Reply to the reviewer comments RC3: 'Review of the manuscript by Schürz et al.' by Anonymous Referee #3

I provide my comments below according to the HESS review criteria. Given some of my major comments below, it does not seem necessary to provide a more detailed line by line annotation at this point.

We want to thank the Anonymous Referee #3 for their detailed review of the manuscript and the valuable comments made to improve the quality of the manuscript. In particular the critical comments on the methodology helped us to reassess the results and the conclusions drawn in this work. The comments made by Anonymous Referee #3 are printed in *serif, italic font* below. Our replies to the comments are written in black, non serif font and our suggestions to revise the manuscript according to a comment are highlighted with the colors blue for insertions and red for deletions.

1. Does the paper address relevant scientific questions within the scope of HESS? Yes. Trying to quantify and attribute uncertainty from various sources in "eco-hydrological" modelling in the context of climate and environmental change.

We appreciate the positive feedback on the relevance of our manuscript.

2. Does the paper present novel concepts, ideas, tools, or data? I found the way figures 2 to 7 very informative. I particularly found figure 2 very appealing in presenting SA results.

We appreciate the positive feedback on the visualization of our findings.

3. Are substantial conclusions reached? Given some of the discussions provided on the methodology below in NO.4, I am not sure if we can say conclusions are substantial.

A detailed reply to the specific comments can be found below (4 a) and b)).

4. Are the scientific methods and assumptions valid and clearly outlined? I very much liked how the manuscript tries to do a systematic and comprehensive approach, step-by-step, to set up the models, define scenarios, conduct SA/UA experiments, visualize (for better communication of) the results, and reach to conclusions. However, I have some major concerns about some of the methods and tools used in this study that I explain below:

a) Discrete PAWN SA: My most fundamental concern is related the way the main SA with PAWN is performed in this work, which also led to main conclusions in the paper. I strongly feel that the PAWN SA results (Figure 2) is largely impacted by the NUMBER of discrete realizations in each category (Table 4) and not by their CONTENT. In other words, it is intuitive that in this design of SA experiments, by default, the category with a higher number of members will always show a higher influence, because parameters sampled here will naturally have a much higher variability with respect to those categories. And this is exactly what we see in SA results and why some results are rather counter intuitive (e.g. negligible or small influence from land use changes or model setup, and very large influence from Climate and parameters). This is a fundamental issue that needs to be addressed by authors as it is the foundation for all conclusions.

We disagree with the argument that the influence of a model input / model setup depends on the number of realizations of that respective input/setup. Indeed, the model parametrization and the climate scenarios had the strongest impact on most of the analyzed processes and were represented by a substantially larger number of realizations compared to the other inputs. A counterexample to the statement that the sensitivity is per design impacted by the number of realizations of an input is illustrated by the influence of the point source scenarios in the Raab catchment for medium and low nitrate-nitrogen ($NO_3^- - N$) loads and $NO_3^- - N$ concentrations for medium and low discharges in this study. The calculated PAWN indices for these measures were substantially larger for the point sources compared to, for instance, the climate scenarios. Yet, only four point source scenarios (and only two industrial emission scenarios that eventually were responsible for the large sensitivities) were used, while 22 climate simulations were implemented.

We want to clearly point out however, that the number of discrete realizations of a model input can affect the calculation of a sensitivity index indirectly. In the case of the PAWN index a distance is calculated between the unconditional and the conditional cumulative distribution function (CDF) of a target variable (Pianosi and Wagener (2015) for example suggest to use the Komogorov-Smirnov test statistics). The unconditional CDF can also be estimated from all simulation that were performed (where all model inputs are perturbed), while to estimate the conditional CDF only simulations are used that used one discrete realization of the input of interest (this means all other inputs are perturbed, while the input of interest is kept constant). The distance measure is calculated for all realizations of a model input accordingly. The calculated distances for all conditional CDFs (keeping the model input constant at every respective realization) do have a certain distribution. To infer the PAWN sensitivity index, the calculated distances are summarized employing any summary statistics (Pianosi and Wagener (2015) for example suggests to use the median or the maximum). The choice of summary statistics can however strongly affect the comparability of the calculated sensitivity indices of the individual model inputs if the distance measure distributions for the model inputs substantially differ. As a consequence, we employed the maximum statistics in this study, as we were primarily interested in the maximum possible

influence an input has on an analyzed target variable. Different summary statistics, but also different methods for global sensitivity analysis (GSA, e.g. the method of Sobol (1993) that analyzes an average influence of a model input) were tested and evaluated during the compilation of this study. The outlined effects were observed in these analyses (yet not shown in this manuscript).
Further, the calculated sensitivities are well supported by the analysis of the simulation uncertainties. Inputs that showed a large influence on an analyzed process also showed a strong effect on the simulation uncertainty bands of that respective process.
Finally, we disagree with the statement that the negligible or small influence from land use changes or model setup are counter intuitive findings. In our opinion these findings were substantially discussed in section 4.1. Other literature cited in section 4.1 strongly supports the findings (e.g. Wagner et al. (2017), Guse et al. (2015), Mehdi et al. (2015a, 2015b), or Bieger et al. (2013) for the impact of land use change, or Jha (2014) for the model setups).

b) Design of Experiments: Authors do a great job particularly in explaining a rather careful and detailed procedure to setup the model, process the required data, define HRUs, and layout future land, pollution, and climate scenarios. This is extensive amount of work. However, I feel that this breadth has caused insufficient scientific depth in places in the manuscript. For example, it is unclear to me why certain various metrics are chosen in the SA analysis with VARS? How are these metrics really different from each other from an SA perspective (in particular, NSE and RSR are directly related, so why both are used?), Why this choice is not consistent with the metrics used in the next steps (e.g. what happened to KGE or RSR)? Perhaps strategically reducing some of the metrics can help in a more efficient way of conducting SA and presenting its results (e.g. some of the quantile classes presented in Figure 2 in each signature measure can be removed).

It is correct that different measures were used as objective criteria in the GSA to identify influential model parameters and in the model calibration (identification of behavioral parameter combinations). The purpose of the GSA was to screen the model parameters. This screening had an inclusive character, which means that the parameter had to be influential for at least one of the selected criteria. Consequently, the similarity of criteria did not affect the results of the parameter screening (if the measures are similar then the same parameters are influential for these objective criteria.). Contrary, the selection of behavioral parameters was exclusive. Thus, only criteria were used that describe the aspects of a simulated time series that we explicitly wanted to evaluate. In the selection of the criteria for the model calibration we referred to literature such as Pfannerstill et al. (2014).
We agree that the measures NSE and RSR are strongly related in their calculation. Yet, both measures differed completely in their application in this study. While the NSE was applied to the simulated and observed times series of a variable, the RSR was applied to various segments of the flow duration curves (FDC). Thus, the resulting NSE values also accounted for the timing of simulated values of a variable, whereas the RSR values of the FDC segments did only account for the distribution of simulated values of a variable letting aside the temporal occurrence of a value.
We fully understand that the Fig. 2 can overwhelm the reader, as we try to present a lot of information in one figure. Nevertheless, we think that all segments of a FDC characterize different processes of the water or the nutrient cycles (in this case). Further, the large number of analyzed segments of the FDCs visually support the gradual shifts of sensitivities

of a target variable from one model input to others. There is a chance that this information is lost, when removing too many of the analyzed FDC segments from the figure.

Or for example, what is the scientific reference or justification for the way UA is conducted here at the end using 7K simulations out of all possible combinations? Wouldn't a Latin Hypercube Sampling be a more effective choice than random sampling? These methods and choices (and other similar ones) must to be clearly justified in the manuscript.

As briefly mentioned above, other methods for GSA were tested as well (while not shown in the manuscript). A preceding analysis employed the Sobol method (Sobol, 1993) for 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).
As we identified issues with the average sensitivity that is expressed by the sensitivities calculated using the Sobol method (see also the reply 4a)) we utilized the random sample that was drawn for the Sobol method to calculate PAWN indices. Pianosi and Wagener (2018) outline how to estimate PAWN indices from any generic sampling. For this study the proposed concept was applied to discrete model inputs in this study.
We see however from this and other reviews on that matter, that the sampling and the confidence in the GSA results require greater attention in the manuscript. Thus, we suggest to revise the section of the input factor sampling in the revised version of the manuscript. Further, as proposed by Francesca Pianosi in her review, we plan to perform a bootstrapping (as presented in Sarrazin et al. (2016)) to calculate confidence intervals for the PAWN indices. This will greatly improve the results of the manuscript.

5. Are the results sufficient to support the interpretations and conclusions? Please see my comments above in NO.4.

We tried to clarify issues raised concerning the methodology that was applied to derive the results illustrated in the manuscript in 4 a) and b). Please find our replies to these comments below the respective sections 4 a) and b).

6. Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)? No. Details of SA/UA experiments are missing. In particular, I found description of the VARS method somewhat short and there are important details that are missing (a more careful description from the original papers or some of newer applications is recommended). Another very important information that is missing is the ranges used for parameters, and an explanation of how these ranges are determined. These ranges can impact all the SA/UA results. Or it is unclear how parameters are tied to HRUs, and how all different setups, with different NO. of HRUs, in different basins have the same number of parameters (42) when doing SA with VARS?

We agree that the explanations concerning the parameter sensitivity analysis are rather short, as we intended to focus on the actual sensitivity study. Yet, you are right that the working steps in the parameter sensitivity analysis and the model parametrization affect the results of the following study.
We suggest to elaborate the parameter sensitivity analysis with greater detail. For the updated version of the manuscript adding a table is planned that provides information on the

initial parameter boundaries, the boundaries of the final behavioral parameter sets and the type of change that was applied to the model parameters (whether the parameters were replaced by a single value globally or the spatially distributed parameter field was changed by a fraction of the parameter value or changed by adding/subtracting an absolute value).

7. Do the authors give proper credit to related work and clearly indicate their own new/original contribution? Yes.

Thank you

8. Does the title clearly reflect the contents of the paper? Yes for the most part.

Based on the assessment of the manuscript title we see no possibility to improve the title to more precisely reflect the contents of the manuscript.

9. Does the abstract provide a concise and complete summary? Yes.

Thank you

10. Is the overall presentation well-structured and clear? Yes for the most part.

Thank you

11. Is the language fluent and precise? I feel the language needs to be modified a bit. Both in terms of English grammar (double check usage of "the" and "comma"), and in terms of being scientifically more precise (e.g. using "pollution" instead of "emission"; or using "most influential input" instead of "most relevant"; or page 3 line 4; or page 3 line 26). I recommend a more careful review of the manuscript in this regard.

Thank you for the feedback on the language of the manuscript. Based on this comment and comments made by other reviewers we plan to carefully review the language in a revised version of the manuscript.

12. Are mathematical formulae, symbols, abbreviations, and units correctly defined and used? Yes.

We appreciate your evaluation.

13. Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated? Some of the quantile classes presented in Figure 2 in each signature measure can be removed.

We outlined our thoughts on reducing the number of quantile classes in our reply on comment 4 a). Please see our reply above.

14. Are the number and quality of references appropriate? Yes.

We appreciate your evaluation.

15. Is the amount and quality of supplementary material appropriate? Yes.

We appreciate your evaluation.

References

Bieger, K., Hörmann, G., and Fohrer, N. (2013). The impact of land use change in the Xiangxi Catchment (China) on water balance and sediment transport, Regional Environmental Change, 15, 485–498.

Guse, B., Pfannerstill, M., and Fohrer, N. (2015). Dynamic Modelling of Land Use Change Impacts on Nitrate Loads in Rivers, Environmental Processes, 2, 575–592.

Mehdi, B., Lehner, B., Gombault, C., Michaud, A., Beaudin, I., Sottile, M.-F., and Blondlot, A. (2015a). Simulated impacts of climate change and agricultural land use change on surface water quality with and without adaptation management strategies, Agriculture, Ecosystems & Environment, 213, 47–60.

Mehdi, B., Ludwig, R., and Lehner, B. (2015b). Evaluating the impacts of climate change and crop land use change on streamflow, nitrates and phosphorus: A modeling study in Bavaria, Journal of Hydrology: Regional Studies, 4, 60–90.

Pfannerstill, M., Guse, B., and Fohrer, N. (2014). Smart low flow signature metrics for an improved overall performance evaluation of hydrological models, Journal of Hydrology, 510, 447–458.

Pianosi, F. and Wagener, T.(2015). A simple and efficient method for global sensitivity analysis based on cumulative distribution functions, Environmental Modelling & Software, 67, 1–11.

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.