Comments from the Editor:

Dear authors

As you can see from reviews, two of them are quite positive (one suggest minor updates), but the one of Referee 3 (report 2) is not - he/she recommends rejection.

I also think this paper presents an interesting approach worth attention of scientific world. So in this respect I am with the two reviewers out of three. Here there is some disagreement, but it is just normal in the world of researchers.

At the same time - I would still strongly suggest to take on board (some of) the comments of the referee who suggests rejection. He/she also has a point - and this opinion can be shared by more people (please also see comments, including mine, at earlier stages - on clearly stating the difference with the traditional approaches.) In this current version, these differences are much better explained already, but still I would ask to address the raised points again. I think it is not too difficult to do.

Good luck.

We thank the Editor for this positive feedback.

Please find our replies to the comments from the reviewers below.

We like to highlight that as main modification we revised the introduction of the manuscript according to the comments from reviewer#1. Moreover, a map of the catchment sites was added as recommended by reviewer#2.

Our text is coloured as follows:

- Reviewer‘ comments
- Answer
- „Text directly taken from the revised version of the manuscript“
- **„Modified parts of the text directly taken from the revised version of the manuscript“**

First reviewer:

In the revised version of the manuscript, the authors have provided clarifications for study area and methodological choices. In addition, introduction has been edited to account for comments by the referees and editor. The manuscript has overall improved, but the following minor issues still remain unaddressed:

1. Unclear reasoning in the introduction: Some claims in the introduction are not substantiated by past literature. For instance:

a. Lines 10-13, Page 2 ('In parameter identification ……… Identification'): It is not true that parameter identification implicitly assumes that model parameters are precisely identified by selected performance criteria. Most sensitivity analysis studies have already shown that this is not true, and that parameters can affect performance in interaction with other parameters within the model. In fact in some cases, the interaction effects dominate the sensitivity indices. In fact, the assumption that an accurate parameter identification can be done is questionable given the interaction effects and uncertainties.

We agree that this part of the introduction needs to be improved. Thus, we revised the second paragraph of the introduction carefully. During the revision, the marked sentence was removed.

b. Line 25-28, Page 3 ('To our …. Task.'): There is a misinterpretation of sensitivity analysis here. If one applies sensitivity analysis with multiple performance measures, a matrix similar to Figure 4 can be obtained, where gray scales can be used to plot the first or higher order contributions of parameters to variation in each measure. A comparison across performance measures can then highlight which parameters uniquely identify which parameter measures and vice-versa. In fact, applying sensitivity analysis in this manner will give more information than the framework proposed in this study as they can also highlight higher order interaction effects. I think this study is not an advancement on sensitivity analysis but an alternative. Both methods can be used to attain the overall goals of the analysis. The only advantage of using CART over SA is that CART identifies the values of parameters that lead to good performance while SA does not. If used only for parameter/performance ranking, both are equally acceptable methods.

We certainly agree that our approach is an alternative to sensitivity analyses. And we also agree that the impact of parameter interactions can be better assessed in a sensitivity analysis method which also considers parameter interactions.
However, we think that the bijective approach in our study is new. The classical approach of identifying the best parameters for multiple performance criteria was enhanced by looking from the side of the model parameters. In our approach we used the entire set of model simulations in RT to identify the most appropriate performance criteria for a given model parameter. Here, to construct RTs only the set of performance criteria and the selected model parameter is used. Thus, all other model parameters are not directly included. They were only indirectly considered since the variation of their values also impacts the modelled discharge time series and thus the performance criteria. Nevertheless since only the selected model parameter is included in RT, the most appropriate performance criteria for handling this parameter can be precisely identified. And moreover, it is can identified whether one of the selected performance criteria is suitable to identify an adequate value for the selected model parameter.
We agree that in a sensitivity analysis, it is possible (and helpful) to compare the sensitivity values of model parameters for multiple performance criteria. However, with a sensitivity analysis it is not possible to detect the most appropriate performance criteria for a given model parameters (in particular for model parameters of low relevance). The performance criteria for which the highest sensitivity value was computed for a given model parameters is

not undoubtedly the most appropriate performance criteria. It could be that this performance criterion is not strongly impacted by variations in model parameters.
When comparing the results of both RT approaches (Figs. 5 and 6 in the revised version of the manuscript), it becomes apparent that the performance criterion with the highest percent contribution for a given model parameter in Fig. 5 is in several cases not identical with the performance criterion with highest percent contribution in Fig. 6. Thus, the analysis from the side of the model parameters provides different results and new knowledge about the interrelationship between model parameters and performance criteria. While the interpretation of the results in Fig. 5 is possible in a similar way with a sensitivity analysis, the Fig. 6 provides results which cannot be derived in a sensitivity analysis. This aspect is added as a new paragraph in the discussion (see P. 15 L. 13-16).

2. Choice of 2000 parameter sets: this choice is still poorly defended. If one has 12 parameters and each parameter range is divided into 10 equally spaced values, one still needs 10^12 values to represent the parameter space. 2000/10^12 is a very sparse set and cannot be claimed to represent the parameter space accurately. Therefore, a convergence analysis is the only realistic way to defend this choice. It is hard to say whether 2000 is far better than 1000. It is unclear why the results from a randomly chosen subset of 1000 (or 1500, or 1800) cannot be checked against the results presented in the main analysis, as a proof of concept.

In general we agree that 2000 model simulations are not enough to identify parameter values such as it is realized in calibration studies. However, our intention was to investigate the relationship between model parameters and performance criteria. Thus, our study is as discussed in the previous comments more related to a sensitivity analysis than to a model calibration. In a sensitivity analysis, the number of required model simulations can be lower such as in our experiences with the FAST model approach. Thus, to investigate the relationship between model parameters and performance criteria, we think that this number of model simulations is appropriate.
Nevertheless, as also discussed in the first review round, we checked how the patterns of our results are impact by a different number of simulations (500 and 3000). While the results for 500 model simulations are largely different from ours, the results for 3000 are similar. Thus, we think that our statements are supported by 2000 model simulations.
In this context, we like to mention that we had previous experiences with the SWAT model in both catchments. A better knowledge of how the model parameters react helps to select reasonable model parameters and reasonable parameter ranges. Based on this, the amount of unrealistic parameter combinations was already reduced. Thus, we think that the number of 2000 model simulations is also justified since more realistic parameter combinations are included as it would be in a first application in a new catchment. Certainly, the number of required model simulations might be higher if using a new model or an unknown catchment.

According to this comment, we extend the paragraph on model simulations in the revised manuscript for a better justification of our procedure:

The paragraph reads now as follows (P.5, L. 31 to P. 6, L.10):

Based on the physically meaningful selection of these twelve model parameters, their values were varied within a set of model simulations. The intention of these model simulations was to derive the interrelationship between model parameters and performance criteria. For this, model simulations for the period from 2000 to 2010 were carried out based on 2000 different parameter sets that were generated with the Latin Hypercube sampling approach as it is implemented in the r-package FME (Soetaert and Petzoldt, 2010). In the Latin Hypercube sampling, all model parameters were changed simultaneously within the whole parameter space. For a more detailed description, readers

are referred to Pfannerstill et al. (2014b).
All parameters values were already in a hydrologically plausible range according to prior modelling experience with the study sites. These constrained parameter ranges allowed selecting an efficient but appropriate number of simulations to perform our analyses. Please note, that the intention of the presented study was not to identify the parameter values exactly, which allowed us to keep the sampling of the parameter space relatively sparse. Instead, we aimed to test and suggest the new connective strength approach. For this purpose, the number of 2000 model runs ensured a sufficient number of combinations at each node of the RTs.

Moreover, we added a sentences to the description of regression trees (P.8, L.26-27):

In our case, the set of 2000 parameter sets and performance criteria are large enough to ensure a sufficient number of results at each node of RT.

Second reviewer:

After reading the revision and the responses to comments of all three reviewers. I'm kind of disappointed for several reasons. Therefore, I would rather suggest reject from publication at this point.

1. Authors mentioned in the response that the "focus of the manuscript is to investigate the relationship between model parameters and performance measures. Thus, we show which model parameters impact which performance measure and which performance measures are influenced by the different model parameters". I agreed, that's what I thought too. In my definition, that's also a part of parameter uncertainty, sensitivity analysis, and model calibration. One cannot investigate parameter uncertainty, sensitivity analysis by not doing what you did right (maybe not use LHS but other methods)? And, in most cases, that's a part of the work for model calibration. Then, scientists can use the calibrated model for further purposes, such as ANSWERING SCIENTIFIC QUESTIONS. That's why, I mentioned previously that the novelty of this manuscript is questionable.

We understand your remark as a general remark to the overall classification of our manuscript.

Following up on our earlier response, we do not aim with our current manuscript to "work" with the calibrated models and answer scientific or applied questions in the respective modelled watersheds. Instead our work is meant to be a contribution in the ongoing struggle of improving the modelling procedure itself, in particular to achieve a better handling of model parameters towards precise parameter value identification. We see our presented method as an additional and helpful approach to obtain knowledge about the suitability of different performance criteria that may be best to identify appropriate values of model parameters. Thus, following on our approach, the identification of the parameter values themselves could be the next step.

It seems that you agree with us, that there are other methods around that are likewise contributing to those issues. Against this background, we contribute with our work technically by adding a new approach (RT-analysis of the LHS simulations) and conceptually by introducing 1) the concept of bij*ective identifiability of parameter and performance measures (connective strength) and subsequently 2) the differentiation of* global sensitivity analyses and the connective strength to the discussion (Thanks for this nice and brief summary to the third reviewer (Richard Arsenault)). Thus, to our knowledge, there are no other approaches published, that consider explicitly the bijective relationship of model parameters and performance measures as we did in our manuscript.

To clarify this, we added this aim at a couple of places to our manuscript (P.3, L.30-32; P.9, L.15-16; P. 13, L. 13-17; P.15, L. 13-16).

We want to emphasize ones more, that we ourselves do not see our work as the final solution of parameter identification in hydrological modeling. Moreover, we see our approach as an alternative to sensitivity analysis and not as replacement. Rather, we see our work as contribution in an ongoing discussion as we stated in the last paragraph of the introduction of our manuscript (P.4, L.5-6):

“We present a way to detect the appropriateness of performance criteria **that are most helpful in the** identification of **hydrologically sound** model parameter **values**.”

We hope that we got your point and that we could clarify the overall classification of our manuscript and where we see specifically the contribution of our work to the discussion of hydrological model evaluation.

2. In this study, some parameters were identified to be more influential than others. If you change the targeted watersheds, you'll definitely have different sensitive parameter sets and values. It is for sure technically-oriented, and I don't see the solid scientific values here.

We agree that for (the modeling of) different watersheds, different parameters will be more influential than others due to different dominances of different hydrological processes. The same holds for the suitability of performance measures to assess the model performance for different watersheds and their characteristics and the respective relationships of performance measures and parameters, which we termed “connective strength” in our manuscript.

Thus, our work had not the intention to present relationships of parameters and performance measures that are generally valid independent of the watershed, instead we want to contribute to the task of identification of most influential parameters in modeling with a generally applicable framework, that allows identifying and comparing characteristic parameter-performance measure relationships of different watersheds. However, we like to mention that some statements such as the strong bijective relationship between soil and evaporation parameters to performance criteria related to water balance, e.g. KGE_beta or RSR for mid flows are expected to be reproducible in other studies.

3. By the way, I just don't understand why you just don't add a map for both watersheds? How difficult is that?

We added a map of both catchments.

4. For what you did, I also think that's a failure. The reason is simple, and I didn't catch that in the first round but the reviewer#1 did. It was mentioned by reviewer#1 that "It is surprising that no appropriate performance criteria is found to relate to CN2, which is generally a sensitive parameter in SWAT". You justified that "CN2 was varied in these HRUs from 40 to 60 which is assumed to be sufficient to maintain the landscape heterogeneity". It seems the range of CN2 has been allocated in values that's already been validated in your previous work. In this case, of course the value of CN2 is not sensitive (it's in close to calibrated values anyway). Then, why are we even doing the proposed work in the first place? Are we supposed to explore the potential ranges (and the relationships between performance measures and parameters) of the parameters to validate the following steps? How can we know if other parameters are not being handled the same way?

We are not quite sure, whether we get the point, but we did not aim with our study to provide a global sensitivity-analysis of the parameter set of the SWAT model itself, but to investigate the parameter-performance measure relationships of the hydrologically realistic model runs. Thus we did not randomly vary all parameters of the SWAT model. Instead we restricted our analysis to ranges of parameter values which are according to our knowledge of the study areas hydrologically plausible. Therefore, the ranges of all parameters were carefully selected with the aim to represent the hydrological behaviour in a realistic way. E.g. CN2 was initially set to different values according to the land use type in the HRU, i.e. higher values for urban areas, lower for forest etc.. Based on these initial settings, CN2 was varied. Thus, certainly spatial heterogeneity is included and it does not make sense from hydrological point of view to increase or decrease CN2 dramatically. In addition it is not useful to increase CN2 above 100 or below 25 which since is beyond the possible values of CN2. Thus, if aiming to maintain spatial heterogeneity in landscape we have to set the parameter variation so that CN2 is within these ranges for all landscapes.

There are still several other issues, but I would like to stop it here. Hope you understand that I love your work, but the issues of novelty presented in this study cannot be resolved in easy ways.

Thank you for this final positive feedback. We interpret your statement like this that although you are not really convinced from our work at the current stage you still see some glance of novelty in it and that your main concern is the classification issue raised in your first remark. We hope that we could clarify this issue with our response to your first remark

Third reviewer (Richard Arsenault)

First, I commend the authors for their work on the revised manuscript. I believe the authors responded adequately to the reviewers comments and that the paper should be published in HESS. While I think that there might not be many uses of the proposed technique in the short-term, the idea of bijective identifiability is indeed novel and scientifically sound, and could open the door to further improvements. The differentiation between global sensitivity analyses and the connective strength approach is clearer. Finally, the limitations of the paper are better defined. Overall a good job by the authors.

Thank you very much for this very positive feedback and the very subtle and precise summary of our work!