy to reduce this error is the use of a higher order of approximation, but this would require a larger ensemble with respect to the second-order methods and consequently larger computational costs, given that a higher order of approximation implies a larger number of ensemble members". As the authors show a larger ensemble is not required when weights are introduced, which contradicts the sentence. The statement is probably meant to hold only for the case if all ensemble members have the same weight.

15. We thank the reviewer for pointing out this possible source of misunderstanding. Ensembles with order higher than 2 require more members than second-order ensembles, also when weights are used. Weights, however, help increase the order and mitigates the ensemble size. Our proposed method only achieves a higher order in specific subspaces, without modifying the overall number of ensemble members.
Lines 45-60 have been rephrased in order to clarify the text. Changed sentences include: "In the present work, we propose a novel weighted ensemble method based on a new high-order sampling, that identifies subspaces of larger uncertainties and provides ensemble mean estimates of order higher than 2 in those specific subspaces, without increasing the overall number of ensemble members, and therefore without major impacts on the computational cost."

Lines 87: Please avoid 'can be proven'. There is no need to appear fancy, since there is no formal proof necessary. One could simply write that this approximation has the correct mean, variance and it is symmetric and hence has the correct zero skewness.

16. We removed “proven” in a number of occurrences. Moreover, following reviewer’s suggestions also included in a previous comment (answer number 5) and in the following ones, we reworked all the text around Fig.1, in order to improve clarity and consistency with the figure.

Figure 1: The rotation of the tetrahedron looks inconsistent with the projections, because the projection on the xy-plane should be a triangle. This seems to have lead to irritations also for other reviewers. I recommend to plot this with the correct rotation.

17. See answer 16 to a previous comment.

Lines 88-90: It is stated "In summary, Fig. 1 represents a weighted ensemble with approximation order 2 in the xyz 3-dimensional space, approximation order 3 in the xy plane and 5 along the x axis". This statement is not true. Shown is a weighted ensemble of 3 members along the x axis, but in 3D and 2D there are four ensemble members with equal weights.

18. See answer 16 to a previous comment.

Lines 90/91: Here the text refers to the 'GHOSH third-order approximation". However, the GHOSH sampling is only introduced later. At this point it would be meaningful to explain the aspects without referring to GHOSH. The relation to the spread "if the z component of the distribution was less relevant than the x and y components" is irritating because this is written in direct relation to figure 1, but there a standard normal distribution is discussed.

19. See answer 16 to a previous comment.

Line 116: Here "... the normalized uncertainty pdf" is mentioned. However, this pdf was never introduced.

20. The text around the moment-matching system in equation (2) has been reworked to improve clarity and consistency (see also answer 2).

Line 136/137:
Here is another example where an insufficient explanation is augmented by an example: "The system solution is used to define \Omega_h, as the (r+1) \times s matrix with coordinates u_m,j (i.e., u_m,j is the entry of the matrix \Omega_h at row and column j),...". If 'coordinates u_m,j' is unclear as indicated by my review comment on the original manuscript, why is this kept? The issue might be the term 'coordinates', which usually define a location in space, while a matrix is not thought to be related to locations in a space. With the experience of having read many papers in Earth science and modeling journals, I think a usual jargon would be '...with entries u_m,j where m denotes the row and k the column'. This concerns

likewise other places where 'coordinates' is used (e.g. "with the coordinates of w as weights" in line 200; weights are obviously not coordinates)

21. We thank the reviewer for the help in identifying the correct jargon for the geoscience community. We removed "coordinates" from the manuscript when not referring to a spatial coordinate.

Line 140:
Where does the covariance matrix L_i L_i^t come from and why is it factorized in this form? (The motivation is obviously from it's use in data assimilation, but the authors miss to explain the motivation and just imply it.)

22. Added a clarificatory sentence to introduce L_i (line 180): "Since an explicit covariance matrix might be intractable for large N, the N x r matrix L_i is adopted as r-rank factorization of the covariance matrix and its columns represent the base of the uncertainty subspace spanned by the ensemble members".

Line 146/ Equation 4:
Why can than random transformation be used here? At the beginning of Sec. 2 it is explained that only particular rotations of the tetrahedron result in the proper projection. This seems to contradict that there is a random rotation applied of \Omega_h.

23. This is an important point that needs to be better clarified in the text, and we thank the reviewer for pointing out. Only particular rotations of the ensemble (i.e., the tetrahedron in sec. 2) result in the proper projection, but possibly still some degrees of freedom remain (i.e., there could be more than one proper rotation). To avoid introducing biases, random rotation (or symmetries) are needed. In the manuscript, an introductive sentence has been added before equation (4): "In order to avoid inducing a bias in the available unexploited degrees of freedom of the sampling, the matrix Omega_i is built from Omega^h by a procedure that includes random symmetries and rotations. These random transformations explore only degrees of freedom that do not compromise the high-order approximation on the principal components of the uncertainty subspace."

Line 178, Eq. 12: Please clarify that the full P^a never needs to be computed, but is only shown in this form for convenience. This also holds for other equations (e.g. 19, 22, 31, 33).

24. After eq. 12, the following sentence has been added: "As in other ensemble-based filter schemes, the uncertainty covariance never needs to be explicitly computed. However, equation (12) and the like (e.g.,19, 22, 31, 33) are still shown in this form for convenience".

Line 223: "uncertainty covariance matrix P". So far all publications on ensemble Kalman filters that I've read agree on calling this matrix 'error covariance matrix'. It is unclear why the authors think that this is not adequate and that it is better to use this non-standard term. This

is again a point where the authors make it unnecessarily difficult for the readers to follow the text.

25. We agree with the reviewer that P is usually called “error covariance matrix”, and changing this convention can be difficult for the reader. However, we introduced “uncertainty” instead of “error” after the first round of revisions, since some reviewers pointed out the difficulty in understanding what was the meaning of “error” in the GHOSH context. Indeed, in the manuscript we take into account the error in the forecast mean, and many arguments involve the word error, making it difficult to understand which error we are referring to, if the error due to lack of knowledge (uncertainty) or the error due to filter design and to model nonlinearities. Furthermore, GHOSH aims to reduce the error where uncertainty is bigger, leading to ambiguities very hard to be removed without changing this name convention. The same holds for the “error subspace”, which in this context leads to similar ambiguities.
It is also worth noting that the word “uncertainty” is already partially substituting the word “error” in the field, indeed is pretty common now speaking about “uncertainty estimation”, instead of “error estimation”, following the newer probabilistic paradigm.
For these reasons, we think that the present wording (that is explained in a footnote at page 2 of the manuscript) is helpful for the reader, even if it requires changing convention.

Lines 287: On the GHOSH sampling it is written "... it takes into account the Q_i effects in equations (31) and (32)". Please explain what is meant. Equations 31 and 32 already take Q_i into account. What does the sampling achieve in addition?

26. We are grateful to the reviewer for pointing out the lack of clarity in this passage. We therefore rephrased the sentence in line 335 adding clearer explanations:
“[...] it takes into account the effects of Q_i and rho in equations (31) and (32) that, augmenting the uncertainty after the forecast, modify the second-order moments of the uncertainty pdf. Without this resampling, the forecast ensemble is no longer a high-order ensemble nor a second-order sampling, since the ensemble does not match the changed covariance anymore. However, if Q_i is supposed to be not significant (i.e., Q_i=0) then the covariance is only modified by the forgetting factor rho and the forecast high-order sampling phase can be widely simplified: in that specific case, indeed, the ensemble members can be safely obtained by inflating the ensemble anomalies, i.e., [...]”

Line 340, equation 41: I commented in my first review on the 'sound choice' of the localization function. Actually, one wonders why the authors introduce here a new localization function. It might be the simplest one, but there is simply to reason to introduce it. This just adds to the complexity of the manuscript without any further purpose. In this regard it is also irritating to read the statement 'other options are viable (... Gaspari and Cohn, 1999)'. It is good to read that the authors think that the most widely used localization function is 'viable', but one wonders why the authors feel qualified to give such statements given that they have not been involved in the development of localization. To avoid such irritations, I recommend to omit Eq. 41 and simply state that the function by Gaspari and Cohn is a common choice.

27. Equation 41 has been removed and now the text at line 393 only mentions Gaspari and Cohn function.

Line 438: "...GHOSH filter produce a successful assimilation...". Please define 'successful'. For me it would mean that the filter converges, but this would require to compare the estimated and true errors.

28. We clarified in the text (line 491) the meaning of "successful assimilation": "i.e., achieving a RMSE smaller than the climatological standard deviation".

Line 446: "GHOSH proves to be always at least as good as SEIK". Please avoid 'proves' here. Obviously, numerical experiments are not suited to 'prove' anything.

29. We substituted "proves" by "shows" as suggested by the reviewer.

Lines 447-449: Is this list of cases then the 'largest improvements' of GHOSH occur mean to be 'and' or 'or'? Thus, does it mean that all conditions have to be fulfilled or any of them?

30. We thank the reviewer for helping to improve the clarity of the manuscript. We meant "and", which has been specified in the list at line 502.

Section 5.3
The discussion shows that 'numerical complexity' as discussed in Sec. 3.1.6 can be misleading because constant factors are not included (if one filter would have a cost of '10N^3' while the other has '1N^3', they would still have the same complexity). The major additional cost of GHOSH seems to be the eigenvalue decomposition in the sampling and the previous matrix-matrix computation. Actually, in my experiments with the provided Python code, I saw that the eigenvalue decomposition lead to a time increase of the GHOSH filter by 50% for ensemble size 31. This cost should also be significant in case of real high-dimensional model, in particular for the matrix-matrix computation part. These aspects should be clarified.
Further, it does not look meaningful to provide a time information over all experiments "SEIK and GHOSH schemes used 12% and 16% of the total time respectively, while the rest was committed to model integration." The run times can depending on the configuration vary significantly. Actually, doing experiments using an observation time interval 0.15 and ensemble size 31, the GHOSH filter had a similar runtime as the model if I removed the eigenvalue decomposition, and with eigenvalue decomposition, the time for GHOSH was higher than the runtime of the model. The runtime of SEIK was similar to that of GHOSH without eigenvalue decomposition, but the model runtime was higher. I have done these tests using an Apply MacBook with M3-processor and a Linux workstation with Intel i5 CPU, both using a Python installation using miniconda and they are consistent over both processor types. For ensemble size 63, the runtime of the filter was clearly higher than that of the model and only for the ensemble sizes 7 and 15 for which the filters diverge the filters show lower run time than the model. For the converged filters, the time fractions required for

the filters were much higher than what is reported by the authors. The time variations with ensemble size should clearly be taken into account when reporting run times.

31. We thank the reviewer for pointing out topics that need clarification with respect to the numerical and computational aspects. The following points have been improved:
(i) Section "3.1.6 Asymptotic computational complexity" has been introduced by a sentence informing the reader that the aim of the section is assessing the scaling capability of the GHOSH filter. Since the scaling is independent of the leading coefficient, it has not been considered.
(ii) In Section "5.3 Computational cost", misleading sentences have been removed and a new paragraph has been added in order to clarify that the provided percentages are averages, are specific to our experiment, and greatly vary from case to case. Furthermore, the eigenvalue decomposition has been referred to as the reason for the GHOSH being more expensive than SEIK.
(iii) In the discussion section, when speaking about the computational complexity, the following paragraph has been added:
"In our experiments (Section 5.3), GHOSH showed an increased computational time with respect to SEIK (on average, the 4% of the total computational time), mainly due to the GHOSH filter's eigenvalue decomposition. However, it is worth noting that the cost related to the eigenvalue decomposition only depends on the number of ensemble members, affecting the O(r^3) part of the full GHOSH complexity O(r^3 + (N+n)r^2). As such, the eigenvalue decomposition is an important term if N+n (i.e., the sum of the dimensions of model and observations) is small (in our experiments on the Lorenz96 model it is smaller than 100), but becomes order of magnitude less relevant in a realistic geoscience application, where degrees of freedom can be of the order of the millions and above. Moreover, while the model integration time and the filter computational time are still comparable in our experiments (12% and 16% for SEIK and GHOSH respectively), in a realistic application this will not be the case. Indeed, the filters computational complexity shows that filter computational time grows linearly with model dimension and the same holds for the Lorenz96 model that, however, is orders of magnitude less complex and computational expensive than a realistic geoscience model. This is consistent with results obtained in a companion paper (Spada at al., 2024), where the GHOSH filter computational time in a realistic application accounts for less than 1% of the total computational cost."

Lines 509/510: "GHOSH filter showed a lower need for inflation with respect to SEIK, removing one of the instability causes observed in the twin experiment". It doesn't look evident from the experiments that the inflation increased instability. Please clarify this aspect.

32. The sentence has been better explained by adding (line 567): "Indeed, in the twin experiment, a larger number of instability cases were observed at lower inflation values both for SEIK and GHOSH (one example with diverging SEIK is shown in Fig. 5)."

Line 548: "with the ESTKF approach ... the T matrix should change at every forecast step". I don't see why the T-matrix should change. The T-matrix used in the ESTKF is very similar to that used in the manuscript. It just that the matrix does not only subtract the ensemble mean

and drop the last column, but it is distributing the information from the last column over all other columns. This should not require any change over time.

33. To preserve ESTKF consistency, it is important to take into account GHOSH projections in the principal components of the uncertainty subspace, which are changing at every forecast. Thanks to the reviewer's comment, this clarification is now part of the text at line 606: "differently from the ESTKF case, the T matrix should change at every forecast to take into account GHOSH projections in the principal components of the uncertainty subspace".

Line 578: "The order of an ensemble method is one of the most relevant proxies of a filter skill". This is a bold statement and I don't think that the statement on SEIK/ETKF/ESTKF vs. SEEK is sufficient to show this. For the SEEK filter, the deficiency seems to be the linearized propagation of the modes defining the covariance matrix. This is still intended to represent order 2.

34. We removed the reference to the SEEK filter and changed our statement at line 651: "Based on the results of the present work, the order of an ensemble method can be one of the proxies of the filter skill, as shown by comparing SEIK (order 2) with GHOSH (order higher than 2)."

Lines 586-588: Please avoid appraisals like 'outstanding' and 'remarkable' since they are not suited to a scientific publication.

35. We removed the suggested words.

Line 590: Here again the misleading statement on "70% RMSE reduction" appears. The statement contains "when tuning with the same forgetting factor", but using the same forgetting factor is not 'tuning'.

36. We removed the misleading sentence.

As before, I recommend to be careful with citations and as far as possible cite original papers. Actually Carrassi et al. (2018) is cited 10 times, for various aspects like 'deterministic sampling' (lines 30 and 74), forgetting factor (line 264), hybrid filters (line 493), but actually, none of these methods have been introduced in this review paper. I recommend to also check this for other references, in particular when review papers are cited.

37. We agree that not only review papers should be cited, and we thank the reviewer for pointing it out. A number of citations have been updated.

Please do a careful proofreading. There are various places where singular and plural are used inconsistently, but also other places with incorrect grammar (missing 'the', etc.). There

is also a 'witch' hidden in the text (but easy to find), which is certainly not the word the authors wanted to use.

38. The manuscript has been carefully proofread and many typos have been corrected. We thank the reviewer for pointing out.

Regarding the provided code:
When testing the Python code, I found that for the GHOSH, the options 'order=2' and 'order=3' do not work. It would be nice to have in-line comments in the code to be able to understand what is done in different parts. For example, the authors coded the initialization of GHOSH (weights and matrix Omega) in some way that supports orders 3 and 5. Unfortunately, the idea behind e.g. matrix omega2 is unclear.

39. We thank the reviewer for reporting the bugs for order 2 and 3, that have been corrected and some in-line comments added. However, it is worth noting that the current implementation is only meant to support order 5, since all the experiments in the manuscript refer to GHOSH at this order. A more general implementation of GHOSH supporting any order would require a numerical nonlinear solver to find solutions to the moment-matching system in equation (2). A complete implementation of the GHOSH filter in the PythonDA framework will possibly be the focus of a future dedicated manuscript.