Authors' Response to Reviewers of "Guidance on evaluating parametric model uncertainty at decision-relevant scales" by Smith et al.

Reviewer comments are in black

Author responses are in blue

Proposed text edits are in red

RC1

I have read with interest the manuscript entitled 'Guidance on evaluating parametric model uncertainty at decision-relevant scales'. The study examines the sensitivity of the simulations of a spatially distributed ecohydrological model to the model parameters for calibration purposes. Sensitivity is considered with respect to different model output metrics that correspond not only to performance metrics assessed at the basin outlet, but also statistics of the model output calculated for the different hillslopes of the basin where no output observations are available.

The study is a welcome contribution to the field of sensitivity analysis of spatially distributed models, which is challenging due to the high dimensionality of the parameter space of these models and which requires further investigations. The study discusses the issue of calibration and uncertainty estimation in absence of output observations, in particular at internal locations of a river basin where model-based information is critically needed to support decision making.

Overall, the manuscript is well written, the analyses were performed with care and the experiments are well documented in the supplements. However, I have a number of suggestions and I think that a number of points need clarification, in particular regarding the choice of (output) sensitivity metrics and the analysis of the multipliers, as detailed below.

Thank you for reviewing our manuscript and for your suggestions. We will adopt many of them into our revised version, as explained below.

p1 L8 'parameter multipliers': I suggest adding 'for spatially distributed parameters' for clarity.

We propose changing this sentence to:

We use global sensitivity analysis to screen parameters for model calibration, and to subsequently evaluate the appropriateness of using multipliers to adjust the values of spatially distributed parameters to further reduce dimensionality.

p1 L15-16 ‘for some parameter multipliers […] reducing dimensionality.’: This needs clarification.

We agree this was confusing as written. Including the word “adjust” in the edit above should help to interpret this sentence. Additionally, we propose revising this sentence to directly reference the SA results, and also propose to add a second key point that we make about the use of multipliers.

3) for some multipliers, calibrating all parameters in the set being adjusted may be preferable to using the multiplier if they have significantly different parameter sensitivity values, while for others, calibrating only a subset of the parameters in the set may be preferable if they are not all influential.

p2 L31 ’sensitivity metrics’: I suggest specifying what this term refer to for clarity (e.g. performance measure or statistics o the simulated model output).

We propose removing metrics from this sentence. The next paragraph details what is meant by sensitivity metrics.

For studies that aim to use the resulting model to spatially optimize decisions, sensitivity should be defined for the decision objective values.

p2 L34-35: Could you provide some references/examples for this?

We propose citing studies from within the existing reference list as examples at the end of the sentence. Gupta and Razavi (2018) in section 1.2 describe different model performance metrics, and they cite many more studies as examples of each type of metric.

(e.g., Herman et al., 2013; van Griensven et al., 2006; Chen et al., 2020)

p2 L37-39 ‘Matching […] with controlling extremes.’: A link between this sentence and the rest of the paragraph is missing.

We propose changing the previous sentence to:

Common calibration performance measures are used to quantify model performance across all flow magnitudes, yet some measures like the Nash-Sutcliffe Efficiency (NSE) lump several features together (Gupta et al., 2009) and specific features can govern the resulting performance value (e.g., peak flows for NSE in Clark et al., 2021).

The abuse of popular performance metrics in hydrologic modeling - Clark - - Water Resources Research - Wiley Online Library

Because our finding that the parameters selected by NSE more closely resemble 5th-95th percentile flow parameters than the 95th percentile parameters, contrary to what might be expected by the results of the Clark et al. study, we propose also adding this sentence to the discussion:

For the Baisman Run watershed, parameters selected by the NSE may more closely resemble 5th - 95th percentile flow parameters because the flows >95th percentile are relatively small, and so the model residuals are a similar order of magnitude for peak flows and other flows.

p3 L92-93 'will be evaluated […] error model': The authors should clarify whether they refer here to future studies that may use the guidance presented in the manuscript.

Sorry for the confusion; we should have used the present tense as we evaluate sensitivity at ungauged locations in this paper. We propose re-writing this sentence:

We do not consider stochastic methods because we evaluate sensitivity in ungauged locations where no data are available to inform an error model.

p4 L107 'performance measure': Consider revising the terminology. The metric of Eq. (2) is not really a performance measure, as it does not use observed values.

That is correct that Eq. 2 is not actually error, so SAE is not an appropriate name for this. This should be sum of absolute median deviation (SAMD). We propose changing this terminology throughout the text. We also propose calling this a "relative variability measure" instead of a performance measure. For completeness, we propose computing the SAMD for the basin outlet as well, and comparing to the SAE in Eq 1.

p5 L130-131 'Therefore, […] the TN estimation method': This sentence is not clear to me. Observed data are also used for the streamflow objectives (Eq. (1)). In addition, does the water quality objective only consider the basin outlet or also hillslopes?

We do not use an error model for streamflow, but WRTDS as a regression model does have an error model (normal distribution for log[TN]). So, we could simulate the quantile estimates of TN in addition to the regression-predicted mean, and do sensitivity analysis on those quantiles. We propose modifying the existing text and added a clarifying sentence to explain this more clearly (starting in Line 127):

As described in Section 3.1, we used a linear regression model with normal residuals to estimate the log-space TN concentration at the outlet as a function of streamflow at the same location. As such, we could compute water quality sensitivity metrics for estimated quantiles from the regression error model, in addition to the regression-predicted mean. The water quality sensitivity metrics corresponded to 1) the 95th percentile of the distribution of estimated TN concentration, 2) the 5th percentile, and 3) the log-space mean (real-space median) on each of the days on which TN was sampled.

Yes, the water quality objectives are only evaluated at the basin outlet. This is solely because the WRTDS regression is valid only for the outlet. With a better spatial prediction model for TN, sensitivity could and should be evaluated across the catchment, as we do for streamflow.

p5 Section 2.3 A justification for the choice of the four calibration performance measures is missing (e.g. a justification could be that these metrics are typically used in previous studies, and in this case some references to some of these studies should be provided).

These are typical calibration metrics in hydrology. We propose changing the first sentence to:

Four performance measures that are typically used to calibrate hydrologic models are used...

And propose citing the following for NSE, pBias, and log-likelihood metrics:

Moriasi, D., Arnold, J., Van Liew, M., Bingner, R., Harmel, R., & Veith, T. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE, 50, 885.

Smith, T., Marshall, L., & Sharma, A. (2015). Modeling residual hydrologic errors with Bayesian inference. Journal of Hydrology, 528, 29-37. https://doi.org/10.1016/j.jhydrol.2015.05.051.

In addition, I think that the metric of Eq. (1) could also be used for calibration purposes. Therefore the difference between the decision-relevant and calibration-relevant sensitivity metrics is fuzzy.

Equation 1 is a performance measure, and that could be used for calibration. Decision relevance vs. calibration relevance is determined by which metrics are used (e.g., using only the lower 5th percentile of flows). We propose describing this more clearly by making section 2.2 "Decision-Relevant and Calibration-Relevant Sensitivity Metrics", and have sub-sub-sections for decision-relevant and calibration-relevant metrics (these are currently sub-sections 2.2 and 2.3). Some of the first paragraph of current sub-section 2.2 would be used in new sub-section 2.2, along with the following paragraph that defines performance measures, and decision and calibration relevant metrics:

Decision relevance is determined by sensitivity metrics that are evaluated on subsets of the timeseries that are relevant to decision objectives. While these metrics could be used for model calibration, that is an uncommon choice because the model would be unlikely to perform well on other data subsets (e.g., Efstratiadis and Koutsoyiannis, 2010). Calibration relevant metrics therefore use the full timeseries in this paper. Performance measures are a way to temporally aggregate a timeseries into a single value indicative of model fit to the observed data (e.g., Moriasi et al., 2007). The performance measures that we selected could all be used for model calibration, but the selected measure for decision-relevant metrics is not commonly used for calibration.

Efstratiadis, A. & Koutsoyiannis, D. (2010) One decade of multi-objective calibration approaches in hydrological modelling: a review, Hydrological Sciences Journal, 55:1, 58-78, DOI: 10.1080/02626660903526292

p6 L170 'EEs for each parameter […] in the parameter domain.': Please add a reference for this (e.g. Pianosi et al. (2016)).

We cited the original methods paper (Morris, 1991) in the paragraph above the equation.

p6 L178 'Step changes': Does this refer to the quantity Delta_{s+1,s,p} of Eq. (7)? Please clarify.

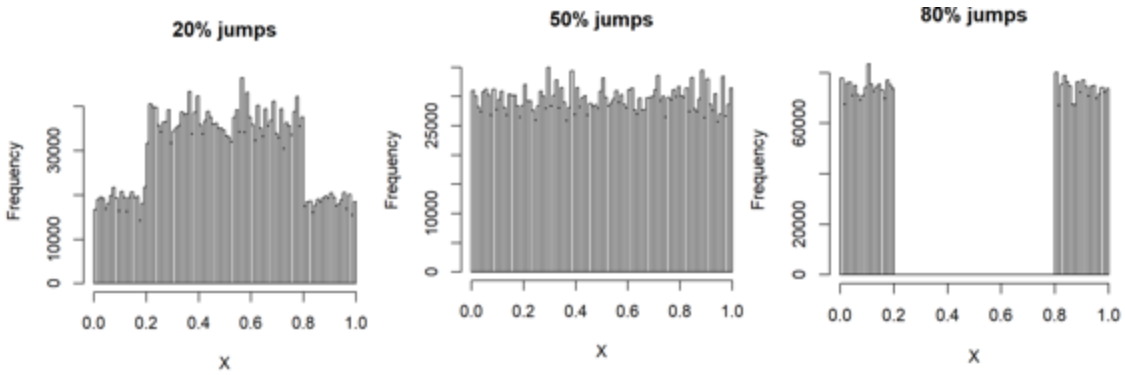
Yes, these are the deltas. We propose changing to:

Step changes, Δ,

p6 L178-179 'to allow for a uniform […] within the specified bounds.': This needs to be better explained.

We do not know of a reference that explains this numerical method, so we propose adding a supplementary figure that illustrates this. We also propose changing the sentence to:

Step changes in parameter values were set to 50 levels, i.e. 50% of their range. For each parameter, this allows for a uniform distribution of parameter values across all samples (cite SI figure below).



Histograms for the Morris trajectory samples for parameter X as a function of the delta step size. Percentages are relative to the range of X.

p8 L217 'mean EE value': Clarify whether this refers to the metric defined in Eq. (8).

We propose changing to:

mean EE value (Eq. 8)

p8 Sect. 2.6: I do not understand this point. The use of a multiplier for a certain parameter type is based on the assumptions that the value of the parameters of this type vary in the same proportions in different locations of the basin, and I do not understand why these parameters should have similar sensitivity (here EEs).

If the parameters truly have a proportional adjustment, then it is still fine to use the multiplier even if the sensitivities are significantly different. We propose editing this section and parts of the results to say that parameters with significantly different sensitivity values *are candidates for* being calibrated individually. More investigation on the cause for their difference in sensitivity could inform the decision to calibrate individually or using a multiplier (e.g., is the difference in sensitivity caused by the parameters acting in vastly different proportions of the watershed area?)

p11 L300-303: This is based on a rather strong assumptions that the C-Q relationship obtained using the WRTDS regression method can be extrapolated to other flow conditions. This may not be appropriate for in-depth nitrogen study, in particular in agricultural areas with highly varying nitrogen inputs. I understand that it is not the point of this study to have a sophisticated nitrogen routine. However, I think that some comments should be added to further highlight the simplifications/limitations of the data-driven nitrogen routine used. My comment also refers to the sentence p21 L503-505.

In line 303, we propose modifying to:

Simulated flows that were outside of the observed range of values were assigned the parameters for the nearest flow value in the table. Extrapolation of the concentration-flow relationship to more extreme flows than were historically observed may provide inaccurate TN estimates, which is a limitation of this statistical prediction method. We expect the error from extrapolation in this basin to be low, as N loads appear to be dominated by effluent from septic systems as evidenced by isotopic sourcing (Kaushal et al., 2011, p. 8229), and septic effluent supply should be fairly steady over time.

Kaushal, S.S., Groffman, P.M., Band, L., Elliott, E.M., Shields, C.A., and Kendall, C. (2011). Tracking nonpoint source nitrogen pollution in human-impacted watersheds. Environmental Science & Technology, 19(45), pp. 8225-8232. https://doi.org/10.1021/es200779e

p13 L359 'healthy ecosystem': I suggesting replacing this expression by something like 'more humid ecosystem' for clarity.

We propose changing "compared to a healthy ecosystem" to:

compared to non-drought conditions

p14 L373-374: 'This result demonstrates […] to calibrate.': This sentence needs clarification.

We propose changing to:

The only reason for differences in which parameters are selected for calibration using the three TN metrics is uncertainty in the mean EE. EE error bars tend to be larger for the upper 95th percentile TN estimate, which results in the selection of more parameters to calibrate. This result demonstrates the value of considering both model error (different TN quantile estimates) and uncertainty in sensitivity (bootstrapped EE estimates) when selecting which parameters to calibrate. More parameters are found to be potentially influential when considering these sources of uncertainty.

p21 L511-514: The authors could refer to the study by Cuntz et al. (2016), which also demonstrates the importance of including in the calibration some parameters that are typically set to fixed values and in particular hard-coded parameters, using the NOAH-MP land surface model.

We agree this is a good paper to cite here.

p22 L553-555: I think that it should also be emphasized that, because of the issue equifinality, calibration strategies that identify an ensemble of possible parameter sets (as compared to a unique 'best' solution) and that therefore consider parameter uncertainty are more appropriate.

We propose modifying this sentence to:

Thus, due to equifinality, calibration methods that estimate parameter distributions are preferable to relying upon a single "best" parameter set; considering such parametric uncertainty in optimizations of engineering control measures should help to discover solutions that are robust to it.

Minor edits

p3 L61: I suggest replacing 'will' that 'are typically'.

Propose changing to "are used to inform"

p7 L192: replace 'smaller' by 'smallest'.

Propose accepting suggestion

References:

Cuntz, M., Mai, J., Samaniego, L., Clark, M., Wulfmeyer, V., Branch, O., et al. (2016). The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model. Journal of Geophysical Research, 121(18), 10,676-10,700. https://doi.org/10.1002/2016JD025097

Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B., & Wagener, T. (2016). Sensitivity analysis of environmental models: A systematic review with practical workflow. Environmental Modelling & Software, 79, 214–232. https://doi.org/10.1016/j.envsoft.2016.02.008