Reply to the reviewer comments RC4: 'Review of the manuscript by Schürz et al.' by Francesca Pianosi

The manuscript presents an interesting application of uncertainty and sensitivity analysis to the SWAT model. The aim is to assess the dominant controls of long-term discharge and nitrate-nitrogen load predictions under climate and land use change, while also taking into account the intrinsic uncertainty in the model, i.e. parameter and setup uncertainty. The analysis is solid and provides interesting insights about the model behaviour. Although the specific findings are only relevant to the investigated model and case studies, their discussion is interesting for the wider community of SWAT users and in general users of environmental impacts assessment models, as it demonstrates the type of findings yielded by GSA and their implications for the refinement and use of the model. The visual analysis introduced in Figure 4-8 is a simple and yet effective complement to quantitative GSA approaches.

Overall the paper is well structured and well written, and I think it should be accepted for publication.

Below are some points that could be addressed to improve the manuscript clarity before publication.

We would like to thank Francesca Pianosi for her constructive review and the valuable comments made. We appreciate the general positive feedback on the manuscript. It is a pleasure for us, that one of the developers of the PAWN sensitivity index evaluated this manuscript. Below, we addressed each comment made. We hope to clarify and discuss all points of concern sufficiently in the following document. The referee comments are printed in *serif, italic font*. Our replies to the comments are written in black, non serif font and our suggestions to revise the manuscript are highlighted with the colors blue for insertions and red for deletions.

Major comments

[1] Language is at times unclear - some examples are given below as Minor points. I also have a general comment about the use of the term "sensitive". The authors use it as interchangeable with "influential" however I find this confusing, because "sensitivity" is an attribute of the output, not of the inputs. I would say that "input x1 is influential on the output" or "the output is sensitive to input x1" but I would not say that "input x1 is sensitive" - this is confusing. Some examples of these unclear occurrences are also given below under Minor points, however if the authors accept my remark they should check the entire manuscript.

We highly appreciate this comment and agree that the example provided above identifies the correct use of the terminology. We accept the remark and will improve the updated version of the manuscript to correctly use the terms “sensitive” and “influental” accordingly.

[2] The definition and use of the behavioural parameter sets is slightly unclear. I think the confusion started on P. 10 L. 6-7 with the sentence "For all SWAT model setups of the Schwechat and the Raab catchments we identified non-unique parameter sets that adequately simulated daily observation of discharge and NO3-N loads".

*Does it mean that you identified one behavioural parameter set for **each** model setup, or that you identified one behavioural parameter set to be applied **in all** the setups? If the former, then how is the dependency between parameterisations and model setups accounted for in the GSA? If the latter, then the underlying assumption is that the same parameter values can effectively represent processes at different aggregation scales (ie for different definitions of the subbasins and HRUs)? This should be clarified. On a parallel note, I find it interesting that out of 100,000 sampled parameterisations only 43 and 52 where found behavioural. This is not uncommon in calibration of complex hydrological models but still worth highlighting. It would also be interesting to see whether these behavioural parameterisations are clustered in specific regions of the parameter space or if they are scattered across the sampled ranges, which would indicate a certain amount of interactions between the parameters. This could be illustrated for example through a parallel coordinate plot.*

We agree that the paragraph as stated leaves room for an ambiguous interpretation of the performed simulation and analysis steps. We designed the study the following way: For the different model setups of the Raab (6 different setups) and the Schwechat (4 different setups) we analyzed the model parameter sensitivities employing global sensitivity analysis (GSA). Thus, we performed six individual GSAs for the Raab catchment and four for the Schwechat catchment. The individual parameter sensitivity analyses resulted in the same sets of influential model parameters for the Raab catchment and the Schwechat catchment, respectively. As a consequence, we selected the same model parameters for all model setups of the Raab catchment and for all setups of the Schwechat catchment. For each case study, we drew 100 000 realizations of parameter combinations from the influential sets of model parameters. The simulations were performed with all model setups involving all drawn parameter combinations.

To answer your first question, a parameter set was eventually considered as a behavioral parameter set, if the simulations performed with **all** model setups involving that parameter set fulfilled the applied objective criteria stated on p.10L7-10.

This design decision was necessary, to treat the model setups and model parametrizations as individual inputs in the GSA (as you have indicated in the first part of your question). We agree with your comment that the selected layout implies the assumption that a parameter combination represents the analyzed processes at different aggregation scales. The individual models (i.e. a model setup with a specific spatial aggregation together with a selected model parametrization) were not analyzed and compared at the subasin or the HRU level. Yet, all models are capable of adequately simulating discharge and nitrate-nitrogen loads at the catchment outlets in the reference period. A drawback of this design decision is that it does not consider parameter combinations that that would result in satisfactory simulations when employed in one or several model setups but do not give good results with **all** model setups. Thus, the influence of the model setup and the model parametrization combined can be greater than the effect resulting from the illustrated study design. Nevertheless, we think that the presented results reveal relevant insights in their

current form, as they isolate the effect of the model setups that were calibrated for a reference period and are now applied for future changing conditions (the same applies to the parametrizations).
The selected experimental design is also a major reason for the low number of resulting behavioral parameter sets in the two case studies.
All of the explanations outlined above are in our opinion not sufficiently addressed in the current version of the manuscript. In particular, we think that the design decisions we made have to be conveyed and highlighted clearly in the methodology. Thus, we suggest to revise the concerning sections p.9L9 to p.11L2 accordingly.

According to the suggested visualization of the model parameters in a coordinate plot, we propose to add the figure below in the Appendix of the manuscript. We omitted any axes and tick labels due to very limited plotting space and the large number of parameters to visualize. The figure however illustrates any clustering and interaction of model parameters. We additionally suggest to add parameter ranges and the type of change of the model parameters in Table 3 on page 10.

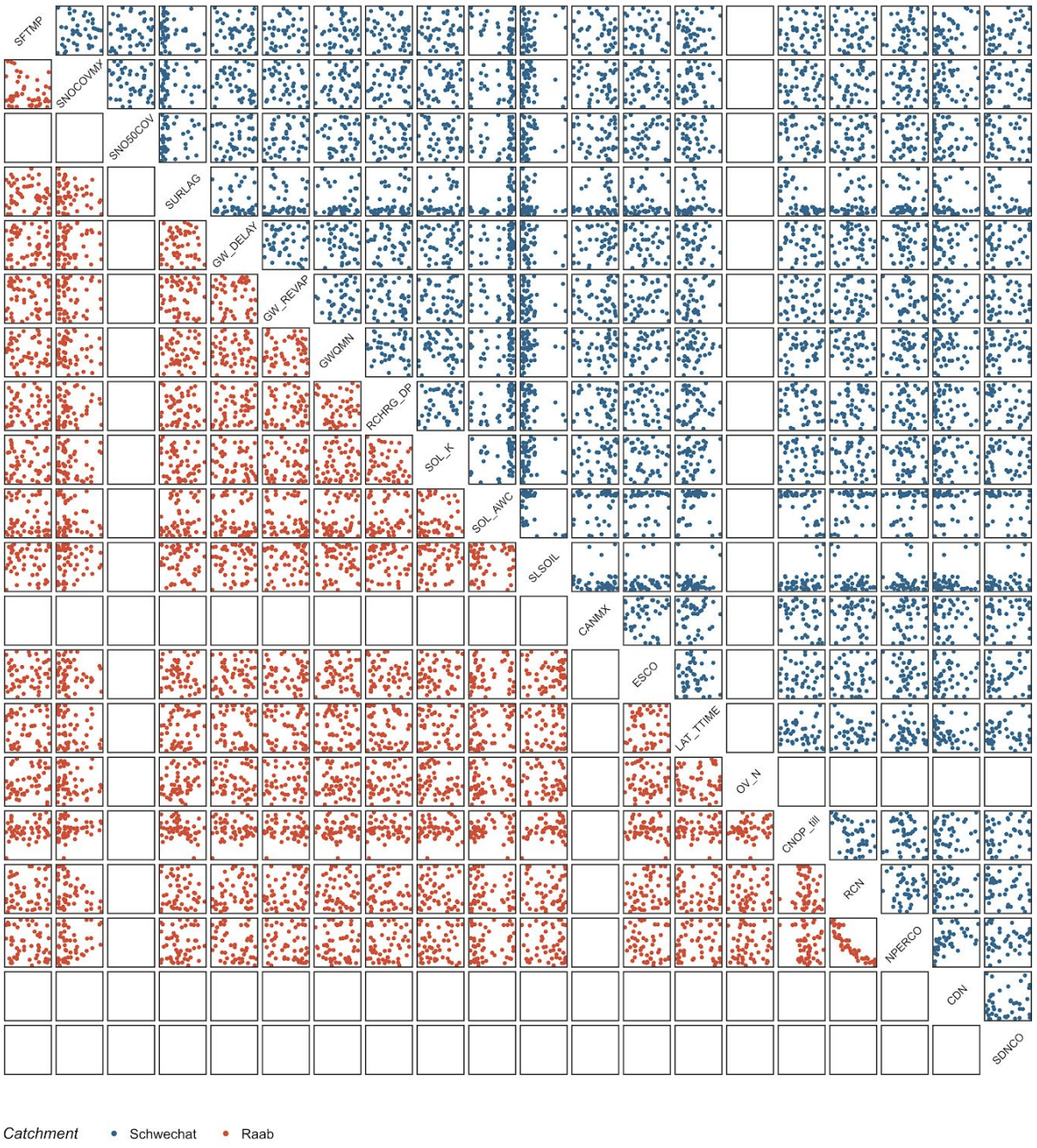

Figure caption:
Coordinate plot of the 43 and 52 behavioral SWAT model parameters that were used with the model setups of the Schwechat and the Raab respectively. Each panel illustrates the connection of two model parameters for the Schwechat in red (below the diagonal) and the Raab in blue (above the diagonal). The x and y axes of each panel show the range of the respective parameter plotted along the x or y dimension. The corresponding parameter ranges for all illustrated parameters are provided in Table XX (Reference to table below).

Table caption:
SWAT model parameters calibrated in the model setups of the Schwechat and the Raab catchments. The type of change indicates whether the model parameters were replaced by absolute values, modified by adding absolute values to the predefined model parameters or, changed by a relative fraction of the predefined model parameter. Illustrated are the initial ranges of the model parameters and the ranges of the final behavioral parameter sets of the model setups of the Schwechat and the Raab catchments.

			Parameter change range	
Parameter	Type of change	Initial parameter range	Schwechat	Raab
SFTMP	replace value	[-1.00, 1.00]	[-0.69, 0.93]	[-0.98, 0.88]
SNOCOVMX	replace value	[100.0, 500.0]	[0.9, 177.0]	[100.8, 447.5]
SNO50COV	replace value	[0.20, 0.50]	[0.21, 0.49]	
SURLAG	replace value	[0.00, 0.50]	[0.02, 0.99]	[0, 0.1]
GW_DELAY	replace value	[0.0, 300.0]	[5.5, 25.0]	[2.1, 283.3]
GW_REVAP	replace value	[0.02, 0.20]	[0.05, 0.15]	[0.02, 0.20]
GWQMN	replace value	[0, 3000]	[566, 2472]	[109, 2925]
RCHRG_DP	replace value	[0.00, 1.00]	[0.31, 0.69]	[0.13, 0.97]
SOL_K	relative change	[-0.90, 10.00]	[0, 0.97]	[-0.79, 9.76]
SOL_AWC	relative change	[-0.90, 2.00]	[-0.86, 1.49]	[0, 1.98]
SLSOIL	replace value	[0.0, 150.0]	[0.9, 27.6]	[14.7, 148.2]
CANMX	relative change	[0.00, 0.25]	[0.34, 2.40]	
ESCO	replace value	[0.00, 0.90]	[0.05, 0.90]	[0.05, 0.89]
LAT_TTIME	replace value	[0.0, 180.0]	[0.8, 6.8]	[5.5, 176.3]
OV_N	absolute change	[-0.09, 0.60]		[0.07, 0.58]
CNOP_till	relative change	[-0.20, 0.10]	[-0.29, -0.06]	[-0.18, 0.01]
RCN	replace value	[2.00, 10.00]	[5.05, 9.97]	[2.3, 8.45]
NPERCO	replace value	[0.00, 1.00]	[0.24, 0.99]	[0.18, 0.7]
CDN	replace value	[0.00, 1.50]	[0.01, 1.44]	
SDNCO	replace value	[0.00, 0.50]	[0.02, 0.49]	

[3] GSA was applied using 7000 samples of the input factors. How was this number chosen? Did the authors checked the adequacy of this sample size? The fact that the ranking based on the sensitivity indices in Figure 2 is confirmed by the visual analysis of Figure 4-8 is reassuring, yet formal methods exist to assess the robustness of the GSA results to the chosen sample size (for example, using bootstrapping confidence intervals as in Sarrazin et al. 2016 or a dummy parameter as in Zadeh et al 2017, both cited in the manuscript). It would be good to include more discussion of this point in the manuscript.

The number of input factor samples used in this study results from previous analyses performed using the present input factor data basis. A preceding analysis employed the Sobol method (Sobol, 1993) for global sensitivity analysis (GSA) using a sampling design proposed by Saltelli (2002) that requires $N(k+2)$ samples, where N is the "base sample" (Saltelli, 2008) that was defined with 1000 in this study and k is the number of inputs (in this case 5).

Similar to the generic sampling strategy that you suggested in a presentation at the EGU 2018 (Pianosi and Wagener, 2018a, that is now published in Pianosi and Wagener, 2018), we utilized a model input sample that was initially drawn for a GSA applying a different method (in our case the Sobol method) and employed it to estimate PAWN indices.
Besides the PAWN Index that is presented in this manuscript, we also tested the Sobol method and a modified version of the STAR-VARS method (Razavi and Gupta, 2016a, 2016b) in the course of this work (results for the latter two analyses are not shown in the manuscript). All experiments expressing the maximum sensitivity of the used target variables to the analyzed input factors showed strongly overlapping results. Thus, we were confident regarding the soundness of the GSA results. We agree however that it would be beneficial to the reader of the manuscript to provide any measure of confidence with the results of the GSA. Thus we suggest to perform bootstrapping (as demonstrated in Sarrazin et al. (2016)) to provide confidence intervals together with the calculated PAWN indices in section 3.2. p.15ff.

[4] The PAWN method was applied using a sampling scheme different from the one originally presented in Pianosi and Wagener (2015), in order to handle discrete-valued input factors. I understand the idea is to consider as fixed points x_i^j all the possible values that the discrete input factor x_i can take. Hence, for each input factor, the number of fixed points coincides with the number of possible values (n_i) that the input can take. If my interpretation is correct, then the text is misleading when it says (P. 13 L. 28) that "a generic random sample of the size N was drawn and subsetted with N/n_i subsets for all x_i^j" as the generic sample is divided into n_i (and not N/n_i) subsets. Is this right?

The term "generic" was used in the present context, as the sampling in all input factor dimensions was done randomly, although restricted by the number of fixed values each input can have. The separation of the total sample into N/n_i subsamples is then a required step to calculate the Kolmogorov-Smirnov distance for the input factor x_i at each location x_i^j. We understand however, that the term "generic sample" might be interpreted as a random sampling of continuous variables that is not the case here. Thus, instead of using the term "generic random sample" we suggest to specify the performed sampling in the following way (p.13L22-25):

~~The sampling scheme of PAWN index (Pianosi and Wagener, 2015) was initially introduced designed for continuous model parameters, which requires a modification for discrete model inputs.~~ Pianosi and Wagener (2015) introduced the PAWN sensitivity method using a specifically tailored sampling design to infer the PAWN indices T_i for continuous model inputs x_i. The ~~initial~~ proposed sampling strategy suggests to draw~~n~~ N_c conditional samples at n randomly sampled points of each influencing variable x_i, where xi is fixed at a value $x_i = x_i^j$ while all others are perturbed. Recently, Pianosi and Wagener (2018) extended the applicability of the PAWN sensitivity method to estimate T_i from a generic random sample of continuous model inputs. To approximate T_i the generic sample N is split into n segments along each model input dimension resulting in conditional samples N_c with an approximate size of N/n. We employed the proposed updated sampling strategy and adapted it for the use with discrete model inputs. ~~Instead, a generic random~~ A sample of the size N was

drawn. For each model input combination every model input was sampled randomly from its discrete realizations. ~~and subsetted with~~ N/n_i ~~subsets for all~~ x_i^j ~~t~~ To infer ~~the values for all~~ $KS_j(x_i)$ for all discrete values x_i^j of a model input x_i the sample N was split into subsets for all n_i discrete values, resulting in subsets of the size N/n_i on average. It is important to consider, that the subset size depends on the number of discrete values n_i of a model input x_i, while the subsets of the sampling scheme proposed by Pianosi and Wagener (2018) were on average N/n for all model inputs x_i.

Also, if I understand the strategy correctly, then the inputs with small number of possible values (for instance the land use scenario) are associated with conditional distributions based on a very large number of samples (around N/n_i =7,500/2 in the case of land use scenarios), while the inputs with large number of possible values (for instance the parameterisation) are associated with conditional distributions based on much smaller number of samples (around 7,000/43). Do you think using such different sample sizes could have had an impact on the estimation of the KS values and hence of the PAWN sensitivity indices?

Your assumption concerning the subset sizes is correct. We admit that the manuscript does not convey this information clearly. Thus, we suggest to update this section of the manuscript as proposed in the reply above. In the current version of the manuscript we did not analyze the effect of the strong differences in the subset sizes on the confidence intervals of the calculated sensitivities. We did however compare the results derived with the PAWN method to the results inferred from an adapted version of the STAR-VARS method (Razavi and Gupta; 2016a, 2016b) that was not affected by the different numbers of discrete values for each model input due to its sampling design. We observed only minor differences between these methods and hence assumed that the effect of the different subset sizes is low. We suggest however for the updated version of the manuscript to consider that point in the bootstrapping. If that assumption is correct, we expect that if the impact of the different subset sizes is low when the confidence intervals remain in a comparable range for different numbers of subsets of the individual model inputs.

Finally, a new sampling strategy was recently proposed for PAWN (Pianosi and Wagener, 2018). While this new strategy is still designed for continuous inputs, and hence could not be used here, it would be good to mention its existence for readers who may want to apply PAWN in the future (as for the case of continuous inputs this would be recommended over the strategy in the 2015 paper).

The publication will be considered in the updated version of the manuscript, as suggested in the updated section above.

[5] I think the discussion in Section 4.2 is interesting but potentially slightly misleading. The authors clarify that "several assumptions were made in the development of scenarios that are highly subjective". I understand the importance of highlighting the subjectivity inherent in the scenario definition if the goal of this study was to make projections of the future evolution of the two catchments. However, this is not the objective when doing GSA. GSA answers the question: "how much output variation do we get if we vary the inputs within certain ranges?" The answer yielded by GSA (i.e. the sensitivity indices, the input ranking, etc.) is certainly conditioned upon the chosen ranges, however this is "intrinsic" to the question asked, regardless of how the choice is made - be it an "objective" calibration exercise (as done for the parameterisations) or a "storylines" approach. In other words, I think the point is to justify why certain scenarios are considered for the impacts assessment study; once they have been selected for that purpose, it follows that they would be used in the GSA too if one wants to know their relative influence with respect to other input factors of the model. So, I do not agree with the sentence (P. 26 L. 13-14) "For the SA of the simulated variables the diversity of the developed scenarios is essential.": diversity may be important for the impacts assessment (is it?) but not necessarily for the GSA. If a limited set of scenarios were selected for the impacts assessment, I would use that set for the GSA even if it is not diverse.

We agree with the comment, that the goal of a GSA is to attribute the variations in simulated outputs to variations in model inputs, rather than simulating possible futures for a catchment. This was not the message we wanted to convey with this study. We intended to point out how GSA and an analysis of the uncertainty bands as illustrated can complement any impact study in understanding the sources of the uncertainties in simulating future conditions.
Maybe this is again, an issue of terminology. What we specifically wanted to address in this section was the fact that the subjective decisions we as modelers make in developing future scenarios will, no doubt, affect the simulation of a variable of interest. Further, when the developed discrete scenarios for a model input result in a wide range of a simulated output this will also affect the sensitivity of the output variable to the respective model input.
The analogue example for a single continuous model parameter would be to change the interval of that parameter in which it can vary for an assessment of the its influence on the model output. The selection of the parameter range is apparently highly subjective as well. Yet, while increasing the interval of a continuous property to cover more extreme regions of the model input space is a simple concept, the impact on the simulation of an output variable caused by any assumptions made in the development of model input scenarios is not always entirely clear in the scenario development. As this issue is not always addressed appropriately in environmental impact studies (e.g. by only using a few climate scenarios in an impact assessment), we saw a high need for this important discussion.
With the term “diverse” we wanted to express to represent a wide range of possible future representations of a model input. The addressed sentence seems however to be redundant as the following sentence repeats the argument. Thus, we suggest to change this section as follows:
~~For the SA of the simulated variables the diversity of the developed scenarios is essential. Thus, s~~Scenarios must cover a broad range of possible futures...

Minor points

P. 1 L. 15: "scenario inputs" should be "input scenarios"

This will be changed accordingly in the updated version of the manuscript.

P. 2 L. 5: "the precipitation of the climate scenarios" sounds a bit odd, maybe "precipitation projections"

We agree with this comment and suggest to change the sentence as follows:
Additionally, the visual analysis of the uncertainty bands illustrated that the anomalies in precipitation of the different climate scenarios dominated the changes in simulation outputs, ~~rather than changes~~ while the differences in air temperature in both case studies showed no considerable impact.

P. 3 L. 3-4: "An assessment is only as good as the dominant contributors of uncertainty in such a modeling chain." Unclear. Something seems to be missing in this sentence: an environmental impacts assessment is only good if dominant contributors of uncertainty are... what? Identified? removed? ...?

The sentence actually does not contribute much information. Thus, we rather suggest to delete it in the updated version of the manuscript.

P. 3 L. 11-12: "model computations" should be "model evaluations" (or "runs" or "executions")

The phrase will be changed to: ... from certain a number of model ~~computations~~ evaluations.

P. 3 L. 19: "Most applications utilize GSA to identify and rank continuous model parameters". Unclear: GSA does not "identify parameters" at most "identify influential parameters"

Will be changed to:
Most applications utilize GSA to identify influential model parameters and to rank ~~continuous~~ model parameters according to their influence on model outputs. Model parameters are usually continuous model inputs.

P. 3 L. 21: "Although," Comma should be removed

This will be changed accordingly.

P. 3 L. 26-27: "An OAT analysis however presumes linear models and non-correlated inputs". Not sure OAT requires a linear model, for instance the Morris method uses a OAT approach and yet is typically applied to nonlinear models. More generally, why should GSA be applied to a linear model at all? If the model is linear than the effect of each input on the model output is simply proportional to the input variation, no need to do GSA to know that.

In this section we highlighted the equivalency of the standard procedure performed in impact assessments an **local** "one-at-a-time" (OAT) analyses. We agree that in this specific sentence we used the more general term "OAT" instead of referring specifically to "local OAT". The presumptions of OAT such as linearity of the model or independence of model inputs was also addressed by Baroni and Tarantola (2014) or Saltelli (2010). To infer the (global) sensitivity of a model output from a delta change of a model input presumes that the

same delta change of the model input at another position in the input space has the same effect on the model output and is not influenced by any other model input (linearity and independence). We further did not suggest to apply sensitivity analysis to a model where a linear relationship between the inputs and outputs is a-priori known. Contrary to that, we state that applying an OAT analysis to infer sensitivities of model outputs implies model linearity and the independence of model inputs.
Concerning terminology we referred to the terms as they are used in Saltelli and Annoni (2010), where OAT was considered to be performed from the same “nominal point” and the analogous analysis performed at various points in the input space was termed “radial” elementary effects (EE). We do however agree that methods such as the Morris method or EE also employ OAT sampling designs while inferring global estimates of the output sensitivities.
Thus, we suggest to modify the commented sentence and add “local” to specifically address the issues with local OAT analysis.

P. 4 L. 4: "complex". Unclear. What is the definition of a "complex" input?

The term “complex” was already used earlier in the manuscript (e.g. p.3L21, p.3L30). Yet, we do not provide an explanation of that term at any point in the manuscript. Further, the term “complex” apparently does not clearly convey what is meant here, where maybe “composite” might be a more precise term to use. Thus we suggest to change the term “complex” to “composite” and further add examples in p.3L21:
Although, it is possible to implement ~~more complex~~ composite model inputs (e.g. climate scenarios that affect several climate variables at the same time, or land use scenarios that can impact the model setup) in GSA...

P. 4 L. 5: "No study is known to us that takes advantage of GSA in the scope of environmental impact studies." What is the definition of "environmental impact studies" here? I would say that GSA has been applied to such studies before, e.g. Anderson et al (2014); Butler et al (2014); Le Cozannet et al (2015)

We agree that if you consider “environmental impact studies” in their actual broad context that GSA has been applied in several studies. Thus, the sentence is misleading and will be deleted. The publications mentioned here as examples should rather be acknowledged and mentioned in the introduction. We will adapt the introduction accordingly

P. 4 L. 13-16: Very long sentence, consider splitting into two.

Based on the GSA and the visual analysis of the simulated uncertainties we are able to draw conclusion on the simulation of discharge and $NO_3^- - N$ loads as impacted by the model setup, model parametrization and the future scenarios of land use, point source emissions and climate. These conclusions are of course limited to assumptions made in the model setup and in the development of the scenarios.

P. 8 L. 3: "Although," Comma should be removed.

This will be changed accordingly.

P. 8 L. 14-15 "The SWAT model setups for the Raab and the Schwechat involved decisions for the selected number of subbasins of a model setup and the definition of the HRUs." Convoluted sentence.

This section will be modified as follows:
~~The SWAT model setups for the Raab and the Schwechat involved decisions for the selected~~
~~number of subbasins of a model setup and the definition of the HRUs. Both modifications are necessary decisions for any SWAT model setup.~~
A SWAT model setup requires the modeler to determine an “appropriate” number of subbasins and to make decisions for the HRUs (such as eliminating “insignificantly” small HRUs from the setup). The SWAT model setups for the for the Raab and for the Schwechat had different numbers of subbasins and defined HRU differently.

P. 8 L. 15: "Both modifications": which modifications? Unclear

Please see the changes suggested in the comment above.

P. 9 L. 2: "involving". Unclear. Maybe "which requires"?

The text will be changed accordingly.

P. 9 L. 11: "to define of the thresholds". Remove "of"? In general, the entire sentence is a bit unclear. How is the "aggregation error" defined? Error in which variable, and with respect to what "correct" value?

Thank you for finding the typo. As this section requires further explanations this section will be modified as follows:
In total, we set up four SWAT models, two with 3 and two with 14 subbasins for the Schwechat catchment and six setups for the Raab catchments with two each of 4, 29, and 54 subbasins. We kept the resulting HRUs of full HRU setups unmodified. The numbers of HRUs in the reduced HRU setups were modified by applying thresholds for land use, soil, and slope classes. HRUs with an area below the defined thresholds were eliminated from a model setup. ~~To maintain a comparable aggregation error with the number of subbasins for the different model setups, w~~ We employed the R package topHRU (Strauch et al., 2016) to ~~define of the~~ determine optimum thresholds for land use, soil, and slope classes ~~and accepted~~ that result in a maximum aggregation error of 5% of the total area of the HRUs when comparing the changes of land use, soil, and slope classes of the full HRU setup and the reduced HRU setup with the same numbers of subbasins. Table 2 gives an overview of the final baseline model setups for both case studies.

P. 9 L. 14: "In a pre-analysis step," In the GSA literature, this kind of "pre-analysis step" is often called a "screening" analysis, as it aims at screening out the non-influential parameters. Maybe worth mentioning the term as it would be familiar to many readers.

We appreciate the suggestion. The text will be changed accordingly.

P. 9 L. 14: "relevant parameters". Relevant to what? Maybe better "influential"

The term will be changed accordingly.

P. 9 L. 21: "FDCs". Explain the acronym

The acronym “FDC” was introduced on p.4L11:
...as well as flow duration curves (FDCs) of daily discharge and daily $NO_3^- - N$ loads...

P. 13 L. 5: "To identify the impact of" maybe better "To measure the relative importance of"

We prefer your suggestion. Thus, we will implement it accordingly.

P. 13 L. 7: "PAWN involves". Unclear what "involves mean. Maybe better "PAWN uses"

The text will be changed to PAWN employs...

P. 13 L. 11: "the sensitivity of a model input x for a target variable y". Sensitivity is an attribute of the output, not the input. I would rephrase as "the sensitivity of a target variable y to a model input x".

As mentioned in our reply to the major comment [1] the entire manuscript will be modified to meet this suggestion. Thus, we also implement this suggestion.

P. 13 Eq. (1) and (2). The mathematical notation could be made clearer. I find it odd that in Eq. (1) KS takes as subscript the index of the fixed point (j) while its argument remains the generic input x_i. This choice also makes it more difficult to understand how maximisation occurs in Eq. (2). I think using the notation $KS(x_i^j)$ in both equations would make things much clearer.

We appreciate this comment and will change the equations accordingly.

P. 13 L. 23: "possible states". Why "states"? The term was never used with this meaning before. I would rather say "possible values".

This is a remnant of a previous version of the manuscript and will be changed accordingly.

P. 13 L. 24: "a lower sensitivity of the input xi on the target variable y". Again, rephrase as either "a lower sensitivity of the target variable y to the input x_i" or "a lower influence of the input x_i on the target variable y".

This suggestion will be implemented.

P. 13 L. 28: "subsetted with" Not sure "subset" can be used as a verb. Maybe better "divided into"

Your statement is correct. The verb “subset” is not listed in any dictionary. Thus, we will rephrase it as suggested.

P. 13 L. 29. "... were used for the sensitivity assessment". I would link this to the mathematical notation just introduced above and say: "... were used as target variable y".

We appreciate the suggestion and will implement it accordingly.

P. 14 L. 10-12: "In this study, we consider all execute model setups to be plausible..." I do not understand this clarification. What other approach would have been possible? To discard some simulations because deemed not plausible? And how would you define then what is plausible and what is not? Please clarify.

We agree that the phrasing sounds odd. Thus, this sentence will be rephrased to:
~~In this study, we consider a~~ All executed model setups ~~to be~~ represent plausible realizations of the future conditions in both catchments to simulate future discharge and $NO_3^- - N$ loads.

P. 14 L. 17: "low number of each input". Unclear. Do the authors mean "low number of inputs" (i.e. 5 inputs) or "low number of possible values taken by each input"?

Here we meant the latter. We will update that phrase to:
~~The low number of each input included in the study~~ The low number of possible values taken by each input allowed...

P. 14 L. 24-28. This sounds like a repetition of what just said in the methodology section, I do not think is needed. I would rather use this opportunity to explain how to read

Figure 2 (what is the difference between the panels and how to read each circle plot).

We will consider to replace this section with an explanation of Figure 2 in the updated version of the manuscript. As further modifications will be added to the figure (e.g. confidence intervals, etc.) we do not suggest modifications to the text at this point.

P. 15 L. 17: "highly sensitive" replace by "highly influential"

This will be changed accordingly in the revised manuscript.

P. 15 L. 25-26: "their overall sensitivities follow the general trend of the climate scenarios to a large extent". Unclear, please rephrase.

We will rephrase this section as follows:
~~For most of the analyzed signature measures t~~ The model setups yielded insignificantly low PAWN indices for the majority of signature measures with values below 0.1 in the Raab case study, indicating that the model setup ~~was not sensitive~~ had a low influence on most analyzed processes. ~~Although the Raab case study shows low sensitivities for the model setups, their overall sensitivities follow the general trend of the climate scenarios to a large extent.~~ The pattern of the resulting PAWN indices of the model setups closely follows however the pattern of the PAWN indices that were calculated for the climate scenarios.

P. 15 L. 33: "difference that is visible for the two". Unclear, do you mean "difference that is visible between the two"?

This sentence will be rephrased to:
~~A substantial difference that is visible for~~ A notable difference between the two case studies is how ~~the reference period relates to the uncertainty bands in amplitude~~ the simulations of long term monthly discharges and $NO_3^- - N$ loads in the reference period compare to the ranges of future simulations.

P. 15 L. 33-34: "how the reference period relates to the uncertainty bands in amplitude". Unclear what this means.

See our suggestion for the modification of that phrase above.

P. 16, caption of Fig. 2: "Model input sensitivities for signature measures". Replace by "sensitivities of signature measures to model inputs". And later on "sensitivities of" should be replaced by "sensitivities to"

The caption will be changed accordingly in the updated version of the manuscript.

P. 17, text and Figure 3: what does "specific discharge" mean? Why "specific"?

I am not sure how well established this term is used in the hydrologic community. The specific discharge relates a discharge (given in e.g. m^3s^{-1}) to the catchment area that produces the runoff and sums it for a specific time interval (in this case on monthly basis). We decided to use the specific discharge for a better comparison of the two catchments with substantially different catchment sizes.

P. 17 L. 6: "show a difference". Does this mean "show an increase"? If so, I would use "increase", it makes it easier for the reader to follow.

We prefer your suggestion. This will be changed accordingly in the updated manuscript.

P. 19 L. 12: "While a grouping...". Remove "While".

We think that “while” is essential for the meaning of the sentence. To consider other reviewer comments the section will be modified as follows:
While a grouping of the individual climate scenarios with respect to their temperature anomalies shows a more indefinite picture~~.~~, ~~A~~all climate scenarios simulated an increase in temperature.

*P. 24 L. 20: caused future land use change" Maybe "caused **by** future..." ?*

This will be changed in the updated version of the manuscript.

P. 26 L. 31-32: "The application of sampling strategies for SA usually do not account for the circumstances that one model input constrains any other model input". I do not fully agree. There is an increasing literature on GSA methods applicable to the case of dependent inputs, see for instance Mara and Tarantola (2012; 2017).

We appreciate the comment and I think it is worth to mention and acknowledge these publications. While it is a more straight forward approach to constrain a continuous property by another continuous property it might be not a straight forward procedure to identify all plausible scenario combinations for multiple model inputs (e.g. some future climate settings might make future agricultural practices implemented in a land use scenario impossible). In the context of the present work we were referring to the latter case. We suggest to clarify this in the manuscript and to acknowledge the substantial work that was done to constrain dependent continuous variables in GSA.

P. 27 L. 17-18: "by a factor of up to 5... up to 8". Do these numbers come out of a comparison of PAWN indices? If so, I am not sure I would draw such quantitative comparison. PAWN indices are (maximum) KS values: what is the practical interpretation of "a factor of 5" between KS values? I find it difficult to imagine.

We appreciate your feedback on that section. As a consequence we will remove that comparison in the updated version of the manuscript.

P. 28 L. 3: "the lack of tool that allow the practitioners access to such methods". Not sure I understand what the authors mean here. Several GSA software tools are available (some are reviewed for example in Pianosi et al 2015). So what is the problem here? That they are not "friendly" enough for practitioners to use them? Or that they are not sufficiently tailored to SWAT applications? Pls clarify.

We agree that the sentence is too vague. Software, toolboxes and libraries to perform GSA are available for many different programming languages, for instance the SAFE toolbox (Pianosi et al., 2015) for matlab, SPOTPY (Houska et al., 2015) for python, or R packages such as sensitivity (Iooss et al., 2015), or fast (Reusser, 2015).
From a practitioner's perspective the challenge is to assemble such a large number of models and to perform thousands of model simulations for a large number of model input combinations, instead of performing the status quo procedure of implementing single scenarios into a calibrated model. To generalize such analysis for the application in environmental impact studies we suggest to come up with frameworks that support the practitioner in this laborious working steps of a case study.
Thus we suggest to specify the section p.28L2-4. We suggest to mention the tool boxes, packages, etc. that are available to perform GSA, but highlight the challenge of of assembling, executing and evaluating such a large number of model setups instead.

References

Anderson, B., Borgonovo, E., Galeotti, M., Roson, R., 2014. Uncertainty in climate change modeling: can global sensitivity analysis be of help? Risk Anal. 34 (2).

Butler, M.P., Reed, P.M., Fisher-Vanden, K., Keller, K., Wagener, T., 2014. Identifying parametric controls and dependencies in integrated assessment models using global sensitivity analysis. Environ. Model. Softw. 59.

Le Cozannet et al (2015) Evaluating uncertainties of future marine flooding occurrence as sea-level rises, Environmental Modelling Software, 73.

Mara and Tarantola, 2012, Variance-based sensitivity indices for models with dependent inputs, Reliability Engineering and System Safety

Tarantola and Mara, 2017, VARIANCE-BASED SENSITIVITY INDICES OF COMPUTER MODELS WITH DEPENDENT INPUTS: THE FOURIER AMPLITUDE SENSITIVITY TEST, International Journal for Uncertainty Quantification

Pianosi and Wagener (2018), Distribution-based sensitivity analysis from a generic input-output sample, Environmental Modelling Software

Pianosi, F., Sarrazin, F., Wagener, T. (2015), A Matlab toolbox for Global Sensitivity

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

Houska, T., Kraft, P., Chamorro-Chavez, A., Breuer, L. (2015). SPOTting model parameters using a ready-made python package. PloS one, 10(12), e0145180.

Iooss, B., , Janon, A., Pujol, G., (with contributions from Boumhaout, K., Da Veiga, S., Delage, T., Fruth, J., Gilquin, L., Guillaume, J., Le Gratiet, L., Lemaitre, p., Nelson, B., Monari, F., Oomen, R., Rakovec, O., Ramos, B., Roustant, O., Song, E., Staum, J., Sueur, R., Touati T., and Weber, F. (2018). sensitivity: Global Sensitivity Analysis of Model Outputs. R package version 1.15.2. https://CRAN.R-project.org/package=sensitivity

Pianosi, F. and Wagener T. (2018b). A new implementation of the PAWN method to perform density-based sensitivity analysis from a generic sample, EGU General Assembly 2018, Geophysical Research Abstracts Vol. 20, EGU2018-10784.

Razavi, S. and Gupta, H. V. (2016a). A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. Theory, Water Resources Research, 52, 423–439.

Razavi, S. and Gupta, H. V. (2016b). A new framework for comprehensive, robust, and efficient global sensitivity analysis: 2. Application, Water Resources Research, 52, 440–455.

Reusser, D. (2015). fast: Implementation of the Fourier Amplitude Sensitivity Test (FAST). R package version 0.64. https://CRAN.R-project.org/package=fast

Saltelli, A. (2002). Making best use of model valuations to compute sensitivity indices. Computer Physics Communications 145, 280–297.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). Global Sensitivity Analysis. The Primer, John Wiley & Sons, Ltd, Chichester, UK.

Saltelli, A. and Annoni, P. (2010). How to avoid a perfunctory sensitivity analysis, Environmental Modelling & Software, 25, 1508–1517

Sarrazin, F., Pianosi, F., and Wagener, T. (2016) Global Sensitivity Analysis of environmental models: Convergence and validation, Environmental Modelling & Software, 79, 135–152.

Sobol, I. M. (1993). Sensitivity analysis for nonlinear mathematical models, Mathematical Modelling and Computational Experiments, 4, 407–414.