Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-563-RC3, 2020

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Quantifying input uncertainty in the calibration of water quality models: reshuffling errors via the secant method" *by* Xia Wu et al.

Anonymous Referee #3

Received and published: 17 December 2020

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:

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.

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.

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.

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 pre-treatment 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.

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.

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.

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.

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).

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.

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).

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

Line 76-77: A critical thing here is that ε 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 εp 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.

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.

Line 81-83: I would be very careful with this statement. This only happens if the model parameters θp and the input error residuals εxp 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 εxp yield 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.

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.

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.

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.

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.

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.

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.

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.

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

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

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

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.

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).

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.

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.

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.

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.

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

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.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-563, 2020.

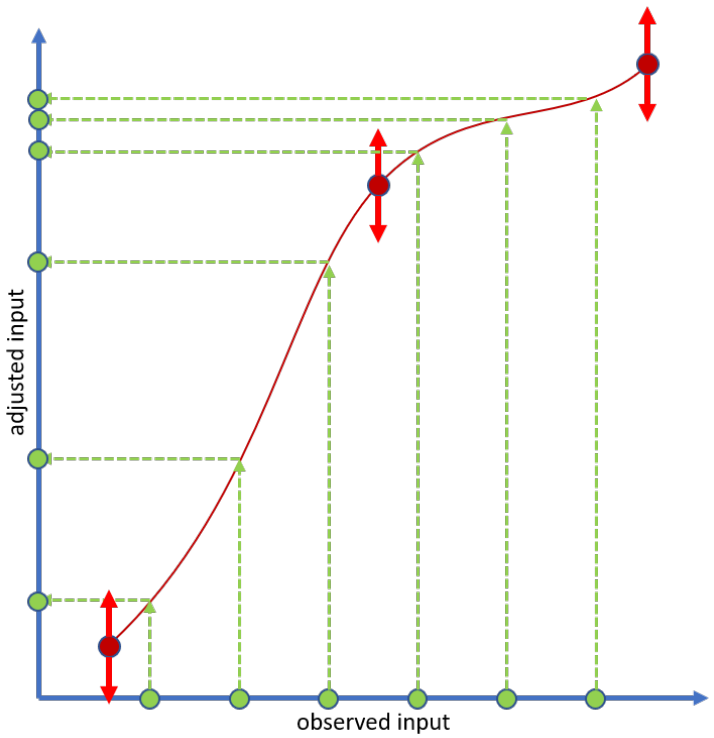

Fig. 1. Example of a non-linear re-scaling with a spline defined by three control points (red dots). You could use such a spline to scale the input values non-linearly.