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

Editor and Reviewer comments are in black. Note that many of the actual changed text edits are the same as proposed in our replies to reviewer comments. We have noted new edits with {highlighted text} preceding the reviewer comment. Line numbers in our replies correspond to the revised manuscript PDF without tracked changes in it.

Author responses are in blue

Text edits are in red

Editor:

Dear authors,
The manuscript has now received two reviews. Both reviewers see merit in this manuscript, though both have raised several points needing clarification.

I've thoroughly read the responses and proposed revisions from the authors, and find them to be a thoughtful treatment of planned changes. I appreciate that the authors separated their responses and their proposed text edits.

A majority of the comments from the reviewers, especially reviewer 2, encourage being explicit around the framing of the manuscript and the major outcomes from the manuscript. I encourage the authors to place particular effort to clarify these points of confusion raised by the reviewers. Overall, the manuscript is well written, and minor clarifications encouraged by the reviewers will improve its impact.

I do think that the major points outlined by the authors in their abstract and throughout the manuscript make a notable contribution to the literature, particularly the literature connecting sensitivity analysis and decision-making. However, I agree with the reviewers that it is worth investing time to ensure that these key messages are clear, and can be interpreted through your figures. As pointed out by one reviewer, it can be challenging to interpret the figures. One option may be to add a visual guide within each figure on how to read/interpret some of your figures (e.g., Fig 4, Fig 5), to reduce the amount of information shown (e.g., reduce the numbers included in Figure 5 to highlight key information), and/or to simply show larger figures or break subplots up (e.g., Figure 2, Figure 5). I really want to dissect Figure 2 and Figure 5, but am struggling to extract information from these given there is so much information packed in, as noted by one of the reviewers.

I encourage the authors to proceed with a thorough revision and I look forward to reading an updated manuscript.

Thank you for your review of our manuscript. We have implemented the proposed changes in response to reviewers and documented them in this response. There are other minor text edits as tracked changes in the PDF. Additionally, we adopted several figure changes to improve their presentation:

Figure 2: Font size for the numbers and legend labels have been increased by 50%, and the figure is now on a full page. The legend has been moved outside of the plot area. A table has been added for the parameters corresponding to the numbers above the error bars.

Figure 4: y axis labels were removed for all but one sub-plot. Increased image size.

Figure 5: The original Figure 5A and 5B were split into Figure 5 and 6. Both figures were enlarged relative to 5A and 5B. The new Figure 5 x and y axes' labels were improved.

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 adopted many of them into our revised version, as explained below.

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

We changed this sentence to:

L7-9:

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 revised this sentence to directly reference the SA results, and also added a second key point that we make about the use of multipliers.

L15-17:

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 removed “metrics” from this sentence. The next paragraph details what is meant by sensitivity metrics.

L32-33:

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 now cite 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.

L39-40:

(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 changed the previous sentence to:

L40-44:

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 of the hydrologic time series 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 also added this sentence to the discussion:

L490-492:

Another possibility is that in the Baisman Run watershed, flows greater than the 95th percentile are still 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 modified this sentence to:

L97-98:

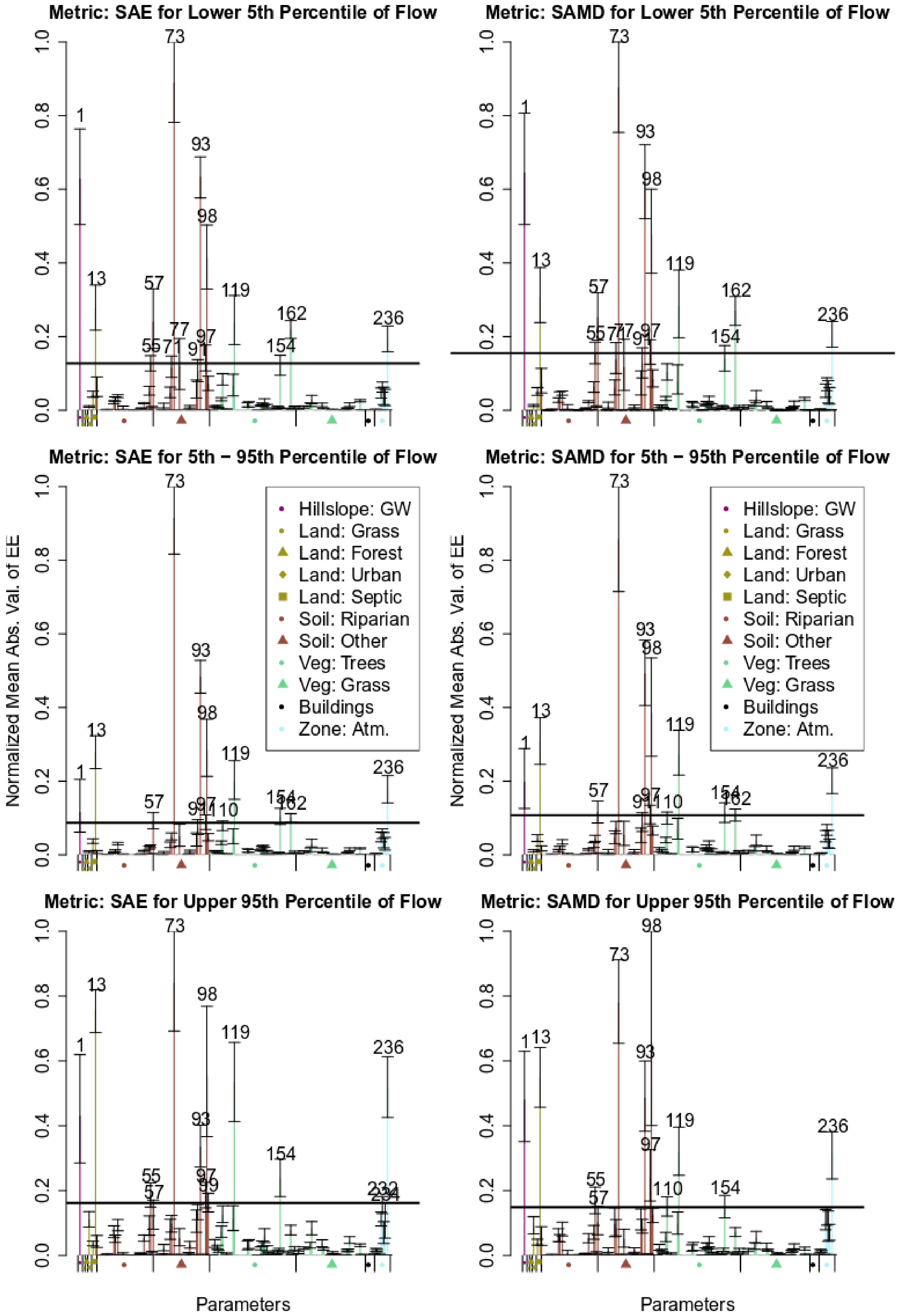
We do not consider stochastic methods because we evaluate sensitivity in ungauged locations where no data are available to inform an error model.

{new analysis, paragraph, and supplementary figure} 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 changed this terminology throughout the text. We also called this a “relative variability measure” instead of a performance measure. For completeness, we computed the SAMD for the basin outlet as well, and compared results to the SAE (supplementary figure below in low res).

L117-124:

For the basin outlet, we used the sum of absolute error (SAE) as the performance measure for decision-relevant sensitivity metrics. Because performance measures require an observation time series to compute, we needed a different approach to measure relative variability for hillslope sensitivity analysis. We used the sum of absolute median deviation (SAMD), where the median value was computed across all model simulations of each hillslope. For completeness, we also used the SAMD for the basin outlet and compared to the SAE results in supplementary material (item S9). We found similar parameter selection and sensitivity ranking results for each method, which demonstrates that an observation time series is not necessary to obtain the parameter set to calibrate, although observations help to check that SA model simulations are reasonable. In this paper, we present basin outlet results for the SAE. The SAE and SAMD expressions are shown in Equations 1 and 2.


Metric: SAE for Lower 5th Percentile of Flow
Metric: SAMD for Lower 5th Percentile of Flow
Metric: SAE for 5th − 95th Percentile of Flow
Metric: SAMD for 5th − 95th Percentile of Flow
Metric: SAE for Upper 95th Percentile of Flow
Metric: SAMD for Upper 95th Percentile of Flow
Normalized Mean Abs. Val. of EE
Normalized Mean Abs. Val. of EE
0.0
0.2
0.4
0.6
0.8
1.0
73
1
93
98
13
57
119
162
236
55
71
77
91
97
154
110
99
232
234
Parameters
Parameters
• Hillslope: GW
• Land: Grass
▲ Land: Forest
◆ Land: Urban
■ Land: Septic
• Soil: Riparian
▲ Soil: Other
• Veg: Trees
▲ Veg: Grass
• Buildings
• Zone: Atm.

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 modified the existing text and added a clarifying sentence to explain this more clearly:

L141-147:

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 time, season, and 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. Later in the paragraph, we state that these metrics are used to represent different features of the hydrologic time series. We changed the first sentence to:

L151:

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

And added the following citations 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.

{new paragraph}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 have described this more clearly by making section 2.2 "Sensitivity Metrics" with two sub-sections for decision-relevant (2.2.1) and calibration-relevant (2.2.2) metrics (these were formerly sections 2.2 and 2.3). The following paragraph composes Section 2.2, which defines performance measures and decision- and calibration-relevant metrics:

L105-115:

In many hydrological studies, sensitivity analysis is used to understand how input parameters influence model performance measures (Jackson et al., 2019), such as the Nash-Sutcliffe efficiency. Performance measures are a way to temporally aggregatea time series into a single value that is indicative of model fit to the observed data (e.g., Moriasi et al., 2007). Gupta and Razavi(2018) note that using such performance measures as sensitivity metrics amounts to a parameter identification study to discoverwhich parameters may be adjusted to improve model fit. Therefore, the calibration-relevant sensitivity metrics in this paper usesuch performance measures on the full time series. Evaluating performance measures for subsets of the time series that describe specific features of interest (Olden and Poff, 2003) should identify those parameters that control processes that generate thosefeatures (e.g., timing vs. volume metrics in Wagener et al., 2009). Therefore, decision-relevant sensitivity metrics are evaluatedon subsets of the time series 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 andKoutsoyiannis, 2010). The following subsections present the decision- and calibration-relevant sensitivity metrics.

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 changed this to:

L196:

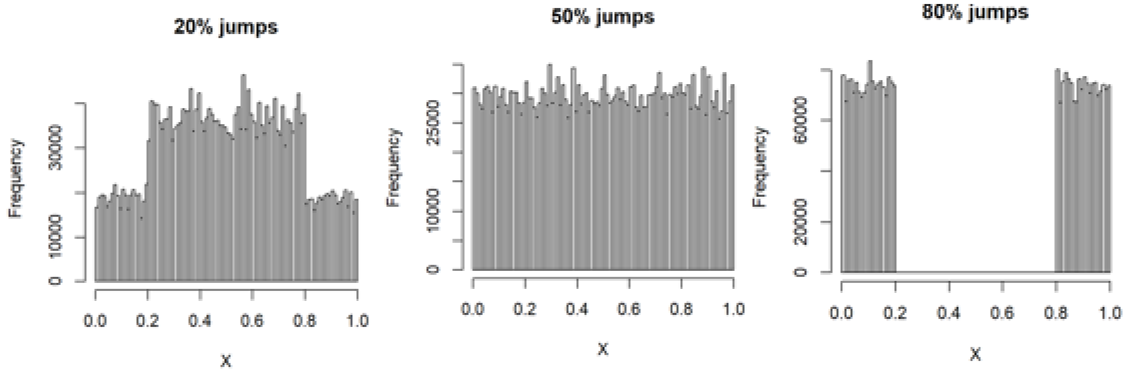
Step changes, Δ,

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 added a supplementary figure that illustrates this. We also changed the sentence to:

L196-198:

Step changes, Δ, 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 (example sampling distributions for other percentages are provided in supplementary item S8).



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 changed this to:

L236:

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).

Now Section 2.5

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 edited 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.

Note that our study watershed is not agricultural and N loads appear to be dominated by septic systems. We edited this to:

L322-327:

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 changed “compared to a healthy ecosystem” to:

L383:

compared to non-drought conditions

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

We changed this to:

L396-401:

The 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 have cited this paper 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 modified this sentence to:

L585-587:

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’.

Changed to “are used to inform”

p7 L192: replace ‘smaller’ by ‘smallest’.

Accepted 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

RC2

1. L8-10: “We evaluate six sensitivity metrics that align with four decision objectives; two metrics consider model residual error that would be considered in spatial optimizations of engineering designs.” This sentence is confusing -- do the six sensitivity metrics add up to the four decision objectives + two model residual error metrics?

   Yes, that is correct. We modified the sentence to:

   L9-10:

   We evaluate six sensitivity metrics, four of which align with decision objectives and two of which consider model residual error…

2. L39: extremes or high flows?

Engineering controls can work to lower high flows and raise low flows. We changed this to:

L45:

Extreme high and low flows

3. It would be helpful to somewhere define “engineering controls”.

   We added examples to the first sentence in the introduction.

   (e.g., green and gray infrastructure)

4. L 116: here are you write flooding, low flow, reservoir water supply objectives, but earlier you wrote flooding, low flow, and all other flows. If these are the same, that should be explicitly stated.

   We changed L131 to:

   flooding, low flow, and all other flow objectives

5. L320: “The goal of this sensitivity analysis is to inform the selection of parameters to calibrate a RHESSys model that could be used in such a reforestation optimization.” is this the overall goal of this paper? If so, this should be stated in the introduction section much earlier.

   We added the following sentence at the end of the first paragraph of the introduction:

   L35-37:

   In this paper, we evaluate the influence of decision-relevant and calibration-relevant sensitivity metrics on parameter selection for calibration, and discuss the potential implications on subsequent model calibration and optimization of water management decisions.

6. L 328: are the elementary effects for all the parameters normalized on a percentage basis? Why compare the 95th percentile for the elementary effects to the overall mean of all

parameters' elementary effects, if that is what is being explained in this sentence? What does the 95th percentile estimate for the elementary effects mean?

As explained in Section 2.4, we completed bootstrapping of the elementary effects to generate a distribution of mean absolute values of elementary effects for each parameter. This recognizes that there is uncertainty in our estimate of a parameter's mean absolute elementary effect. From that distribution, we obtain the 95th percentile estimate of the mean absolute value EE for each parameter. Normalization in each panel of figure 2 is completed by taking the maximum 95th percentile EE value across all parameters and setting that to 1 (i.e., all EEs are divided by this value). The minimum is 0. So, the figure normalization is not on a percentage basis. It should also be noted that the EE is a normalized metric to begin with, as it is the absolute value of the change in the output metric per change in the input parameter, where the change in the input parameter is 50% of its range in our study.

Also stated in Section 2.4, we do not compare 95th percentile EEs to the overall mean across all parameters EEs. We sort the mean absolute EE values across parameters from largest to smallest and find the top X%. We then compare each parameter's 95th percentile estimate of its mean absolute EE to the X%-ile. Any parameter whose 95th percentile mean absolute EE estimate is above that threshold is selected for calibration.

7. Figure 3 seems to be referred to before figure 2 (L335).

   This reference to Figure 3 is to state that we will discuss a point in more detail later in the same section. We plan to leave this as-is unless asked to change.

8. L349: does an elementary effect value of exactly 0 mean that this parameter has no effect on the stream flow or hillslope metric? It would be helpful to state this explicitly.

   Yes, it does. We added the following to the end of the sentence:

   L373-374:

   (i.e., these parameters do not affect model-predicted streamflow)

9. {new table} The text discussing figure 2 is useful, (line 348 in the rest of this paragraph), but without knowing what the specific parameter numbers are in figure 2, I'm not sure what to take from this graphic.

The supplementary material provides the full list in number order in a spreadsheet. We discuss the parameters with the largest elementary effects within the text that you mention.

We have also included a table of the numbered parameters within Figure 2.

Number	Parameter Name
1	H: GW Loss Coef.
13	L4: Septic Water Load
55	Riparian Soil (S8+S108): m
57	S9: Soil Thickness
70	S9: Pore Size
71	S9: Air Entry Pres.
73	S9: Sat. to GW Coef.
77	S109: Soil Thickness
91	S109: Air Entry Pres.
93	S109: Sat. to GW Coef.
97	Other Soil (S9+S109): Ksat
98	Other Soil (S9+S109): m
99	Other Soil (S9+S109): Poro.
110	Tree: Day Leaf On
118	Tree: Leaf Cuticular Cond.
119	Tree: Max Stomatal Cond.
154	Tree: Stomatal Fraction
162	Tree: Rainwater Capacity
232	Z: Avg. Temp. Coef.
234	Z: Atm. Trans. Coef. 2
236	Z: Wind Speed

10. Line 480: I see now that engineering designs are not explicitly evaluated in this paper. My earlier comment (comment 3), asked about what engineering controls meant. The focus on engineering controls in the introduction section led me to believe that this paper would be about engineering controls. Rather it seems that this paper has implications for where to locate engineering controls but does not directly investigate this placement. If this is accurate, then I would suggest deemphasizing engineering controls from the introduction section.

    This is correct that we use siting of engineering controls as a motivating reason for doing a spatially distributed sensitivity analysis. Calibrated models are used to optimize these and other water management decisions, and parameter screening is used to reduce the dimensionality of the search to make the calibration more tractable. We propose broadening the introduction to say "engineering controls and water management decisions" to be more applicable. We changed the first sentence of the introduction to:

Spatially distributed hydrologic models are commonly employed to inform water management decisions across a watershed, such as the optimization of locations of engineering control measures.

11. From what I understood of this article, the first main finding was that parameters describing watershed characteristics are sometimes important for modeling hillslope hydrologic response even though they do not affect the streamflow at the model outlet much. The authors state that this might be important in the spatial location of engineering controls. There are many other reasons why getting the hydrology right within the watershed is important (modeling of spatially distributed soil moisture, etc.), but a major limitation is that we don't normally have data to compare to within the watershed, so in practice it would be hard to calibrate these parameters that don't affect streamflow much. The second main finding was that commonly used metrics (e.g., NSE) are not as sensitive to the decision relevant streamflows that we would want them to be. These are both important findings and points to make, but I found the article overall hard to read and understand. The authors may be served by focusing the text on the main findings and reducing discussion of peripheral topics.

These are two key points of the article. Because there are so few papers on decision-relevant sensitivity metrics, we thought it would be useful to provide an extended discussion that describes the importance of having the decision objectives in mind when completing a sensitivity analysis and subsequent calibration and optimization. Spatial sensitivity analysis is also often limited by data in sub-catchments and can lead to calibration challenges. We think it is important to discuss how the resulting parametric uncertainty can be used in robust optimizations.

We propose a better framing of the intent to use SA results to inform calibration of a model that is used to optimize decisions. We propose adding this sentence to the end of the first paragraph (same sentence as mentioned in reply to comment 5):

In this paper, we evaluate the influence of decision-relevant and calibration-relevant sensitivity metrics on parameter selection for calibration, and discuss the potential implications on subsequent model calibration and optimization of water management decisions.