Authors’ response to Anonymous Referee #3

Italicized text: comments made by Referee #3.

Blue text: Authors’ responses. The line numbers mentioned below correspond to those in the revised version of the manuscript.

The article by Yang and Chui shows the results of a study on the prediction of the hydrological response of sustainable urban drainage systems (SuDS) using Machine Learning algorithms.

Through an in-depth examination of both the manuscript and the authors 'responses to other reviewers' comments, I was able to appreciate the effort made by the authors to adequately address the constructive comments of colleagues.

However, I believe that the article is not yet ready for publication and should be re-evaluated after further review.

Thank you very much for reviewing our paper and providing insightful suggestions. We have revised our paper carefully according to your comments and the comments from the editor. The main changes to the manuscript are as follows.

1. Improved presentation. (a) The paper is shortened by around 25%, i.e., about 3,000 words have been removed. (b) We reduced the use of equations and jargon from the machine learning (ML) field. 5 equations have been removed. (c) We removed the detailed introduction to explanation methods, which are too basic and not essential for understanding the study results. (d) We added three new figures (Figures 1, 3, and 10) to graphically illustrate the model training processes, the numerical experiments to be conducted, and the processes of inferring the physical processes being modeled from ML modeling results. (e) The figures are easier to read due to the removal of redundant results.
2. More in-depth discussions on the practices of inferring the physical processes being modeled from ML modeling results. In section 3.6, we presented a high-level overview of the practices of making inferences based on the model explanations. We show that there are large uncertainties involved in every step of the inference process, which are often overlooked in current hydrological ML studies. We consider this to be an important message of this paper.
3. Removal of non-essential findings. (a) In the previous submission, four feature engineering methods were used; however, they are overall similar. In the updated manuscript, only one method is used. (b) We removed the content on testing whether overfitting occurs in the optimization process. (c) The content on the estimation of “water age” has been remove as they may be considered redundant to the catchment response time estimation results.
4. Better organization of the results. In the previous submission, we did not raise an interesting research question that connects different parts of the paper, so that the results seem to be random. In the updated manuscript, we shift our focus from building more accurate models to testing whether the explanations derived from ML modeling results are consistent with our hydrological knowledge. The four numerical experiments to be conducted are described

clearly in the method section, and all of them are closely related to the main questions proposed in the introduction section.

5. New numerical experiments on analyzing the correlation between a model's prediction accuracy and the physical realism of the inferred physical processes. This experiment is added because we intended to show that an ML model does not have to make right predictions for the right reasons, so that the explanations obtained by analyzing model structures can be certainly unreliable. We then present an in-depth explanation in Section 3.6, where the results of all four experiments are connected.

Please find our responses to each of your comments below. More details about the major changes can be found in the manuscript document with major changes explained. Please let us know if you have further suggestions. Thank you again for reviewing our paper.

(General comments)

I expect the Authors to further improve the article in order to overcome my main concerns:

1. First, I believe that a Machine Learning-based approach is less suitable for addressing a SuDS problem than a physically based systemic approach. If it is true that in some cases the geometric and hydraulic characteristics of these systems are not well known, it is even more true that the presence of experimental field data on inflows and outflows represents a rare exception. For this reason, a physically based modeling is in most cases to be preferred and allows to address a wider variety of problems, while in this case a model based on Machine Learning algorithms would be limited to the specific case. Authors should give more convincing reasons for the choice of approach.

We fully agree with your point that ML methods are less suitable for modeling SuDS. In fact, we also mostly use process-based models, such as SWMM and GIFMod, in our SuDS-related studies; and we have written papers to improve these models. Please find below our justifications for conducting this research and the modifications made to the manuscript.

In this paper, we apply ML methods to SuDS studies based on the following considerations.

1. ML as a solution to benchmarking process-based models in SuDS-related studies. In our own research, we sometimes encounter SuDS sites that are so uniquely designed that we have to significantly twist the model presentations in the commonly used process-based models. However, we do not know whether such adjustments have led to satisfactory prediction accuracies if a "satisfactory" accuracy level is not defined. This study demonstrates that ML models can be set up quickly with little effort so that the prediction accuracies derived using ML methods can be used as a reference.
2. SuDS catchments as representations of small urban catchments. We certainly agree that monitoring data for SuDS are limited, and if such data are available then the other types of field measurements normally should also be available. We intended to use SuDS catchments as representations of small-scale urban catchments, and for which more data are available. For instance, the United States Geological Survey (USGS) has monitored many small streams near cities and the data are publicly available. The Shayler Crossing Watershed (SHC) studied

in the paper is in fact a small urban catchment (around 1 km^2) with SuDS features, which is of a much larger scale than a SuDS site.

3. It is easier to verify the physical realism of the inferred physical processes in small catchments compared to large natural catchments. For example, runoffs of a small catchment are expected to be affected mostly by recent rainfalls. However, in large natural catchments, it is difficult to identify the forcing factors that contribute the most to runoff generation, as the snow melting and baseflow processes that are related to the long-term climate patterns may be the dominant processes.

We made the following changes in the updated manuscript to make the justifications for conducting this research clearer.

1. We explained that ML models should be used as a reference to evaluate process-based models so that ML models should not be considered as a replacement to process-based models.

   "The resulting statistical models may be adopted for solving various prediction tasks and used as references to assess the prediction accuracy of process-based models." (line 51 to 52)

   "Thus, in future SuDS studies, it can be useful to quickly train some ML models based on available data and used them as a reference to evaluate the prediction accuracies of process-based models." (line 421 to 423)

   "ML models can be set up relatively easily provided that observation data of the variables of interest are available and thus are recommended to be used as a reference to evaluate process-based models." (line 585 to 587)

2. We commented that the SHC study is conducted to test how well the proposed method performed for small urban catchments.

   "SHC represents a typical residential area in the U.S. and is thus selected to test the applicability of the proposed ML methods in modeling small urban catchments." (line 326 to 327)

   "The proposed ML model training methods can be potentially extended to study other small scale urban catchments that have similar configurations to SHC." (line 423 to 424)

3. We explained that it can sometimes be difficult to come up with expected catchment hydrological behaviors, which are used for assessing the consistency of the inferred processes.

   "The inability to draw a definitive conclusion can also be caused by the lack of knowledge of the processes being modeled. For instance, it is difficult to identify the expected hydrological behaviors for ungauged natural catchments." (line 305 to 306)

Finally, we also brought up the issues of trustworthiness of ML models in the introduction section, so that we remind the readers that we should be critical against ML models as they are opaque.

> *"Most of the current hydrology literature uses post-hoc explainability techniques to test whether an ML model makes right predictions for the right reasons, where a model is*

generally considered more trustworthy if it can generate predictions in a way that is consistent with our knowledge of the system being modeled." (line 91 to 93)

2. The novelty of the work is not relevant from a methodological point of view. The XGBoost algorithm is widely used in literature, as well as the other tools (Nested cross validation, Bayesian optimization, etc.) used in modeling. Authors should better highlight in the introduction section why this work would represent a significant upgrade over existing literature, worthy of publication in HESS.

Thank you very much for raising the concerns about the novelty of this paper. We fully agree that the application of XGBoost and other ML tools is not innovative. The innovative part of this paper is on evaluating the physical realism of the hydrological processes inferred from ML modeling results, i.e., we examined whether ML models make the right prediction for the right reasons. However, in previous submissions, we emphasized too much on prediction accuracy estimation and model training method, and the content on the physical realism estimation is hard to find due to the lack of logical connections between different results. The previous submissions are more like "engineering" papers rather than research papers because the interesting findings are reported along with the ML prediction accuracy evaluation results without a proper definition of the research questions.

To address the issues mentioned above, we rewrote most parts of the paper, with a shift in focus from building accurate prediction models to testing the physical realism of inferred hydrological processes. We present below the innovative points and main findings, and our attempt to improve readability according to the updated research objective is presented in our response to your next comment.

1. A framework to assess the physical realism of the inferred physical processes is defined in the methods section, which emphasizes the importance of assessing the reliability of the inferences.
2. A high-level overview of the activities of inferring physical processes being modeled from ML modeling results is shown in Section 3.6.
3. We showed both conceptually and empirically that large uncertainty exists in every step of the inference processes. However, many current studies overlooked these uncertainties, and the inferred processes are regarded as "semi-truths" or findings. We consider that it is important to broadcast this information to hydrological communities.
4. We showed empirically that a model's prediction accuracy is not necessarily correlated with its ability to provide consistent explanations to the physical processes it models. That is, an ML model can make right predictions for the wrong reasons. Thus, we should not consider explanations produced by more accurate models as closer to reality.

3. The article is very long, and in some places still unclear. It could certainly be shortened without compromising its contents. This is the most important aspect on which Authors should focus their efforts in order to obtain a more concise and clear manuscript.

Thank you very much for the comment on the readability of this paper. We consider the lack of readability was mainly caused by an improperly defined research question in the previous submissions. We were too critical about process-based models in the introduction section, and the goal is to build more accurate ML models. However, in the results section, we reported that ML models can generate inconsistent explanations to the physical processes, which lower their credibility and are confusing results. However, too much effort was spent on introducing an ML model training method and assessing the prediction accuracy, and we did not clearly explain why and how to assess the consistency of the explanations.

To address this issue, we redefined the objective as,

"Therefore, in hydrological studies, it is meaningful to ask whether post-hoc explainability techniques and ML models can provide physically plausible explanations to the processes of the system being modeled and whether a model's abilities to provide accurate predictions and plausible explanations are correlated." (line 103 to 105)

To accommodate the updated objective to improve readability, we made the following changes to the manuscript.

1. The introduction and methods sections have been modified according to the updated objective. We reduce the content on the shortcomings of process-based models, introduction to detailed ML methods, and added more explanations on testing the physical realism of the explanations of ML models.
2. We removed 3,000 words from the manuscript, which is about 25% of the main content of the previous submission. The writing is more precise in the updated manuscript.
3. We removed 5 equations and some detailed introduction to various ML methods.
4. The detailed introduction to ML model explanation methods has been removed, as they are not essential for understanding the study results.
5. Three graphical illustrations are added. Figure 1 shows briefly the raw input, the features, and the output of ML models, so the readers can more easily relate the equations to hydrological modeling. Figure 3 shows concrete examples of the processes of explaining the basis of model prediction, the process of inferring physical processes being model, and procedures to evaluating physical realism of the inferred processes. The concrete examples are discussed when each process is introduced in the method sections. We believe this can help the readers understand the methods better. Figure 10 is added to illustrate the process of inferring physical processes from ML modeling results from a higher level. To facilitate understanding, we use a cake metaphor, where we compare the ML explanation-based inference process to the process of identifying the ingredients used in a cake. This metaphor allows easier explanations of the uncertainties involved.
6. We removed much content on testing ML models' prediction accuracies and collectively analyzed the results of the two catchments. (1) In the previous submission, we used 4 similar feature engineering methods. And in the updated manuscript only one method is presented, as they are overall similar. (2) The results on evaluating whether overfits occur were removed. (3) The results on estimating the "water age" of stormwater runoffs were removed. (4) The physical realism of both study sites is now examined together, in the same figure, and described using the same paragraph.
7. We shortened the conclusion section significantly; the new conclusions are presented using bulletin points for clarity. Many non-essential findings are omitted.

(Specific comments)

Furthermore, I would ask the Authors to consider the following additional comments:

4. P3 L67 - "Machine learning methods, also referred to as data-driven modeling, predictive modeling, and statistical learning": These terms are not strictly equivalent.

Thank you for pointing the issue related to terminology. In the updated manuscript, we state that "Terms that are closely related to ML include data-driven modeling, predictive modeling, and statistical learning." (line 54), which are more precise than the original statement.

5. P4 L126 – Section 2.1.1 Local and global methods: This section is very basic and not essential for the manuscript. Sections 2.1.1 and 2.1.2 should be merged and shortened.

We have removed section 2.1.1, which is a basic introduction to ML explanation methods. Section 2.1.2, which is on the SHAP method, has now been shortened by 45% (from 900 words to 400 words). For more details, please see our responses to your comment #3.

6. P14 L183 – "The water levels are converted into discharge measurements using stage discharge rating curves": How were the curves obtained? Do they come from an appropriate calibration?

We removed this sentence on water level-discharge conversion from the updated manuscript for more concise presentations. The conversion was performed by USGS. In the updated manuscript we just state "The rainfall-discharge data collected between 2010 and 2013 by USGS are used in this study." (line 353 to 354)

7. P16 L426 – "The feature engineering and XGBoost hyperparameters are automatically optimized using the Bayesian optimization": Authors should report the optimal values of the hyperparameters, possibly in a table.

Thank you for your suggestion. We added Table 1 to report the range of the considered values. However, the optimal values are not shown because more than 400 optimal models are trained in this study. The optimal hyperparameters and the intermediate modeling results (together with all the source code) can be found in the link provided in the Code availability section.

8. P18 L486 – Nash-Sutcliffe efficiency coefficient and coefficient of determination R^2 are very similar metrics, it is not useful to consider both.

Thank you for sharing your opinion regarding the model performance metrics. R^2 is used because it is a well-known metric used in statistics and ML, so readers from different fields may look for R2. In some recent hydrological papers, both NSE and R^2 are used, e.g., in Sahour et al. (2020) and Yang et al. (2020).

However, we do agree that discussing two similar metrics is redundant. Therefore, in later experiments, such as experiment #4 and Figure 9, only NSE metrics are used. Also, in the abstract and conclusion section, only NSE metrics are reported.

9. P28 L668 – Conclusions: This section should be much more concise and effective in summarizing the main findings of the study.

Thank you for your comments. We reduced 500 words (i.e., a 75% reduction) from the conclusion section. The main conclusions are now presented as bulletin points for clearer presentations.

References

Sahour, H., Gholami, V. and Vazifedan, M.: A comparative analysis of statistical and machine learning techniques for mapping the spatial distribution of groundwater salinity in a coastal aquifer, J. Hydrol., 591, 125321, doi:10.1016/j.jhydrol.2020.125321, 2020.

Yang, W., Yang, H. and Yang, D.: Classifying floods by quantifying driver contributions in the Eastern Monsoon Region of China, J. Hydrol., 585, 124767, doi:10.1016/j.jhydrol.2020.124767, 2020.