Review of: "Guidance on evaluating parametric model uncertainty at decision-relevant scales"

Summary and recommendation

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.

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. 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).
    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.

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
Decision - relevant metrics				
Flow	High flows	Basin	SAE	Basin $_{\text{high flows}}$
		Hillslope	SAMD	Hillslope $_{\text{high flows}}$
	Low flows	Basin	SAE	Basin $_{\text{low flows}}$
		Hillslope	SAMD	Hillslope $_{\text{low flows}}$
	Other flows	Basin	MAE SAE	Basin $_{\text{other flows}}$
		Hillslope	SAMD	Hillslope $_{\text{other flows}}$
Water quality	High TN concentration	Basin	SAE	TN $_{\text{high}}$
	Low TN concentration	Basin	SAE	TN $_{\text{low}}$
	Mean TN concentration	Basin	SAE	TN $_{\text{mean}}$
Calibration – relevant metrics				
Flow	All flows	Basin	NSE	?
			LNSE	?
			pBias	?
			Log likely-hood	?

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.

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….

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

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.

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…”
- 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.
- 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.”
- 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?

- 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).
- 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).
- 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.
- 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 seems 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!).
- 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…”
- 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).”
- Lines 143 – 145: Somewhere in here the authors should state which performance measure they used here (SAE?).
- 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.”
- Line 237: “Then, we flagged…” Does “flagged” mean “selected”?
- 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.
- 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.

- 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).
- 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).
- 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.
- Line 375: top row should be left column
- Line 393: bottom row should be right column
- 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)
- 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."

Figure comments

- Suggest adding a conceptual figure to the beginning of the methods to describe overall approach
- 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.
    - 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
    - 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.

- 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.