

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. 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,

I have now received three anonymous reviews in response to your revised article. One reviewer is happy with the changes you've made, while two other reviewers encourage further revision of your article. The majority of the reviewers were positive and commented on the strong substance of your article, but encourage some reframing of your article (particularly in the introduction). The reviewers also noted a few minor changes and clarifications needed as well.

Overall, I encourage you to undertake the recommendations provided by the reviewers. I look forward to receiving your revised manuscript.

[We have revised the manuscript introduction and made other clarifying suggestions to the text. We also revised the figures and added a summary table, as suggested by Reviewer 4.](#)

Reviewer #1

The authors have done a good job addressing my comments and I do not have remaining comments.

[Thank you for reviewing our revised manuscript. Your comments in the first round greatly improved the paper.](#)

Reviewer #3

The idea behind the manuscript is interesting however, I find that the overall organization of the paper should be improved to favor the comprehension of researchers not familiar with GSA methods. Diverse parts of the manuscript are difficult to read and the quality of the figures still needs to be improved.

As a second major point, I invite the authors to revise the literature carefully because there are some relevant studies on sensitivity-based parameter calibration, for the identification of the relative importance of the parameters, and the locations where measurements should be

collected being the model most sensitive to its parameters.

If the authors will address the two major points above, the novelty that the paper brings will be clarified as well as the possibility to replicate the methodology will be facilitated.

Thank you for reviewing our manuscript.

For your first point, we cite many recent review papers on GSA and spatial SA. Because these review articles are available, we do not see a need to be more introductory to GSA methods in this paper.

We also modified the figures and added a summary table, as discussed in our response to Reviewer 4.

For your second point, the purpose of this paper is to provide decision-relevant guidance, not general guidance, on sensitivity-based parameter calibration. We agree there are many studies that provide general guidance, and we already cite some examples in the manuscript. However, we are not aware of studies that focus on a decision-informed SA for model calibration.

Reviewer #4

This paper aims to evaluate the variability in model parameters that result from sensitivity analyses based on (1) different spatial scales; and (2) different objective functions/ metrics. The paper is well written overall, and I think makes very salient and clear arguments for the importance of considering spatial scales and metrics relevant for management. I appreciate that the authors clearly put a lot of really hard work into this research, and think that the message is important for the larger scientific community. However, I think that the important message that this research is trying to get across is getting lost in the details, some of which are only peripherally related to the research. I think that the authors can do more to clarify the message, and remove non-essential points (or move to the SI). I further found some of the figures difficult to decipher, and the framework in the introduction a little confusing/ jumbled. Overall I believe that this paper will be a fine contribution to HESS once these issues are addressed.

Thank you for thoroughly reviewing our manuscript. We think your suggestions help to improve the clarity of our work. Please see our responses to your comments below.

Major comments

1. The text is dense and overly detailed in some places. I believe the take home messages are important but are getting lost in the details. I suggest the authors remove any methods, results, and discussion section that isn't relevant to the main points to the SI.

We have shortened and merged some sentences and paragraphs in the revised version. Please see the tracked changes for these minor edits.

For example:

- a. I think the entire section 2.3.1 (Elementary Effects for Parameters with Relational Constraints) is not essential to have in the main text. Rather, the authors could briefly state that the 271 parameters in RHYSSes were summarized into 237 parameters by combining those that are structurally dependent (and then refer to the SI for additional detail).

We accepted this suggestion and added the following sentence in the Section 3 model description.

Some parameters are structurally dependent, so we aggregated EEs for these parameters, resulting in 237 unique EEs for each sensitivity metric (supplementary information item S0 describes the aggregation method)

- b. Since the authors did not conduct a stochastic modeling approach, leave any discussion related to stochastic modeling (i.e., much of Section 2.1) to the discussion.

We mention stochastic modeling in Section 2.1 because it motivates considering model error for TN. We think that the uncertainty sources associated with an error model make more sense in Section 2.1 where other uncertainty sources are described.

2. Terminology should be simplified throughout.

- a. Objectives vs. sensitivity metrics vs. objective functions vs. performance measure. These terms are used throughout and should be clarified. The term “objectives” is particularly confusing as the reader may associate this term with “objective function”, which I believe the authors refer to as “sensitivity metrics” (?). I suggest that the authors read through the MS carefully and think about where it is possible to simplify and reduce these terms.
 - i. For example, lines 130 – 135, the authors use “objectives” to refer to overarching management goals for water quality and quantity (“We consider water quantity and quality objectives as they are among the most common for hydrological modeling studies”), then sensitivity metrics to refer to flooding, low flow, and other flow objectives (“We evaluate three streamflow sensitivity metrics relevant to flooding, low flow, and all other flow objectives, respectively”), and then define these again as objectives (“These mutually exclusive objectives are respectively quantified as 1) flows greater than the historical 95th percentile, 2) flows less than the historical 5th percentile, and 3) flows between the historical 5th and 95th percentiles”). This is only one example of excessive/confusing use of these terms.

- ii. It seems that there is a lot of overlap between these terms, or maybe they are the same. In any case, I suggest creating a table that describes the different levels and defines the metric names, and then use these metric names consistently throughout, see example below.

Applied to		Scale	Performance measure	Sensitivity metric name
<i>Decision - relevant metrics</i>				
Flow	High flows	Basin	SAE	Basin _{high flows}
		Hillslope	SAMD	Hillslope _{high flows}
	Low flows	Basin	SAE	Basin _{low flows}
		Hillslope	SAMD	Hillslope _{low flows}
	Other flows	Basin	MAE SAE	Basin _{other flows}
		Hillslope	SAMD	Hillslope _{other flows}
Water quality	High TN concentration	Basin	SAE	TN _{high}
	Low TN concentration	Basin	SAE	TN _{low}
	Mean TN concentration	Basin	SAE	TN _{mean}
<i>Calibration – relevant metrics</i>				
Flow	All flows	Basin	NSE	?
			LNSE	?
			pBias	?
			Log likely-hood	?

Thank you for pointing out the potential for confusion among these terms. They are all different, with the exception of “sensitivity metrics” and “objective function”, the latter of which we have removed from the paper to avoid confusion. We already had a paragraph dedicated to the definition of sensitivity metrics vs. performance measures (Section 2.2). The definition of objective is defined in the introduction for calibration and decision objectives. In most cases, we have rephrased as “decision maker’s objectives” to be clear this is not an objective function.

We say in Section 2.2.2 that the performance measure is as you list in the table and the sensitivity metric is the application of the performance measure to all flows. We do not use MAE for other flows or anywhere in the manuscript.

For the first use of performance measure, we now distinguish it from sensitivity metric:

Common calibration performance measures are employed as sensitivity metrics by evaluating performance across all flow magnitudes

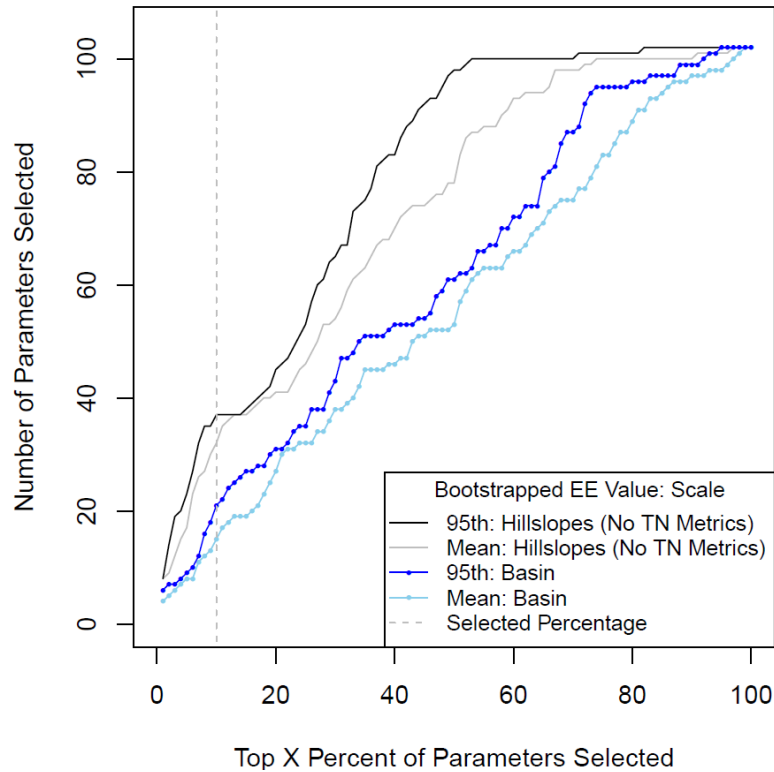
We have also added a new table in line with what you suggested:

Table 1. Table of decision-relevant and calibration-relevant sensitivity metrics.

Sensitivity Metric		Scale	Performance Measure
Decision-Relevant Metrics			
Streamflow	High Flow Days Low Flow Days Other Days	Basin	SAE
	High Flow Days Low Flow Days Other Days	Hillslope	SAMD
TN Concentration	High TN, All Days Mean TN, All Days Low TN, All Days	Basin	SAE
Calibration-Relevant Metrics			
Streamflow	All Flows, All Days	Basin	NSE LNSE pBias LogL

- a. The 95th percentile terms are confusing on Figure 1; I was getting the lines on the plot confused with the 95th percentile flow sensitivity metrics, 5th percentile flows sensitivity metrics, etc. I think the authors could simplify and clarify by referring to the 95th percentile flows simply as “high flows”, and 5th percentile flows as “low flows” after defining in table.

We edited the legend title to say “Bootstrapped EE value: Scale”



3. Introduction could be improved to better frame the two issues this paper is tackling. From my understanding, the main two issues are: (1) spatial scales of calibration do not match spatial scales relevant for management; and (2) calibration performance metrics do not represent hydrologic outcomes relevant for management. I think the issue of equifinality – which is relevant for issue (1) -- is getting mixed up in issue (2) in lines 38 – 52. Below I suggest an outline for first few paragraphs of the introduction.
 - a. Management controls are spatially distributed throughout a watershed, and therefore modeling management approaches often call for spatially explicit models (i.e., distributed models)
 - i. Distributed models require calibration of many parameters, some of which are not even observable
 - ii. This calibration is challenging since observations are rarely available at scales needed to constrain all of these parameters; watershed outlets are gauged only so calibration is performed at the watershed scale.
 - iii. This leads to the well-known issue of equifinality when unknown parameter values are not constrained many ways to get to the same answer at the outlet large uncertainty in parameter values at local (finer) scales.
 - b. This presents two major challenges for modeling studies that aim to evaluate impacts of decision making....
 - i. Spatial scales of calibration do not match spatial scales relevant for management. Equifinality is particularly problematic for watershed

models that aim to predict effects or optimize locations of management controls, since these are sensitive to local scale parametrization (which is highly uncertain when the model is only calibrated to a single location)

- ii. Calibration performance metrics do not represent hydrologic outcomes relevant for management. A further, even more basic issue faced by modeling management decisions is that the majority of calibration performance metrics (e.g., NSE) are not necessarily, or explicitly, sensitive to hydrologic outcomes relevant for decision making (e.g., high flows, low flows).
- c. The combination of these two issues have consequences....

Thank you for this suggested reorganization of the introduction. We have taken your suggestion to restructure the first 2 paragraphs of the introduction by presenting equifinality in the first paragraph.

4. Figures should be simplified and made more legible. I provide detailed suggestions below.

Thanks for these suggestions. See our comments below for how we addressed them.

5. If the authors found similar parameter selections using SAMD and SAE at the basin scale (lines 120 – 124), why did they proceed with SAE for the basin outlet scale? It seems like for the purposes of comparing hillslope scale to basin scale selected parameters, it would be more defensible to use SAMD for both.

We were asked in the first revision to evaluate SAMD for the basin outlet and found that the results were similar. Keeping SAE in the main text was largely a time consideration. We point to the supplementary material for a comparison of SAMD and SAE for decision-relevant flow sensitivity metrics.

Minor comments

- Many opportunities to simplify language by re-arranging sentences and avoiding passive voice, for example:
 - Line 23: “such as the optimization of locations of engineering control measures...” could be simplified to “such as the optimal locations of engineering control measures...”
 - Line 24: “Accurate simulations of streamflows and nutrient fluxes in ungauged locations are desired to estimate the impact of control measures...” could be simplified to “Quantifying the impact of control measures requires accurate estimates of streamflow and nutrient fluxes in ungauged locations...”

We accepted these edits and simplified sentences in other parts of the manuscript.

- Line 25: “...on multiple objective functions” This is not essential to the point this paragraph is trying to make, and also introduces a new term that hasn’t been defined. If the authors think that mentioning multiple objective functions is necessary in the second sentence of the MS, I suggest defining it first. Otherwise, remove from this sentence.

We removed objectives from this sentence.

- Line 30 – 32: “Reviews of sensitivity at the outset of a study.” These two sentences could be shortened and simplified: “Recent reviews of sensitivity analysis methods for spatially distributed models (e.g., Pianosi et al., 2016; Razavi and Gupta, 2015; Koo et al., 2020b; Lilburne and Tarantola, 2009) emphasize the need to consider, at the outset of a study, the definition of sensitivity within the study context.”

We simplified these sentences in line with your suggestion.

Recent reviews of sensitivity analysis methods for spatially distributed models \citep{Pianosi2016,Razavi2015,Koo2020position,Lilburne2009} emphasize the critical need to answer, at the outset of a study, “What is the intended definition for sensitivity in the current context?” \citep{Razavi2015}.

- Line 33: “decision objective values” is a confusing term that has not been defined yet. What are “decision objectives” and how are they different from “decision objective values” in this sentence?

We changed this sentence to be more general:

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

- Line 35-37: “In this study, we evaluate...water management decisions”. This sentence seems like it would fit better towards the end of the introduction – I was a little thrown off that the authors describe the objectives of the paper that at the end of the first paragraph, but then go on to provide further motivation/ background (P2), and then go back to the objectives of the paper again (P3).

We have removed this sentence because we have a similar sentence later in the introduction.

- Line 50-51: “This would suggest there is equifinality...across the watershed.” I’m not sure that the fact that distributed stormwater control outcomes are affected by different parameters than watershed scale outcomes suggests that there is equifinality. Equifinality exists regardless of whether a stormwater control is being simulated in the model. I think I would suggest the authors use the fact that equifinality is a rampant issue in distributed models and poses unique challenges for simulating stormwater control measures, which are often distributed across a watershed. In other words, introduce equifinality earlier on in the introduction (i.e., in P1 where the authors

describe the fact that these models have hundreds of parameters that need to be calibrated).

We edited the introduction to introduce the concept of equifinality in paragraph 1.

- Line 78-72: “the results we obtain...impact on sensitivity metrics.” These two sentences are confusing as they are written in a passive voice; it is unclear whether the authors “provide general guidelines for spatially distributed models” and “inform prioritization of data collection efforts”, or whether this was done separately/ by another study/ in practice.

We have rewritten to say:

We use the results of a comprehensive sensitivity analysis of all non-structural model parameters to provide general guidelines for spatially distributed models and some specific recommendations for RHESSys users.

- Line 93-95: “If employing a stochastic modeling approach...could be considered in a sensitivity analysis”. again, since this paper focuses on parametric uncertainty and assumes a static model, this does not seem relevant and could be removed. Moreover, these lines include terms that are not (a) defined previously, like error model shape, and (b) are not used again in the manuscript – this additional information detracts from the main point of the paper by distracting the reader (or, at least me!).

The TN model considers residual error, and we think this paragraph helps to motivate considering residual error in SA. We provide citations to papers that explore each of the concepts in more detail. We revised the error model sentence to say:

...additional uncertainty sources include the choice of residual error model shape (e.g., lognormal) \citep{Smith2015}...

- Line 118 – 119: “Because performance measures require an observation time series to compute, we needed a different approach to measure relative variability for hillslope sensitivity analysis. At the hillslope scale, we use...” I suggest rephrasing and simplifying: “At the hillslope scale (where observation time series are not available), we use the sum of absolute median deviation...”

We edited this sentence:

For hillslopes (where observations are not available) we used the sum of absolute median deviation (SAMD), where the median value for each hillslope was computed across all model simulations.

- Line 130 – 133: “We consider water quantity and quality objectives historical 5th and 95th percentiles.” These sentences are a little confusing because there are so many different terms used and it’s not clear what they all refer to (see major comment 2a above). Suggested revision: “We consider sensitivity metrics related to decision-making for water quantity and quality outcomes as they are among the most common for hydrological modeling studies. For water quality, we quantify SAE (basin scale) and SAMD (hillslope scale) separately for (1) high flows (flows greater than the historical 95th percentile), (2) low flows (flows less than the historical 5th percentile), and (3) all other flows (flows between the historical 5th and 95th percentiles).”

We simplified these sentences based on your suggestion.

We consider sensitivity metrics that are relevant to water quantity and quality outcomes because they are among the most common for hydrological modeling studies. For water quantity, we compute SAE (basin) and SAMD (hillslopes) for three mutually exclusive flows: 1) high flows greater than the historical 95th percentile, 2) low flows less than the historical 5th percentile, and 3) all other flows between the historical 5th and 95th percentiles.

- Lines 143 – 145: Somewhere in here the authors should state which performance measure they used here (SAE?).

Added.

The water quality sensitivity metrics are the SAE for...

- Lines 165 – 167: “We selected the likelihood model based on...which is a generalized normal distribution.” Suggest simplifying: “We selected the skew exponential power model (a generalized normal distribution) as the likelihood model due to its ability to fit the wide range of residual distribution shapes that result from random sampling.”

We edited this sentence:

We selected the skew exponential power (generalized normal) distribution \citep{Schoups2010} as the likelihood model due to its ability to fit a wide variety of residual distribution shapes that could result from random sampling of hydrological model parameters.

- Line 237: “Then, we flagged...” Does “flagged” mean “selected”?

We revised this sentence to remove flagged.

All of the parameters whose estimated 95th percentile EE values were greater than this cutoff value would be selected for calibration for that metric.

- Lines 269 – 272: “While authors Lin and Band...unrealistic mortality).” This sentence isn’t essential for the point of the paragraph. I suggest moving this to the discussion or SI.
We think it’s important to keep part of this sentence to justify why we didn’t simulate nitrogen from RHESSys.

We found that randomly sampling non-structural growth model parameters within their specified ranges commonly resulted in unstable ecosystems (e.g., very large trees or unrealistic mortality).

- Section 4. Case study site description. The order of the sentences in this paragraph are a little disjointed. I suggest moving lines 341 – 344 (“The Baisman Run watershed...reforestation optimization.”) to before the sentence starting on line 337 (“After a five year spin-up period...”). This would make it so first you present all of the background info on the watershed, and then you discuss your modeling approach. As it is, you describe the watershed, discuss your modeling approach, and then describe the watershed again.

We accepted this suggestion.

- Line 334 – 345: “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.” This was surprising to me, since the introduction really focused on stormwater control measures, not reforestation. If this truly is the goal of the paper, the introduction needs to be revised to focus on reforestation efforts. Also, this is a strange place to put the goal of the paper – it should be in the introduction (and it is, in fact, but the introduction states that “The goal is to discover to which parameters the decision objectives are most sensitive across the watershed”, which is different than that stated in lines 334 – 345).

Thanks for pointing out that we say reforestation here. An in-prep paper based on this work focuses on reforestation. We have changed this to say “stormwater infrastructure optimization” in this paragraph. We also edited the short non-technical summary of our paper

Watershed models are used to simulate streamflow and water quality, and to inform siting and sizing decisions for runoff and nutrient control projects. Data are limited for many watershed processes that are represented in such models, which requires selecting the most important processes to be calibrated. We show that this selection should be based on decision-relevant metrics at the spatial scales of interest for the control projects. This should enable more robust project designs.

The use of “goal” here is a poor word choice. We replaced “goal” with “hypothetical motivation” so it’s not confused with the goal of our paper.

The hypothetical motivation for this sensitivity analysis is to inform the selection of parameters to calibrate a RHESSys model that could be used in an optimization of stormwater infrastructure that aims to control flooding and reduce nutrient exports.

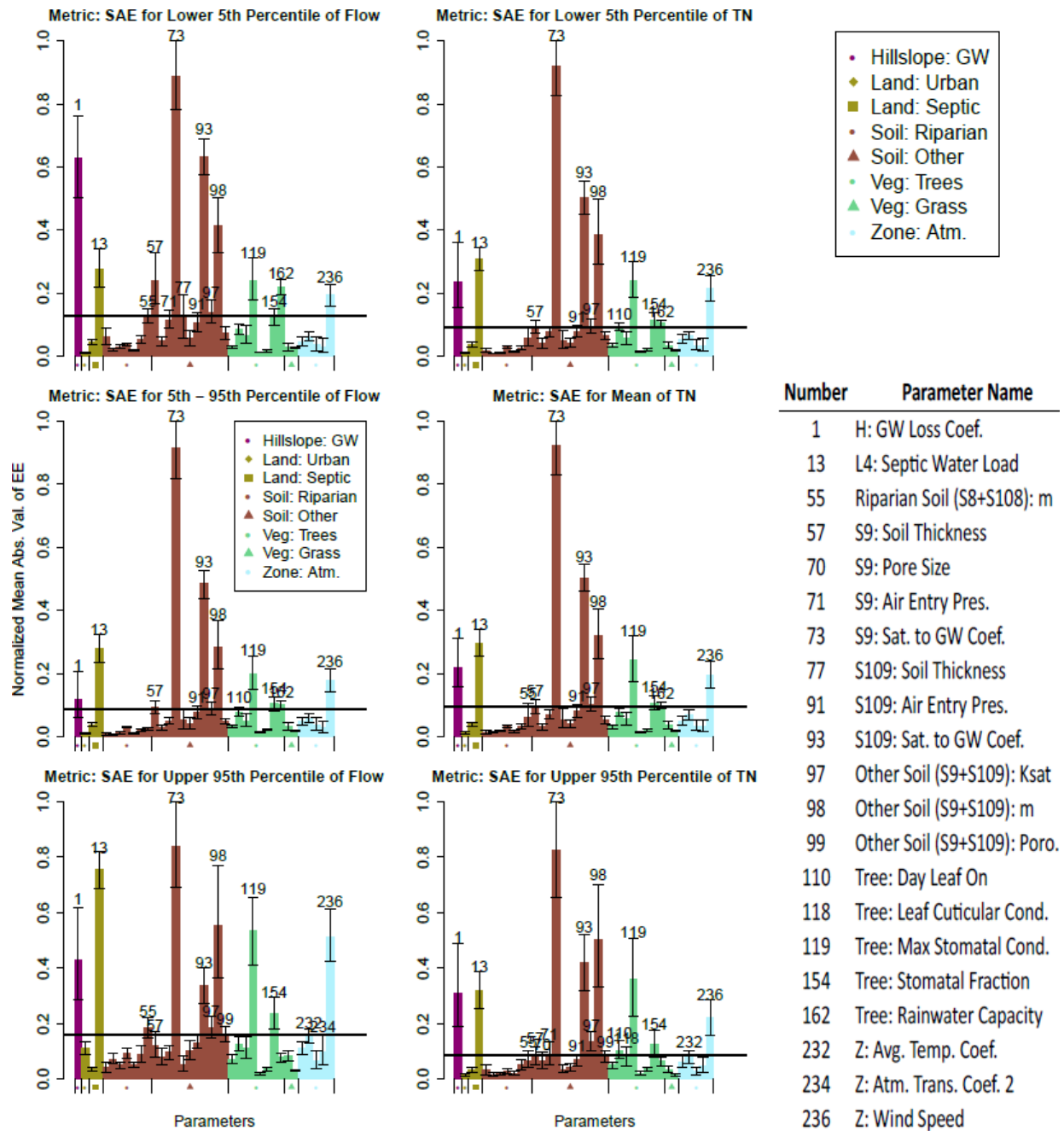
- Lines 301 – 307: This paragraph might fit better at the end of a section (i.e., end of the intro, methods or case study site description).
The paragraph is currently at the end of the section describing the RHESSys model, and that seems like the best place to us. We do not want to end the Introduction or Methods with this paragraph, as it describes a less general contribution of our work to understanding parameter sensitivities in RHESSys, specifically; however, we do not want to end the Case Study Site Description with this paragraph, as it describes a more general contribution than to just our model site.
- Lines 369 – 271: If I am interpreting this correctly, these lines are saying that 21 parameters were selected for basin outlet, 18 of which were based on streamflow metrics, and 19 based on TN metrics. This would imply that out of the 21 parameters selected, only 5 are not overlapping between the streamflow and TN metrics. This, to me, does not necessarily support “using sensitivity metrics for each output variable or objective” since there is actually a lot of overlap between the parameters that were selected.
You’re interpreting correctly. If a set of sensitivity metrics for streamflow or TN were used, then there would be either 2 or 3 parameters missing from the calibration that are statistically significantly important for the other set of metrics. Choosing to not include them wouldn’t be justifiable. While this may seem like a small number of parameters to miss, excluding them could have decision-relevant implications for stormwater design. We would also like to note that it is somewhat surprising the parameters are not fully overlapping since our TN modeling is based on a regression with streamflow.
- Line 375: top row should be left column
- Line 393: bottom row should be right column
We made these edits.
- Line 409 – 411: “The majority of the watershed is forested...correspond to power lines.” This seems like watershed background that should be moved to the case study site description (Section 4)
We moved these sentences to Section 4.

- Line 581 – 582: “If we select all parameters...that will lead to more parameters compared to using only the gauge location.” This sentence is confusing, suggest revising: “More calibration parameters result from sensitivity analysis at local scales (i.e. ungauged hillslope) than do from sensitivity analysis at watershed scales.”
We accepted this edit.

Figure comments

- Suggest adding a conceptual figure to the beginning of the methods to describe overall approach
We added the table that you suggested and we think that serves well as a conceptual overview of what sensitivity metrics and scales we compare.
- Figure 2.
 - Suggest transposing the subplots so that the flow metrics are all along a single row, and TN metrics are in the second row. This would make it easier to compare across the different flow and TN metrics.
We had the figure in this orientation in the previous version of the manuscript, but the figure panels were too small to be useful.
 - Suggest only showing those that meet the 10% threshold (very hard to distinguish between lines as is, lots of the numbers overlap)
 - This could free up some space along the x-axis for parameter names, rather than symbols/ numbers
We edited the figure to show only those parameters that are selected by any of the metrics in Table 1. We kept the symbols because they are used in Figure 3
 - The caption says this provides the EEs for “the six sensitivity metrics”, but I only see SAE, which would imply this is only for the basin scale decision-relevant metrics? What about SAMD (hillslope scale), and all calibration relevant metrics? The text (line 372) says Figure 2 shows “basin scale EEs”, but still this doesn’t explain why calibration relevant metrics aren’t included. Again, I think this is an issue of terminology and should be clarified throughout, but I point it out specifically here since the caption of the figure is incorrect, or the text is misleading.
We have edited the figure caption to:

Mean absolute value of elementary effects for RHESSys model parameters evaluated for the six decision-relevant sensitivity metrics at the basin outlet.



- Figure 3
 - Separate into two figures: one with land cover maps and hillslopes (currently A and B), and one with EE ranks and indicators (currently C and D).
 - Make the land cover maps Figure 1, move up to be with the case study site description (Section 4), where they are already referenced
 - To further simplify this figure, consider grouping the hillslopes based on relevant properties (i.e, forested/ non-forested/ impervious) and using the mean EE across hillslopes in that group. This would be more meaningful for the reader

(and would support the points the authors make in lines 412 – 439), and would simplify the figure a lot.

Thank you for these suggestions. We have modified the figure in line with these suggestions. Panel A is now the former panel C, and panel B is the former panel D with an additional aggregation across hillslopes 1-8 and 9-14.

