Response to Reviewer #3:

We have noted the reviewer's comprehensive remarks, and the following provides our response to each point. We appreciate the level of detail that the reviewer has gone into, and would like to thank them for several of their suggestions which will significantly improve how we convey the proposed approach.

However, we think many of the comments have arisen as the reviewer may not be familiar with existing approaches to estimate errors in inputs for these types of models (BATEA, IBUNE), just as the reviewer has mentioned in the comments. We hope our response and the proposed changes strike the right balance between addressing the reviewer's comments and recognising the past work in this area which we don't wish to recreate in our manuscript.

Response to Reviewer #3:

Complex patterns in input uncertainty such as spatial or temporal error correlations are an important topic in environmental science. In their present study, the authors seek to explore the ubiquitous issue of complex input uncertainty structures by proposing a novel method called Bayesian error analysis with reshuffling (BEAR). The proposed method is based on sampling an estimated input error and subsequently sorting the resulting realizations in an order which reduces residual mismatch to the observations. The authors then proceed to demonstrate the performance of their algorithm for a synthetic and a real case and compare its performance to a number of alternative setups.

I find the approach a very interesting and creative idea, and always appreciate it if someone takes the risks inherent to exploring a new methodological idea. Unfortunately, I have some reservations concerning its theoretical justifiability, which I hope the authors can address. Failing that, there might also be alternative ways to achieve similar effects which might stand on more robust theoretical foundations. Concerning these suggestions and the method itself, I have the following (major) comments:

Major comments:

1. Theoretical foundations: A key step of the approach is sorting input error realizations to reduce the residual mismatch between model predictions and system observations. I fear that this compromises the randomness of the error realizations, with potential consequences for the validity

of the Bayesian inference that are difficult to predict. Rigorously deriving this algorithm from more basic theoretical foundations might better illustrate the consequences of the authors' assumptions for Bayesian inference. If the authors do not have this expertise themselves (not to worry: few people in the environmental sciences do), I recommend seeking the help of the local statistics department – they are often keen to help. It seems an unfortunate truth of Bayesian statistics that interesting ideas for algorithms which sound nice on paper often tend to violate the Bayesian framework in unforeseen ways.

Thanks for your comments. We propose including a section that will derive this algorithm from the theoretical foundation of Bayesian inference (see Appendix B). We believe the remarks of the reviewer here are because the original manuscript implements an approximate Bayesian approach to model calibration which confuses the main contribution of the paper (which is to optimize input error ranks, rather than input error magnitudes). To address this, we propose recasting the implementation of our algorithm via SMC. The BEAR algorithm could be implemented via SMC, or GLUE, or SCE-UA, or any common model calibration approach, simply by optimising the latent-input errors on the error rank rather than error magnitude to realise the deterministic relationship between the residual error and the input error. It should be noted that we don't get the posterior distribution of the error ranks, we aim to optimise the error ranks by finding the deterministic relationship between the input error rank and the model parameters (see Appendix B).

2. Improving future performance: While I am not intimately familiar with the alternative methods (BATEA, IBUNE) referenced by the authors, improving future model performance is a common motivation for learning complex uncertainty structures in hydrological models. The approach in this manuscript, however, requires a concurrent time series of observations. As a consequence, it does not really improve the error model itself and hence offers little value for improving future predictions. On the other hand, attempting to learn time-varying bias or correlation structures in the input errors – admittedly a sometimes computationally formidable task – can increase the likelihood of future predictions substantially. Such approaches would also be significantly easier to justify. I would recommend mentioning this limitation for the BEAR algorithm in the manuscript, or – better yet – either explore or propose ways on how this limitation could be circumvented.

We think the reviewer's comments here may be due to a misunderstanding of how errors manifest in these types of hydrologic-water quality models, and how these models are used for predictions/scenarios vs in calibration. Additionally, we must clarify that water quality model applications such as the ones illustrated here rely on observed inputs alone, with observed response data being used only in model calibration. A better calibration of the water quality model is what is the intended outcome from our study.

The reviewer is correct in that the approach we propose does not 'improve' the error model but it improves the water quality model specification to now have parameters closer to what would be achieved under no error conditions. This is because it is assumed that the error model is known a priori and the overall objective is not to improve on this. We believe this is not an unreasonable assumption, as the model inputs are derived using a hydrologic model or rating curve whose error can be derived independently of the water quality model implementation. These independent observations provide insight to the distribution of input errors that can then be leveraged in the model calibration.

By identifying the order of a sample of input errors from this distribution, the model calibration results in model parameters that are closest to the 'truth', as can be seen in our synthetic case studies where the true parameters are known (Figure 3). **The overall goal in identifying proper input errors is then to ultimately improve our estimate of the model parameters**, so that the model can be more effectively used for scenario analysis (where we may know the hydrologic regime of a catchment in a hypothetical future), for forecasting under the assumption of perfect inputs (where the driving hydrologic forecast is independently obtained via a numerical weather prediction and a hydrologic model) or for regionalisation of the water quality model (where the model is transferred to a catchment without data). In all of these cases, an ideal model has unbiased parameter estimates. This is our goal in identifying the optimal ranks of input errors via BEAR, not to use the model for predictions with input data suffering the same errors.

3. The influence of other error types: It seems the authors focussed most of their attention on input uncertainty. While I agree that input uncertainty can play a very important role, the influence of

observation errors and model structural uncertainty plays a substantial role as well. In this study, the authors assumed both the model and the output observations were error-free – and derived their algorithm accordingly –, but in practice these assumptions are virtually never met. I would encourage the authors to explore how (if at all) their algorithm can avoid surrogacy effects in the presence of observation and model errors. (What I mean by “surrogacy effects” is that the algorithm’s adjustments to the input error realizations also ‘soak up’ [and consequently mask] errors in the output observations and the model itself in a bid to reduce the output residuals. This is of course undesirable.) See also my penultimate minor comment below.

The reviewer has raised here a very important point, and we completely agree with the issue of compounding sources of error and potential ‘surrogacy’ effects.

We have aimed to address this in our case by implementing a real data case which likely has issues of error surrogacy, to identify how our approach works in a ‘real’ setting. We believe this helps balance our discussion of the approach and its potential limitations. To additionally ensure we do not oversell our method given other sources of error, we also propose including a synthetic example in Appendix C that examines how the method performs under increasing error on the model calibration data. We hope this will help properly identify the usefulness of our approach but its potential limitations depending on the settings in which BEAR might be implemented.

4. Deterministic functions as an alternative: If the authors find that their algorithm might be based on flawed assumptions (following a more detailed theoretical derivation or investigation of the distribution it effectively samples from), the authors might wish to explore possible alternatives. If reducing the output residuals through adjustments to the input data remains the goal, a safer route might be to couple a deterministic input pre-treatment routine with the WQM and add its parameterization to the WQM parameter vector. Functionally, this pre-treatment routine can simply be interpreted as part of the deterministic model. Choices for this pre-treatment routine could be, for example, a one-dimensional spline which re-scales input magnitudes non-linearly (see the attached Figure 1, for an example with three extra parameters). More complex function choices might allow the consideration of lag, temporal or spatial correlations, etc. This would have the additional advantage over the BEAR framework that this pretreatment routine could also improve future predictions, assuming that it compensated true bias and is not overfitted. This

comment is not a request for change, but a hopefully constructive suggestion for alternatives so that the authors might salvage some of their work in case it would turn out theoretically indefensible.

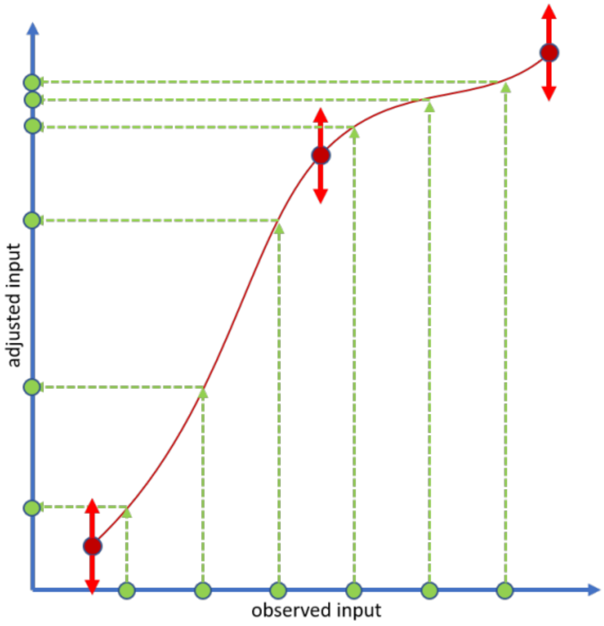


Fig. 1. Example of a non-linear re-scaling with a spline defined by three control points (red dots). You could use such a spline to scale the input values non-linearly.

We appreciate the thought that the reviewer has put into this comment, and we have considered this concept carefully, however, respectfully we do not see it as a viable pathway for the types of errors that we see in inputs to WQMs. The approach you suggested here is a non-linear bias correction to the input data which will correct systematic biases. In our case, it is the randomness of the observational error that means individual errors can't be explicitly identified prior to modeling.

This random error does lead to biased model parameters in calibration. We present here a synthetic case to demonstrate such, where the model inputs are specified as $\boldsymbol{X}^{o} = \boldsymbol{X}^{*}+\boldsymbol{\varepsilon}_{X},\boldsymbol{\varepsilon}_{X} \sim N(0,0.2^{2})$, the observed outputs are specified as $\boldsymbol{Y}^{o} = \boldsymbol{Y}^{*}=M(\boldsymbol{X}^{*},\theta^{*})$, and the objective function is selected as the MSE of the residual errors in log space. The mean of Gaussian distribution of the input error is set as 0, where the observed inputs are not systematically biased from the synthetic true inputs (Figure R1). According to the approach suggested by the reviewer, the observed inputs do not need modification, or it makes no difference before and after the suggested approach. However, in the comparison of different methods in Figure R.1, if we do not deal with the stochastic errors in input data, like the traditional method (the same as the suggested approach in this case), the model parameters will be biased from their reference values. In contrast, the BEAR method can effectively identify the stochastic errors, beyond the systematic bias of the input data, as the BEAR method reorders the error rank according to the residual error, and alleviates the impacts of both stochastic and systematic errors on the model parameters.

We additionally note that we are not proposing a method for rescaling errors, we are simply re-ordering their positions in the data to better correspond to deviations from model fits.

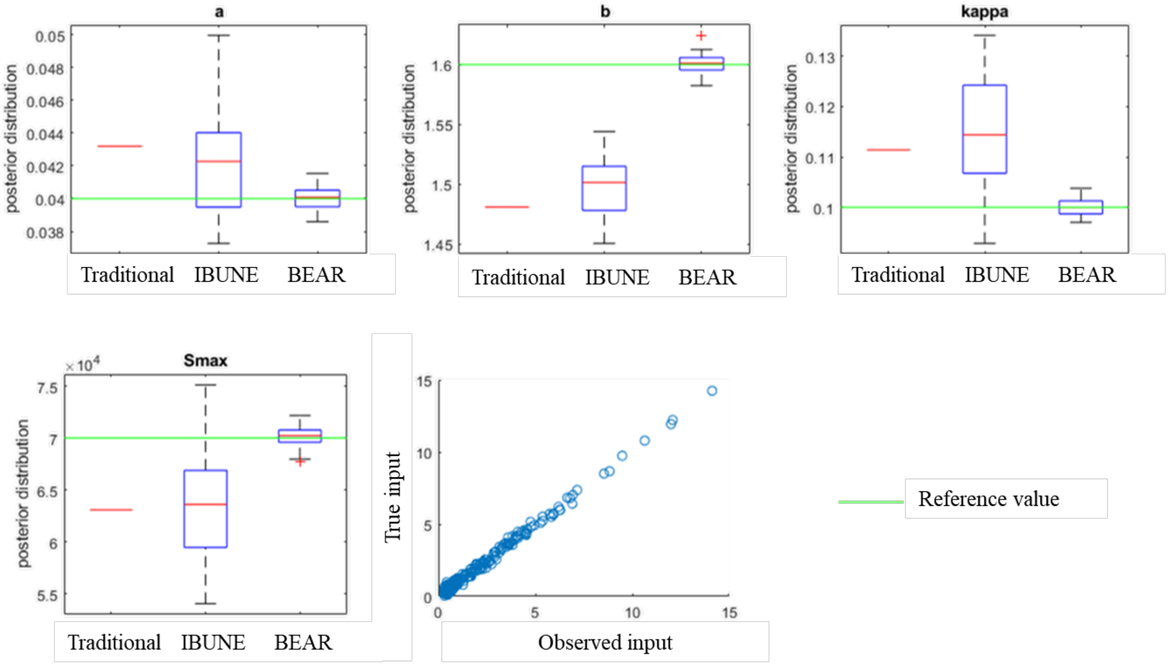


Figure R1 The posterior distributions of the model parameters estimated via the different methods and the scatter plot of the observed inputs vs true inputs in the synthetic case

5. BEAS instead of BEAR: This could be filed under nit-picking, but since the algorithm's name features so prominently, I chose to raise this to a major comment instead. The use of the word 'shuffling' implies randomness in the re-ordering. If I understood the authors' algorithm correctly, though, the re-ordering itself is entirely deterministic. As such, changing the name to something along the lines of "Bayesian error analysis with sorting" (BEAS) or "Bayesian error analysis with

re-ordering" (if the authors like to retain their – admittedly very nice – acronym) might better reflect its deterministic nature.

We very much appreciate the valuable suggestion here, as the reviewer is completely correct. Yes, "reordering" better reflects the deterministic nature of error quantified via this new method. Therefore, the method name will be changed into "Bayesian error analysis with reordering" to keep the BEAR acronym for short.

6. Why ABC: In the manuscript, the authors use an "Approximate Bayesian Computation via Sequential Monte Carlo" (ABC-SMC) approach. While I am not personally familiar with this approach, I struggle to see why it is necessary to resort to ABC, aside from any potential (forgive me) self-inflicted complications induced by the BEAR algorithm. The model and output variables seem pretty simple, so to my untrained eyes it is difficult to see why the formulation of an analytical likelihood should be impossible in this case. One could also cast the procedure the authors presented in Figure 1 with very few changes in terms of an MCMC routine, provided the re-ordering or error realizations ends up being statistically justifiable, of course. I would encourage the authors to provide a bit more detail on why ABC was necessary.

Thanks for the helpful suggestions here. The reviewer is completely correct that the ABC method is not necessary, and the core idea of the BEAR method (the reordering strategy in the rank estimation) can be easily applied in any other calibration algorithm, for example, MCMC and SMC algorithms. In order to clarify this point, we propose modifying the manuscript to make this clear in the algorithm.

7. Focus on a good fit: A key idea which seems to permeate the present manuscript is that it is desirable to obtain error realizations, if necessary by force (i.e., re-ordering), which match the observations as closely as possible. This is of course true, but not at all cost. Even assuming a severely mischaracterized prior input error distribution (e.g., a Gaussian with a standard deviation of 1E-10 and a mean of -1E8), one could theoretically obtain an error realization which causes the model to fit the observations perfectly if we only drew sufficiently (read: infinitely) many samples. The challenge, then, is not to find such a realization within our prior distribution (it will exist in

any distribution with sufficiently broad support), but to find a distribution from which there is a high probability to obtain such a sample. Crucially, such a distribution should be independent from future observations, and I fear that this may not be the case for the approach proposed in this manuscript. In this approach, after re-ordering, the realizations are no longer i.i.d. samples from the input distribution (similarly to how one could interpret correlated Gaussian samples merely re-ordered independent Gaussian samples, but would nonetheless be wrong in claiming that an independent Gaussian distribution is identical to a correlated Gaussian distribution). If the authors decide to pursue my request for a derivation of a theoretical foundation for their approach (see comment 1), I recommend focussing on investigating from what effective distribution they are really sampling. Considering BEAR's ability to yield a reliably good fit with seemingly arbitrary prior input error realizations, I fear that the distribution you effectively sample from may be well approximated with (for example) a Gaussian with a mean inversely obtained from the observation residuals. If this turns out to be the case, the method would be more or less equivalent to just calculating the input error residuals through an inverse method of choice by minimizing the output residuals. This would not be very useful, and I would recommend exploring one of the approaches I suggested in comment 4 instead. See also my penultimate minor comment.

The reviewer has raised an interesting point here, that the reordered errors may include features (such as autocorrelation or dependency) that are not present in the original error sample. We propose including statistical analysis of the reordered errors (e.g. ACF, PACF, correlation of the errors to modelled TSS magnitude) as a diagnostic for the real case on the potential interactions between observation error and model structural error.

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). However, we do recognize that the reordering process provides insight into how the input errors may be ideally distributed across time that can inform how the surrogacy effect is manifested (see comment 3). For example, if the errors are highly autocorrelated this could highlight potential model structural errors if the WQM has not properly represented the storage characteristics of TSS in the catchment. We believe that including further diagnostics will help elucidate this effect.

Note that the approach does in fact rely on an appropriate prior distribution being identified describing the input error distribution (see Appendix C). If this error distribution is mischaracterized and the whole statistical features are different from the real input, the resultant outputs will not converge to the calibration data, and will lead to larger residual errors. Subsequent updates of the error values will aim to find a more appropriate sample of the input error to get a smaller residual error. In some cases, the parameter error will compensate this mischaracterized error distribution, leading to an overfitting problem. This compensation should occur on not only a single input error realization, but also the whole error distribution. In Figure C.1 of Appendix C, the standard deviation of the input error is estimated accurately, converging to the reference value when there is no interference of other sources of errors. However, if there is interference of other sources of errors, the impacts can not be avoided in any of the methods unless the method can deal with all sources of errors together.

Specific comments:

Line 22-25: You mention the importance of complex interactions of different error sources directly in the first paragraph but proceed to largely ignore their influence in the remaining manuscript. I think this part is important and should be discussed in greater detail in the remainder of the manuscript (particularly also the methods/theory section).

Thanks for your suggestion, we agree this was not being properly highlighted. We propose including an explanation in the methodology that emphasizes this:

"It should be noted that the derivation of the BEAR method is based on the assumption that the model only suffers from the input error and parameter error, but other sources of error (i.e. model structural error and output observational error) can also impair the estimation of the model parameters and are inevitable in the WQM. The ability of the BEAR method is tested in the real case where the interference of other sources of error have been considered."

Line 49-51: During this review, I have briefly glanced into the corresponding methods BATEA and IBUNE, and apparently there was quite a commentary battle between the authors over these methods (see doi:10.1029/2007WR006538 and https://doi.org/10.1029/2008WR007215), and Renard et al. 2009 noted that IBUNE may in fact not reduce dimensionality. The choice is of course ultimately up to the authors, but it might be useful to add a small comment noting that the claim of dimension reduction by IBUNE is also challenged.

In the comments from Renard et al. 2009, there are two different implementations of the IBUNE framework. The first is the same as the method D described in this manuscript, and Renard et al. 2009 argue the randomness in this approach will lead to an underestimation of the input error variance. The second implementation is the same as the BATEA approach, which can not reduce the computational dimension.

According to the IBUNE paper (Ajami et al., 2007), the authors argue that IBUNE can circumvent the high dimensionality issues by randomly sampling the multiplier to each time step from an assumed normal distribution. Therefore, we interpret the first implementation as the 'true' version of IBUNE, and provide discussion on this in the manuscript.

Line 69-75: I would be careful here. In reality, there are many different error sources, certainly not the least of which is model structural error. Calibrating (i.e., simply minimizing the residuals between simulated and observed output) in the presence of other error sources is prone to surrogacy effects, so you can never really be sure you recovered the 'true' parameters. Even making the (unrealistic) assumption that we could manage to completely remove all error sources in the model and its input, we could still only retrieve the 'true' model parameters if our inverse problem is unique. I would mention these restrictions here (only works if all error sources can be removed completely, unique inverse problem).

Thank you for pointing out the equifinality problem. We propose modifying the description here as follows:

"*where the optimal parameters* $\boldsymbol{\theta}^p$ *will lead to the same simulation corresponding to the true values* $\boldsymbol{\theta}^*$ *and the model residual* $\boldsymbol{\varepsilon}^p$ *will decrease to zero. Due to equifinality, the parameter* $\boldsymbol{\theta}^p$ *may not converge to the true parameter* $\boldsymbol{\theta}^*$ *if the inverse problem is not unique, but both may result in the same simulation. In this study, these parameters are called the ideal model parameters, also denoted as* $\boldsymbol{\theta}^*$ *due to the same impacts of true model parameters.*"

Line 70-73: In Equation (2), the variables Y^o and Y^s are used without any introduction. I assume both variables stand for the observed and simulated output. Please introduce these variables.

Thanks for pointing this out. We propose clarifying this as follows:

"where $\boldsymbol{Y}^s$ is the output simulated from the model M corresponding to the observed input $\boldsymbol{X}^o$ and model parameter θ^c, and the observed output $\boldsymbol{Y}^o$ is assumed without observational errors in the derivation, thus can be denoted as $\boldsymbol{Y}^*$."

Line 76-77: A critical thing here is that $\boldsymbol{\varepsilon}$ is previously introduced as an error, which implies that you consider it to be a random variable. However, subtracting a (say) Gaussian random variable from another Gaussian variable with the same properties does not reduce variance to zero but actually doubles it (if both random variables are independent). What you seem to have in mind here only works if ε and $\boldsymbol{\varepsilon}^p$ have identical properties and are perfectly correlated (note that this implies a lot more than just sharing the same statistical moments!), or if you are talking about error realizations. You should clarify this. This relates to major comment 7. You can only create this perfect correlation if you can somehow extract the error realizations of ε (which only works under the assumption that you already have the input samples, that there are no other errors, and that the inverse problem is unique). Consequently, I fear that you may create/mimic this perfect correlation by implicitly solving an inverse problem, which would make the proposed method not very useful.

We agree with your argument that "Subtracting a (say) Gaussian random variable from another Gaussian variable with the same properties does not reduce variance to zero but actually doubles it (if both random variables are independent)". This is also the reason why IBUNE (method D in this manuscript) does not improve the accuracy of the error identification. The claim in this manuscript is "*If the equivalence between $\boldsymbol{\varepsilon}_X$ and $\boldsymbol{\varepsilon}_X^p$ can be ensured for each data point, the modified input X^p then becomes the same as the true value X^*. The proposed calibration (Eq. (4)) will result in an ideal calibration (Eq. (1)), where the optimal parameters $\boldsymbol{\theta}^p$ will lead to the same simulation corresponding to the true values $\boldsymbol{\theta}^*$ and the model residual $\boldsymbol{\varepsilon}^p$ will decrease to zero. Thus, the precise identification of the input error series will result in the ideal model parameters and minimized residual error, which is the aim of model calibration considering the input error quantification.*" As the reviewer notes, we are referring here to the error realizations, not the distribution of errors.

We also agree that the perfect identification of input error series and model parameters only happens when "there are no other errors, and that the inverse problem is unique". We propose adding clarification as follows: *"It should be noted that the derivation of the BEAR method is based on the assumption that the model only suffers from the input error and parameter error, but other sources of error (i.e. model structural error and output observational error) can also impair the estimation of the model parameters and are inevitable in a WQM. The ability of the BEAR method is tested in the real case where other sources of error have been considered."*

Considering the unique inverse problem, we propose changing the description here to "*Due to equifinality, the parameter $\boldsymbol{\theta}^p$ may not converge to the true parameter $\boldsymbol{\theta}^*$ if the inverse problem is not unique, but both may result in the same simulation. In this study, these parameters are called the ideal model parameters, also denoted as $\boldsymbol{\theta}^*$ due to the same impacts of true model parameters.*"

Line 79: In Equation (4), you also mark the parameters – on which this entire exercise should be conditional on – as changing due to your proposed approach. The equations you have shown so far imply that the procedure you describe here is applied after parameter calibration. During a first reading of this paper, it is not immediately clear why the calibrated parameter values should change with your proposed approach. On a second reading, it becomes evident that you do not really calibrate but sample the parameter posterior, but that this sampling process is inter-woven with your BEAR routine, hence the parameter values are also affected. I recommend commenting on this already here to save your readers some confusion. Maybe it would also help not to talk about calibration at all in this context.

Apologies for the lack of clarification here- in fact, the model calibration is the ultimate goal in properly identifying the model errors. The reordering step is implemented after one set of the model parameters has been sampled. Corresponding to this set of model parameters, the optimal ranks of input errors are inferred via the secant method. Although the sampled errors have been reordered, they still follow the assumed error distribution and the overall statistical characteristics remain unchanged. In Equation (4), with the constraint of the input error distribution, if the calibrated model parameters are far from the true parameters, the residual error cannot be reduced effectively. If the model parameters are close to the ideal model parameter, the input error can be identified precisely via the secant method, and the model residual error will approach zero.

We propose clarifying the model calibration goal in the paper so that the intent is clear.

Line 81-83: I would be very careful with this statement. This only happens if the model parameters θp and the input error residuals ε_X^p cannot compensate each other, i.e. if there is only a single, unique combination of parameters and errors which yields zero residual. If you have a Pareto front along which different values of θp and input error residuals ε_X^pyield zero residual, you cannot be certain that you have correctly identified the 'true' parameter values and the 'true' input error, even in a scenario where no other errors/uncertainties exist. In addition to this, my reservations concerning other error types raised in major comment 3 also apply. I would recommend changing this statement accordingly and exploring its consequences for your algorithm in greater detail.

We appreciate your comment. Yes, the input error and model parameter error might compensate each other, which leads to a Pareto front along which different values of θp and input error residuals yield zero residual. In the manuscript, we compare two types of input error models: additive formulation ($\boldsymbol{X}^{o}=\boldsymbol{X}^{*}+\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}\sim N(0.2,0.5^{2})$, *add* in Table 3) and multiplicative formulation ($\boldsymbol{X}^{o}=\boldsymbol{X}^{*}\exp(\boldsymbol{\varepsilon}),\boldsymbol{\varepsilon}\sim N(0.2,0.5^{2})$, *mul* in Table 3). According to synthetic cases, we find the multiplicative formulation is more likely to compensate the model parameter error, but the compensation effect will reduce with accurate prior information on the input error model. Probably because the impacts of error in multiplicative formation is similar as *b* in Equation 9, or the nature of multiplicative formulation is more likely to change with the model parameter than additive formulation, due to having more flexibility. In sum, the compensate effect can be reduced through the proper selection of the input error model since it affects the identification accuracy. But this problem is common in all the methods considering input error, and is not special in the BEAR method.

Line 104-106: I would rephrase this a bit, because following the procedure you outlined in Figure 1 (a very nice schematic, by the way), it is not only two steps: you sample the error once, then iterate over a large number of re-ordering steps until you find an order which minimizes your output residuals. This could do with some clarification.

Thanks for your suggestion. We propose rephrasing the sentence as:

"*This new approach, referred to as the Bayesian error analysis with reordering (BEAR) method, contains two parts: sampling the errors from the estimated error distribution to maintain the overall statistical characteristics of input errors and reordering these sampled errors via the secant method for a few iterations until the input error order can be optimized to achieve the defined target about the residual error.*"

Line 115-116: If I understood your explanations here correctly, maybe an easier way of explaining what you are doing is that you sort your updated error ranks, then assign to each of them a new integer rank based on its position in the sorted list. This might be easier than trying to explain this procedure with scaling.

Thanks for your suggestion. We propose modifying this sentence as follows:

"*Sorting* $k_{i,q}$ *in all the ranks* $k_{i,q}(i = 1,...,n)$ *can address this problem by effectively assigning to each of them a new integer rank based on its position in the sorted list*."

Line 125-131: As mentioned in major comment 6, please devote some space to explain why the models you use in the following necessitate the use of ABC. Even after going through the manuscript a few times, I struggle to see why standard Bayesian approaches would be impossible to use. At the risk of evoking the anger of our ABC-focussed colleagues: direct is usually better than approximate. It also does not become clear in the manuscript why an ensemble-based approach is used – couldn't the same procedure be implemented in an MCMC-style acceptance/rejection algorithm? If there are some ABC reasons for requiring an ensemble, it you might also want to explain why and how it is used.

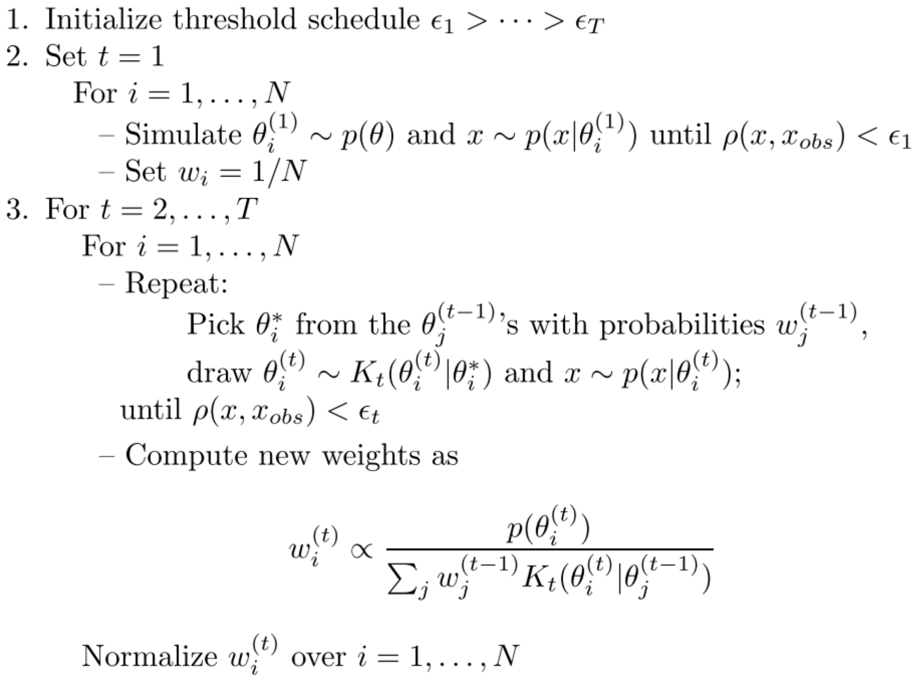
We apologize for the lack of clarification. The ABC is not necessary for the BEAR method to be implemented. We admit there is no difference in the calibration results between the standard Bayesian approaches and the ABC approach for this case study. In order to clarify this point, we propose modifying the manuscript to make this clear in the algorithm, please also see the proposed Appendix B.

Line 132-136/Figure 1: This explanation of the method is very short, and essentially only explains that you approach the posterior distribution iteratively through a number of intermediate steps, but not how this is achieved exactly. Figure 1 provides more information and suggests some sort of acceptance/rejection scheme depending on whether your procedure can reduce the error residuals below a certain threshold, but the nature of the posterior distributions from which new parameter values are drawn remains undefined. You also seem to update an input error parameter ηx, which seemingly contradicts statements you made suggesting the input errors are sampled from a pre-estimated distribution (Line 10, Line 193-194). This step is also never mentioned in the text itself up to this point – you only mention that you estimate the input error distribution's hyperparameters much later. The text also frequently mentions 'populations', which evoke the idea of an ensemble-based method, but none of the steps mentioned in the text so far actually seem to require an ensemble. Please provide some more (written) detail about how your algorithm functions exactly.

Thanks for your comments, please see the following:

1) "The nature of the posterior distributions from which new parameter values are drawn remains undefined". The steps about generating a new set of the model parameters from the previous posterior parameters are demonstrated in the below figure describing the SMC algorithm. Also, we propose adding clarification in the manuscript, as follows:

"*The details of sampling a new set of model parameters from the posterior distribution of the previous population can refer to the study of (Bonassi and West, 2015).*"



```
1. Initialize threshold schedule $\epsilon_1 > \cdots > \epsilon_T$
2. Set $t = 1$
     For $i = 1, \ldots, N$
       – Simulate $\theta_i^{(1)} \sim p(\theta)$ and $x \sim p(x|\theta_i^{(1)})$ until $\rho(x, x_{obs}) < \epsilon_1$
       – Set $w_i = 1/N$
3. For $t = 2, \ldots, T$
     For $i = 1, \ldots, N$
       – Repeat:
             Pick $\theta_i^*$ from the $\theta_j^{(t-1)}$'s with probabilities $w_j^{(t-1)}$,
             draw $\theta_i^{(t)} \sim K_t(\theta_i^{(t)}|\theta_i^*)$ and $x \sim p(x|\theta_i^{(t)})$;
         until $\rho(x, x_{obs}) < \epsilon_t$
       – Compute new weights as

             $w_i^{(t)} \propto \frac{p(\theta_i^{(t)})}{\sum_j w_j^{(t-1)} K_t(\theta_i^{(t)}|\theta_j^{(t-1)})}$

     Normalize $w_i^{(t)}$ over $i = 1, \ldots, N$
```

Figure 1 ABC-SMC algorithm (Bonassi and West, 2015)

2) A description about "update an input error parameter ηx" will be added, as follows.

"*In the real-life applications, the input error model can be approximated based on the studies described in the introduction, but the error parameter might not be estimated as a deterministic value accurately. In those cases, these error parameters should be considered as the hyperparameters and calibrated with the model parameters together (denoted as η_x in Fig.1).*"

3) Here the term "population" is used as in the SMC algorithm, and it is like an ensemble-based method. We will clarify this by a better description of how the algorithm is implemented via SMC.

Line 145-148 and Line 159-162: This is just a comment towards the general "Why ABC?" discussion. It seems to me that a classic MCMC procedure would avoid the need for adjusting the acceptance threshold dynamically, as proposed parameters are always compared to the previous entry in the chain.

Yes, ABC is not necessary for the BEAR method and any MCMC procedure will avoid the need for adjusting the acceptance threshold. We propose recasting the implementation of our optimization algorithm via SMC.

Line 154-157: I confess that this explanation is quite impenetrable, and probably causes more confusion here than it does good. I recommend restructuring this explanation or removing it altogether. A good alternative would be to visualize this with a small figure, possibly added to the supporting information if length limitations to not permit embedding it into the main text.

Thanks for your suggestion. We will integrate an example and its illustration in the Appendix to explain the specific steps involved (see an illustration of this in the following Appendix A).

Line 192: I do not see from Equation 8 or 9 or their surrounding text how the spatial scale factors into this model. Through the Sa variable? Please clarify this.

Thanks for your comments. "*The two formulations in Bwmod were developed in a small-scale experiment (Sartor and Boyd, 1972), while in applications at the catchment scale, the conceptualized parameters largely abandon their physical meanings and the formulations can be considered a "black-box" (Bonhomme and Petrucci, 2017).*" This study attempts to simulate the sediment dynamics in the catchment scale. Thus, we say "the spatial scale is set as the catchment in this study". This will be clarified as follows:

"*This study will test the BEAR algorithm in a case of simulating the daily sediment dynamics of one catchment, thus, the time scale is typically set as daily and the spatial scale is set as the catchment.*"

Line 200: In Equation 8, you do not introduce Smax and κ. Please introduce these variables as well.

The introduction of Smax and κ is shown in Table 2. We propose adding the clarification as follows:

"*where the descriptions of κ and Smax are shown in Table 2*"

Line 203: In Equation 9, you do not introduce a and Qt^b. Please introduce these variables as well.

Qt^b should be $(Q_t)^b$, we propose modifying this as follows:

$$s(S_{a,t}) = a \cdot (Q_t)^b \cdot S_{a,t} \qquad (9)$$

"*where the descriptions of a and b are shown in Table 2, and* Q_t *is the streamflow at the catchment outlet at time t.*"

Line 205: In Equation 10, you do not introduce Qt. Please introduce this variable as well.

See the above reply.

Line 209-211: For this section, there are a few assumptions which could warrant greater discussion. If the errors are normally estimated in advance based on a rating curve, why is there a constant offset of 0.2? Couldn't this systematic bias be corrected through the rating curve itself? Alternatively, if the offset is necessary because your errors are asymmetrically fat-tailed, wouldn't a different distribution (such as a scaled beta or gamma distribution) be a better choice? It is commendable to make the synthetic test case more challenging by introducing bias as well, but how would this be recognized a priori in a real test case if it wasn't already considered in the rating curve? Some more information might clarify the authors' choice of distribution for the audience.

Thanks for your comments. Yes, we agree with your statement that if the errors are normally estimated in advance based on a rating curve, the mean of the normal distribution should be 0. By including the systematic bias of 0.2 in this synthetic study we aim to test the BEAR method in wider applications, even if these are not necessarily representative of true errors. Figure 3(1) and (2) indicates that in the synthetic case only suffering from the input error and parameter error, the mean and standard deviation of input error can be estimated precisely together with the model parameters. Therefore, in situations where the input errors have no systematic bias and only a standard deviation must be calibrated, the BEAR method also works. According to your comments, we propose clarifying this as follows:

"*If the input errors are estimated based on a rating curve, like the procedure in the following real case, the mean of input error should be 0. But in order to test the ability of the BEAR method in wider applications, the systematic bias 0.2 has been considered in the synthetic case.*"

Line 224-226: This part here is a bit unclear. What I deduce from the context is that you looked at two scenarios – one, where you left the prior input error fixed, and one where you estimated the input error hyperparameters as well. I would not talk about 'conditions' in this context, but rather about 'scenarios'. If I understood your drift here correctly, I would also add a comment which puts more emphasis on the fact that you subvert one of the principal assumptions you made earlier in the second scenario (namely, that the input error distribution is a prior/pre-estimated).

Thanks for your comments. We propose modifying the description as follows:

"*Given the description in the introduction, the input error can be pre-estimated in some studies. While in other cases, the prior information about the input error cannot be estimated or the accuracy is in question. Therefore, two scenarios concerning the prior information of error parameters (i.e. σ and μ) have been considered: one is fixed as the reference values (denoted as 'fixed' in Table 3), the other one is estimated as the hyperparameters with the model parameters (denoted as 'inferred' in Table 3.*"

Line 232-234: I would recommend to critically re-examine this part in the light of major comment 7. The high correlation of scenario R with the realizations of the synthetic true error series – which are supposed to be realizations from an independent Gaussian distribution – might be reason for concern, as they suggest that you might be implicitly solving an inverse problem for the input error residuals. This would have little to do with a Bayesian framework.

Please the reply for comment 7.

Figure 4: Unfortunately, this figure is really hard to read. If possible, I would recommend splitting this up into several figures and providing the figures for individual scenarios in the supporting information. The choice of colors also makes it very difficult to see what's going on (especially the neon green and the soft peach color). I am not familiar with the HESS compiler, but I would also recommend either a significantly larger resolution and a different image format such as .tiff or .gif, as the current figure is in quite a low resolution and has serious compression artifacts. For graphs such as this one, vector-based formats such as .svg or .pdf (if saved straight from Python with pyplot.savefig) might also allow readers of the electronic version to zoom in arbitrarily close for details. This could be particularly valuable here, since most of the relevant details are quite small.

Thanks for your suggestion. We will improve the quality of the figures, including splitting this into several figures, improving the resolutions, using colors that are better distinguishable, and modifying the placing of legends.

Line 296-297: I would remove this statement, as you have not experimentally backed this statement up and it is not immediately obvious. I see little reason why inverting the observation residuals to find optimal input error realizations would be a more difficult task than re-ordering a pre-existing set of realizations. Quite the opposite, in fact.

The approach suggested by the reviewer, "inverting the observation residuals to find optimal input error realizations", is BATEA, which calibrates all the input error series together with model parameters. This leads to a high dimension problem (Renard et al., 2009). The statement here "*It is far more efficient to estimate the error rank than estimate the error value.*" is based on the explanation "*In a continuous sequence of data, the potential error values have an infinite number of combinations, while the error rank has limited combinations, dependent on the data length.*".

Besides, directly optimizing the input error value according to the model residual is more likely to change the assumption about the error distribution. Therefore, in the BEAR method, the sampling and reordering strategy can not only ensure the error distribution assumption but also avoid the high dimension problem. The explanation has been already stated in the manuscript.

"*Note that modifying each input error according to the 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 with the aim of minimizing the objective function and subsequently be overfitted. There are two ways to impose restrictions. One is to regard errors and model parameters as a whole in calibration, resulting in the high dimensional computation (Kavetski et al., 2006). The other is to sample error randomly from the assumed error model IBUNE (Ajami et al., 2007), whose precision cannot be guaranteed. While in the rank domain, the value range of the sampled errors can be effectively limited by the assumed error model.*"

Line 301-307: This is a very important paragraph. As I suggested in comment 7, through re-ordering you are no longer sampling from the prior input error model, which makes the protection from perfect fits you mention here somewhat arbitrary. As an illustration, I would like you to consider the behaviour of this re-ordering for longer time series: For a single observation, re-ordering can yield no improvement, and the residual fit depends exclusively on the realization you drew. For a few observations, re-ordering will induce moderate improvements, and the residual fit depends somewhat on the realization you drew. However, in the limit of infinitely many observations (assuming the statistical moments of the prior input error distribution are correctly characterized), re-ordering your error realizations should allow the residuals to be compensated completely at every single observation, irrespective of what specific realization you drew. This makes the protection against overfitting (and the expected residual error) dependent on the length of the observation time series and seems to converge towards deterministic (over) fitting. At the same time, the effective input error uncertainty decreases to zero. In a conventional Bayesian framework, even if the correlations in the input errors are perfectly identified, this would never happen.

The reviewer is correct that reordering will not be possible if a single observation is present, as the rationale behind reordering involves not changing the magnitude but the position of the error. As the number of samples increases, the fit will improve (assuming the error distribution has been identified properly a priori). However, an infinite number of observations will not lead to a good model fit if the prior distribution is seriously mischaracterized. For example, if the input error distribution underestimates the frequency of large streamflow errors (say with a mis-specified positive skewness) the model parameters will attempt to compensate for the reduction in overall TSS concentration and the model fit will be poorer than if a correct distribution of input errors was assumed, even with infinite observations. This highlights the need for the approach to properly identify the input error distribution ahead of the model implementation, and future work will demonstrate how model selection techniques could be used when there is limited information on input errors a priori. Please also see the reply for the comment 7.

Line 314: The dot after (Fig 1.) should probably be a comma

This will be modified as flows:

"*Considering these two points, the BEAR method set q iterations in the algorithm (Fig. 1), and q increases until a defined target about the residual error is achieved.*"

Summary:

In summary, I find the approach an interesting and ambitious idea, but have reservations concerning its theoretical validity, which I hope the authors can address in their revision. If my fears concerning it solving an implicit inverse problem for the input error residuals happen to be confirmed, the authors might consider the following alternative avenues:

a) The approach might be re-interpreted as a diagnostic tool for input error residuals; there is some value in identifying input error residuals and the correlations between them. In this case, however, it may be worthwhile to investigate whether the re-shuffling strategy is needed, or whether a more straightforward inverse method might be more efficient.

b) If predictive improvements are desired, following the suggestions in major comment 4 could be a viable and interesting alternative avenue.

I wish the authors the best of luck with the manuscript and hope that my comments are useful.

We thank the reviewer for their thought-provoking comments. We hope our responses above have helped clarify some of the concerns the reviewer has raised. We intend to assess the remaining issues mentioned in our revision.

Appendix A: The illustration of the BEAR method

Table A 1 An example illustrating the BEAR method

		1st iteration (random sample)																			
row	time step i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	sampled input error	0.07	-0.12	0.07	0.16	0.05	.0.07	0.07	-0.03	0.03	-0.08	0.09	-0.11	-0.11	-0.08	-0.29	0.14	0.03	-0.08	0.14	-0.17
2	input error rank k	13	3	14	20	12	17	15	9	10	7	16	4	5	6	1	19	11	8	18	2
3	residual error ε	-0.29	0.49	-0.58	-0.98	-0.78	0.29	-0.66	0.59	-1.31	-0.31	-0.87	0.76	0.46	0.54	0.25	-0.80	-0.07	0.56	-0.23	0.40
	MSE										0.40										
		2nd iteration (random sample)																			
4	sampled input error	-0.01	-0.02	0.03	0.03	-0.09	0.00	-0.02	0.06	0.11	0.11	-0.09	0.01	-0.12	-0.11	0.00	0.15	-0.08	0.04	-0.02	0.11
5	input error rank k	9	6	14	13	3	10	8	16	17	18	4	12	1	2	11	20	5	15	7	19
6	residual error ε	-0.13	0.23	-0.43	-0.41	-0.21	0.70	-0.23	0.09	-1.88	-1.52	0.20	0.17	0.53	0.60	-0.43	-0.72	0.36	0.12	0.47	-0.82
	MSE										0.47										
		3rd iteration (updating the error rank via the secant method)																			
7	calculated pre-rank K	5.8	8.7	14.0	8.0	-0.3	22.0	4.3	17.3	-6.1	4.2	6.2	14.3	31.3	42.0	4.7	29.0	10.0	16.9	14.4	7.6
8	ranked rank k	6	10	12	9	2	17	4	16	1	3	7	14	19	20	5	18	11	15	13	8
		3rd iteration (reordering errors according to the updated error ranks)																			
9	reordered input error	-0.02	0.00	0.01	-0.01	-0.11	0.11	-0.09	0.06	-0.12	-0.09	-0.02	0.03	0.11	0.15	-0.08	0.11	0.00	0.04	0.03	-0.02
10	residual error ε	-0.23	0.20	-0.34	-0.24	-0.12	0.19	0.14	0.08	-0.40	-0.31	-0.22	0.03	-0.17	0.26	-0.09	-0.55	0.11	0.14	0.27	-0.23
11	MSE										0.06										

The implementation of the BEAR method contains two main parts: sampling the errors from an assumed error distribution and reordering them with the inferred ranks via the secant method. An example is illustrated in **Error! Reference source not found.** and the explanation about the specific steps is presented in the following contents.

(1) In the 1st iteration (q=1), the errors are randomly sampled from the assumed error distribution (row 1), and then they are sorted to get their ranks (row 2). This error series is employed to modify the input data, which corresponds to a new model simulation and model residual (row 3).

(2) Repeat step (1) in the 2nd iteration (q=2) as two sets of samples are prerequisites for updating via the secant method. The results are shown in row 4, 5 and 6. **Error! Reference source not found.** demonstrates that the ranges of the error distribution are the same between the true input errors (black line) and the sampled errors (blue and green lines) as they come from the same error distribution under the condition that prior knowledge of the input error distribution is correct. However, the value at each time step is not close.

(3) At the 1st time step (i=1) in the 3rd iteration (q=3), the pre-rank $K_{1,3}$ is calculated via the secant method (illustrated as the following equation). The details are demonstrated in red boxes.

$$K_{1,3} = k_{1,2} - \varepsilon_{1,2}^{p} \frac{k_{1,2} - k_{1,1}}{\varepsilon_{1,2}^{p} - \varepsilon_{1,1}^{p}} \text{=9-(-0.13)} \frac{9-13}{-0.13-(-0.29)} = 5.8$$

(4) Repeat the step (3) for all the time steps. The calculated pre-ranks are shown in row 7.

(5) Sort all the pre-ranks to get the error rank (row 8).

(6) According to the updated error ranks (row 8), the sampled errors in the 2nd iteration (row 4) are reordered. The example for the 1st time step is demonstrated in black boxes. The error rank at the 1st time step is updated as 6, and the rank 6 corresponds to the error value -0.02 in the 2nd iteration. Therefore, -0.02 is the input error at the 1st time step in the 3rd iteration. Following this example, the sampled errors at all the time steps are reordered. The results are shown in row 9. **Error! Reference source not found.** demonstrates that after reordering the errors with the inferred ranks, the estimated errors are much close to the true input error.

(7) The reordered input error will lead to a new input data, a new model simulation and a new model residual. The residual error is shown in row 10.

(8) If a defined target about the residual error is achieved, the input error estimation is accepted; Otherwise, q=q+1, repeat step (3)~(7) until q is larger than the maximum numbers of iteration Q.

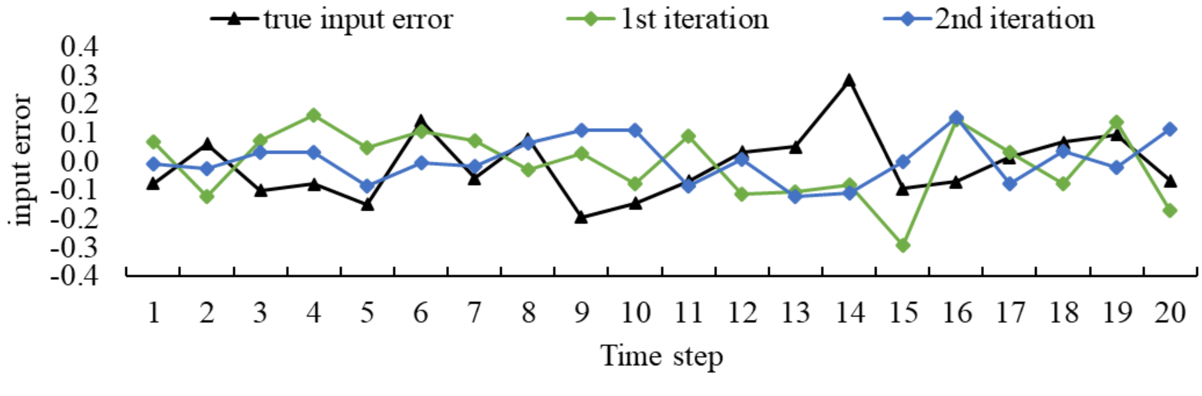


Figure A 1 Demonstration of the input error estimation in **Error! Reference source not found.** at the 1st and 2nd iteration where the input errors are randomly sampled

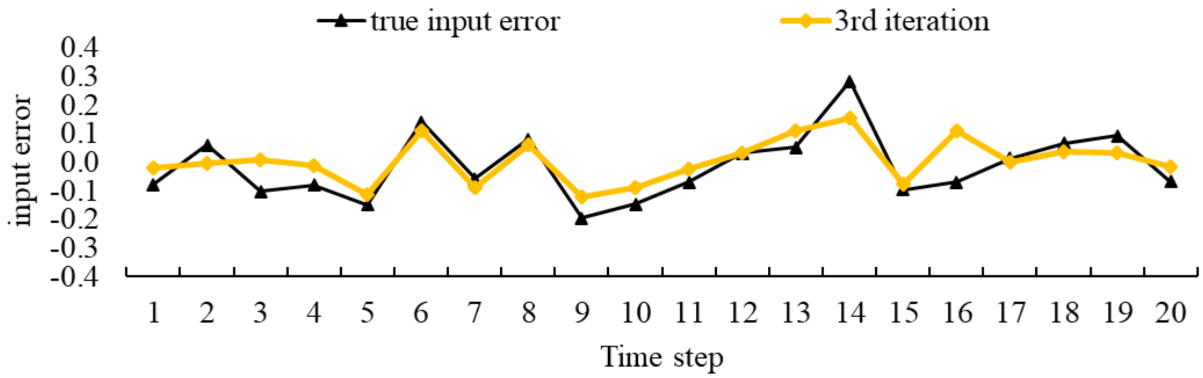


Figure A 2 Demonstration of the input error estimation in **Error! Reference source not found.** at the 3rd iteration where the input errors are reordered according to the updated error ranks

Appendix B: Theoretical foundation

1) Basic notation

In general, a model *M*() simulates the output $\boldsymbol{Y}^{\mathrm{s}}$ given the observed input $\boldsymbol{X}^{\mathrm{o}}$, model parameters $\boldsymbol{\theta}$, as follows:

$$\boldsymbol{Y}^{\mathrm{s}} = M(\boldsymbol{X}^{o}, \boldsymbol{\theta}) \tag{1}$$

Here and in the following, s represents the simulated value, o represents the observed value, and * represents the true value.

2) **Input errors**

Assume the input errors are represented by input multipliers sampled from an uncorrelated lognormal distribution, and the observed input $\boldsymbol{X}^{o}$ can then be related to the true input $\boldsymbol{X}^{*}$ by the following equation:

$$\boldsymbol{X}^{o} = \boldsymbol{X}^{*} \exp(\boldsymbol{\varepsilon}_{X}), \boldsymbol{\varepsilon}_{X} \sim N(\mu_{X}, \sigma_{X}^{2}) \tag{2}$$

where the $\boldsymbol{\varepsilon}_{X}$ are assumed to follow a Gaussian distribution with mean μ_{X} and variance σ_{X}^{2}.

3) Output observational errors and model structural errors

In the derivation, these two parts are assumed to be error-free, therefore,

$$\boldsymbol{Y}^{o} = \boldsymbol{Y}^{*} \tag{3}$$

$$M() = M^{*}() \tag{4}$$

4) Remnant errors

Based on the previous assumptions, the observed output equals the true output, and the difference between the simulated output and the observed output, $\boldsymbol{\varepsilon}$, will be equal to the difference between the simulated output and the true output, as follows:

$$\boldsymbol{Y}^{\mathrm{s}} = \boldsymbol{Y}^{\mathrm{o}} + \boldsymbol{\varepsilon} = \boldsymbol{Y}^{*} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim (0, \sigma^{2}) \tag{5}$$

where the remnant errors $\boldsymbol{\varepsilon}$ are assumed to follow a Gaussian distribution with mean 0 and variance σ^{2}.

5) Bayesian inference

Following the study of Renard et al. (2010), the posterior distribution of all inferred quantities is given by Bayes' theorem as follow:

$$\begin{aligned} & p(\boldsymbol{\theta},\boldsymbol{\varepsilon}_X,\mu_X,\sigma_X,\sigma \mid \boldsymbol{Y^o},\boldsymbol{X^o}) \propto \\ & p(\boldsymbol{Y^o} \mid \boldsymbol{\theta},\boldsymbol{\varepsilon}_X,\boldsymbol{X^o})\, p(\boldsymbol{\varepsilon}_X \mid \mu_X,\sigma_X)\, p(\boldsymbol{\theta},\mu_X,\sigma_X,\sigma) \end{aligned} \tag{6}$$

The full posterior distribution comprises the following three parts: the likelihood of the observed output $p(\boldsymbol{Y^o} \mid \boldsymbol{\theta},\boldsymbol{\varepsilon}_X,\boldsymbol{X^o})$, the hierarchical parts of the input multiplier $p(\boldsymbol{\varepsilon}_X \mid \mu_X,\sigma_X)$ and the prior distribution of deterministic parameters and hyperparameters $p(\boldsymbol{\theta},\mu_X,\sigma_X,\sigma,)$.

Renard et al. (2009) argue that in the IBUNE method, the $\boldsymbol{\varepsilon}_X$ are different in different iterations due to random sampling, therefore, cannot be updated effectively due to breaking the theoretical foundation of Bayesian inference. In the BEAR method, the secant method is applied to find a deterministic relationship between the $\boldsymbol{\varepsilon}_X$ and the model parameters $\boldsymbol{\theta}$ and hyperparameters of the multipliers μ_X,σ_X. Therefore, $\boldsymbol{\varepsilon}_X$ can be determined by $\boldsymbol{\theta},\mu_X,\sigma_X$, as follows:

$$\boldsymbol{\varepsilon}_X = f(\boldsymbol{\theta},\mu_X,\sigma_X) \tag{7}$$

Considering $\boldsymbol{\varepsilon}_X$ are sampled from $N(\mu_X,\sigma_X^2)$, $p(\boldsymbol{\varepsilon}_X \mid \mu_X,\sigma_X)$are the same when $\mu_X,\sigma_X$are determined and do not need to be considered in the algorithm. Therefore, the posterior distribution of all inferred parameters in the BEAR method are as follows:

$$\begin{aligned} & p(\boldsymbol{\theta},\boldsymbol{\varepsilon}_X,\mu_X,\sigma_X,\sigma \mid \boldsymbol{Y^o},\boldsymbol{X^o}) \propto \\ & p(\boldsymbol{Y^o} \mid \boldsymbol{\theta},\mu_X,\sigma_X,\boldsymbol{X^o})\, p(\boldsymbol{\theta},\mu_X,\sigma_X,\sigma) \end{aligned} \tag{8}$$

The problem of parameter estimation and error identification can then be interpreted as the calibration of $\boldsymbol{\theta},\mu_X,\sigma_X$ when the relationship between the errors and parameters are determined. In the value estimation of the errors, only estimating the parameters and errors together can achieve this, however, in the rank estimation of the errors, the secant method can be applied to realize the rank between each error and its corresponding residual error (depending on the parameters).

Appendix C: The results of the synthetic cases with the increasing standard deviation of output observational errors

In order to explore the ability of the BEAR method with the interference of other sources of errors, synthetic cases have been built as the following table.

Table C. 1 Description of the synthetic cases

subfigure	Input error model	Output error model	Prior information of input error model in calibration
(a)	$X^{o} = X^{*}\exp(\boldsymbol{\varepsilon}_X)$, $\boldsymbol{\varepsilon}_X \sim N(0.2, 0.3^2)$	$Y^{o} = M(X^{*}, \theta)\exp(\boldsymbol{\varepsilon}_Y)$, $\boldsymbol{\varepsilon}_Y \sim N(0, \sigma_Y^2)$ $\sigma_Y = 0, 0.1, 0.2, 0.3$	$X^{o} = X^{*}\exp(\boldsymbol{\varepsilon}_X)$, $\boldsymbol{\varepsilon}_X \sim N(0.2, 0.3^2)$
(b)(c)			$X^{o} = X^{*}\exp(\boldsymbol{\varepsilon}_X)$, $\boldsymbol{\varepsilon}_X \sim N(0.2, {\sigma_X}^2), \sigma_X \in (0,1)$

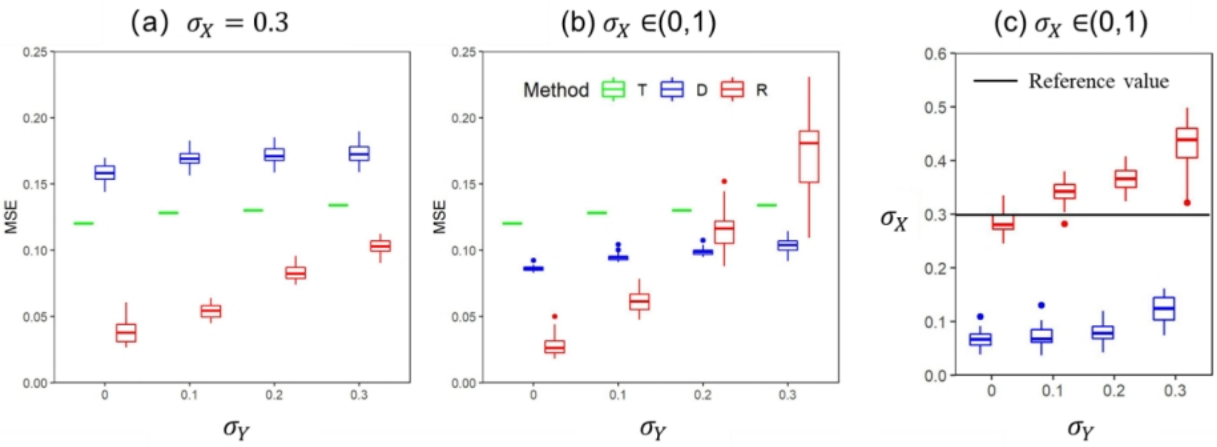


Figure C.1 Mean square error of modified input vs synthetic input(a,b) and the estimated standard deviation of input error (c) for different synthetic cases (T represents the traditional method without considering the impacts of input error, D represents the IBUNE method, and R represents the BEAR method)

Figure C.1(a) demonstrates if the prior information of standard deviation of input errors is accurate, the BEAR method will always bring a better improvement of the input data, while the IBUNE method leads to a worse modified input than the input data without modification in method T. Figure C.1(b) illustrates that if the standard deviation of input errors is estimated in a wide range, whether the BEAR method can improve the input data or not depends on the significance of the other sources of errors. The more dominant the input error is, the more effective the BEAR method is. When there are no other sources of errors($\sigma_Y=0$), the BEAR method can isolate the input error and parameter error perfectly as the standard deviation of input error converges to the reference value. However, the results of the IBUNE method are consistent with a much smaller estimation of the standard deviation of input errors in Figure C.1(c).

To sum up, the ability of the BEAR method depends on the accuracy of the prior information of the input error model and the significance of the input error in the residual error. It makes sense that other sources of error will interfere with the identification of the input error and its prior information can constrain these negative impacts. While the IBUNE method can slightly modify the input data only when the standard deviation of the estimated input error is much smaller than the true value. It is most likely to make use of the stochastic errors to improve the original input data, but not really identify the input error.

References

AJAMI, N. K., DUAN, Q. & SOROOSHIAN, S. 2007. An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water resources research,* 43.

BONASSI, F. V. & WEST, M. 2015. Sequential Monte Carlo with adaptive weights for approximate Bayesian computation. *Bayesian Analysis,* 10**,** 171-187.

RENARD, B., KAVETSKI, D. & KUCZERA, G. 2009. Comment on "An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction" by Newsha K. Ajami et al. *Water Resources Research,* 45.

RENARD, B., KAVETSKI, D., KUCZERA, G., THYER, M. & FRANKS, S. W. 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research,* 46.