Review of "GHOSH v1.0.0: a novel Gauss-Hermite High-Order Sampling Hybrid filter for computationally efficient data assimilation in geosciences" by S. Spada et al.

Reply on Reviewer comment RC3

The manuscript introduces a higher-order ensemble sampling scheme and a corresponding ensemble Kalman filter analysis step. The sampling scheme utilizes ensemble weights to represent higher-order moments of the ensemble distribution without necessarily increasing the ensemble size. The analysis step is motivated by the error-subspace Kalman filter SEIK and utilizes the ensemble weights. Additional sampling steps are introduced during the forecast phase and after the analysis step in order to maintain the higher-order sampling to the prescribed order. After the introduction of the sampling and assimilation algorithm, the method is first assessed in comparison to the SEIK filter in twin experiments using the chaotic Lorenz-96 toy model, which is frequently used to assess data assimilation methods. Subsequently, the new method is used in a realistic model application in which satellite chlorophyll observations are assimilated into the OGSTM-BFM biogeochemical model. The manuscript concludes that the method can result in better state estimates than the SEIK filter for the same ensemble size in case of the Lorenz-96 model. Further it is concluded that using higher-order sampling is feasible in a realistic application and can result in better state estimates compared to using only second-order sampling in the biogeochemistry data assimilation application.

Before I come to scientific assessment, I like to point out that this manuscript does not seem to fit well into the scope of GMD. Presented is the development of a sampling and data assimilation method. This is neither a new model development, description, assessment or development of new assessment methods, which are the scope of GMD. The data assimilation is implemented and used with a toy model and a realistic ocean-biogeochemical model, but the focus is on the algorithm and no particular model-related developments are described. To this end, I recommend to transfer the manuscript to a journal with a better fitting scope. Within the journals of the EGU, this is Nonlinear Processes in Geophysics (NPG).

From the scientific viewpoint, the proposed algorithm is new and the experiments show that it has the potential to improve the skill of ensemble data assimilation by accounting for higher order moments in the ensemble distribution and hence improving the mean and covariance estimates used in the analysis step of the ensemble filter. However, the manuscript needs to be improved at different places. The description of the algorithm has to be refined and it has to be better explained. Further the numerical experiments have to be organized in a way that the benefit of the new algorithm becomes clear in typical settings. The application to the 3D ocean-biogeochemical model seems to overload the manuscript while it does not even compare the results to a standard algorithm. Overall, this requires a major revision of the manuscript.

We really thank the Reviewer for the very accurate reading of our manuscript and for having tested the algorithm. The comments and suggestions provided by the

Reviewer have been very helpful to revise the manuscript contents, and we think that the manuscript quality will be significantly increased after the completion of the revision. Hereafter, our point-by-point responses to the Reviewer's comments are provided in bold green.

Firstly, we would like to clarify the main objective of the work, since it appears this was not clear enough. We are introducing a new method (GHOSH filter) that is focused on the use of a higher order of approximation to evaluate the ensemble mean in data assimilation applications. We are not evaluating standard deviation, kurtosis and higher statistical moments of the ensemble in the present work. On the other hand, it is true that the equality of the statistical moments up to order h implies an approximation of the mean up to order h, as demonstrated in Appendix A.

Concerning the option to move the manuscript in another journal, we chosen to submit the manuscript in GMD since "papers focussing on data assimilation are welcome" as "Development and technical papers" (https://www.geoscientific-model-development.net/about/manuscript_types.html#item2, last visit 27 March 2024). Indeed, we firstly submitted the manuscript as a "Development and technical paper" but the Topic Editor has then adjusted the manuscript type to "Model description paper", but did not raise concerns about the relevance of the manuscript with respect to the GMD scope. More recently, when discussing the option to split the manuscript in two parts (see a later comment response), the Topic Editor confirmed that the work falls into the scope of GMD. Given the above reasons, being the review process in progress and having already taken quite a long time, we would prefer to stay in GMD.

Major comments:

On the derivation and mathematical treatment: At least for GMD, with is not a mathematical journal, the manuscript explains far too less of the algorithmic aspects. At several instances the authors just state 'can be proven', but omit attempts to explain why some equation holds. For me this is clearly insufficient and it is likely that readers cannot follow the descriptions. I do not recommend to include proofs (given that GMD is not a mathematical journal), but the authors should sufficiently explain why some relationship or equation holds. This does not need to be long, but will help the readers tremendously, while currently the authors force the readers to 'just believe' what they have written. (This comment will likely also hold should the manuscript be transferred for NPG, since also this is not a strictly mathematical journal).

Thank you for this comment that can help to make the manuscript more accessible to GMD readers. Some of the "can be proven" statements are explicitly addressed in the following point-by-point responses but other instances will be addressed in the new version of the manuscript.

Experiments with the Lorenz-96 model: The first set of numerical experiments is conducted using the Lorenz-96 model. The experiments show that the newly proposed GHOSH filter can yield smaller RMSEs by up to 66%. Given that this error reduction of the GHOSH filter relative to the SEIK filter is very large (I don't think that I have seen such magnitude even in comparisons of fully nonlinear particle filters with EnKFs), I recommend to be particularly careful in ensuring that the experiments are representative and comparable to previous studies. In particular, one has to ensure that this advantage of the GHOSH filter is not an artifact of the particular model and data assimilation configuration. In this respect I see several weaknesses as the model configuration and implementation are rather untypical, the initialization of the ensemble is different from other studies even for the SEIK filters, and the experiments are far too short. I will comment in more detail on this further below.

We addressed all the suggestions on the Lorenz-96 model experiments in the point-by-point responses, however we would like to thank the Reviewer, since the manuscript results will be certainly improved by enlarging the set of toy-model numerical experiments.

Experiments with the realistic OGSTM-BFM model: In these experiments, in which real satellite chlorophyll data are assimilated, only different configurations of the GHOSH filter are assessed. Thus, a comparison to an EnKF like SEIK is missing. The result of the experiments is that using order 5 yields comparable RMSD for chlorophyll and phosphate but lower RMSD for nitrate, compared to using a lower order. The magnitude of this effect depends on the ensemble inflation, and for too strong inflation also the GHOSH filter leads to a deterioration of nitrate compared to the control run. To this end, the higher-order sampling can yield a better result in the multivariate update for a particular choice of the forgetting factor. Unfortunately, the manuscript does not provide insights into why this does happen.

Further, maps of the chlorophyll and phosphate as well as Hovmoeller plots of chlorophyll and phosphate are discussed. However, these discussions do not provide any further insight into the GHOSH filter, but merely show that this filter can be successfully applied to this model. These results might be interesting for researchers interested in chlorophyll assimilation, but for a methodological study on this new filter method, the experiments provide very little insight. Only the first part including Table 2 seems to be relevant, but here a comparison to the SEIK filter is missing.

Given the complexity of the model and the expert knowledge a reader needs to appreciate the discussion on the effects on chlorophyll, nitrogen and phosphate, I recommend to remove this experiment from the manuscript. In order to assess multivariate assimilation (which cannot be done with the Lorenz-96 model) I rather recommend to utilize some other idealized multivariate model. Perhaps, the 2-scale Lorenz-2005 model could be a possible choice. For the methodological study it would be very important to actually assess the reasons for why multivariate updates are better (if they are). Just showing that they are better in one particular model configuration without understanding the reasons has little value because one cannot generalize the result. Below I will not further comment on details in the sections on the OGSTM-BFM model.

Also Reviewer #1 commented about the option to remove the realistic 3D application, moreover we received a positive answer by the Topic Editor about a request we sent in the past weeks on the possibility to split the manuscript in two parts. Thus, we will revise the manuscript considering the twin experiment and realistic application separately.

Detailed comments:

Title: It is not clear why the authors give their algorithm a version number, and this is very untypical. While I know that GMD requires version numbers when a certain model version is presented, this manuscript is not about a particular implementation of a model but about a numerical algorithm. This issue might again point to the fact that the manuscript is not well suited for GMD as I commented on before. Apart from this, I recommend to add to the title the word 'ensmeble' to clarify that an ensemble filter is introduced.

As briefly discussed above, the Topic Editor adjusted the manuscript type to "Model description paper", and on this occasion the Topic Editor explicitly asked for a versioning number. Indeed, the versioning number request seemed a sound request, since the GHOSH algorithm could be modified and improved in the future. Concerning the title, we agree with the Reviewer on the need to add the word "ensemble" in the title, and we propose the following novel title:

"GHOSH v1.0.0: a novel Gauss-Hermite High-Order Sampling Hybrid ensemble filter for computationally efficient data assimilation in geosciences"

Abstract: The abstract includes unnecessary details. E.g. information how many experiments have been conducted with the Lorenz-96 model ('two thousands', line 9) and which particular parameters have been varied (line 12) are not suitable for the abstract. The abstract is also irritating in that first 'GHOSH' is mentioned as a sampling method while in line 14 it is then named a 'filter'. Perhaps, one can better formulate this as introducing a filer and the corresponding sampling method.

The abstract will be improved according to the Reviewer's suggestions. In particular, the word "thousands" and details on the varied parameters will be removed while at line 8, "GHOSH sampling method" will be changed to "GHOSH filter's sampling method".

The statement 'use of a higher order of convergence substantially improves the performance of the assimilation with respect to nitrate' (lines 20/21) is not valid in this form. Valid is that the higher-order filter improved the nitrate for the particular experiment, but this cannot be generalized in any way. The final sentence of the abstract on the computation time is

superficial. The situation that most time is spend in computing model forecasts does not need to hold in general. Instead, the execution time has to be compared for the actual filter analysis including the resampling steps of the GHOSH filter. Please revise the abstract accordingly.

We propose to change “improves" at line 20 to “improved” (and similarly at line 18). Concerning the final abstract sentence on the computational time, this part will need to be deeply revised since the 3D application will no longer be included in the manuscript (part I - Toy model application): “In view of potential implementation of the GHOSH filter in operational applications, it should be noted that GHOSH and SEIK filters have the same asymptotic computational complexity, and that in Lorenz-96 applications the computational costs are dominated by the model integration, and GHOSH and SEIK differ by only 4% of the total computational time.”

Moreover, we will better comment on computational costs also for the 3D application (part II - Mediterranean Sea biogeochemistry application):

“In view of potential implementation of the GHOSH filter in operational applications, it should be noticed that GHOSH and SEIK filters have the same asymptotic computational complexity. Further, in our three-dimensional realistic experiment, the time-to-solution was widely dominated by model evolution and I/O operations, since, as in all ensemble Kalman filters, the integration of a realistic model is by far more computationally expensive than the assimilation scheme.”

Introduction section:

- Unfortunately, this section contains various small issues and it is difficult to comment on all of them. I focus here on the most relevant:

The section includes invalid referencing. E.g. the list of references in lines 28-29 appears to be an arbitrary selection of studies that applied DA. Since there are many publications about applications of data assimilation it can be a better choice to cite review papers that summarize the state of research instead of picking 'some' article for which it is then unclear why the authors consider these as particularly relevant.

We thank the Reviewer for the suggestion. We will revise the citations at the Introduction beginning to provide a wider view of DA methods and applications:

“(among the others, see the methods and applications reviewed by Carrassi et al., 2018; Houtekamer and Zhang, 2016; Lahoz and Schneider, 2014; van Leeuwen et al., 2019; Martin et al., 2015; Roth et al., 2017; Vetra-Carvalho et al., 2018)”

- With regard to citing review papers, I strongly recommend to cite those explicitly as reviews, e.g. using 'see, e.g.'. For example, Carrassi et al. (2018) is cited for 'modified

implementations have been proposed for application to non-linear ones' (line 35). This review paper did however not introduce the 'modified implementations' but reviews what has been published in other original studies. This likewise holds for the citation of Bannister (2017b) which is cited for 'EnVar' (line 38), but is a review of variational methods. (BTW: The second paragraph of the introduction does not seem to be relevant for this study on an ensemble filter method and it could be omitted without loss of relevant information). Further, for books it is common practice to cite them providing a chapter where the particular information can be found (Readers should not be required to read a whole book). Another case of invalid reference is e.g. citing Bocquet et al. (2010) for 'particle filter methods'. The paper discusses methods for non-Gaussian data assimilation, but it is not an original paper for particle filters. In any case, there are more recent review papers, like the cited review by van Leeuwen et al. (2019) which better cover the state of research.

The correct citation of reviews will be adopted in the revised version of the manuscript. We thank the Reviewer since this change will improve the use of references.

Moreover, we will remove the second paragraph of the introduction according to the Reviewer suggestion; and we will adopt van Leeuwen et al. (2019) instead of Bocquet et al. (2010).

- It is also important that the authors provide references for claims they include in the text. E.g. in lines 30/31 it is written 'the use of DA has been steadily increasing in the last decade and it is now ubiquitous in earth sciences, also thanks to the unprecedented -and always increasing- availability of both data and computational capability.' I'm not sure from where the authors got the insight that the increasing computational capability and data availability lead to an increased use of DA. However, I would suppose that there should be some review paper examining this. Actually, most data types that are currently assimilated in the ocean and biogeochemistry (e.g. sea surface hight, sea surface temperature, chlorophyll) have been available for more than a two decades and have already been assimilated more than 10 years ago. Also the computing power was sufficient more than 10 years ago even to perform ensemble data assimilation as is evident from data assimilation studies that were published from 15-20 years ago. The major changes appear that the model complexity and resolution was steadily increased. These facts obviously contradict the authors statement.

- Related to this, the authors state that recently the use of ensemble algorithms has been proposed 'thanks to the scalability of parallel implementations' (line 37). Then, Evensen (1994) is cited for the EnKF. I cannot really see that a paper that was published 29 years ago is 'recent'. It is further not evident that these methods were proposed because of the scalability of parallel implementation since Evensen (1994) proposed the EnKF as a method to use a sampled covariance matrix as a dynamic estimate of the uncertainty in order to advance the extended Kalman filter that was used by Evensen before (see Evensen, 1992, 1993), but not as a method solving a scalability issue. Please revise the text, to avoid such potential misleading statements.

Thanks to the Reviewer suggestions the first part of the introduction will be revised to be more consistent with literature. In particular, we will introduce references to the review by Vetra-Carvalho et al. (2018), where the computational costs and the increase of observations are proposed as two of the factors that motivated recent EnKF developments toward more complex and non-linear applications. In the text rephrasing the reference to Evensen (1994) will be removed since no more necessary:

"Thanks to the scalability of parallel implementations, the use of ensemble algorithms (see e.g., Vetra Carvalho et al., 2018; Houtekamer and Zhang, 2016; Bannister, 2017) have been proposed to estimate uncertainty and improve assimilation skills in Kalman filters and variational methodologies. On the other hand, some of the strong points of the ensemble and variational have been merged in hybrid filters (e.g., Hamill and Snyder, 2000). Moreover, recent developments of EnKF have been conceived to face increasing non-linearities and model complexity that are also related to the expanding availability of observations and computational resources (see e.g., Vetra Carvalho et al., 2018)."

- There are also further questionable statements.

-- E.g. in lines 49/50: 'second order sampling methods provide only a second order approximation of the model'. Actually, the ensemble in ensemble KFs is not meant to approximate the model, but it is used to represent model uncertainty, i.e. the probability distribution. One should be very clear about this.

We thank the Reviewer for raising this point about the aim of the ensemble KFs. Indeed, we meant to highlight that, when second-order exact EnKFs are applied with models that cannot be represented by a second order polynomial, the mean estimation provided by the ensemble is affected by an approximation error. The error in the mean is strictly related to the error made in approximating the model by a second-order polynomial. For sake of clarity, we recall that second-order sampling methods are based on the fact that, assuming that the model m is a second order polynomial, the expected value of the two random variables X and Y is the same, i.e. E[m(X)] = E[m(Y)], where X represents the system uncertainty and Y its ensemble approximation produced by the sampling method (i.e., the ensemble describes the discrete probability mass function of Y). When m is not a second order polynomial, the error on the expected value || E[m(X)] - E[m(Y)] || can be estimated by

$$||E[m(X)] - E[m(Y)]|| < ||E[m(X)] - E[p(X)]|| + ||E[p(X)] - E[p(Y)]|| + ||E[p(Y)] - E[m(Y)]||,$$

where p is the best second order polynomial approximating m. The second term in the r.h.s. vanishes because the second order sampling is exact on second order polynomials. What remains are the errors of the second order approximation of the model m with the polynomial p.

We propose to modify the sentence 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, which is affected by an error strictly related to the error made in approximating the model by a second-order polynomial.”

-- Then in lines 50/51: 'the second order approximation is more effective the closer the ensemble members are to each other'. Here, I suppose the authors imply that the nonlinear effect are smaller if the perturbation of the ensemble members is small and hence the second-order approximation should hold better. While this seems to be natural, Rainwater and Hunt (2013) found that a smaller ensemble spread did not improve the results of an EnKF. Thus, the statement does not seem to be valid in this form.

We fully agree with the Reviewer that the nonlinear effects are smaller when the perturbation of the ensemble members is small, and hence the second-order approximation holds better. In other words, if the true uncertainty is small, then a second-order approximation is affected by a smaller error. The statement is not meant to contradict Rainwater and Hunt (2013), who propose an empirical technique to reduce the analysis error by inflating the forecast spread and deflating the analysis spread with the aim to improve EnKF results under certain circumstances. We propose to modify the sentence focusing on uncertainty rather than the ensemble spread (also in line with comments from Reviewers 1 and 2):

“Furthermore, the second order approximation is more effective the closer the ensemble members are to each other (i.e., small uncertainty), 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.”

-- Related to this is the statement "a relatively large ensemble spread is often required in data assimilation applications," (line 52). I'm not aware of that this is true and the authors would need to provide a reference for this statement.

The sentence will be modified according to the previous comment.

-- Also the statement "second order methods use an ensemble of r + 1 members to span an r-dimensional error subspace" (line 57) is not supported by the literature. Actually the perturbed-observations EnKF by Evensen (1994) is second-order (because it uses the Kalman filter analysis step), but it is not aimed at spanning an r-dimensional subspace. Using a particular sampling with r+1 members was introduced by Pham with the SEEK and SEIK filters, but not widely adopted. Also the ensemble initialization used with the (L)ETKF scheme is not explicitly 2nd order. This also holds for the notion of the 'error subspace',

which was further discussed by Nerger et al. (2005) while other filter methods, like the ETKF or newer variants of the EnKF (Evensen 2003/2004), were developed without this notion.

We thank the Reviewer for this comment that highlights that the definition of KF method order was unclear. We will make the definition of "order" more explicit, in particular we will clarify that we are referring to second-order exact deterministic methods (like ETKF or SEIK while Evensen's EnKF is a stochastic method which is not included in the reported definition of second-order methods):

L. 43: "The number of ensemble members can be reduced by adopting deterministic sampling methods (as opposed to stochastic EnKF methods, see e.g. Carrassi et al., 2018). Examples of deterministic EnKF using second-order sampling methods are SEIK (Pham (2001)) and ETKF (Bishop et al. (2001)). Extending the definition of second-order exact methods in Pham (2001), with the term "order" we refer to the polynomial order of approximation of the filter or of its sampling method. Namely, in case of hth-order, the ensemble has the property of providing a forecast mean with no error as long as the evolution function used for forecasting (i.e., the model) is a polynomial of order h. As proven in Appendix A, this is equivalent to sampling an ensemble that preserves (before applying the model) the first h statistical moments."

Moreover, the correct reference to Nerger (2005) will be added near the first occurrence of "error subspace":

L. 54: "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 case and consequently bigger computational costs, given that a higher order of approximation implies a larger number of ensemble members to represent the same error subspace (see Nerger 2005 and 2012 for an introduction to the error subspace concept)"

- Please note that the reference Pham (1996) is incomplete, but it is likely also not a valid reference according to citation standards. It is a technical report and was never officially published in a peer reviewed journal. While I'm aware of this report, I also could not find it any more on the internet. To this end, I recommend to replace it by a valid reference to a peer-reviewer publication. The first real publication of the SEIK filter seems to be Pham et al. (1998b), different from what is cited in the manuscript. Since this article is mainly in French and only has a shorted text in English, a better reference would be Pham (2001).

Thanks, the citation will be corrected.

- The introduction misses to explain what the authors mean by 'high order' by 'order' at all. Given that GMD is not a mathematical journal this should be explained.

We thank the Reviewer for underling points that could be improved in order to help the reader to better understand the presented work. The definition of order in introduction will be improved and will be more explicit following the Reviewer comments.

- Please be careful when discussing a sampling in contrast to discussing filtering. E.g. lines 65/66 state "we propose a novel weighted ensemble method based on a new high-order sampling, that provides ensemble mean estimates of order higher than 2". Actually, while the existing ensemble Kalman filter method use the second-order analysis scheme (treating only mean and covariances), they are not required that the ensemble was sampled with second order (in fact unless one uses second-order exact sampling, the ensemble sampling does not consider the sampling order.)

- Also related to lines 65/66, I a bit irritating about the implied meaning of 'order'. I understand it as accuracy in sampling the PDF. However, then the mean is the first order moment of the PDF. How can this moment be accurate by 'order higher than 2'. I'm likely missing something here, but I think that it is likely that many readers will also wonder. To this end, I see the clear need to explain the terms 'high order' and 'higher order' so that the readers knwo how to understand the discussions throughout the manuscript.

- In this respect, please also check whether the distinction in 'high order' and 'higher order' is required and if so, please explain the difference.

Based on the comment Reviewer on L 57, we will clarify the definition of "order" in the Introduction. This improvement will help to clarify the last three points raised by the Reviewer. We will carefully verify that misleading wording of "order" will be avoided in the Introduction and through the whole manuscript.

- I will stop with details on the Introduction at this point. Overall, the authors should carefully revise the introduction. In this they should ensure that the statements are true and that they are sufficiently supported by the literature. This implies to provide references to claims as described before. It further implies to be careful when citing review papers so that it clear that no original works are cited (This also holds in other sections, e.g. the reference to Carrassi et al. (2018) in line 90 and line 264 are also not valid - Carrassi is neither an original reference for deterministic square-root sampling methods not for the forgetting factor).

Thank you for your careful reading of the Introduction. In addition to the changes listed above, we will take care of other similar occurrences in the manuscript.

Figure 1 and lines 93-104: The figure is difficult to understand and I'm not sure if it is really sufficient to visualize the 3rd order sampling in 2 dimensions or 5th order in 1 dimension.

Considering this and the following Reviewer's comments and comments from Reviewer #1, we propose to describe Fig. 1 in more detail in a novel appendix (Appendix C) adding the description (including equations and calculations) of the concepts summarised in Fig. 1 panels (e.g., computation of the moments). Indeed,

this part of the manuscript is too concise and a reader could really benefit from additional details.

- Please include a clearer explanation for the shaded circular regions (in the top panel) in the caption (the caption states 'isosurfaces', but of what?).

- In the 2D and 1D views I cannot see that a 3rd or 5th oder approximation is shown. It is also not obvious that the 2D view is a projection of the tetrahedron. In the way this is plotted, the lower triangle looks like it's parallel to the y-x plane so that the projection would be a triangle with one sample in its center.

- With regard to lines 102/103, I cannot follow why 3 weighted members should be able to represent an approximation of order 5. With weights one can obviously shift the mean, variance and skewness. But it is unclear how kurtosis could be expressed with 3 members (perhaps even with 4 members). Perhaps this irritating is due to the fact that it is not clear how the moments would be computed in this case.

- It would be useful for the readers if the figure is replotted with a better perspective. Likewise the text should better explain the how the orders are expressed in the sampling. Instead of stating 'it can be proven' in line 98 it would be useful to explain why this is the case and perhaps explain how a skewed (e.g. non-Gaussian) distribution would be represented (I don't think that this is a projection of the tetrahedron representing the distribution, but a sampling of the projection of the actual distribution). Since the example uses a Gaussian distribution whose higher moments are generally zero (for odd orders) or have a constant value (for even), it seems to be difficult to actually sketch the representation of higher moments. Showing how higher moments could be represented for a non-Gaussian case could be useful. It might also be useful to show the example on a Gaussian case that has different variances per direction.

- Linked to this is the sentence in lines 103/104 stating 'what we have obtained is a 3-dimensional second order weighted ensemble which is a third-order approximation in the xy plane and a fifth order approximation along the x axis.' For me, this statement is not evident from the figure. It might be that the sketched samplings in the xy-plane and along the x-axis are of this higher order - perhaps because they represent zero higher orders, but this is not visible from the dots that are shown in the figure.

All the above comments will be addressed in the novel Appendix. For instance, the Appendix C will include a practical example of a 5th order weighted ensemble in 1D composed by 3 points P1, P2 and P3, i.e.:

$P1: x1 = \sqrt{3};\ w1 = 1/6,$

$P2: x2 = -\sqrt{3};\ w2 = 1/6,$

$P3: x3 = 0;\ w3 = 2/3.$

The first moment (mean) is:

$w1 \cdot x1 + w2 \cdot x2 + w3 \cdot x3 = 1/6 \cdot \sqrt{3} + 1/6 \cdot (-\sqrt{3}) + 2/3 \cdot 0 = 0,$

the second moment (variance) is:

$w1 \cdot x1^2 + w2 \cdot x2^2 + w3 \cdot x3^2 = 1/6 \cdot 3 + 1/6 \cdot 3 + 2/3 \cdot 0 = 1,$

the third moment (skewness) is:

$w1 \cdot x1^3 + w2 \cdot x2^3 + w3 \cdot x3^3 = 1/6 \cdot \sqrt{3}^3 + 1/6 \cdot (-\sqrt{3})^3 + 2/3 \cdot 0 = 0,$

the fourth moment (kurtosis) is:

$w1 \cdot x1^4 + w2 \cdot x2^4 + w3 \cdot x3^4 = 1/6 \cdot 9 + 1/6 \cdot 9 + 2/3 \cdot 0 = 3,$

the fifth moment is:

$w1 \cdot x1^5 + w2 \cdot x2^5 + w3 \cdot x3^5 = 1/6 \cdot \sqrt{3}^5 + 1/6 \cdot (-\sqrt{3})^5 + 2/3 \cdot 0 = 0.$

Thus, the ensemble P1, P2, P3 is compliant with the statistical moments of a standard normal distribution up to order 5.

The property of this 1-dimensional 5th-order sampling, can be extended to a 2-dimensional ensemble, aimed to keep order 5 along the x-axis while achieving order 3 in the xy plane. Such 4-members ensemble is:

$P1: (x1, y1) = \left(\sqrt{3}, 0\right);\ w1 = 1/6,$

$P2: (x2, y2) = \left(-\sqrt{3}, 0\right);\ w2 = 1/6,$

$P3: (x3, y3) = \left(0, \sqrt{3/2}\right);\ w3 = 1/3,$

$P3: (x4, y4) = \left(0, -\sqrt{3/2}\right);\ w4 = 1/3.$

Note that all the moments along the x-axis are already computed above up to order 5, since this ensemble and the previous one are indistinguishable looking only at the x-coordinate. In fact, considering the projection on the x-axis, P3 and P4 behave as a unique member with weight w3+w4 because they have the same x-coordinate x3=x4. It remains to be checked if the moments involving y match the moments of standard normal distribution up to order 3, i.e.:

the 1st moment associated to y (i.e., the mean along y) is:

$w1 \cdot y1 + w2 \cdot y2 + w3 \cdot y3 + w4 \cdot y4 =$

$= 1/6 \cdot 0 + 1/6 \cdot 0 + 1/3 \cdot \sqrt{3/2} + 1/3 \cdot \left(-\sqrt{3/2}\right) = 0,$

the 2nd moment associated to y^2 (i.e., the variance of y) is:

$$w1 \cdot y1^2 + \ \ldots \ + w4 \cdot y4^2 = 1/6 \cdot 0 + 1/6 \cdot 0 \ + 1/3 \cdot 3/2 + 1/3 \cdot 3/2 = 1,$$

the 2nd moment associated to xy (i.e., the covariance between x and y) is:

$$w1 \cdot x1 \cdot y1 + \ldots + w4 \cdot x4 \cdot y4 =$$

$$= 1/6 \cdot \sqrt{3} \cdot 0 + 1/6 \cdot \left(-\sqrt{3}\right) \cdot 0 + 1/3 \cdot 0 \cdot \sqrt{3/2} + 1/3 \cdot 0 \cdot \left(-\sqrt{3/2}\right) = 0,$$

the 3rd moment associated to x^2 y is:

$$w1 \cdot x1^2 \cdot y1 + \ldots + w4 \cdot x4^2 \cdot y4 =$$

$$= 1/6 \cdot 3 \cdot 0 + 1/6 \cdot 3 \cdot 0 + 1/3 \cdot 0 \cdot \sqrt{3/2} + 1/3 \cdot 0 \cdot \left(-\sqrt{3/2}\right) = 0,$$

the 3rd moment associated to x y^2 is:

$$w1 \cdot x1 \cdot y1^2 + \ldots + w4 \cdot x4 \cdot y4^2 =$$

$$= 1/6 \cdot \sqrt{3} \cdot 0 + 1/6 \cdot \left(-\sqrt{3}\right) \cdot 0 + 1/3 \cdot 0 \cdot 3/2 + 1/3 \cdot 0 \cdot 3/2 = 0,$$

the 3rd moment associated to y^3 is:

$$w1 \cdot y1^3 + \ \ldots \ + w4 \cdot y4^3 = 1/6 \cdot 0 + 1/6 \cdot 0 \ + 1/3 \cdot \sqrt{3/2}^3 + 1/3 \cdot \left(-\sqrt{3/2}\right)^3 = 0.$$

The dimensions can be increased once more by building in a similar way a 3-dimensional ensemble, as shown in Fig. 1. This ensemble will be indistinguishable from the previous one in the xy plane projection and capable of achieving 2nd order on the xyz space. Such ensemble is:

$$P1\colon (x1, y1, \ z1) = \left(\sqrt{3}, 0, -\sqrt{2}\right); \ w1 = 1/6,$$

$$P2\colon (x2, y2, \ z2) = \left(-\sqrt{3}, 0, -\sqrt{2}\right); \ w2 = 1/6,$$

$$P3\colon (x3, y3, \ z3) = \left(0, \sqrt{3/2}, \sqrt{2}/2\right); \ w3 = 1/3,$$

$$P3\colon (x4, y4, \ z4) = \left(0, -\sqrt{3/2}, \sqrt{2}/2\right); \ w4 = 1/3.$$

All the moments involving x and y are already checked in the previous calculation up to order 5 along x and up to order 3 in the xy plane. It remains to be checked if the moments involving z match the moments of standard normal distribution up to order 2, i.e.:

the 1st moment associated to z (i.e., the mean along z) is:

$$w1 \cdot z1 + w2 \cdot z2 + w3 \cdot z3 + w4 \cdot z4 =$$

$$= 1/6 \cdot \left(-\sqrt{2}\right) + 1/6 \cdot \left(-\sqrt{2}\right) \ + 1/3 \cdot \sqrt{2}/2 + 1/3 \cdot \sqrt{2}/2 = 0,$$

the 2nd moment associated to z^2 (i.e., the variance of z) is:

$$w1 \cdot z1^2 + \ \ldots \ + w4 \cdot z4^2 = 1/6 \cdot 2 + 1/6 \cdot 2 \ + 1/3 \cdot 1/2 + 1/3 \cdot 1/2 = 1,$$

the 2nd moment associated to xz (i.e., the covariance between x and z) is:

$$w1 \cdot x1 \cdot z1 + \ldots + w4 \cdot x4 \cdot z4 =$$

$$= 1/6 \cdot \sqrt{3} \cdot \left(-\sqrt{2}\right) + 1/6 \cdot \left(-\sqrt{3}\right) \cdot \left(-\sqrt{2}\right) + 1/3 \cdot 0 \cdot \sqrt{2}/2 + 1/3 \cdot 0 \cdot \sqrt{2}/2 = 0,$$

the 2nd moment associated to yz (i.e., the covariance between y and z) is:

$$w1 \cdot y1 \cdot z1 + \ldots + w4 \cdot y4 \cdot z4 =$$

$$= 1/6 \cdot 0 \cdot \left(-\sqrt{2}\right) + 1/6 \cdot 0 \cdot \left(-\sqrt{2}\right) + 1/3 \cdot \sqrt{3/2} \cdot \sqrt{2}/2 + 1/3 \cdot \left(-\sqrt{3/2}\right) \cdot \sqrt{2}/2 = 0.$$

All the above proves that the weighted ensemble P1, P2, P3, P4 is a second-order ensemble achieving order 3 in the subspace of the x and y directions and order 5 along the x direction.

We are confident that these step-by-step calculations will help the Reader in following the manuscript arguments.

- For the non-mathematical readers of GMD it might also be useful to avoid the term 'standard' for the normal distribution but instead mention the variance explicitly.

Thank you for spotting this issue. We will explicitly refer to the variance of the standard distribution. In particular we will modify L. 356:

"Observations are generated for each truth period (truth_20, truth_40, truth_60, truth_80) by extracting from the state vector, the values of the even indexed variables (i.e., x2, x4, ..., x62) every Δt time units and adding to each observed variable a random number sampled from a standard normal distribution (i.e., with variance equal to 1)."

Lines 110/111 mention that a higher-order sampling of a Gaussian distribution can be obtained by the Gauss-Hermit quadrature rule. At this point reader might likely wonder why a higher-order sampling of a Gaussian is required, when one can fully represent it by its mean and covariance and the second-order exact sampling provides this sampling. On the other hand, it might be obvious that a higher-order sampling helps for non-Gaussian distributions, but in this case the Gauss-Hermite quadrature doesn't seem to hold. It would be good if this aspect, which closely linked to the motivation of the method, is better explained. Perhaps, it becomes clearer when the authors explain 'high/er order' as commented on before.

Given the improved definition of "order" that will be provided in the introduction, this part of the manuscript should be clearer. Indeed, the order of the sampling is not related to the Gaussianity of the pdf but to the order of approximation of the mean, which, as mentioned above, is closely related to how good the approximation of the nonlinearity of the model is. In addition, the novel Appendix C will use simple examples to guide the reader in understanding how moments of order higher than two

can affect a sampling of a Gaussian pdf. In fact, even if its first two moments completely characterise the distribution, the sampling needs to also match the higher moments to produce an ensemble capable of achieving an order of approximation higher than two.

Lines 112-114 describe that the higher-order sampling can be performed for a limited number of directions, while for the other directions second-order sampling can be used. Here I'm wondering if this holds after the application of the model. If we have, e.g., a polynomial function that includes terms that 'mix' (e.g. by multiplication) the directions used for higher-order sampling with those with second-order sampling, one obtains mixing effects. With this the clear separation in higher-order and second-order directions should no longer hold. Does the re-sampling during the forecast corrects these effects? Please explain these effects in a clear way.

Since each re-sampling computes its own PCA, the re-sampling during forecast chooses new principal directions taking into account the model application and its mixing effects on the previous principal components. It is also worth mentioning that, as shown in Appendix A (eq. A2 and following), any multiplicative term is already taken into account exactly up to the considered polynomial order. In more detail, if there are mixings in the form of multiplications between two directions, this is a second order term and it is exactly taken into account by any second-order sampling. If the mixing includes multiplications between more than two directions (some from the second-order and some from the hth-order directions), then the polynomial function has a higher (then two) polynomial order, and an error appears, but, since the directions sampled with order 2 were the directions where the spread is smaller, the error is comparably small because it is attenuated by at least one low-spread factor. Instead, the higher order terms (up to order h) composed by high-spread factors only, which would represent a bigger source of error for a second-order ensemble, do not add any error to the mean computed from a GHOSH ensemble, since the large-spread directions are sampled with order h. In this sense, the GHOSH sampling can be seen as a method to remove the principal sources of error in the mean estimation.

The details provided in Appendix A together with the clearer definition of "order" provided in the manuscript, should help the reader to understand the mixing effects.

Equation (1): I'm somewhat lost here. There are many indices which influence each line of the equation system which I cannot really disentangle. I think it would be useful to show some more lines, maybe the first 3, in addition to the general one. Is the subscript 'j_1,...,j_\xi' just one number or a list of indices. How many equations are in the system? (This relates to line 137 where it is stated that a large value of h required a smaller s or larger r, which is qualitative but does not provide an indication of actual counts for r/s/h)

Based on this Reviewer comment and also on a comment from Reviewer #1, we propose to modify the system equations and their description to improve readability, as in the attached document "system1.pdf"

Line 133: Here the 'Gaussian case' is mentioned. Usually we consider a Gaussian to be a distribution that is fully described by the first two orders. Why should a higher-order sampling be relevant for this?

We think that the changes made in the Introduction about the meaning of "order" and the previous responses will help to clarify why the "Gaussian case" is considered.

Line 130: It is stated: "Such probability distribution must be uncorrelated, normalized and have 0 mean". Please explain why this 'must be'.

The statement will be expanded to give an insight of why those properties are required, as proposed in the reply to the previous Reviewer comment about equation (1).

Line 138: I cannot follow the explanation how \Omega_h is defined. It is described as a "matrix with coordinates"? What does this mean. Is it possible to provide a proper definition, e.g. in form of an equation?

The definition of \Omega_h will be clarified as follows:

"matrix with elements u_ij (i.e., u_ij is the entry of the matrix Omega_h at row i and column j)"

Equation (4): This equation is also difficult. It combines \Omega_h with an orthogonal random matrix of size (r+1)x(s+1) and some other orthogonal (r+1)x(r-s) matrix. The construction is unclear to me. Since the second matrix has r-s columns, does one need to ensure that it is orthogonal to the other matrix. Further, Pham (1996) does not seem to provide a scheme to generate such a matrix, but only a matrix of size (r+1)x(r). The readers are here left alone by speculating what the correct matrices might be and how they might be generated. In doubt it should be possible to provide a method to generate the matrices in the Appendix.

Thank you for the suggestion. We will add an Appendix (Appendix B) to give details on the construction of the orthogonal matrices of Eq. (4). Lines 141-152 will be modified accordingly.

Equation (6): This equation shows that the filter relies on the covariance information contained in the matrix L. This looks particular, since usually we consider higher order to reach beyond the covariance matrix, whic his the second moment of the PDF. For me it is unclear how higher orders can be taken into account if here only the first two orders are used. Please provide an explanation for this particular formulation of the algorithm.

As observed by the Reviewer, the filter relies on the covariance information contained in the matrix L. On the other hand, moments of order higher than 2 are encoded in the matrix Omega and are not dynamically computed from the ensemble.

Line 178 and following: Line 185 states that the forecast resampling is performed at each time. It it is unclear whether this sampling is only done once after a whole forecast period or whether it is done after each time step of the time stepping scheme (The definition of 'time t_i' is not clear). Figure 2 might indicate that it is each time step, but it is not clear. Further, we know that resampling likely violates balance constraints that are contained in the model states of physical models. Does the resampling performed here also have such effects?

We thank the Reviewer for giving us the opportunity to clarify this aspect. The forecast resampling can be done at chosen pre-defined time steps that are not those of the time-stepping scheme. The sentence at L. 167-168 will be modified accordingly:

“the GHOSH filter provides an estimate of the state of a system at some pre-fixed times ti in terms of the state vector and the covariance matrix that represents the error estimate of the state vector”.

Concerning the balancing constraints, violations cannot be excluded at GHOSH forecast resampling times. As highlighted by the Reviewer, these violations occur generally in resamplings. However, Nerger et al. (2012) proposed an approach that aims at minimizing those violations, and it can be also applied in GHOSH by choosing an opportune T matrix (equation (18)). The answers to the next Reviewer comments will further expand on this subject.

Line 224: The statement "other T can be explored without affecting the algorithm (e.g., see Nerger et al. (2012))." is incorrect. Actually the cited study shows that the projections on the SEIK filter are inconsistent and that for consistency a different projection is required. This does obviously 'affect the algorithm'.

The wording of the sentence was misleading, we meant that other Ts would modify the algorithm results but its main structure would be unchanged.

We propose to modify the sentence to clarify its meaning: “other Ts can be explored without affecting the main structure of the algorithm (e.g., Nerger et al., 2012).”

In addition, see the response to the next Reviewer comment.

Equation (26): The GHOSH filter seems to use the same back-projection from the error-subspace to the state space as the SEIK filter. Nerger et al. (2012) have shown that this is inconsistent. Given that the authors are aware of the results of Nerger et al. (2012), I'm wondering why they decided to develop a new filter scheme on the basis of a filter that was shown to be mathematically inconsistent. A consistent form could be built easily following the ESTKF introduced by Nerger et al. (2012), which is likewise an error-subspace formulation.

The response to this comment would probably help to further clarify the two previous points issued by the Reviewer.

Nerger et al. (2012) showed that the SEIK is "inconsistent" in a specific sense:

*** starting from an ensemble X, the SEIK extracts an error subspace base L = X T and a covariance matrix A in that subspace.**

*** If the SEIK sampling is applied to L and A, the obtained ensemble is different from X.**

*** In this context, "inconsistent" means that "going from an ensemble to an error subspace" and "going from an error subspace to an ensemble" are not one the inverse of the other.**

This kind of inconsistency may be not always problematic, indeed the same error pdf can be represented (in terms of mean and covariance) by infinite different ensembles. ESTKF (in its deterministic version) is more consistent in the sense that its sampling strategy provides exactly the same ensemble if no observations are available, and, if the assimilation does not change dramatically the error subspace, then the ensemble after assimilation would be "not too different" from the ensemble before assimilation. This feature is desirable if, for example, there are problems with physical balance constraints, as mentioned above by the Reviewer. On the other hand, in the Lorenz-96 applications presented by Nerger 2012 it has been also shown that SEIK with random rotations (applied also in GHOSH) is as good as ESTKF. Further, a GHOSH filter builded with the ESTKF approach presents a further layer of complexity, since the corresponding T matrix should change at every forecast due to the fact that the sampling depends on the principal components of the ensemble. Taking all this into account, we choose to propose the present version of the GHOSH filter.

According to the responses given to the present and the two previous Reviewer comments, we propose to add a paragraph in the manuscript discussion about possible improvements in balancing constraints that could be obtained by applying an ESTKF approach in the GHOSH filter.

Equation (31): Using a model error covariance matrix Q in this form can be inconsistent for nonlinear models unless it is applied at each time step. This relates to me earlier comment requesting a clarification whether the resampling is performed at each model time step or only at those time steps at which also an analysis step is computed. If it is not applied at

each time, please provide additional explanation why this linearized form of applying model errors is used.

The Q matrix in equation (31) is based on the idea of parametrizing with an additive Gaussian noise (with zero mean and covariance matrix Q) some of the sources of uncertainty in the forecast estimation not already accounted for by the ensemble. Following this interpretation, Q contributes to P only at the pre-fixed forecast times during the GHOSH forecast phase, and not at each model integration time step. We acknowledge that this strategy implements a relatively simple approach and that also other strategies could be applied (e.g., model parameters perturbation, as underlined by the Reviewer in a later comment).However, the strategy illustrated in equation (31) (adopted in Pham et al. (1998a,b), as the reviewer points out in the next comment) seems to be quite suitable for the Kalman filter equations, which also imply a Gaussian approximation at each assimilation. Further, the proposed strategy encompasses the concept of hybridization of P by adding a static covariance matrix representing ensemble-non-dependent uncertainties. Please, see also the responses to the next Reviewer comments, as they further expand this subject.

We propose to change the manuscript to better describe the meaning of Q:

line 260: "where Q_i is the N × N covariance matrix of an additive unbiased noise parametrizing some of the sources of uncertainty in the forecast estimation not already accounted for by the ensemble, while A_i […]"

Lines 264/265, Eq. (32): It is stated "The last term in equation (32) is one of the novel element of the present work.". This statement is not fully true. Actually the original papers about the SEIK and SEEK filters (Pham et al., 1998a,b) include the model error covariance matrix Q. There, also a projection with L is used. A difference is that Pham et al. use a projection on Q while here a projection on the inverse of Q is used. (Given that Q is a large matrix in realistic applications, the projection by Pham et al. is likely computationally more efficient.)

Thanks to this Reviewer comment, the Q projection in the GHOSH filter will be better explained. Indeed, we will clarify that our approach is different from the one adopted in Pham et al. (1998a and b) since it is non-orthogonal and it is not equivalent to the inverse of the projection proposed by Pham et al. (1998a, b):

"While the use of Q in equation (31) follows Pham et al. (1998a and b), equation (32) projects Q in a novel non-orthogonal way that is induced by the scalar product defined by Q-1. This approach can be interpreted as a form of hybridization that aims at focusing on the effects of the ensemble-non-dependent uncertainty in the ensemble error subspace."

Line 262: I cannot follow the argumentation that the addition of Q is a 'hybridization' (the manuscripts states this already in the introduction and further argues for it around line 603 in

the discussion section). The hybridization as, e.g., explained in the reviews by Bannister (2017a, 2008) is a combination of a time-dependent covariance matrix with a covariance matrix representing climatology. The combination of both is used to represent the state error covariance matrix P. This is different from adding a model error covariance matrix which represents dynamic model uncertainties. Further, the addition of the model error covariance matrix is neither new nor a requirement of the GHOSH filter. Likewise it could be applied in the SEIK filter as the publications by Pham et al. (1998a,b) showed, even though this approach seems to be hardly used nowadays. To this end I recommend to avoid the potentially misleading term 'hybrid'.

We fully agree with the Reviewer that hybridization is a combination of a time-dependent (i.e., ensemble-dependent) covariance matrix with a parametrized covariance matrix, often derived from climatology. As discussed in the previous responses, we also acknowledge that the approach to include Q proposed in Eq. (32) is similar to the one followed by Pham et al. (1998a and b). On the other hand, it is worth noticing that Eq. (31) describes the covariance matrix P as the sum of a dynamical part derived from the ensemble and a parametric part named Q. Indeed, Q is a pre-computed ensemble-non-dependent matrix that parametrizes other sources of error (e.g., model errors, undersampling errors, nonlinearity errors), and a climatological covariance matrix is a suitable candidate for Q. In this sense , both the SEIK (as proposed by Pham et al., 1998a and b) and GHOSH can be seen as "hybrid" filters, even if Pham et al. (1998 a and b) never used the 'hybrid' term.

In the manuscript, we will clarify that Q in Eq. (31) and (32) is a parametric part in the definition of P, as proposed in a previous response.

Lines 284/285: On the forecast resampling of the GHOSH filter it is stated "it takes into account the model error effects in equations (31) and (32), which are otherwise neglected". The claim that without the addition of the model error covariance matrix Q in the sampling the effect of model errors would be neglected is not fully true. In modern applications of ensemble Kalman filters, the model error is typically represented by stochastic perturbations during the model integration instead of adding a model error covariance matrix at the end of the forecast phase. (The perturbations are expected to allow for a better representation of model nonlinearity on the model errors compared to the linearized effect of adding a model error covariance matrix)

We agree with the Reviewer: the forecast resampling is not the only way to account for model errors. The sentence is indeed misleading and it will be changed in: "it takes into account the Q effects (Eq. (31) and (32)), which would be otherwise neglected."

Lines 314/315: "A reasonable choice is a polynomial weight function" - what makes the particular choice 'reasonable' and why was this form chosen? Actually, the proposed function in Eq. (41) seems to be uncommon (Most common is the 5th order polynomial of Gaspari and Cohn (2006), which is also a covariance function). Is the definition in Eq. (41) complete (it seems that f>0 for d>d_l, which should not happen)?

We meant “reasonable” in the sense that Eq. 41 is the simplest (i.e., lowest order) polynomial function with the required properties (it is equal to 1 in p and goes smoothly to zero out of the localization radius). We will change the text in order to improve the definition of f (including that f=0 for d>dl), and we will cite Gaspari and Cohn (2006) as another possible option.

“A reasonable choice is to rescale by a function that is equal to 1 in p and goes smoothly to zero out of the localization radius, e.g.,

f(d) = [...],

with d being the distance from p and dl the localization radius. Equation (41) is the simplest (i.e., lowest order) polynomial function with the required properties, but other options are viable (e.g., the 5th order polynomial of Gaspari and Cohn (2006))”,

Section 4.1 Lorenz-96 experiments

As mentioned before there are several weaknesses which should be corrected. In the current form the configuration seems to be particularly problematic for the SEIK filter, e.g. due to too short experiments and too less inflation, so that the strong improvement by the GHOSH filter relative to the SEIK filter is caused by the particular configuration.

- Perhaps it would be useful to include 'Lorenz-96 model' in the section title. Given that this is a toy model, the experiments are necessarily twin experiments.

The title will be changed according to the Reviewer suggestion.

- The model is configured with state dimension 62. While this is a valid size, this value is very unusual and I'm not aware of other publications using it. Can the authors give a reason for this particular choice which makes it difficult, if not impossible, to compare the results to previous studies? The properties of different ensemble Kalman filters when applied to the Lorenz-96 model have been assessed before; thus they can be considered as known. The SEIK filter was however not often used, but it was applied with the Lorenz-96 model e.g. in Nerger et al. (2012). It would support the results of the manuscript if the authors could present consistent result to such previous studies.

The number of variables (state dimension) N=62 has been chosen according to the maximum ensemble size of 63. Indeed, in the twin experiments, the different ensemble sizes have been chosen (lines 368-369) in order to make it possible to test different values of s, i.e., the number of principal components approximated with order h higher than two. In particular, we chose to test h=5 for s ranging from 2 to 5, corresponding to N ranging from 7 to 63. As explained at L 368, the choices on s and N are made in order to guarantee the maximum number s of components

approximated with order h=5 for each ensemble size, resulting in N=63 for h=5 and s=5.

A more common choice would have been to use N=40 (as done in Nerger et al., 2012), however N=40 would have been too small to test the s=5 case. In addition, we are interested in studying the behaviour of the filters on non-assimilated variables, thus our observation operator is different from the one adopted in Nerger et al. (2012) where all variables are observed. The setting differences preclude the quantitative comparison between Lorenz-96 experiments in Nerger et al. (2012) and in the present study, however qualitative comparisons are possible since both studies provide results that are internally consistent. Finally, thanks to the Reviewer comments, the range of twin experiments will be enlarged making the results more comparable with previous studies (see later comments). We propose to enrich the discussion about comparison of the results with other Lorenz-96 studies with the motivations and considerations proposed in the present comment response.

- The duration of the experiment is only 20 time units. This results in experiments with observations being available in intervals between 0.1 and 0.3 time unit result in only between 200 and 66 analysis steps. As it is known that the filters sometimes converge slowly, the experiments likely mainly assess how fast the filters converge compared to how well they perform after convergence. In addition, the cases with longer observation intervals are expected to be quite unstable and one needs to average over a sufficient number of analysis steps to get interpretable results. To this end, length of the experiments is much too short. E.g. Sakov and Oke (2006) run over 11000 time units while Nerger et al. (2012) run over 2500 time units. Using such longer experiments would yield comparability to previous study with the Lorenz-96 model. Actually, I experimented with the code provided in Zenodo and the case EnsSize=31, delta_obs=0.1, forget=0.7 and t_span=[20.0, 100.0] yields the same RMSE level for both filters in the second half of the experiment. This also happens for delta_obs=0.15, forget=0.7, but the spinup for the SEIK filter is slower. Thus, with longer experiments at least the asymptotic difference in the filter results becomes very small also for the shorted foreast periods in contrast to what is shown in Fig. 4. This seems to change for delta_obs>=0.2 where the SEIK filter yields larger RMSE and tends to diverge for EnsSize=31 and lower. For EnsSize=63, the RMSE seems to be nearly identical (this might be indicated by the pink color in Fig. 4, but the colorscale only allows a qualitative estimate because it scales with 1/3, 1/2, 1 so that the values at the ticks for <1 are unclear)

We really thank the Reviewer for having tested the code provided in Zenodo. Moreover, the comment about the experiment length can give us the opportunity to enlarge the experiment setups presented in the manuscript adding experiments with a longer time window. On the other hand, we believe that the results for 20 time units are relevant to evaluate the GHOSH filter performances. In particular, it is worth noticing that:

- **In the Lorenz 96 model 0.2 time units are comparable to 1 day of atmospheric dynamics (see, e.g., Grooms 2022), thus a 20 time units long experiment is comparable to 100 days.**
- **The GHOSH filter has been developed with the aim of realistic applications in oceanographic operational simulations. In this framework, an initial fast convergence of the filter is strongly beneficial to improve the simulation accuracy on temporal scales of interest. Moreover, in the slower-than-atmosphere ocean dynamics, 20 time units can be compared to even more than 100 days.**
- **In the evaluation of the RMSEs, we excluded the first 10 time units helping to exclude the initial transient phase in the comparison between the two filters.**

Concerning the Reviewer's concerns about possible results instability in the cases with longer observation intervals, we agree that this issue can affect experiments with relatively low number of analysis steps. To avoid this kind of instabilities, in the present study we used a large number of experiments for each tested DA setting. Indeed, 400 experiments have been carried out for each setting (i.e., each square in Fig. 4) and results of Fig. 4 are obtained averaging over all the 400 experiments. Having used this relatively large number of experiments ensure an adequate stability of the results, since the expected standard deviation of the results obtained over N experiments is $\sqrt{N}$ smaller than the standard deviation of a random single experiment (i.e., 20 times smaller in our study).

Based on the motivations discussed above and the suggestions proposed by the Reviewer, we propose to

- **A comment about the exclusion of the first 10 time units for the RMSE evaluation will be added.**
- **We will further comment about the stability thanks to the large number of experiments used to compare SEIK and GHOSH.**
- **We will add to the manuscript the results of a longer time window (150 time units, i.e., 2 years, based on 0.2 time units = 1 day) comparing SEIK and GHOSH tuned with their respective best forgetting factors in each setting (according to the previous shorter experiments). Results on the 150 time units are provided in Fig. R1, and show that GHOSH performances are at least as good as those of SEIK, with largest reductions of the RMSE for observation frequency=0.2 and ensemble size=31.**
- **In Fig. 4, tick values will be corrected to be more readable.**

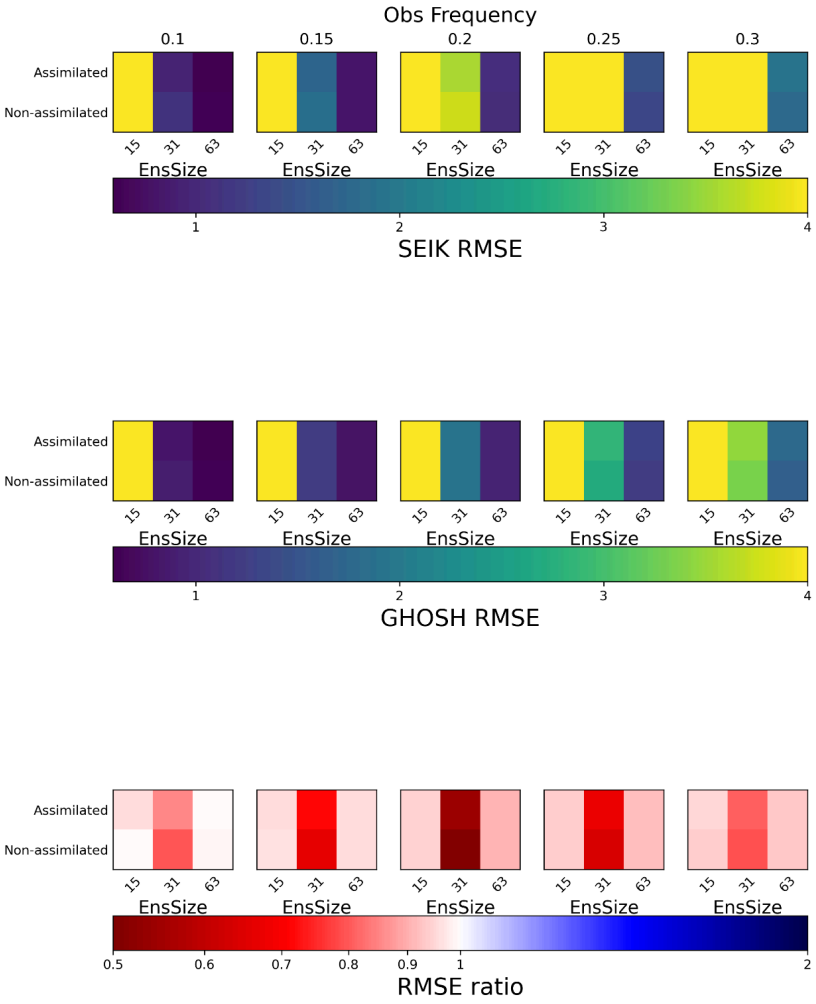

Fig. R1. Result summary of twin experiment on 120 time units. RMSE for SEIK (top), GHOSH (middle) and their ratio (bottom) are shown for different observation frequencies (columns). RMSEs are calculated for assimilated and non-assimilated observations (rows in each colour map) and for different numbers of ensemble members (columns in each colour map).

- The initial ensemble is sampled using a diagonal error covariance matrix. This is very unusual since typically second-order exact sampling is applied to a covariance matrix that is estimated from the model dynamics. I like to point out that usually when we apply a toy model like Lorenz-96, we attempt to make the experiments as realistic as possible, but a diagonal matrix would not be used in real applications. If a diagonal matrix is used to generate an ensemble of r+1 states only r elements of the state vector will be perturbed (usually this would be the first r elements of the state vector, but in the provided Python code the order is changed due to the use of a sorting routine (resorting the prescribed value 5 in all elements) and the perturbation is distributed. This distribution was not changed when repeating the experiments, e.g. elements 16 and 62 were perturbed in case of ensemble size 3), which might also influence the results). Using the common approach of a matrix sampled from a model trajectory would yield a matrix that is not diagonal. Thus, the

eigenvectors will not be the unit vectors and more elements of the state vector will be perturbed. This should lead to more realistic data assimilation experiments which are more representative. Hence, I strongly recommend to use this common approach.

In agreement with the Reviewer suggestion, we re-run all the experiments initialising the filters according to the climatological variability of the system. We will change accordingly the parts of the manuscript relevant to the filter initialization, the figures and the exact numbers in the result section. Other descriptive parts will remain untouched since the overall comparison between the two filters remains the same as in the submitted manuscript. For instance, Fig. R2 results are very similar to those shown in Fig. 4 in the submitted manuscript.

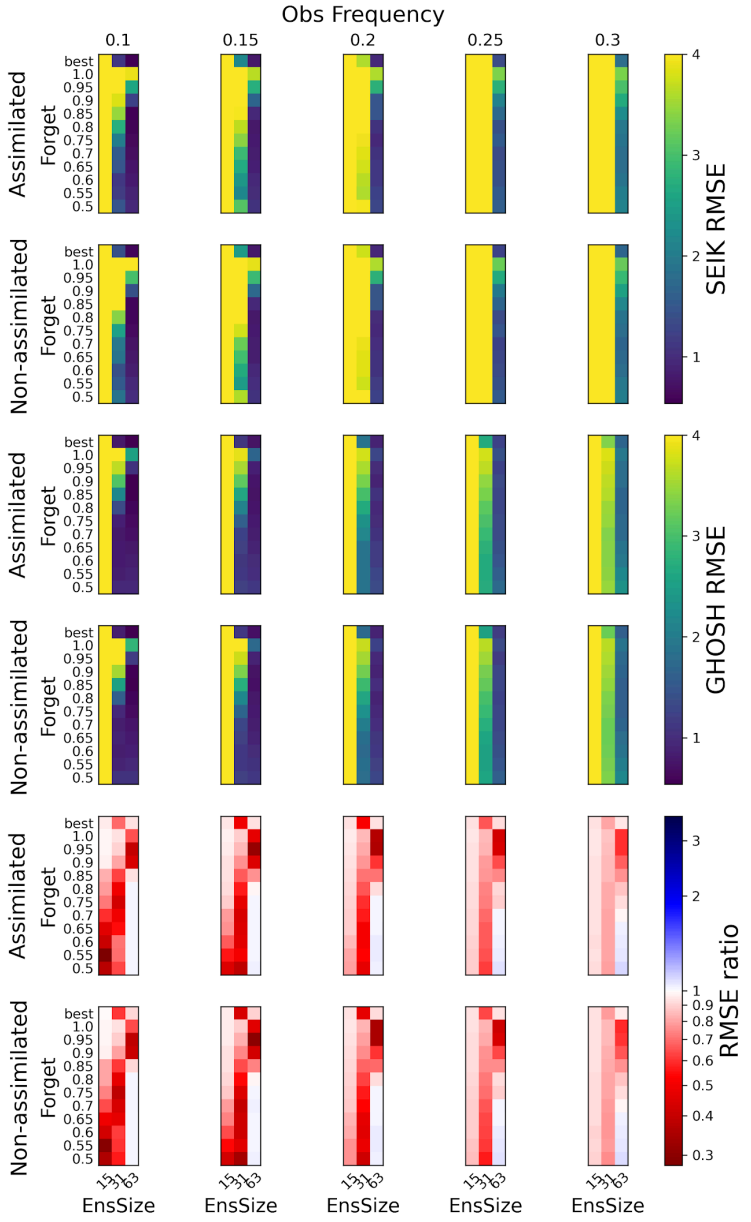

Fig. R2. Result summary of twin experiments: each square in the colour maps represents the aggregated results of 400 twin experiments, changing truth,

observations and initial conditions. The results are summarised with colour maps aggregated in six different rows, from top to bottom: SEIK RMSE of assimilated variables, SEIK RMSE of non-assimilated variables, GHOSH RMSE of assimilated variables, GHOSH RMSE of non-assimilated variables, the ratio of GHOSH RMSE over SEIK RMSE of assimilated variables, the ratio of GHOSH RMSE over SEIK RMSE of non-assimilated variables (red colour implies that GHOSH is better than SEIK). Each column of colour maps has a different observation frequency, with the numbers on the top indicating the time elapsed between each observation/assimilation. Each color map shows different forgetting factors (Forget, along the y axis) and ensemble sizes (EnsSize, along the x axis).

- The model is run using the Python library solver 'solve_ivp'. As also described by the authors as a possible reason for the long run time and instability with the SEIK filter, this solver does not use a fixed time step size. Given that the Lorenz-96 model is deterministically chaotic, its behavior will vary when the time step size is varied. To this end, it should be better to use e.g. a classical Runge-Kutta 4th order implementation with fixed time step size (which 0.05 is a typical value). This is also the typical implementation as e.g. used by Nerger et al. (2012).

As noticed by the Reviewer, the solver used in the twin experiments is the scipy solve_ivp method RK45 that makes use of a non-fixed time step. Detailed information on the solver are provided in the scipy official documentation:

- **“Explicit Runge-Kutta method of order 5(4) [J. R. Dormand, P. J. Prince, “A family of embedded Runge-Kutta formulae”, Journal of Computational and Applied Mathematics, Vol. 6, No. 1, pp. 19-26, 1980]. The error is controlled assuming accuracy of the fourth-order method, but steps are taken using the fifth-order accurate formula (local extrapolation is done). A quartic interpolation polynomial is used for the dense output”.**
- **“The solver keeps the local error estimates less than atol + rtol * abs(y). Here rtol controls a relative accuracy (number of correct digits), while atol controls absolute accuracy (number of correct decimal places)” (the state vector is referred as y).**

We chose this solver since the non-fixed time step size is an advanced feature that increases/decreases the time step size in order to keep the error under a fixed user-defined threshold. We will correct lines 519-522 in order to avoid the misleading interpretation of being the non-fixed time step that causes SEIK instabilities and longer run time. Instead, it will be clarified that it is the SEIK instability that leads to longer run time thanks to the non-fixed time step. In fact, the time step is reduced automatically by the solver to keep the error under the prescribed tolerance even in presence of stiffness induced by the SEIK instability.

“This difference in computational time is more evident and occurs more often when the forgetting factor is lower (i.e., when the inflation is more pronounced). The reason can be understood by looking at Fig. 5: in this experiment with theforgetting factor equal to 0.75, the SEIK filter presents an unstable and diverging behaviour. When it happens, the SciPy’s solve_ivp integrating routine reduces the time step size to grant the required accuracy, which comes with longer integration time. Remarkably, we did not observe any diverging behaviour in the GHOSH filter runs.”

- For ensemble size 15 the filters likely diverge (From Fig. 4 the RMSE appear to be at around 4, and Fig. 1 of Nerger et al. (2012) shows divergence for ensemble size <18, which is related to the value of the forcing parameter in the model). There is no point in comparing the performance of diverging filters. Both filters fail, but one a little less than the other. This likewise holds for the statement on ensemble size 7 in line 495. In fact a filter divergence might also happen for ensemble size 31 for the observation intervals 0.25 and 0.3 for both filters. The RMS error for the SEIK filter seems to be above 3, which might be high enough to indicate filter divergence. For the GHOSH filter, the errors seem to be between 2.5 and 3, which might also be divergence. (Unfortunately, the colorbar makes it difficult to see such details. Please consider to use a colorbar that not only varies brightness, but also the hue). In any case, it is obviously easy to check for filter divergence - if the estimated error, i.e. the ensemble spread, is much smaller that the true error (RMSE), the filter has diverged. I recommend to perform such check on all results. Further, if the RMSE is not less than the long term standard deviation of the model dynamics (which seems to be about 3.6) the data assimilation obviously failed.

We thank the Reviewer for the comments on results shown in Fig. 4, where we would like to give a fair and broad overview of the comparison between the two filters. Indeed, we think that it can be useful to provide RMSE values, even in case of non-convergence of the filters, to highlight possible relevant RMSE differences, and we do not feel that hiding some information would be beneficial for the reader. Hence, also considering the response to the Reviewer comment on line 494, we will keep the SEIK-GHOSH RMSE comparison in the last two rows of the Fig. 4 colormaps. On the other hand, we acknowledge that it is significant to highlight non-convergence cases, and we will rescale SEIK and GHOSH colorbars using a maximum value equal to 4 (i.e., nearly equal to the climatological standard deviation that is 3.7), such that it will be clear for the reader when a set of experiments does not provide any improvement with respect to the climatology. Finally, we will update Fig. 4 with a colorbar that includes hue. Both the mentioned improvements of Fig. 4 have been implemented in Fig. R2.

- Particular cases seem to be the observation intervals 0.2 and 0.25. Here it might be the that SEIK filter diverges, while the GHOSH filter still converges. This is a particular

difference, which could be interesting to analyze. (However, it might be influenced by the effect in the next statement)

We thank the Reviewer for underlining this interesting difference between SEIK and GHOSH, which will be included in the result section:"Tuned with its best forgetting factor, the GHOSH filter with 31 ensemble members improves the state estimation error compared to the climatological standard deviation of the model for every observation interval. The same is not true for the SEIK filter, which converges (i.e., RMSE lower than the climatological standard deviation) only for observation intervals shorter than 0.2."

- In the cases with observations intervals 0.2 and larger, the GHOSH filter shows smaller RMSE which are still at a high level. Since the analysis errors are show, this points to the question whether the sampling in the GHOSH filter leads to a higher ensemble variance than the SEIK filter. This would basically allow for a larger increment in the analysis step. I recommend to check whether this holds, since this effect could be unrelated to a representation of higher orders, but mainly an effect of more randomness.

We would like to address the comment of the Reviewer about the motivation of GHOSH better performances from two perspectives:

- **The set of experiments summarised in Fig. 4 can help to verify that GHOSH performances are not motivated by a larger ensemble spread. Indeed, a larger GHOSH ensemble spread would in some cases also degrade the GHOSH performances with respect to SEIK, but this is not the case. For instance, looking at the RMSDs for non-assimilated variables with 31 ensemble members and observation frequency equal to 0.1 (Fig. R2), SEIK shows best performances when the forgetting factor is equal to 0.6. Below 0.6 (i.e., with a larger inflation), SEIK RMSE increases. For the same settings, GHOSH RMSE is lower than in SEIK for the whole forgetting factor range (bottom panel in Fig. R2), even when the forgetting factor is lower (i.e., the inflation is larger) than the SEIK optimal. Meaning that GHOSH performs better than SEIK independently from the ensemble spread. If the better performances of GHOSH were related to a larger spread, an inflation larger than the SEIK optimal would degrade GHOSH with respect to SEIK (i.e., GHOSH RMSEs would be larger than SEIK RMSEs).**
- **From a theoretical point of view, it has to be kept in mind that the GHOSH sampling has exactly the same spread as SEIK's second order exact sampling, since both of them produce ensembles that preserve variances. Further, the GHOSH forecast ensemble has a smaller error in the forecast mean estimation, as proved in Appendix A. Finally, it is worth mentioning that the same arguments used for the mean prove that also forecast covariance matrix is better estimated by GHOSH with respect to SEIK.**

- For the observation intervals 0.1, 0.15 and 0.2, Fig. 4 leaves the impression that the forgetting factor of 0.7 is still too large for the SEIK filter. The optimal inflation is obviously visible when the RMSE is minimum and when it increases for even smaller values of the forgetting factor and this point should be visible in the figure.

According to the Reviewer's suggestion, we will extend the forgetting factor range down to 0.5 (Fig. R2).

- There are also some important pieces of information missing: Firstly, I suppose that the experiment is run without localization, but this is never stated in the manuscript. In addition, the choice of the actual solver in 'solve_ivp' is not mentioned. Also, the observation error variance should be explicitly stated

We thank the Reviewer for the suggestions to add more detailed information on experiments.The suggested improvements to the text will be added in the manuscript:

L 369:

"The values adopted for s are 2,3,4 and 5, respectively for 7,15,31 and 63 ensemble members. No localization has been applied, since the number of variables N is relatively small and comparable with some of the ensemble size settings."

L 350:

"The equations have been implemented in python and numerically solved using SciPy's solve_ivp routine with its default solver method (i.e., 'RK45')"

The observation error variance will be included according to a previous Reviewer's comment.

- The numerical experiments use random numbers. I recommend to re-initialize the seed for the random number at the beginning of an experiment. Only in this case, one can obtain reproducible results. One can run with different seeds to exclude that a particular interpretation results from these.

Since the results are averaged over 400 experiments per square (Fig. 4), even if they are random, the expected variability is 20 (square root of 400) times smaller than the standard deviation of the single experiment. This is sufficient to make the experiment results robust enough to be reproduced. Indeed, in many previous tests (not shown) we obtained the same (or non-significantly different) ratios between GHOSH and SEIK RMSEs. As proposed in a previous comment, we will be more explicit about the robustness of Fig. 4 results.

Section 5.1, title: As for section 4.1 I recommend to mention the Lorenz-96 model in the section title

The title will be changed as suggested.

Line 490: Please define 'best'. I suppose that the result with the smallest RMSE is meant.

It will be explicitly stated in the text that ‘best’ is the best result in terms of RMSE.

L 490: “The first line in each color map, labelled "best", represents the best result, in terms of lowest RMSE, obtained among the set of tested forgetting factors.”

Line 494: The results for ensemble size 7 were not shown "since in this case SEIK and GHOSH behave very similarly, showing very poor performances". This is the known behavior that without localization the filters diverge if the ensemble is too small. However, this likewise holds for ensemble size 15 which should also be removed. There is no point in discussing RMSEs of diverged filters.

Being GHOSH a new filter featuring a higher order of approximation compared with existing filters, we think that it deserves to be tested in heterogeneous settings. As discussed before, we think that it is fair to report a large range of results, also because they are all showing a better behaviour of GHOSH with respect to SEIK, even in case of non-convergence. We imagine that this information could be interesting for real filters applications: it could be that under certain circumstances (maybe in a small time window) a realistic model is more non-linear, or harder to successfully assimilate. Maybe the number of ensemble members could be not big enough under those circumstances, and the assimilation performances would be relatively poor. Our results suggest that, also under these circumstances, the GHOSH filter would provide a better state estimation than the second-order SEIK filter.

Fig. 3: The figure shows that the first two variables are very well estimated by both filters (and partly better by the SEIK and the GHOSH filter) in between the times ~1 to 3.5 for the first variables and even longer for variable 2. In contrast the RMSE is still large at around a value of 3. This indicates that most of the other variables of the model have a much larger error. This likely supports my earlier statement on the ensemble initialization: For an ensemble of size r+1 only r elements of the state are perturbed in the ensemble and can hence not be corrected by the data assimilation. Only if the model dynamics act for long enough time, ensemble spread will build up, but at this time, the ensemble spread might already have become too low to achieve a convergence of the filter. At this point, experiments with the Python code show that a larger inflation (e.g. forget=0.6) helps to obtain convergence. The resampling performed by the GHOSH filter might distribute the

spread faster due to its random effects. If this happens this advantage is the GHOSH filter would mainly be induced by the particular initial sampling and might not a an effect of higher orders. Please check if such an effec tis present or can be excluded.

Thanks to a previous Reviewer comment, we verified that the ensemble initialisation does not affect the GHOSH performances with respect to SEIK. Moreover, Fig. 3 will be modified accordingly.

Lines 498-501, Fig. 4: It is described that the advantage of the GHOSH filter is particularly large if the observations are very frequent. This result looks surprising for me. We know that the nonlinearity of the DA problem increases if the forecast phase is longer, but here the higher order sampling, which should be relevant for non-Gaussian ensemble distributions, is particularly good if the forecast phase is short. Also the experiments with the Lorenz-96 model show a smaller improvement of the GHOSH filter relative to the SEIK filter for longer forecasts despite the fact that these increase the nonlinearity of the DA problem. Further it looks surprising that the 'best' result of the GHOSH filter in case of the largest ensemble size of 63 is only slightly better than that of the SEIK filter, while for ensemble size 31 the difference is larger. Here, the statement on lines 500/501 that the largest improvement is obtained when "the filter can take into account a high dimensional error subspace (i.e., the ensemble size is large)" does not seem to hold. If so, why is the improvement of the GHOSH less for the largest ensemble? The larger ensemble reduces the sampling errors and this should be the case for all orders. Here it looks like that the reduced sampling errors in the second-order sampling have a stronger effect that the higher order sampling. This effect should be carefully discussed. For the figure it would be useful to clearly mark cases in which the filters diverge.

Please, note that lines 498-501 refers to the largest improvements among all the experiments, and they are not referring to the "best" case. In that context, the statement "the filter can take into account a high dimensional error subspace (i.e., the ensemble size is large)" refers to the darkest squares appearing in the 63-EnsSize columns. Lines 502-506 instead, discuss the "best" inflation case, pointing out that the darkest squares correspond to a "moderate number of ensemble members (31)".

We will specify that considerations at L. 498-501 refer to varying forgetting factors:

L. 498 "In the range of explored forgetting factors, the largest improvements (dark red, RMSE ratio 0.31) occur when [...]"

Concerning the other questions raised by the Reviewer, we can argue some explanations. Looking at the "best" inflation configuration, in the case of 31 ensemble members, the RMSE ratio between GHOSH and SEIK reaches its maximum for observations every 0.15 time units and then decreases monotonically. In the case of 63 ensemble members, instead, the ratio increases monotonically up to 0.25 time units and then decreases. Thus, in both cases, the GHOSH filter shows its better capability to manage non-linearities (since the max is not at the higher observation frequency), but at some point, the RMSE ratio stops improving. On the other hand, it

is worth mentioning that the RMSE ratio is not a perfect proxy of the filter capability of managing non-linearities. Indeed, when the error has a filter-independent part and a relatively small amount of information is available, neither of the filters can achieve high accuracy. Actually, having a part of the error that does not depend on the chosen filter (an information-dependent error) is highly expected, and it is the part of the error intrinsically dependent by the amount of information. As the information-dependent error becomes dominant with respect to the improvement given by the capability of managing non-linearities, the RMSE ratio between GHOSH and SEIK increases consequently.

Considering the ensemble size effects, in our opinion, the fact that a better RMSE ratio is obtained with 31 ensemble members instead of 63, is understandable by taking into account that the GHOSH algorithm reduces the error of the mean (other moments are positively affected, but with lesser magnitude) by exploiting higher order moments (the majority of which are imposed as hyperparameters, not estimated by the ensemble). We fully agree with the Reviewer on the fact that a bigger ensemble reduces the sampling error for all moments, but this does not affect GHOSH more than how it affects SEIK, since both of them estimate only mean and covariance from the ensemble. The point is that the GHOSH is capable of reducing the sampling error even with a smaller ensemble size. But when the error is small enough (big ensemble size), the advantage of the GHOSH is not as impressive as in the case of a smaller ensemble size. In other words, if the performances of a second-order-exact filter (i.e., SEIK in our tests) are close to the best possible performance given a certain amount of information, then it remains only a small part of the error that is possible to improve by the use of a higher order filter (such as the GHOSH filter).

Considering these last Reviewer comments, we propose to enrich and clarify the discussion of the results adding comments on possible effects of observation frequency and ensemble size.

line 512: The RMSE of the GHOSH filter is mentioned to be slightly more reduced that those of the SEIK filter. Are these differences statistically significant?

Based on the Reviewer comment, we would like to clarify that “slightly larger” refers to the RMSE reduction in the non-assimilated variables with respect to the assimilated ones. Meaning that the RMSE reduction happens in both the assimilated and in the non-assimilated variables, but slightly more in the non-assimilated ones. While the RMSE reduction from SEIK to GHOSH is quite clear, since it occurs in the majority of the settings (each tested on 400 simulations), we did not quantify the statistical significance of the RMSE ratio difference between assimilated and non-assimilated variables, and we think that it can be out of the scope of the present work. However, the fact that in all the 15 cases (5 observation frequencies for each of the 3 ensemble sizes considered) the RMSE ratio was lower in the non-assimilated variables makes it hard to believe in a random chance. In addition, also the new tests in the larger time window (Fig. R1) show the same behaviour.

Computing cost: The computing cast is discussed shortly for the OGSTM-BFM model, but no actual numbers are provided. However, it should be possible to provide timings for the Lorenz-96 model case where one can compare the time of the SEIK filter analysis step with the timing of the GHOSH analysis step and sampling. This cost should be considered separately from the cost of the model. (The model is likely faster than the model in this case since the forecasts are only between 2 and 6 times steps of a Runge-Kutta scheme)

We thank the Reviewer for this comment that will help us to a more complete vision of the GHOSH computational costs. The GHOSH asymptotic computational complexity will be added to the manuscript as well as the time to solution for the Lorenz96 experiments, with timings for filter operations and model integration.

L. 514: "From the computational point of view, the whole experiment set needed around 9 computational hours. SEIK and GHOSH schemes used 12% and 16% of the total time respectively, while the rest was committed to model integration. The GHOSH filter executes more operations than SEIK (e.g., an eigenvalue decomposition) resulting in more computational time even if the asymptotic computational complexity of the two methods is the same. However, even in the case of a relatively simple model like the Lorenz96, the time to solution is dominated by the model integration and the difference between GHOSH and SEIK only accounts for 4% of the total time. Unexpectedly, the model integration time (averaged every 100 twin experiments) when applying the SEIK filter sometimes is longer than the GHOSH filter case, up to twice the time, depending on some settings and random factors."

Lines 613-618: This paragraph contains several over-statements. The error reduction that was achieved mainly showed that the filter method works successfully, but it does not show an advantage compared to e.g. the SEIK filter (which was not applied here). The statement that non-assimilated variables are not degraded cannot be generalized. In addition, it is incorrect as Table 2 shows: For daily assimilation there is degradation of a forgetting factor of 0.8 is used. For weekly assimilation, there is degradation for h=2 and h=3 if a forgetting factor of 0.5 is used. (This comment is no longer relevant if the 3D experiments are removed as recommended)

We thank the Reviewer for having provided this comment, which will be taken into account for the manuscript on the 3D application (Part II). In particular, it will be clarified that: i) we are referring to our 3D experiment, without necessarily claiming that our results are generalizable, and ii) the observation error is optimised for one of the 3D simulations only. Thus, it can be expected that the others suffer from degradation. In this framework, we will better discuss the results of the sensitivity analysis in the 3D application.

Line 623: "it has been shown that the GHOSH increased accuracy reduces instabilities and numerical divergence". This statement is not valid in this form. For the Lorenz-96 model in the particular implementation used here, the statement holds. However, this finding might be specific for the time stepping methods used here and it cannot be generalized in any form.

We will modify the sentence to highlight that this finding has been obtained in the tested configuration: "it has been shown that the GHOSH increased accuracy reduces instabilities and numerical divergence in our Lorenz-96 implementation".

Lines 656-659: Here implications the 'higher polynomial order' or the GHOSH sampling is discussed. Linked to the methods section 2, the relevance of the polynomial order was never explained, but only mentioned. As such it is unlikely that many reader can follow the discussion. Given that neither GMD nor NPG are mathematical journals, I recommend to explain the relevant aspects already in Sec. 2.

We thank the Reviewer for highlighting this issue. In the Discussion, we will add a reference to Appendix A, where the relevance of the polynomial order is proved:

“The advantage of a higher order is not limited to a better estimation of the mean state (Appendix A) but it extends to the covariance matrix of the error probability distribution.”

Moreover, the novel Appendix C containing explicit calculation in a simple case, will help the reader to better understand the idea behind the high order sampling.

Lines 663-667: " this strategy might lead to inaccurate estimations because the model error covariance matrix Q_i is not taken into account in the interpolation". This statement does again not take into account that there are other possibilities to apply model errors in the ensemble integration. When stochastic perturbations are applied, model errors is taken into account in the 'interpolation' that is done at the analysis time. This then also holds for nonlinear observation operations mentioned in line 666.

The words “model error” will be removed, in order to avoid misleading interpretation of the statement, which refers to Q_i (representing climatological covariance, model error in form of covariance matrix, or other forms of parametric error covariance) and not to any strategy used to take into account the model error.

“this strategy might lead to inaccurate estimations because the covariance matrix Q_i of equation (31) is not taken into account in the interpolation. For this reason, if Q_i is not zero, GHOSH applies the observation operator to a new ensemble, resampled between the forecast and analysis phase.”

Lines 671-672: Here a relation of the constant ensemble weights in the GHOSH filter to the weight in particle filters a drawn. I cannot see any relation because the weights have a different meaning. In the GHOSH filter, the weights are used to represent higher order moments, while in the particle filter the weights represent likelihoods. I recommend to remove this statement. (Please note that the reference to Bocquet (2010) is invalid as was discussed for the Introduction)

We agree with the Reviewer on the fact that particle filter (PF) weights represent likelihoods, and that the scope of using weights are different in the two approaches. We will modify the sentence to highlight those aspects and to avoid unnecessary parallelism. Moreover, we will include references suggested in a later comment:

"Similarly to other ensemble filters, a weighted ensemble is used in the GHOSH filter. However, the GHOSH filter differs substantially from the particle filter (van Leeuwen et al., 2019) and from other types of weighted ensemble filters (e.g., Hoteit et al., 2008, Stordal et al., 2011). In particular, in these filters weights (representing likelihood) change over time, while particles remain the same (and strategies need to be applied to resample the particles when weights "collapse", (e.g., van Leeuwen et al., 2019). In the case of GHOSH, the ensemble is used to estimate specific moments and then resampled to keep constant the weights, because those weights have special desired properties that help reduce the error of the mean estimation. In this, GHOSH has similarities with the nonlinear ensemble transform filter (NETF, Toedter and Ahrens, 2015), which similarly does a resampling to keep weights constant but with uniform weights and a second-order sampling procedure."

Lines 674-675, 698-699: Here it is argued about the computing cost, which was also discussed in the results section of the OGSTM-BFM model, where no actual numbers were provided. Actually, it should be possible to provide timings for the Lorenz-96 model case where one can compare the time of the SEIK filter analysis step with the timing of the sampling plus analysis step of the GHOSH filter. This cost should be considered separately from the cost of the model. As mentioned before, simply arguing that the time for the model is much higher than that for the filter and sampling is superficial and should be avoided.

Thanks to this and a previous Reviewer comment, the computational complexity of the GHOSH scheme will be mentioned in Section 3, while the timing will be added to Lorenz96 experiment and discussed in the corresponding result section. We also propose the following changes to the text:

L. 673: "Finally, the computational complexity of the GHOSH filter is comparable to other second order deterministic filters (e.g., SEIK, ETKF). The computational cost, as in most ensemble methods, mainly depends on the ensemble size, which multiplies the most demanding model integration cost (Nerger et al. (2005)). The performance of

GHOSH and SEIK implemented in Lorenz96 showed that the cost of the GHOSH is no more a concern than in other second-order data assimilation schemes"

L. 697: "From the computational point of view, GHOSH and SEIK have the same asymptotic complexity. Our implementations proved to be comparable in terms of computational cost, which is dominated by the model integration cost that scales with the ensemble size."

Lines 688-689: "the GHOSH filter showed improved capacity to take into account non-linearities by a lesser need of inflation with respect to SEIK". Here, I don't see how inflation should be related to non-linearity. This was never assessed in the manuscript and I don't think this is a common relationship. As such, the statement should be removed.

The sentence was misleading, since the relationship between nonlinearity and the need for inflation was totally implicit. Indeed, our comment was based on the fact that non-linearity introduces errors not considered by ensemble filters, which becomes overconfident and needs inflation to compensate (see e.g., Raanes et al., 2019; Bocquet et al., 2015; Rainwater and Hunt, 2013). Thus, it can be inferred that the more a filter is capable of tackling non-linearities, the less it needs for inflation (up to some extent, since overconfidence is also related to the ensemble size). The Lorenz 96 experiments showed that GHOSH required less inflation compared to SEIK to achieve comparable performances (Fig. 4), and in this sense we think that the GHOSH filter showed improved capacity to take into account nonlinearity.

We propose to add a sentence in the Discussion at L. 626, to better explain the link between the inflation, nonlinearity and results shown in Fig. 4:

"The lower need of inflation can be also seen as a GHOSH improved capacity to take into account nonlinearity. Indeed, nonlinearity typically introduces errors that are not considered by ensemble filters, which needs inflation to compensate for this error covariance underestimation (see e.g., Raanes et al., 2019; Bocquet et al., 2015; Rainwater and Hunt, 2013)."

Discussion section 6: It would be useful if the authors include a discussing relating the GHOSH filters to other existing filters that are aimed at nonlinear data assimilation and pointing out the differences. Partly this is done by mentioning particle filters in line 672. However, there are filters that are closer to the GHOSH filter. E.g. the Gaussian mixture filter (Hoteit et al., 2008, Stordal et al., 2011) seems to be related, but obviously different. Perhaps, also the nonlinear ensemble transform filter (Toedter and Ahrens, 2015) shares some similarities given that this is also a transform filter.

We thank the Reviewer for pointing out useful references which will be added to the discussion:

“Similarly to other ensemble filters, a weighted ensemble is used in the GHOSH filter. However, the GHOSH filter differs substantially from the particle filter (van Leeuwen et al., 2019) and from other types of weighted ensemble filters (e.g., Hoteit et al., 2008, Stordal et al., 2011). In particular, in these filters weights (representing likelihood) change over time, while particles remain the same (and strategies need to be applied to resample the particles when weights “collapse”, (e.g., van Leeuwen et al., 2019). In the case of GHOSH, the ensemble is used to estimate specific moments and then resampled to keep constant the weights, because those weights have special desired properties that help reduce the error of the mean estimation. In this, GHOSH has similarities with the nonlinear ensemble transform filter (NETF, Toedter and Ahrens, 2015), which similarly does a resampling to keep weights constant but with uniform weights and a second-order sampling procedure.”

Code availability:

I was able to download the codes. Unfortunately, I was not able to find the GHOSH filter in the Fortran implementation. I found an option for the higher-order sampling but neither the place where it is applied and neither the analysis step of GHOSH. Maybe it would also be useful to provide a cleaner code (e.g. in the Python code there are many out-commented lines) and also some in-line documentation (there are essentially no comments in the codes which makes reading them very difficult). In any case, this would just be 'nice', but a paper acceptance would not depend on this.

We thank the Reviewer for having downloaded the codes. Concerning the issues raised by the Reviewer:

- **The GHOSH filter originates from a tailored Fortran implementation of the SEIK filter. After enough modifications, it was no longer recognizable as SEIK, and we named it “GHOSH”. For that reason, the GHOSH routines are still named SEIK in that original version of the Fortran code used for the OGSTM application. We will add a “readme” file in the repository to inform possible users about subroutine naming.**
- **The Python code has been substantially improved and, following the Reviewer suggestion, many out-commented lines have been removed. All the GHOSH filter routines are coded in a total of around 100 pythonic lines that should now be pretty easy to follow after reading the manuscript.**

Small things:

- line 162: 'identical' -> 'identity'

- line 248: \Omega_i is not computed in Eq. (6), but only used there. Please rephrase

- line 335: 'by' -> 'from'

- line 492: 'lines' -> 'rows' (in general in all description of Fig. 4)

- line 611: 'up to 3 times better RMSE': 'better' is not well defined. Its better to quantity as e.g. 'up to X% lower RMSE'

All the “small things” highlighted by the Reviewer will be revised in the manuscript.

References

Pham, D. T., Verron, J. and Gourdeau, L. 1998b. Singular evolutive Kalman filters for data assimilation in oceanography. C. R. Acad. Sci., Ser. II 326(4), 255–260

Pham, D. T. 2001. Stochastic methods for sequential data assimilation in strongly non-linear systems. Mon. Wea. Rev. 129, 1194–1207.

Rainwater, Hunt, Ensemble data assimilation with an adjusted forecast spread, Tellus A, 65:19929, 2013

Hoteit, I., Pham, D.-T., Triantafyllou, G. and Korres, G. 2008. A new approximate solution of the optimal nonlinear filter for data assimilation in meteorology and oceanography. Mon. Wea. Rev. 136:317-334.

Stordal, A. S., Karlsen, H. A., Nævdal, G., Skaug, H. J. and Valles, B. 2011. Bridging the ensemble Kalman filter and particle filters: The adaptive Gaussian mixture filter. Comput. Geosci. 15:293-305.

J. Toedter, and B. Ahrens, 2015: A second-order exact ensemble square root filter for nonlinear data assimilation. Mon. Wea. Rev., 143:1347-1367

Response references

Bannister, R. N.: A review of operational methods of variational and ensemble-variational data assimilation, Quarterly Journal of the Royal Meteorological Society, 143, 607–633, https://doi.org/10.1002/qj.2982, 2017.

Bocquet, M., Raanes, P. N., and Hannart, A.: Expanding the validity of the ensemble Kalman filter without the intrinsic need for inflation, Nonlinear Processes in Geophysics, 22, 645–662, https://doi.org/10.5194/npg-22-645-2015, 2015.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G.: Data assimilation in the geosciences: An overview of methods, issues, and perspectives, Wiley Interdisciplinary Reviews: Climate Change, 9, e535, https://doi.org/10.1002/wcc.535, 2018.

Grooms, I.: A comparison of nonlinear extensions to the ensemble Kalman filter, Comput Geosci, 26, 633–650, https://doi.org/10.1007/s10596-022-10141-x, 2022.

Houtekamer, P. L. and Zhang, F.: Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation, Mon. Wea. Rev., 144, 4489–4532, https://doi.org/10.1175/MWR-D-15-0440.1, 2016.

Lahoz, W. A. and Schneider, P.: Data assimilation: making sense of Earth Observation, Front. Environ. Sci., 2, https://doi.org/10.3389/fenvs.2014.00016, 2014.

Leeuwen, P. J. van: Particle Filtering in Geophysical Systems, Monthly Weather Review, 137, 4089–4114, https://doi.org/10.1175/2009MWR2835.1, 2009.

van Leeuwen, P. J., Künsch, H. R., Nerger, L., Potthast, R., and Reich, S.: Particle filters for high-dimensional geoscience applications: A review, Quarterly Journal of the Royal Meteorological Society, 145, 2335–2365, https://doi.org/10.1002/qj.3551, 2019.

Martin, M. J., Balmaseda, M., Bertino, L., Brasseur, P., Brassington, G., Cummings, J., Fujii, Y., Lea, D. J., Lellouche, J.-M., Mogensen, K., Oke, P. R., Smith, G. C., Testut, C.-E., Waagbø, G. A., Waters, J., and Weaver, A. T.: Status and future of data assimilation in operational oceanography, Journal of Operational Oceanography, 8, s28–s48, https://doi.org/10.1080/1755876X.2015.1022055, 2015.

Raanes, P. N., Bocquet, M., and Carrassi, A.: Adaptive covariance inflation in the ensemble Kalman filter by Gaussian scale mixtures, Quarterly Journal of the Royal Meteorological Society, 145, 53–75, https://doi.org/10.1002/qj.3386, 2019.

Rainwater, S. and Hunt, B. R.: Ensemble data assimilation with an adjusted forecast spread, Tellus A: Dynamic Meteorology and Oceanography, 65, 19929, https://doi.org/10.3402/tellusa.v65i0.19929, 2013.

Roth, M., Hendeby, G., Fritsche, C., and Gustafsson, F.: The Ensemble Kalman filter: a signal processing perspective, EURASIP J. Adv. Signal Process., 2017, 56, https://doi.org/10.1186/s13634-017-0492-x, 2017.

Vetra-Carvalho, S., van Leeuwen, P. J., Nerger, L., Barth, A., Altaf, M. U., Brasseur, P., Kirchgessner, P., and Beckers, J.-M.: State-of-the-art stochastic data assimilation methods for high-dimensional non-Gaussian problems, Tellus A: Dynamic Meteorology and Oceanography, 70, 1–43, https://doi.org/10.1080/16000870.2018.1445364, 2018.