Comments from editor

Thank you for your considerable effort in improving this manuscript. I acknowledge that the quality has improved and major points have been addressed. One of the reviewers suggests only minor revisions at this point, while the other still sees major points that have to be clarified. I suggest that you address their comments, including the effect of reordering the realizations of the error distribution. I believe that this additional illustration would be very useful for the reader and help to understand the mechanics of the presented approach. After having considered these changes, I believe that this manuscript will eventually be a valuable contribution to stimulate new techniques of efficient parameter estimation under uncertainty.

We thank the editor and reviewers for the overall positive assessment of the manuscript.

According to the comments of the 1st reviewer, we have revised all the descriptions related to BATEA and IBUNE and highlighted the contribution of this study on the introduction of rank estimation and the secant method in the input error identification.

Based on the comments of the 2nd reviewer, we have moved the Appendix B into the main text as Section "2.3 Bayesian inference of input uncertainty and the BEAR method" to explain the theoretical basis of the BEAR method in the Bayesian framework and added Section"4.2 The effect of reordering on the error realization" to clarify the mechanics of reordering step in the input error identification.

We appreciate these useful comments, which we believe have helped improve the quality of the manuscript and inspired more understandings of the BEAR method from different angles. We have responded to each point in turn in the following sections. The comments from the reviewer are provided in blue text and our responses are organized point-by-point in black text. The manuscript text after the proposed changes is shown in "*black italics*". The number of the line, equation and section refers to the revised version of the manuscript without track changes, shown in yellow highlight.

Anonymous Referee #1

I think the manuscript clearly improved over the last iterations. Yet, I still struggle with the discrepancy between expectations that are raised in the title, abstract and introduction, and what is presented in the article, i.e. that BEAR is as comprehensive as BATEA and IBUNE and therefore an alternative to these two. This impression is fostered by phrases like L.53-54: "From the above, a new strategy should be developed to avoid high dimensional computation and ensure the accuracy of error identification." or in chapter "2.4 Comparison with other methods", where, first, BATEA is discarded with the plain statement that it might run into dimensionality problems and then BEAR is presented as a modification to IBUNE (Lines 172-173 "The "BEAR" method adds a reordering process into the "IBUNE" method to improve the accuracy of input error quantification") .

To overcome this problem, the authors should make last modifications:

1) From the beginning on present the BEAR method as add-on to SMC or potential extension to existing methods like IBUNE, but not as a stand-alone, equal alternative to BATEA or IBUNE

2) Only keep BATEA in the introduction as benchmark reference and for discussion, but not in the comparison – if BEAR shall be presented as an alternative, the authors should have run BATEA as reference. This includes deleting repetitions like lines 311ff in the Discussion chapter 4: "There are two ways to impose restrictions. One is to regard errors and model parameters as a whole in calibration, like the BATEA framework (Kavetski et al., 2006), resulting in a high dimensional computation." Otherwise, there is the impression as if there was an actual comparison made.

Thanks for your additional review and comments. We agree with this summary that the BEAR method is a modification of the input uncertainty quantification of the IBUNE framework, but not as comprehensive as the full implementation of BATEA and IBUNE. To avoid overselling the approach, we have revised all descriptions related to BATEA and IBUNE, as follows:

1) The descriptions in line 53-56 have been modified as follows:

*"Therefore, a modification should be made in the IBUNE approach to improve the accuracy of input error identification.*

*To complete this goal, this study attempts to add a reordering strategy into the IBUNE framework and names this developed algorithm as Bayesian Error Analysis with Reordering (BEAR).”* (line 54-57)

2) The descriptions in Section 2.2 “Considering the limitations of BATEA and IBUNE framework discussed in the introduction, an improved strategy should be explored to avoid the high dimension challenge and meanwhile promote the error estimation accuracy.” have been modified as follows:

“*Unlike directly estimating the input error value via existing methods, this study attempts to transform the input error quantification into the rank domain*.” (line 108-109)

3) The descriptions about BATEA have been deleted in “2.5 Comparison with other methods” to focus on the comparison between IBUNE and BEAR.

“*In this study, three methods, including the “Traditional” method, “IBUNE” method and “BEAR” method, are compared to evaluate the ability of the BEAR method in estimating the model parameters and quantifying input errors. The “Traditional” method regards the observed input as error-free without identifying input errors (i.e. Eq. (2)), while the other two methods employ a latent variable to counteract the impact of input error and derive a modified input (i.e. Eq.(3)). In the “IBUNE” method, potential input errors are randomly sampled from the assumed error distribution and filtered by the maximization of the likelihood function (Ajami et al., 2007). Although the comprehensive IBUNE framework additionally deals with model structural uncertainty via Bayesian Model Averaging (BMA), this study only compares the capacity of its input error identification. The “BEAR” method adds a reordering process into the “IBUNE” method to improve the accuracy of input error quantification*.” (line 195-203)

4) The discussion in Chapter 4 (repetition in lines 311ff) has been modified as follows:

“*In addition, rank estimation can make better use of the knowledge of the input error distribution. In a direct value estimation, it is difficult to keep the overall error distribution the same when the errors are updated in the calibration. The estimated errors are more likely to compensate for other sources of errors to maximize the* *likelihood function and subsequently be overfitted. By contrast, in rank estimation, the errors at all the time steps are sampled from the pre-estimated error distribution first and then reordered. Whatever the error rank estimates are, they always follow the pre-estimated error distribution, and the compensation effect will be limited. In the IBUNE framework (Ajami et al., 2007), the errors are also sampled from the error distribution, but not reordered. Thus, the error precision at each time step cannot be guaranteed. In the BEAR method, adjusting the sampled errors according to the inferred error rank reduces the randomness of the error allocation in the IBUNE framework (Ajami et al., 2007), which significantly improves the accuracy of the error estimation (as demonstrated by much higher correlations than the IBUNE method in Fig. 2(2)).*

*Unlike formal Bayesian inference, the rank estimation does not update the posterior distribution of the input errors, but optimises their time-varying values through the relationship between the input error rank and corresponding model residual error. The rank estimation is implemented after the model parameters have been updated and the model residual error depends on the input error estimation. Thus, the reordering strategy identifies the optimal input error rank conditional to the model parameters, effectively considering the interaction between the input error and the parameter error. This is akin to calibrating the input errors along with the model parameters in the BATEA framework (Kavetski et al., 2006).”* (line 340-355)

5) The conclusion in Chapter 5 has been modified as follows:

“*The novelties of this algorithm are: (1) The estimation focuses on the error rank rather than the error value, using the constraints of the known overall input error distribution and then improving the precision of the input error estimation by optimising the error allocation in a time series. (2) The introduction of the secant method addresses the nonlinearity in the WQM transformation and updates the error rank of each input data according to its corresponding model residual.”* (line 410-414)

Personally, I find the treatment of the errors still rather arbitrary – by design, the method will minimize residuals between model and observations. I doubt that this is the intention of Bayesian methods. Therefore, it is also no surprise that NSE values from BEAR are often higher (see e.g. Fig. 1 and 2) or the variance of residuals is lower (see Fig. 4).

Thanks for raising this concern. Based on your comments, we included a section that clarifies the BEAR modification compared to classical Bayesian inference, section "2.3 Bayesian inference of input uncertainty and the BEAR method" to explain it. Regarding the issue of minimizing residual between model and observations, the effectiveness of our approach does depend on the assumption that the input error is dominant in the residual error. The explanation is as follows:

"*The secant method in the BEAR algorithm is applied to find the optimal ranks of input errors to minimise the model residual errors towards zero, as characterised by the minimized Residual Sum of Squares (RSS). Minimizing the RSS imposes the same effect as maximizing the likelihood function. The effectiveness of this step in quantifying the input errors is based on the assumption that the input error is dominant in the residual error and then minimizing RSS is the same as allocating the total error into the input errors. Otherwise, other dominant sources of errors will affect the estimation of the optimal input errors leading to poor input error identification.*" (line 156-161)

We also note that "*Unlike formal Bayesian inference, the rank estimation does not update the posterior distribution of the input errors, but optimises their time-varying values through the relationship between the input error rank and corresponding model residual error. The rank estimation is implemented after the model parameters have been updated and the model residual error depends on the input error estimation. Thus, the reordering strategy identifies the optimal input error rank conditional to the model parameters, effectively considering the interaction between the input error and the parameter error. This is akin to calibrating the input errors along with the model parameters in the BATEA framework (Kavetski et al., 2006)*" (line 350-355)

That said, I do see two points why it still might be worth publishing: 1st) The manuscript addresses the problem "input error" that has not been addressed as much over the last years. Yet, it is a very important one, e.g. especially regarding novel data-driven machine-learning models through which input errors might be propagated without regulation since unlike mechanistic models, they do not contain physical relations that might buffer some part of the input error. 2nd) The authors propose the use of the secant method to address the problem and even if I personally do not find the presented procedure to be the "problem-solver" modelers might look out for, readers can make their mind of whether this method still bears more potential (working with ranks of errors, etc.) or not.

So considering the points above, I would consider the publication of this manuscript as a contribution to the discussion about issues and ideas in addressing input errors, but not as a proven framework to tackle them that outperforms existing methods.

Thanks for your suggestion. We agree with that the contribution of this study should focus on the introduction of rank estimation and the secant method in the input error identification. Therefore, we have revised the descriptions to highlight these points and deleted the statements on tackling the limitations of BATEA and IBUNE (see the above responses).

Anonymous Referee #2

I thank the authors for their detailed revision of the manuscript. Overall, their revision helped clarifying a lot of uncertain aspects of the algorithm, and their responses to my concerns were mostly satisfactory, but work still remains to be done. Unfortunately, I fear that my main concerns about the theoretical foundations and consequences of the error re-ordering remain inadequately addressed. The derivation provided in Appendix B does not sufficiently clarify these points eithers: the core question (*how exactly the re-ordering affects the base error distribution's statistical moments or hyperparameters, and what the consequences are in Bayesian terms*) remains unaddressed. Since I take it that the authors would like to stick with the error re-ordering approach, I would argue that this leaves you with two possible pathways:

1) Provide a thorough theoretical derivation and in-depth investigation of what effect the error re-ordering really has, and what this means in Bayesian terms. After experimenting a bit with error re-ordering myself (see the Python code snippet below), I believe that a good start point might be couplings or measure transport (Pierre E. Jacob has a nice online lecture series on that called *Couplings and Monte Carlo*), as you seem to convert one distribution (the raw error distribution) into one with different statistical moments, one somehow moulded to the ideal error realizations.

Thank you for this constructive comment. We appreciated the lectures shared by the reviewer from Pierre E. Jacob and carefully considered the method on *Couplings and Monte Carlo*. However, we believe the goal of coupling is intrinsically different from our method and is not feasible in the rank estimation. *Coupling* aims to gain a posterior distribution with different statistical moments, while the BEAR method does not change the error distribution (by sampling the errors from the same pre-estimated distribution), but aims to adjust the positions of sampled errors according to the inferred ranks via the secant method. In other words, re-ordering won't change the overall statistical moments on the error population (i.e. mean and standard deviation. The details have been discussed in the additional section "4.2 The effect of reordering on the error realization".

2) Alternatively, you could simply drop the "Bayesian" attribute from your study or replace it with "Pseudo-Bayesian". This might require that you adjust the

> acronym. Even without full theoretical justification, your algorithm can still provide a useful heuristic, and maybe that is enough. Even in this case, however, I believe the manuscript would benefit from an isolated analysis of what mechanically happens to the input error distribution when subject to reordering. I have provided a few thoughts on this below. As a consequence, I would recommend another round of major revisions. If you follow option (2), I recommend specifically to add a new section into the theory/methodology chapter which explores and illustrates the effects of re-ordering errors on the raw error distributions in detail (this is important for the reader's understanding of the approach– it should be part of the main manuscript, not the Appendix). To make space for this section, you could absorb some of the practical comments in the discussion (specifically, sections 4.1 and 4.3). There are also a lot of tangential comments addressing reviewer concerns throughout the manuscript which could be removed if their key points are addressed in this new section. This might also support the narrative thread of the manuscript by helping you to avoid the need to go on explanatory tangents. As I don't want to leave you hang out to dry on such a large and amorphous task, I would specifically suggest exploring a simple example case in this proposed section. Specifically:
>
> • Ignore the model (for the purpose of error reordering this is unnecessary); instead, skip straight to positing some hidden "ideal" sequence of error realizations which perfectly compensate the true residual error (similar to your figure A1); derive the corresponding ranks;
>
> • Use a simple, structured residual error sequence to make it easier to read the induced effect. The sequence doesn't have to be realistic, merely insightful;
>
> • Use a residual error distribution perfectly adjusted to the structured error you defined. In a synthetic test case, it's easy to derive an empirical cdf.
>
> • Then explore the consequences of re-ordering for time series of different length or input error distributions of different quality. Discuss the statistical moments after reordering.

We admit the sampling and reordering strategy in the BEAR method itself is not a formal Bayesian approach, but rather an additional step in the existing Bayesian inference methods to find the deterministic relationship between the model residual error and the observational error. We have updated the text to make this point explicit, with an additional section "2.3 Bayesian inference of input uncertainty and the BEAR method" that clearly articulates the relationship between our approach and the classical Bayesian approach (see the following reply). Additionally, according to your suggestion, we have deleted the section "4.3 The extension to other modeling scenarios" and moved the summary of this content into Section 5 "Conclusion and recommendation".

I have provided an example Python script for this below and appended some of its result figures for different time series lengths of 10, 100, and 1000 in Figures 1, 2, and 3. In this example, I arbitrarily assumed the true/ideal error realizations to follow a sine curve. Note that something more realistic (like random samples from a Gaussian distribution) would have also worked, but the simplicity of a sine curve makes it significantly easier to read the effect of the re-ordering.

Some thoughts on the results:

As I suspected in major comment #7 for the first round of revisions, the longer the time series, the more likely the method is to achieve a "perfect fit", so the effect of error re-ordering depends on the length of the time series. You discuss this briefly in the manuscript, but I think that this is among the most important mechanisms of BEAR, so it is worth demonstrating in isolation. For short time series (Figure 1), error re-ordering can already induce some degree of improvements by causing the marginal sample mean to follow the "ideal" error realizations; at the same time, the marginal error standard deviation decreases. This effect is exacerbated for longer time series (T=100, Figure 2, and T=1000, Figure 3). The consequence seems to be that the unordered, raw error distribution is "molded" to the ideal error realizations. I suspect (and you seem to share these suspicions in your responses) that in the limit of an infinitely long time series, error realizations would be compensated perfectly. This is important to discuss for prospective users of your manuscript, as it affects the algorithm's behaviour in somewhat unexpected ways.

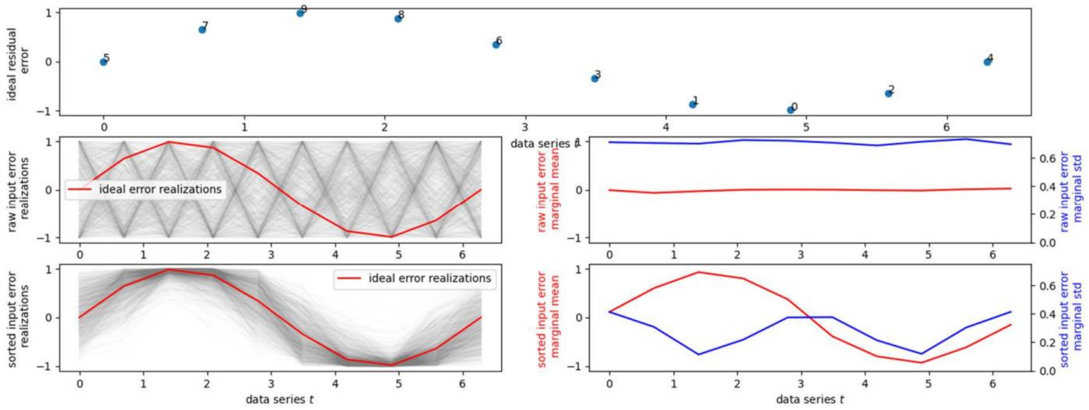

*Figure 1. Effect of error re-ordering for a perfect error distribution, an ensemble size of N=1000 and a time series length of T=10. The upper subplot shows the "ideal" realizations to compensate some residual error. The left centre plot shows N unordered realizations of an error distribution with the correct statistical moments of the ideal realizations (obtained by forming a cdf for a full sine wave). The right centre plot shows the marginal mean and standard deviation at each time step. The left bottom plot shows the N error realizations in the subplot above after ordering, and the right bottom plot shows the corresponding marginal mean and standard deviation.*

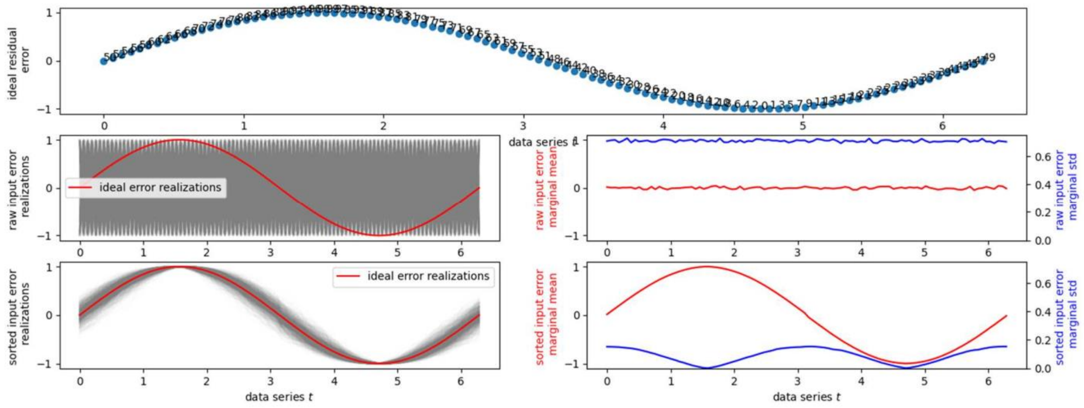

*Figure 2. Effect of error re-ordering for a perfect error distribution, an ensemble size of N=1000 and a time series length of T=100. The upper subplot shows the "ideal" realizations to compensate some residual error. The left centre plot shows N unordered realizations of an error distribution with the correct statistical moments of the ideal realizations (obtained by forming a cdf for a full sine wave). The right centre plot shows the marginal mean and standard deviation at each time step. The left bottom plot shows the N error realizations in the subplot above after ordering, and the right bottom plot shows the corresponding marginal mean and standard deviation.*

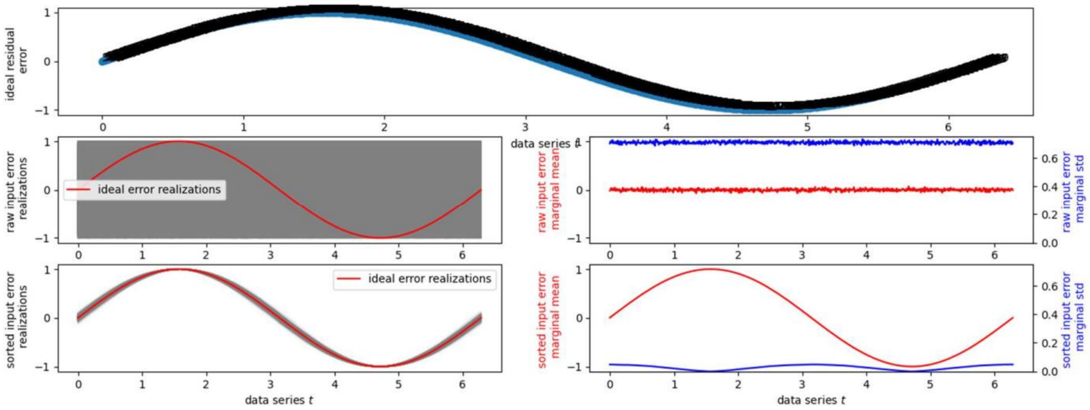

*Figure 3. Effect of error re-ordering for a perfect error distribution, an ensemble size of N=1000 and a time series length of T=1000. The upper subplot shows the "ideal" realizations to compensate some residual error. The left centre plot shows N unordered realizations of an error distribution with the correct statistical moments of the ideal realizations (obtained by forming a cdf for a full sine wave). The right centre plot shows the marginal mean and standard deviation at each time step. The left bottom plot shows the N error realizations in the subplot above after ordering, and the right bottom plot shows the corresponding marginal mean and standard deviation.*

Other interesting things to visualize might be what happens if the error distribution is not perfect (for this, just replace "vals = dist(np.random.uniform(size=(1000, resolution)))" in the code with some other distribution). You already show this indirectly in your models, but demonstrating this effect in isolation rather than through the lens of performance metrics might be a lot clearer. Feel free to take inspiration from my example code or use it directly. I have attached it at the end of this manuscript.

Of course, this code snippet just demonstrates what happens when we are re-ordering error realizations, not how you arrive at the error ranks and their interaction with the parameter inference in the first place (which are potentially additional topics to discuss). hope that even if you decide to follow option (1), this snippet might give you some ideas on where to start with the Bayesian justification. Good luck!

We appreciate your open, detailed analysis of the impact of the time series length, which helped inspire us to analyse the effect of reordering. We added one section "4.2 The effect of reordering on the error realization" and summarized all the results in the following Figure 5 to demonstrate the mechanics of error reordering.

First, we want to point out that we agree that the length of the time series will affect the efficacy of the method. As Figure 5 shows, larger data lengths bring more accurate estimation of model parameter and input errors. The reason for this has been discussed in Section 4.2. Usually, we are operating on the assumption that we have a representative length of sampled values, certainly enough that the model parameters are estimated properly, and that this is not unusual for the applications we are looking at (where we would need at least a year of data to estimate the parameters).

Considering the changes of the marginal mean of the standard deviation, we discuss this in Section 4.2 as follows:

"*Figure 5 demonstrates the mechanics of input error reordering in the BEAR method and input error filtering in the IBUNE method to understand their effects on the input error realizations and model parameter estimation. The 1st sequence represents the situation where the raw input errors are randomly sampled from the pre-estimated error distribution, therefore, their marginal means and standard deviations are the same as the parameters of overall error distribution (demonstrated as the cyan lines in column (c)). In the later sequence. these errors are optimised via different methods. In the IBUNE method, these sampled input error series are selected by the maximized likelihood function and the interval of input errors become a little converged (in (b1)) and their marginal standard deviations reduce slightly (in (c1)). However, in the BEAR method, these input errors are rcordered according to the inferred ranks via the secant method, and the reordered errors gradually converge to the true values (represented by the blue interval are near the red line in (b2)). Therefore, their marginal means are similar to the true values and their marginal standard deviations reduce to zero (in (c2)). In the BEAR method, the promotion of the input error identification in the sequential updating will improve the model parmaeter estimation, represented by the posterior distribution of model parameter b converging to the true value in (a2). While in the IBUNE method, the identification of input errors is not precise and the bias of the model parameter still exists in (a1).*

*The data length can affect the efficacy of the BEAR method but impose little effect on the IBUNE method. The IBUNE method takes advantage of the stochastic errors and keeps the marginal error distribution almost constant. The input error realization at each time step seems independent, only filtered by the overall likelihood function. Therefore, the number of sampled errors does not matter in the IBUNE method.*

*However, in the BEAR method, the input errors at all the time steps are not sampled independently, they are from one sample set. Therefore, before or after reordering, all errors will keep the same statistical features of the input error distribution, and only their marginal distribution changes due to the convergence to the unknown true values. Figure 5 (b2) demonstrates that when the data length (the same as the error number) is small, the input error estimation might be biased from the true values. This likely arises from the above-mentioned sampling bias or the impacts of the model parameter error because the sampling bias reduces with the larger number of error samples and the impacts of parameter error are more likely to be offset when the data length is long.”* (line 357-379)

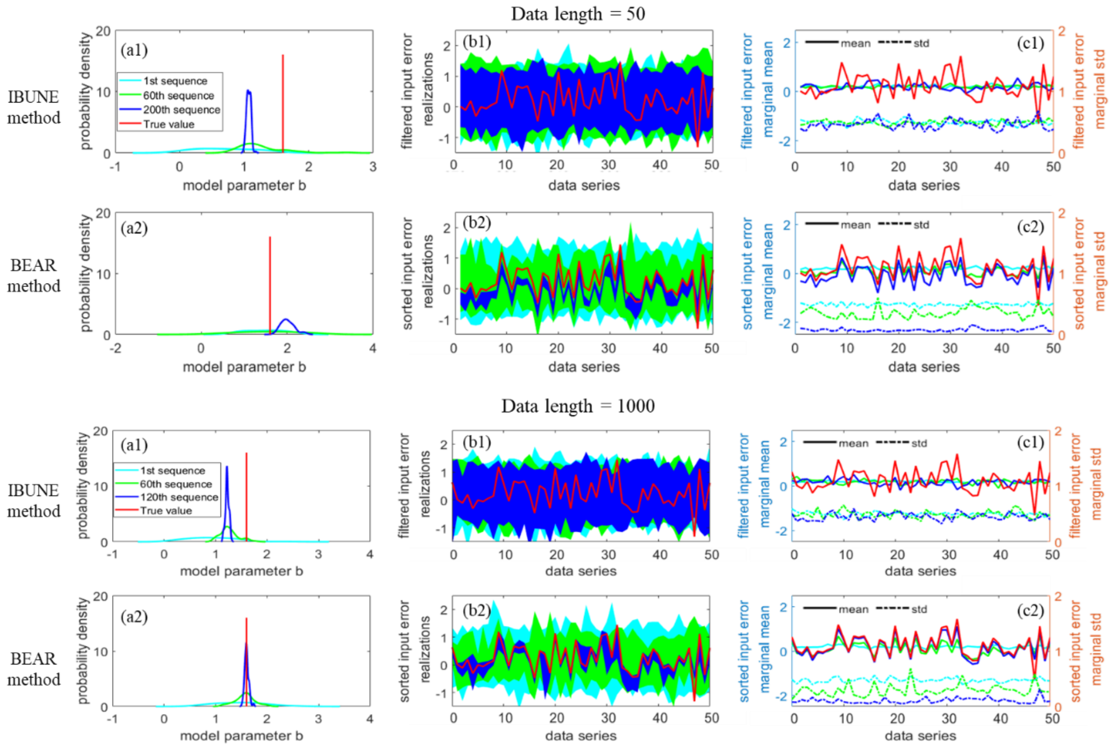

**Figure 5 Comparision of the results for scenario 1 in the synthetic case 1, the ensemble particle size of *N*=100 at different sequences of calibration (represented in different colours), via different methods (row 1: IBUNE method; row 2: BEAR method) and under different data lengths in calibration (the upper group: data length is 50; the lower group: data length is 1000, selects 50 data the same as the upper group to show). Column a shows the probability density of model parameter *b* at different sequences of calibration. The other model parameters have the same pattern of change and thus there is no need to show. Column b shows the value interval of input error realizations of 100 particles after reordering in the BEAR method or filtering in the IBUNE method and Column c shows the corresponding marginal mean and standard deviations at each time step. The 1$^{st}$ sequence (in cyan) shows the raw input errors of random sampling, before reordering or filtering.**

**Specific comments** (any line numbers I list correspond to the non-track-changes manuscript):

Line 47: the same multiplier applied to one storm event

Maybe "the same multiplier applied to all time steps of one storm event" might be better, if I understand this part correctly; using "same" for a singular object ("storm event") sounds a bit strange.

Thanks for your suggestion. This has been changed as recommended. (line 48)

Line 184: the time scale is typically set as daily and the spatial scale is set as the catchment

This is not particularly clear. I assume you simulate in daily time steps, and aggregate the catchment's (presumably) surface area into a single spatial unit? If so, it might be better to replace this with "thus, we use daily time steps and consider the catchment a single, homogeneous spatial unit" or something along these lines.

Thanks for your suggestion. This has been clarified as recommended. (line 214)

Lines 302: From this point of view, it is more efficient to estimate the error rank than estimate the error value, This sentence ends on a comma, not a period.

This has been corrected. (line 332) Thanks.

Line 309-311: Besides, to avoid the high-dimension calculation, modifying each input error according to its corresponding residual error only works in the rank domain. In the value domain, if there is no constraint on the estimated input errors, they will fully compensate for the residual error to maximize the likelihood function and subsequently be overfitted.

This requires more discussion in the revised manuscript, as it seemingly contradicts what you write in the paragraph immediately prior: In the previous paragraph, you recommend sampling error realizations repeatedly and selecting the optimal realization to overcome "sampling bias" and improve the fit to the actual observations. However, in this paragraph you praise this very same sampling bias for preventing overfit. These are contradictory messages: provided you get the ranks right, if you were to resample an infinite number of times, you would eventually get an error realization which compensates the true error perfectly (even if your input error distribution is a really poor approximation to the "true" error distribution), thus negating your protection against overfit. The fact that this protection against overfit depends on the length of the time series might not be that much of an issue if you interpret your approach merely as a heuristic, but even in this case you need some practical guidelines on when to re-sample for short time series. The proposed dedicated section might help clearing some of this confusion up.

Thanks for your concerns. We believe the manuscript needs more clarification on the difference between "sampling bias" and "compensation effect".

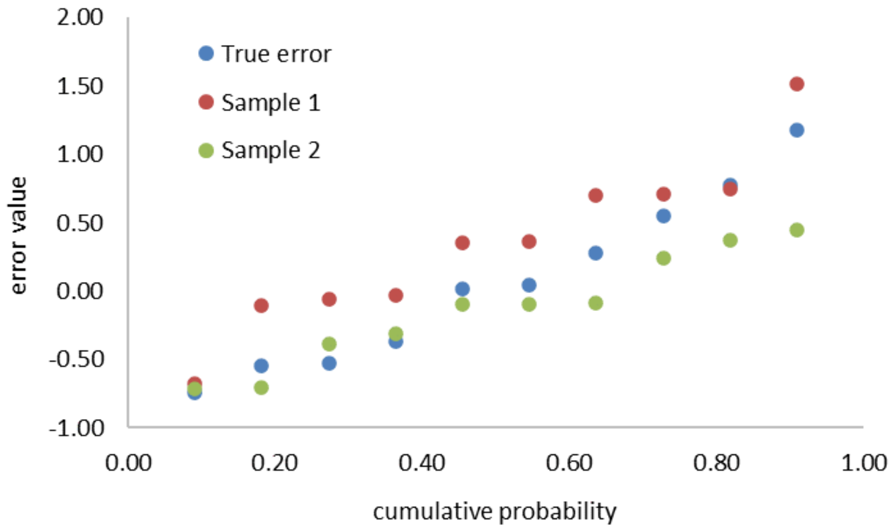

*Figure R1 The cumulative probability of sampled errors and true errors (sample number =10 )*

In Figure R1, three groups of errors are sampled from the same normal distribution (*N(0,0.5$^2$)*), but for the same order (with the same cumulative probability distribution), the sampled errors from different groups (in different colours) are not the same. The value difference at the same order is referred to as "sampling bias". In order words, even if all the error ranks are estimated right, there is still a difference between the reordered error series and the true values, which comes from the sampling step, not the reordering step. The sampling bias is more significant when the error number is smaller or the variance is larger. Repeatedly sampling and selecting the optimal set can reduce the potential impact of this.

It should be noted that sampling bias cannot prevent overfitting. Here, these sentences aim to explain that overfitting (referring to the compensation effect) is more likely to appear in value estimation due to the lack of constraint on the error distribution. In value estimation, it is difficult to keep the overall distribution of the sampled errors the same when the errors are updated in calibration, then the compensation for the residual error is more likely to appear. By contrast, in rank estimation, the errors at all time steps are sampled from the pre-estimated error distribution. Whatever the error rank estimates are, they always follow the error distribution, and the compensation effect will be reduced.

From the above, this has been clarified as follows:

"*For the same error distribution and the same cumulative probability distribution (corresponding to the same error rank), the errors sampled at different times could be largely different, especially for a small sample size (depending on the data length) or a large $\sigma$ of the assumed error distribution. This problem can be addressed by selecting the optimal solution from multiple samples according to the maximum likelihood function.*" (line 333-337)

*"In addition, rank estimation can make better use of the knowledge of the input error distribution. In a direct value estimation, it is difficult to keep the overall error distribution the same when the errors are updated in the calibration. The estimated errors are more likely to compensate for other sources of errors to maximize the likelihood function and subsequently be overfitted. By contrast, in rank estimation, the errors at all the time steps are sampled from the pre-estimated error distribution first and then reordered. Whatever the error rank estimates are, they always follow the pre-estimated error distribution, and the compensation effect will be limited."* (line 340-345)

Line 320-322: Thus, unlike formal Bayesian inference, the BEAR method does not update the posterior distribution of the input errors, but identifies the input error through the deterministic relationship between the input error and model parameter.

As far as I can see, this is the first time you mention that BEAR is not a formal Bayesian inference method, so I suspect you would go for option (2). In any case, mentioning this in the discussion for the first time is a bit late: Something this important should be stated earlier, as early as the introduction or abstract. If you add the section exploring the consequences of error reordering, this would also be a good place to elaborate on this.

Aside from this manuscript structuring argument, the statement in these lines is also unfortunately wrong. Refer to the attached figures for demonstration. I also elaborate more on this two comments below.

Thanks for this insightful comment.

We admit this statement is a bit sudden. We have clarified this as follows to show its connection with the preceding discussion in Section 4.1.

"*Unlike formal Bayesian inference, the rank estimation does not update the posterior distribution of the input errors, but optimises their time-varying values through the relationship between the input error rank and corresponding model residual error. The rank estimation is implemented after the model parameters have been updated and the model residual error depends on the input error estimation. Thus, the reordering strategy identifies the optimal input error rank conditional to the model parameters, effectively considering the interaction between the input error and the parameter error. This is akin to calibrating the input errors along with the model parameters in the BATEA framework (Kavetski et al., 2006).*" (line 350-355)

In addition, we have added the section "4.2 The effect of reordering on the error realization" to better explain this statement.

Line 383: "However, the work in this study still identifies a few areas needing to be explored."

Nit-pick: This sentence is a bit unwieldy, in my opinion. How about "However, this study identifies a few areas which still need to be explored:"?

Thanks for your suggestion. This has been changed as recommended. (line 415)

Response file / Response to major comment #7: "*A point of clarification: in each subsequent iteration of the BEAR algorithm, a new population of input errors are sampled from their a priori distribution. This means that the distribution of errors at each iteration is the same prior to reordering, i.e. the population of errors do not converge to a distribution that has different statistical features (mean, standard deviation, skewness).*"

You are of course correct, but I fear your comment misses the point somewhat: While it is undoubtedly true that the raw input error distribution never changes, BEAR *does* effectively update the input error distribution. In the figures from my code snippet, this can be seen in the bottom right subplots when compared to the centre right subplots. It becomes evident that the both the effective mean and standard deviation (and likely higher statistical moments as well) change dramatically after reordering. In essence, the fact that you are re-ordering transforms your raw error distribution and causes you to sample some different latent distribution instead. What this distribution really is remains the key question of your entire approach. If you can figure that out, you'll be one large step closer to justifying this approach theoretically. =)

Thanks for your comments. We should clarify the difference between the overall error distribution of all the time steps (sampled in a single iteration of the algorithm) vs the error distribution at each time step.

For the overall error distribution of all the time steps, this does not changes before or after reordering and in subsequent iterations of the algorithm. In the BEAR method, the errors are firstly sampled from the pre-estimated error distribution (error number = number of time steps) and randomly distributed on the different time steps. Then all the random samples are reordered according to the inferred error ranks, but the overall distribution stays the same.

For the error distribution at each time step, the aim of error identification in this kind of study is to make it converge to the true value. Just like the demonstration in your figures, the ideal result is that its mean is the same as the true value (the residual error in your cases), and its standard deviation is as small as possible.

Therefore, reordering does not cause us to sample some different latent distribution instead. The errors are always sampled from the pre-estimated overall error distribution. The converged error distribution at each time step after reordering is what we're trying to achieve.

In a sense, what you are doing seems distantly related to ideas in measure transport, see for example Marzouk et al. (2016) for an overview. In measure transport, the ultimate goal is to indirectly sample from an (almost) arbitrary target distribution. This is achieved by sampling a simple reference distribution instead (for example a multivariate standard Gaussian), then converting these reference samples through a deterministic function into samples from the target distribution. Of course, finding the correct transformation function is the key objective of this entire endeavour, and consequently its main challenge. In your study, you approach this from the opposite direction: you have some transformation, now you should find out what distribution you are sampling.

Marzouk, Y., Moselhy, T., Parno, M., & Spantini, A. (2016). An introduction to sampling via measure transport. *arXiv preprint arXiv:1602.05023*; https://arxiv.org/abs/1602.05023.

In summary, I would say the parallels to your approach are as follows: even though the reference distribution (corresponding to your raw input error distribution) never changes, the *pushforward distribution* (corresponding to the latent distribution your re-ordered error realizations are effectively sampled from) changes with the transformation function (in your case, the re-ordering according to different error ranks).

Yes, we totally agree with your summary that "the reference distribution (corresponding to your raw input error distribution) never changes, the pushforward distribution (corresponding to the latent distribution your re-ordered error realizations are effectively sampled from) changes." It should be noted that the reference distribution is for the overall distribution of all the errors, while the pushforward distribution is for the error at each time step. Therefore, in the above response, we differentiate the overall error distribution and the error distribution at each time step, and based on your analysis, we have clarified this in Section 4.2 from the changes of the marginal mean and std.

After learning the paper describing measure transport, we have not found an effective way to apply the proposed method for input error estimation or combine it with the secant method, which we believe needs further investigation of this method. Applying the measure transport approach in this framework is an exploration of a totally new method and we believe the reordering error approach we propose (according to inferred ranks) seems an easier way to identify the transport of the marginal error distribution. However, we agree that the measure transport approach suggests an interesting future approach to integrate or compare with the method we have proposed, potentially building a more solid theoretical foundation in formal Bayesian inference.