Reply to Anonymous Referee #2
Review received and published: 3 September 2020

Dear Reviewer,

Thanks a lot for your thorough review and the valuable suggestions. We will reply below in detail to your comments. Your comments are *italic*; our replies are highlighted **bold**. The **line numbers in red** are referring to the revised draft.

Best regards,
Julie, James, and Bryan

The manuscript focuses on Sensitivity Analysis (SA) of hydrological models. It introduces a more general version of the well-known Sobol' method, designed to operate on groups of parameters instead of on individual parameters. Overall I enjoyed reading the manuscript - its on a topical area and the methods described are sound. I appreciate this work on mathematical model analysis, and the idea of grouped parameter sensitivity is novel at least in hydrology as far as I know. With multi-model/flexible frameworks such as RAVEN and others, analysis of their sensitivity would benefit from such "grouped" analysis.

We thank the reviewer for this positive evaluation and are glad to hear that it was an enjoyable read.

I have the following concerns with the current manuscript form:

1. *The algorithms are not explained in a sufficiently clear way. For example, for the description of Sobol' method on lines 300-307, and the description of the xSSA method on lines 324-330, are in my opinion not sufficient for a paper presenting a mathematical method. Yes, I could probably translate the description there into a procedure/ pseudocode, but: first I would not be quite sure if I got it right, and second I (respectfully) suggest the onus is on the authors to provide such an un-ambiguous description. Appendix B is helpful to a degree, but seems to use a different notation to the main text (where are the matrices A and B and Cm?).*

   We understand that the reviewer might have been confused due to the lack of some information. The Sobol' method is however one of the most cited methods in sensitivity analysis and part of every single package providing implementations to this method. We refer to the most relevant publication, i.e. Sobol' (1993) in the section about the traditional Sobol' Sensitivity Analysis (Sec. 2.2.1) and even explain the method in that section. There is really nothing more to it than sampling a matrix $\mathcal{A}$, $\mathcal{B}$ and construction the matrices $\mathcal{C}_m$ as explained. The user would not even need to do that if using a package. The matrices are of dimension $K \times N$ with K being 1000 in our experiments and K being 11 to derive for example the parameter sensitivities of the shared benchmark problem. We do not think that listing those matrices would help to understand the method. We adjusted slightly the text regarding the sampling of the matrices $\mathcal{A}$ and $\mathcal{B}$ and hope that it is now more clear that those are purely sampled:

   line 323 ff. **For the numerical estimation of the indexes S_{x_i} and ST_{x_i}, one samples two base matrices $\mathcal{A}$ and $\mathcal{B}$ which each contain K parameter sets (rows) of N parameters (columns). The samples are assumed to be independent within one matrix and between the matrices. We used the stratified sampling of Sobol' sequences here to improve convergence speed of the derived indexes compared to a Monte-Carlo sampling.**

In Appendix B we solely list the analytical results of the shared-parameter benchmark as a reference that users could test their implementations in case this is needed. We would also like to emphasize that we are planning to make all codes used here openly accessible once this manuscript gets accepted for publication.

2. *Terms such "uncertainty", "sensitivity", "influence", "importance" are being used in a pretty loose, seemingly interchangeable way. For example, the paragraph on lines 31- 40, which starts with "uncertainty" and then immediately switches to "sensitivity". Then line 104 mentions "sensitive/influential/important" parameters. Are these referring to the same characteristic? Similar confusing usage then carries through later in the manuscript. I suggest the terminology should be much tighter to avoid confusion. Given the mathematically demanding topic, I would suggest giving clear definitions of the various concepts (with links to existing literature where appropriate), and avoiding the alternation of these terms in the remainder of the presentation. There are useful and interesting ideas on lines 100-115, but these are already using the terms above in a way I found unnecessarily confusing because its not clear which terms are used synonymously and which are not.*

   We have revised the final paragraph to the introduction clarifying some general definitions for the entire manuscript. We hope that it is now more clear that we use most of the terms you listed interchangeably. We apologize that this was confusing in our first version of the draft. For example:

   line 141 ff. We also wish to mention that the terms 'sensitive' and 'influential' are used interchangeably throughout this work.

   We agree that it might lead to too much leeway of interpretation for a reader when using too many interchangeable terms. We went through the manuscript and hope that we reduced this ambiguity. Here are some examples:

   line 521 ff. The analysis of model parameters (Fig. 5A) shows that the most sensitive ones are [...]

   line 586 ff. The strong impact of these processes (together with the input adjustments) highlights the sensitivity of streamflow regarding snow and melting processes in this mountainous, energy-limited catchment.

   line 589 ff. This demonstrates that soil and surface processes are of secondary sensitivity regarding streamflow. Their sensitivity may increase if the uncertainty of the snow and melting processes can be reduced, i.e. by narrowing parameter ranges during calibration.

   line 610 ff. Evaporation (dark blue) is [...], expectedly, less sensitive during winter. Snow balance (medium green) and potential melt (orange) are sensitive as long as snow is present (Nov to May).

   The current literature review is heavily focused on sensitivity analysis - which is appropriate given the topic. But if the connection to uncertainty is to be made, I would say the literature review of the latter is currently rudimentary at best.

   We agree which the reviewer that our literature review heavily focuses on the sensitivity topic given that we try to make a contribution to tat field. There is a connection between uncertainty of a parameter (reflected in the range we associate to a parameter and we then use as search space in calibration or to derive the sensitivity of a parameter) and its sensitivity. The range of a parameter is influencing its sensitivity, i.e. if the (uncertainty) range of a parameter gets reduced its sensitivity will also get reduced. But it might be that the parameter overall is insensitive (no matter which range picked). So, it is hard to determine from the range alone a sensitivity of a parameter; hence sensitivity analyses are performed. We thought that this is an obvious connection and did not want to

distract the reader by putting too much emphasize on that topic. We have made the following adjustments in the manuscript:

line 16 ff. **[...] such information may readily inform model calibration and uncertainty analysis.**

line 34 ff. **A key purpose of model sensitivity analysis is to inform model calibration or model uncertainty analysis so as to focus either of these analyses on only the model inputs/model structural choices the model outputs are most sensitive to.**

3. *The aims and key contributions of the study seem to drift over the course of the manuscript/ presentation. For example the Introduction is focused on sensitivity analysis (and to some extent uncertainty) - but in the Conclusions the contribution #1 is listed as formulating model ensembles as weighted sums of process options, with Sensitivity Analysis then being contribution #2. I think the coherence between the introduction/ aims and contributions could be improved, so that there is a clearer set of aims, appropriate background given on each aim, and then a clear set of conclusions that match those aims.*

We totally agree with that and apologize that the key contributions had not been clearly stated in the introduction. We rephrased the following paragraph to match the order and wording we use in the conclusions. We hope that our key contributions are now more clear to the reader.

line 96 ff. **Two main contributions of this work are to (A) reformulate a hydrologic modeling framework so that it can define model structure by weighting or blending of discrete model process options continuously for simulating process level hydrologic fluxes and (B) to propose a technique, the Extended Sobol' Sensitivity Analysis (xSSA) method, based on the existing concept of grouping parameters when applying the Sobol' method (Sobol and Kucherenko, 2004; Saltelli et al., 2008; Gilquin et al., 2015) to derive the sensitivity of a model prediction (here streamflow) to model structural choices.**

A clearer vision of the contributions could also help improve the structure of the manuscript, by putting the important contributions much earlier. This would avoid the multiple forward references to the proposed method and its properties before its actual description is given- e.g., see lines 235-237, which are not really that meaningful before seeing how the xSSA method operates. The new xSSA method in Section 2.2.2 comes after several quite detailed sections on models and case studies - and it was not immediately apparent that this is the main advance being presented.

We indeed started our initial draft with the proposed reversed order (first explaining the xSSA method including the grouping and weights of process options and then explaining the benchmark models, study domain and Raven). The problem then is that most readers will have trouble to understand what we mean with process options and processes (in a real-world) example. We therefore decided to explain first all the "vocabularies" (i.e., benchmarks, hydrologic model, process options and processes) before getting to how a sensitivity analysis would work without defining groups (Sec. 2.2.1) leading to the version that is then based on groups, i.e. xSSA (Sec. 2.2.2). This is then followed by all the experiments we intent to run (Sec. 2.3). We made a few adjustments in the introductory paragraph for the Material & Methods. We are hoping that this will help the reader to navigate through this long section and maybe directly go to the section with the major contribution:

line 144 ff. **The section will first introduce the models and their setups (Sec. 2.1) used to test and validate the proposed Extended Sobol' Sensitivity Analysis**

(xSSA) method as here applied to determine model structure and parameter sensitivities. In section 2.2, we will briefly revisit the traditional method of Sobol' that is so far primarily used to obtain model parameter sensitivities (sensitivity metric A; Sec. 2.2.1) before we introduce the major contribution of this work (Sec. 2.2.2) which supports sensitivity estimates for model process options (sensitivity metric C) and model processes (sensitivity metric D) besides the sensitivities of model parameters (sensitivity metric B). Finally, we present the experiments used to test the proposed method and address the research questions raised in the introduction (Sec. 2.3).

4. *Some lack of clarity in how important new concepts are defined. E.g., is the sensitivity to groups of parameters taken as sensitivity to processes? Or is that something different? Please check wording across manuscript.*

   We have added a final paragraph to the introduction clarifying some general definitions for the entire manuscript. We hope that it is now more clear that all sensitivities we derive are regarding streamflow in this study. The method however is not at all limited to streamflow but can be applied to any model output of interest.

   line 99 ff. **[...] applying the Sobol' method (Sobol and Kucherenko, 2004; Saltelli et al., 2008; Gilquin et al., 2015) to derive the sensitivity of a model prediction (here streamflow) to model structural choices.**

   line 134 ff. **We propose a method for estimating how sensitive a simulated model output is to groups of parameters. We have chosen here streamflow as this model output as it is the fundamental and most important and common output variable in hydrologic studies. The sensitivities of the groups of parameters is hence obtained regarding streamflow. The groups defined here are either individual parameters (metric B) or the set of parameters that is used in an individual process option (metric C) or all parameters used in any available process option for a modelled process (metric D). We acknowledge that the definition of these groups is subjective and has been chosen here to demonstrate a novel approach of how to evaluate process and process option sensitivities, i.e., how sensitive is the simulated streamflow regarding the choice of a specific infiltration process description or how sensitive is the simulated streamflow regarding infiltration in general.**

   The abstract also highlights for the two results 3) and 4) that results are regarding streamflow. But we understand that it should have been more emphasized in the manuscript.

- *Line 115-122 - I suggest this summary of findings would work better in Abstract + Conclusions. It would also help being clearer in the wording on the comparisons that are being made. Is "conventional" approach the xSSA or the Baroni method?*

  The paragraph the reviewer is pointing out here is indeed explaining the outline of the study and the major analyses that will be undertaken. The reviewer is right that the last sentence is containing a result/outcome and has hence been removed:

  removed: **"The method is demonstrated to be more efficient than a conventional approach (see metric A) whereby the standard Sobol' method is repeatedly applied to distinct model structures as in the study by Van Hoey et al. (2014), in addition to providing more useful information regarding model sensitivities."**

  This efficiency improvement had been mentioned in the abstract:

line 11 ff. 2) The xSSA method with process weighting is computationally less expensive than the alternative aggregate sensitivity analysis approach performed for the exhaustive set of structural model configurations, with savings of 81.9% for the benchmark model and 98.6% for the watershed case study.

and the conclusions:

line 630 ff. The method of weighted process options is shown to significantly reduce number of model runs required to run a sensitivity analysis based on model parameters. For the shared-parameter benchmark model 81.9% fewer model runs are required (A: 72 000 vs B: 13 000). For the hydrologic model example, the reduction is greater than 98.6% (A: 3 258 000 vs B: 45 000).

already in the previous version of the manuscript. Besides removing the sentence claiming already results, the paragraph the reviewer mentions, remains unchanged and we hope that the reviewer agrees that an outline of the major components of this study is helpful to get the reader prepared for the remaining part of the manuscript.

- *Line 278, where it is pointed out that a traditional single-parameter SA analysis could produce grouped-sensitivity analysis by aggregating results for individual parameters? In a paper advocating the new "grouped-SA" method - should such comparison receive priority to show the advantages of the new method. The hypothetical scenario where sensitivity is underestimated (line 279) - is this common in practice? As this goes to the motivation for the new method, I think it could receive more attention.*

  We agree. We think that the approach of estimating parameter sensitivities for each individual enumerated model (like the 12 theoretical models) and then using the average of all the sensitivities per parameter (basically the mean of each column in Table B1) would be the most obvious way to come up with a sensitivity for each parameter across multiple models. We derived these average sensitivities $\overline{\mathcal{S}^n_{x_i}}$ of the seven parameters of the shared-parameter benchmark (mean of each column in Table B1) and compare them to the parameter sensitivities $\mathcal{S}_{x_i}$ derived with the weighted model approach (Eq. B1):

$$\begin{bmatrix} \overline{S^n_{x_1}} \\ \overline{S^n_{x_2}} \\ \overline{S^n_{x_3}} \\ \overline{S^n_{x_4}} \\ \overline{S^n_{x_5}} \\ \overline{S^n_{x_6}} \\ \overline{S^n_{x_7}} \end{bmatrix} = \begin{bmatrix} 0.1223 \\ 0.2978 \\ 0.0506 \\ 0.0699 \\ 0.0699 \\ 0.0288 \\ 0.6506 \end{bmatrix}, \begin{bmatrix} \overline{ST^n_{x_1}} \\ \overline{ST^n_{x_2}} \\ \overline{ST^n_{x_3}} \\ \overline{ST^n_{x_4}} \\ \overline{ST^n_{x_5}} \\ \overline{ST^n_{x_6}} \\ \overline{ST^n_{x_7}} \end{bmatrix} = \begin{bmatrix} 0.3238 \\ 0.5052 \\ 0.1321 \\ 0.2562 \\ 0.2562 \\ 0.0288 \\ 0.6506 \end{bmatrix} \textbf{ and } \begin{bmatrix} S_{x_1} \\ S_{x_2} \\ S_{x_3} \\ S_{x_4} \\ S_{x_5} \\ S_{x_6} \\ S_{x_7} \end{bmatrix} = \begin{bmatrix} 0.0230 \\ 0.3806 \\ 0.0022 \\ 0.0002 \\ 0.0002 \\ 0.0000 \\ 0.0393 \end{bmatrix}, \begin{bmatrix} ST_{x_1} \\ ST_{x_2} \\ ST_{x_3} \\ ST_{x_4} \\ ST_{x_5} \\ ST_{x_6} \\ ST_{x_7} \end{bmatrix} = \begin{bmatrix} 0.0753 \\ 0.7709 \\ 0.0045 \\ 0.0010 \\ 0.0010 \\ 0.0000 \\ 0.0524 \end{bmatrix}$$

  The sensitivities of the parameters in both approaches vary drastically. For example, the most sensitive parameter when using the 12 enumerate models is parameter x_7 while the most sensitive parameter with the proposed approach is parameter x_2. The huge differences are mostly explained with the fact that in the first approach not all parameters are active all the time. Another reason is the interaction between parameters. Even though we have these results, we choose to keep this out of the paper so as to not complicate the discussion.

- *Line 352 "limitations of existing Baroni method" - as this comparison is important in this paper - would seem preferable to describe the Baroni method in appropriate detail before discussing its limitations.*

  We now provide a more detailed description of this method in the introduction and hope this helps to understand the method itself, and makes the comparison

to the proposed method easier. The method is now referred to as the Discrete Values Method based on a comment of another reviewer.

line 50 ff. To date, there have been limited attempts to simultaneously estimate model parameter, input, and structural sensitivities. One notable attempt is introduced by Baroni and Tarantola (2014) using a Sobol' sensitivity analysis based on grouped parameter. In that study, groups of soil and crop parameters, the number of soil layers, and a group of parameters to perturb inputs are investigated. These groups of parameters are pre-sampled and a finite set of parameters for each of the four groups is chosen and each set is enumerated. The sensitivity analysis is then based on those enumerated sets. This means, rather than sampling each individual parameter like in a classic Sobol' analysis, an integer for each group acting as a hyper-parameter is sampled. The model is then run with the associated pre-sampled parameter set. While the approach may be generally applicable to arbitrary structural differences, in their testing, Baroni and Tarantola (2014) varied only in how the model was internally discretized (i.e., in the number of soil layers). The soil and crop parameters were always used for the same soil and crop process. The major limitation of this method is, however, that individual parameters need to be mutually exclusive and can only be associated to one type of uncertainty. The method hence limits the groups that can be defined, for instance, overlapping group definitions are not possible. The method will be referred to as "discrete values method (DVM)" in the following and will be contrasted to the method developed here to examine this limitation in more detail.

- *Line 535: "it can be deduced that the potential melt, the quickflow options BASE_VIC and BASE_TOPMODEL, and the evaporation options are most influential upon modeled streamflow". Here the lack of clarity on what is meant by "influential" can cause confusion to a reader. Especially sensitivity to a specific option for a process (eg, BASE_VIC for quickflow) - normally sensitivity is to a range of possible values for a decision - here it is to a single specific value? I don't quite follow this.*

  The reviewer is right that (parameter) sensitivity is regarding the range of possible parameter ranges. The sensitivity of process options is now the sensitivity of the model output (here streamflow) based on a range of different settings of these process options. This means that a process option is picked and a range of different parameterizations for this process option is tested and evaluated which impact (sensitivity) is has on the model output (here streamflow). Compared to the standard parameter sensitivity where the impact of one parameter is evaluated, the proposed method evaluated the impact of a group of parameters. It could technically be any group of parameters but we decided to define each group as either the parameters that belong to a certain process option and compare those groups' sensitivities or to group all parameters of an entire process to estimate the impact (sensitivity) of this group (process) on the streamflow.

  In the paragraph the reviewer mentions, we identify that potential melt, two quickflow options and the evaporation options have the most impact on the simulated streamflow time series. Means that if the parameterization of any of those process options is modified the simulated streamflow would change more than with any other change made in any of the other process options.

- *Section 3.3 - nice sections. Would be improved by providing clearer definitions of sensitivity, influential processes, uncertainty, etc (see earlier comment). Current usage is unnecessarily loose and confusing here.*

  We agree with the reviewer that using too many of those terms interchangeably

might lead to confusion for the readers and might lead to too much leeway of interpretation. We went through the manuscript and hope that we reduced this ambiguity.

line 521 ff. The analysis of model parameters (Fig. 5A) shows that the most sensitive ones are [...]

line 586 ff. The strong impact of these processes (together with the input adjustments) highlights the sensitivity of streamflow regarding snow and melting processes in this mountainous, energy-limited catchment.

line 589 ff. This demonstrates that soil and surface processes are of secondary sensitivity regarding streamflow. Their sensitivity may increase if the uncertainty of the snow and melting processes can be reduced, i.e. by narrowing parameter ranges during calibration.

line 610 ff. Evaporation (dark blue) is [...], expectedly, less sensitive during winter. Snow balance (medium green) and potential melt (orange) are sensitive as long as snow is present (Nov to May).

Many these comments focus on presentation , but given the technically demanding nature of the work, a more targeted presentation would make it easier to digest by an interested reader.

Other comments

1. *Line 4: "apply" or "develop"?*

   We still use apply here since using grouped parameters for a Sobol' analysis has been used in the past. The groups just never have been based on process options and processes. The part of xSSA that is new is the usage of weights to have several process options active in parallel. To avoid confusion and not ovre-selling our new approach, we prefer to stick with "apply".

2. *Line 24 - what is "they" referring to? Also what does "non-unique" refer to here? Is this with regard to many models co-existing in the literature? Or non-uniqueness in their inversion when estimating parameters? I think some clarity would be useful here*

   It refers to the conceptual models that are non-unique (as it depends on the concept and simplifications made by the individual modeler defining the conceptual model). We rephrased the beginning of the sentence to the following:

   line 24 ff. The model descriptions are also non-unique as they depend on the modelers simplifications and choices made during the model conceptualization. A large number of non-unique process algorithms can be found [...]

3. *Line 27 - are these decisions always subjective? Surely there exist studies where model decisions are developed according to sensible strategies?*

   We think that these decisions are made subjectively in most of the cases. In case a modeler picks one hydrologic model (based on experience, availability, expertise in a research group, etc), the choice of the process options is fixed (and hence subjective) from the start as the process option/conceptualization of that model would be chosen. Nearly, nobody would ever question that set of process algorithms (unless the focus of the study is to improve that model). In case a modeling framework is used the modeler would be faced from the beginning of which process algorithm to use. We are not aware of any study that is proposing an objective selection of the most appropriate process definition.

4. *"Sensitivity to model structural uncertainty" - I think studies such as McMillan et al. (2010); Clark et al. (2011) and other have investigated this?*

We added those citations.

5. *"recent" - with references back to 2008 is this still recent?*

   Probably not all of them are "recent" but some are. We anyway just removed that word.

6. *Baroni method - seems an important method in the context of this work. I think it would be helpful to provide the gist of that method at least in an Appendix, in the way that is is applied here.*
   We totally agree. We have added a brief description about the Baroni method (now called "discrete values method (DVM)") to the introduction:

   line 50 ff. To date, there have been limited attempts to simultaneously estimate model parameter, input, and structural sensitivities. One notable attempt is introduced by Baroni and Tarantola (2014) using a Sobol' sensitivity analysis based on grouped parameter. In that study, groups of soil and crop parameters, the number of soil layers, and a group of parameters to perturb inputs are investigated. These groups of parameters are pre-sampled and a finite set of parameters for each of the four groups is chosen and each set is enumerated. The sensitivity analysis is then based on those enumerated sets. This means, rather than sampling each individual parameter like in a classic Sobol' analysis, an integer for each group acting as a hyper-parameter is sampled. The model is then run with the associated pre-sampled parameter set. While the approach may be generally applicable to arbitrary structural differences, in their testing, Baroni and Tarantola (2014) varied only in how the model was internally discretized (i.e., in the number of soil layers). The soil and crop parameters were always used for the same soil and crop process. The major limitation of this method is, however, that individual parameters need to be mutually exclusive and can only be associated to one type of uncertainty. The method hence limits the groups that can be defined, for instance, overlapping group definitions are not possible. The method will be referred to as "discrete values method (DVM)" in the following and will be contrasted to the method developed here to examine this limitation in more detail.

   It is also a little unclear from the abstract that a comparison to this method is made. Eg line 13 "alternative" - if this is Baroni's method - should this be "existing" method? To avoid a confusion the reference algorithm should be clearly described.

   The "alternative" method mentioned in the abstract does not refer to the Baroni method but to the method "performed for the exhaustive set of structural model configurations". We do not mention the comparison to the Baroni method in the abstract as we want to focus on the novelty of using process weights in the abstract and the results accompanied with this novelty. The comparison to the Baroni method is solely to highlight that there might be limitations in the existing approaches that are overcome by the proposed method.

7. *line 49 - "did not change when moving between model structures" - is this for different hydrological models? or models from across multiple disciplines?*

   This does not appear anymore since we rewrote the section about the Baroni method (see reply to comment above) and hope that it is less confusing now.

8. *line 50 - what are "hyper-parameters"?*

   Hyper-parameters can be for example multipliers that are applied to all parameters in a group (e.g., porosity parameters of all soil types) rather than analyzing

every individual parameter independently. The manuscript now explains them a bit more through:

line 54 ff. This means, rather than sampling each individual parameter like in a classic Sobol' analysis, an integer for each group acting as a hyper-parameter is sampled. The model is then run with the associated pre-sampled parameter set.

9. *line 52 - not entirely clear what "form" refers to here. I found the entire sentence a bit confusing when trying to understand exactly what its trying to say*

   We meant "type [of uncertainty]". We changed that in this paragraph (rewritten section about the Baroni method) and throughout the whole manuscript.

   line 58 ff. The major limitation of this method is, however, that individual parameters need to be mutually exclusive and can only be associated to one type of uncertainty.

10. *line 53 - "the method introduced ..." - is an incomplete sentence?*

    Resolved. This paragraph was rewritten entirely to add a better description of the Baroni method (see reply to comment 6).

11. *line 55 - "individual" - maybe clarify that the previous study assessed ONLY combined sensitivities? This is not clear from the current wording. And I thought that combined sensitivities are an advance rather than individual sensitivities? So why is that a limitation of the previous work?*

    Yes, correct. Only the sensitivity of the model regrading all parameters (parameter uncertainty) was determined. But it is not clear which parameter(s) are causing this sensitivity. We do not think (and also do not state) that this is a disadvantage. It might just limit the insights of such an analysis as we mention in the last sentence of that paragraph: "Similar to the discrete values method, parameters were treated in an aggregate fashion which made it impossible to attribute the parameter sensitivity to a certain parameter or model component."

12. *line 62 - "sensitivity of a model" - is this for model simulations? or model parameters? or both? See comment about making sure the key concepts are clearly defined*

    We mean the sensitivity of a model output here. We explain in the second half of that sentence that Van Hoey et al. (2014) is looking at parameter and structural sensitivities. We added "output" to the text now and hope it is less confusing.

    line 71 ff. Van Hoey et al. (2014) is one of the few studies that explicitly examined the sensitivity of a model output to changes in process representation, estimating sensitivities of parameters of various model structures with two or three alternatives per process, e.g., linear vs. non-linear storage; with or without an interflow process.

13. *line 78-79 - "it is therefore ..." - i think these ideas on the utility of SA should be introduced earlier in the presentation, to provide a stronger motivation and a practical context for the work.*

    We think that introducing first the current status of the work in the field of structural sensitivity analysis is a better order. Making the claim that we are limited in the way we can analyze structural uncertainty without having any of these terms and approaches introduced, seems to be much more confusing. We hope that the reviewer agrees- especially since we explain some concepts now (hopefully) more clear in the paragraphs before.

14. *line 88 - this property "structure can vary continuously" / "weighted average". I found this aspect quite interesting in the work. The statement below that xSSA "is made uniquely possible" to RAVEN - do you mean it can only be used by RAVEN? This seems strange as multi-model ensembles where each model has a weight are fairly common (e.g., see the "model averaging" literature).*

    The multi-model averaging community would always average the final model outputs (here streamflow). Raven allows the user to get an internal average of, for example, the amount of snow melting in each time step before this amount of meltwater is then used to derive anything else in the model (e.g., the amount of water infiltrating in this time step). To date this is (to our knowledge) only possible in Raven but could be implemented in any model that is allowing for several process options.

15. *line 96 - "uniquely"?*

    We hope that our response to the comment above (#14) resolves this issue and makes clear that this feature is indeed very unique.

16. *line 105: Metric B - very interesting concept. but without some elaboration seems potentially ill-defined. Eg, how do you determine if a parameter appearing in different model structures is "the same parameter"?*

    The reviewer is totally right that it is in the eye of the modeler (or person setting up the sensitivity analysis) to treat a parameter that appears in several process options as "the same". We treated the parameters the same if they have the same units (and therefore ranges that would be tested) and would be treated the same in a model analysis and interpretation of results. One example would be the depth of the top soil layer (in [mm]). Several processes and process options use this parameter and there is certainly no ambiguity of how to interpret- for example- the optimal value of this parameter. Other examples might be the temperature where snow is melting or the porosity of a soil type. Of course there might be parameters where this is less obvious but the person setting up the study could handle those parameters separately, e.g. soil depth method 1 and soil depth method 2. The method however gives us the huge opportunity to evaluate what some model parameters actually mean and if some parameters across conceptualizations are comparable.

17. *line 120: "conventional approach" - is this the Baroni method? If so best to name it. Also it was referred to as "alternative" in the Abstract*

    No, we refer here to the traditional/conventional/alternative approach as used for Metric A. We agree that this is confusing and rephrased the sentence to:

    line 113 ff. We here pose these as four distinct sensitivity metrics:

    A. Conditional parameter sensitivity: *Which model parameter is most influential given a certain model structure?*
    For example, which model parameter is most influential in the HBV model? (This is the traditional Sobol' metric. This conventional approach would test all possible models and derive parameter sensitivities conditional on the model tested.)

    B. ...

    We also made an adjustment at another place in the introduction and hope that this strengthens the understanding of the link between the proposed method and the "traditional" method:

    line 100 ff. To our knowledge, the method of grouping parameters to derive sensitivities of parameters, process options, and processes without the explicit

necessity of averaging parameter sensitivities after deriving them for individual models (referred to as conventional/ traditional sensitivity analysis) has not yet been applied.

We also would like to highlight that the section explaining this method is called "Traditional Sobol' Sensitivity Analysis".

18. *Section 2 - consider splitting into several sections and place in order of relevance to the contributions of the paper*

    We find it very hard to make that clearer than it is right now. We previously tried to organize the Material and Methods following the metrics A-D we introduce but that leads to a significant duplication of information. On the other hand the contributions are plenty-fold and will likely not be able to be condensed in one paragraph- even though everything comes together in Section 2.2. The introductory paragraph for section 2 is meant to help the reader to navigate through the section. We hope that our earlier response regarding the best structure of the Methods section (major comment #3) also helps to make this more clear.

19. *line 145 - see earlier comment - how do you know it the "same" parameter? It seems a relevant discussion point*

    Yes, it is. Please see our reply above (comment 16).

20. *line 169/ eqn 16 - how do you "decide" in a modeling context what is a shared parameter? Say is x_3 in eq 16 the same as x_3 as in eqn 13? Is this considered determined purely by the choice made by the modeler regarding the parameters to calibrate?*

    We hope our explanation above (comment 16) helps to clarify this. The modeler has always the option to define these parameters separately for each process option but in a lot of cases it is obvious that these parameters are shared.

21. *line 235-236 - I think these are discussion points - would work better in Discussion rather than forward references here - at this point of the paper the new method is not described yet!*

    Agreed. We remove the following sentence from the Material & Methods: "However, in the case of a framework without weights for process options, the application of the method would be much less efficient." We discuss the improvement of efficiency in the Conclusions.

22. *Line 312 - would help clarify here that this is approach is new and introduced in this work. And as mentioned earlier - I think it would benefit from being given more prominence in the paper.*

    We added the following to that paragraph:

    line 338 ff. Although the grouping of parameters has previously been used (Sobol and Kucherenko, 2004; Saltelli et al., 2008; Gilquin et al., 2015), it is- to our knowledge- the first time they have been used to group parameters of process options in the context of examining model structure sensitivity.

23. *line 318 - "depicts"?*

    We think the reviewer points to line 320. There it says "[...] where V depicts variances and E expected values." We do not know exactly what the reviewer means with his/her comment and do not know what to adjust.

24. *line 409 - "hereafter called Baroni method" - already said this earlier on line 48 - but still referring to this method by multiple names*

    We now refer to this method consistently as the "discrete values method (DVM)".

Besides we think that it might be helpful to introduce this abbreviation again at the beginning of the results- just as a quick reminder for the readers.

25. *Appendix A - an extra 1-2 sentences that refer to where in the main text are these weights used would be helpful here*

    We agree. We added the following as an introduction in Appendix A:

    line 659 ff. In this work we define a model that is using the weighted average of a set of process options instead of choosing one fixed process option (Eq. 18). This is enabling to analyze several model structures at the same time by either setting weights to 0 or 1 (which selects exactly one option) or any weight in between which leads to the weighted average of those process option outputs.

    The sampling of such weights needs to lead to independent and identically (not necessarily uniform) distribution for each of the weights w_i.

26. *Appendix B - I am confused why this seems duplicated in the Intro and the Appendix. If this is new - would seem better somewhere in the Theory and then Discussed, where it can be discussed in appropriate detail.*

    We are sorry that this duplication led to confusion. We initially thought that the repetition of the four metrics would be convenient for the reader. We removed this now and hope the readers will remember those metrics (A-D) from the introduction. We also highlight the four different metrics in bold font in the Appendix B now. We hope that further improves the readability.

Figures

1. *Figure 2 - the blue font in panel B is quite hard to read*

   We are using a darker blue shade now for the bars and the left y-axis label.

2. *Figure 5 (and others to various extents) - could be more generous with fontsize, as many labels etc are virtually illegible*

   We increased the fontsize of the tick labels in Figure 5 and the fontsize used in the legend of Figure 6. Figure 5 is now also "rotated" by 90 degrees. That should make it much easier to read all labels.

References

Baroni, G. and Tarantola, S.: A General Probabilistic Framework for uncertainty and global sensitivity analysis of deterministic models: A hydrological case study, Environmental Modelling & Software, 51, 26–34, 2014.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resources Research, 47, 5468–16, 2011.

Friedl, M. and Sulla-Menashe, D.: MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500 m SIN Grid [data set], NASA EOSDIS Land Processes DAAC 10, 2015.

Gilquin, L., Prieur, C., and Arnaud, E.: Replication procedure for grouped Sobol' indices estimation in dependent uncertainty spaces, Information and Inference A Journal of the IMA, 4, 354–379, 2015.

McMillan, H., Freer, J., Pappenberger, F., Krueger, T., and Clark, M.: Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions, Hydrological Processes, 47, 1270–1284, 2010.

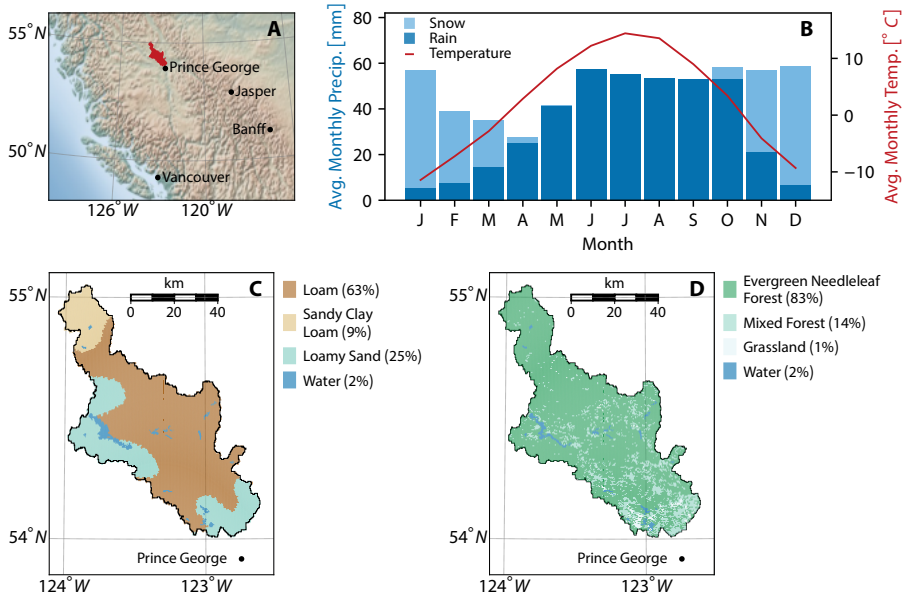

Figure 2: (A) Location of the Salmon River catchment (red polygon) in British Columbia, Canada. The watershed is 4230 km^2 and located around 700 km north of Vancouver. It is located in the Rocky Mountains with an elevation of 606 m above sea level at the streamflow gauge station of the Salmon River (08KC001). (B) The average monthly mean temperatures (red line) and average monthly precipitation is divided into rain (dark blue) and snow (light blue). Maps of (C) the four soil types based on the Harmonized World Soil Data (HWSD; 30") (Nachtergaele et al., 2010) and (D) four land cover types based on the MCD12Q1 MODIS/Terra+Aqua Land Cover (500m) (Friedl and Sulla-Menashe, 2015) of the Salmon River catchment are provided. The colors indicate different soil and land use classes.

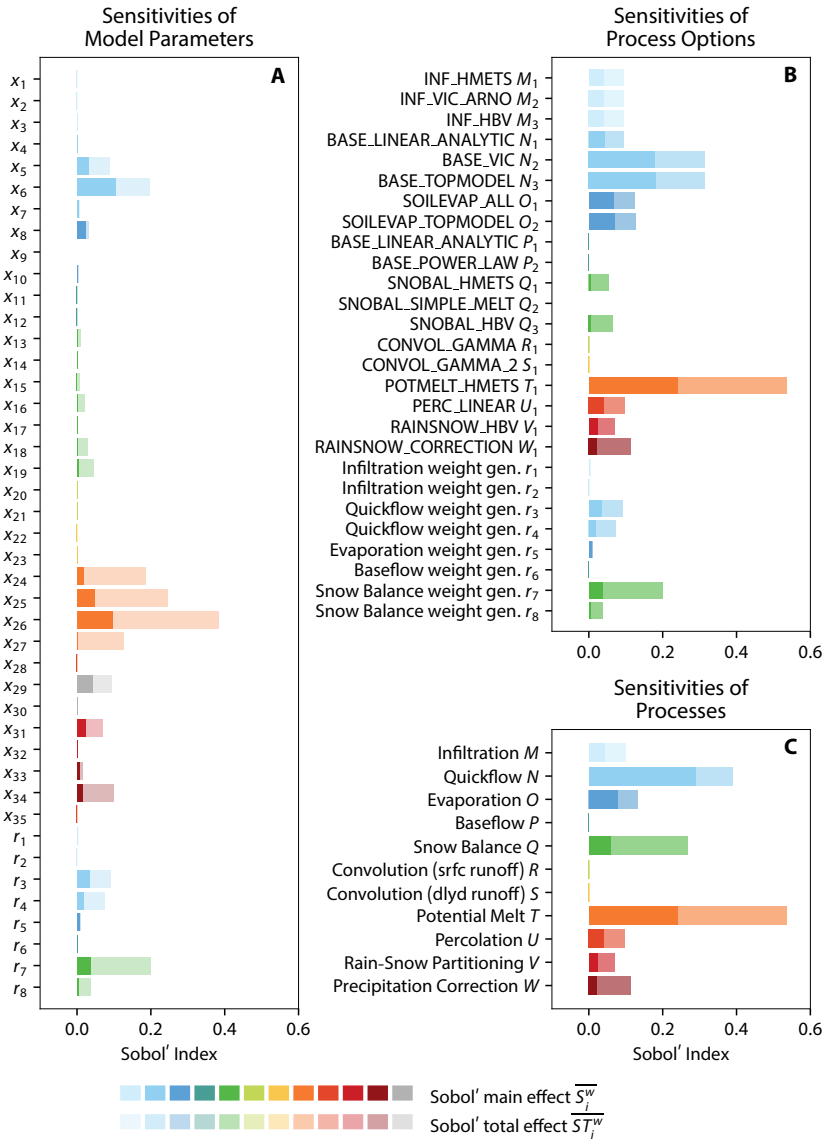

Figure 5: Results of the Sobol' sensitivity analysis of the hydrologic modeling framework Raven. (A) The sensitivities of 35 model parameters (see Table C2) and 8 parameters r_i that are used to determine the weights of process options are estimated. The Sobol' sensitivity index estimates are determined also for (B) 19 process options and (C) the 11 processes. The information which parameters are used in which process option and process can be found in Table C1. The different colors indicate the association of parameters and process options to the eleven processes. Parameters x_{29} and x_{30} are associated with several process options and are not colored but gray. The Sobol' main and total effects are shown (dark and light colored bars, respectively). All sensitivity index estimates shown are originally time-dependent and are aggregated as variance-weighted averages (Eq. 23 and 24). The average weights over the course of the year are shown in Figure 6.

Nachtergaele, F., van Velthuizen, H., Verelst, L., Batjes, N. H., Dijkshoorn, K., van Engelen, V. W. P., Fischer, G., Jones, A., Montanarella, L., Petri, M., Prieler, S., Shi, X., Teixera, E., and Wiberg, D.: The harmonized world soil database, in: 19th World Congress of Soil Science, Soil Solutions for a Changing World, Brisbane, Australia, 1-6 August 2010, pp. 34–37, 2010.

Saltelli, A., Ratto, M., Andres, T. H., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S.: Global sensitivity analysis. The primer, John Wiley & Sons, Ltd., 2008.

Sobol', I. M.: Sensitivity analysis for non-linear mathematical models, Mathematical Modeling & Computational Experiment (Engl. Transl.), 1, 407–414, 1993.

Sobol, I. M. and Kucherenko, S. S.: Global Sensitivity Indices for Nonlinear Mathematical Models. Review, WILMOTT magazine, pp. 2–7, 2004.

Van Hoey, S., Seuntjens, P., van der Kwast, J., and Nopens, I.: A qualitative model structure sensitivity analysis method to support model selection, Journal of Hydrology, 519, 3426–3435, 2014.