We would like to thank the Anonymous Referee #1 for providing valuable and constructive suggestions regarding our manuscript. Please find our responses below.

Italicized text: comments made by Referee #1.
Blue text: Authors' responses.

1. Firstly, this work is innovative for explaining machine learning predictions in hydrology forecasting. With applying AI in various fields and getting excellent results, it is a hot topic to interpret the machine learning. But this manuscript still has some questions needed revised. Generally, it is a good research point, but manuscript is hard to understand.

The referee commented that our manuscript is hard to understand. To improve the readability of the paper, we updated the manuscript in the following aspects: (a) simplification of the methods, (b) removal of non-essential findings, and (c) re-organizing the paper according to the updated research objective. The details are as follow.

(a) Simplification of the methods. We updated the feature engineering method and the hyperparameter optimization method, both of which are used for training machine learning models.

In the revised manuscript, the high-resolution rainfall time series is converted into rainfall depth features using three hyperparameters, m, l, and n. Only the rainfalls recorded between $t-m$ and $t-0$ are considered. Each rainfall depth recorded between $t-l$ and $t-0$ is used for creating rainfall depth features. And n intervals are created for aggregating the rainfall recorded between $t-l-1$ and $t-m$, and the intervals roughly form an arithmetic sequence. See the illustration in Figure 1. The updated method is easier to understand, and the complex equations (Eqs. 5-7) in the original submission can be removed.

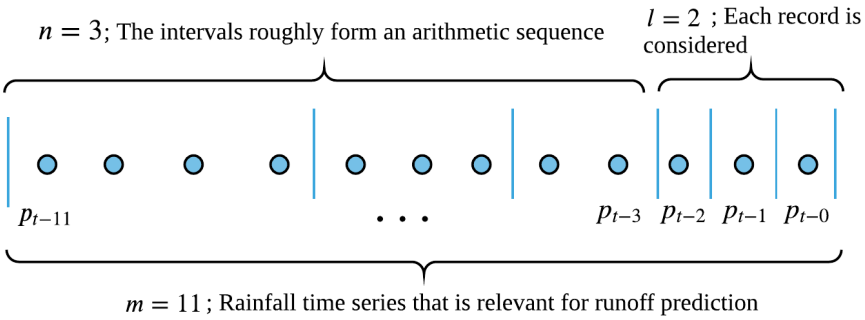

Figure 1. Illustration of the method to derive rainfall depth features. p_{t-i} denotes the rainfall depth recorded at time $t-i$.

We also simplified the hyperparameter optimization method. In particular, in the revised manuscript, we no longer differentiate the feature engineering hyperparameters and the

XGBoost hyperparameters, and all the hyperparameters were optimized together through an automated Bayesian optimization method (Snoek et al., 2012). Figure 2 shows that the Bayesian optimization method can find high-quality solutions (as indicated by low inner cross-validation (CV) errors) in a few optimization steps.

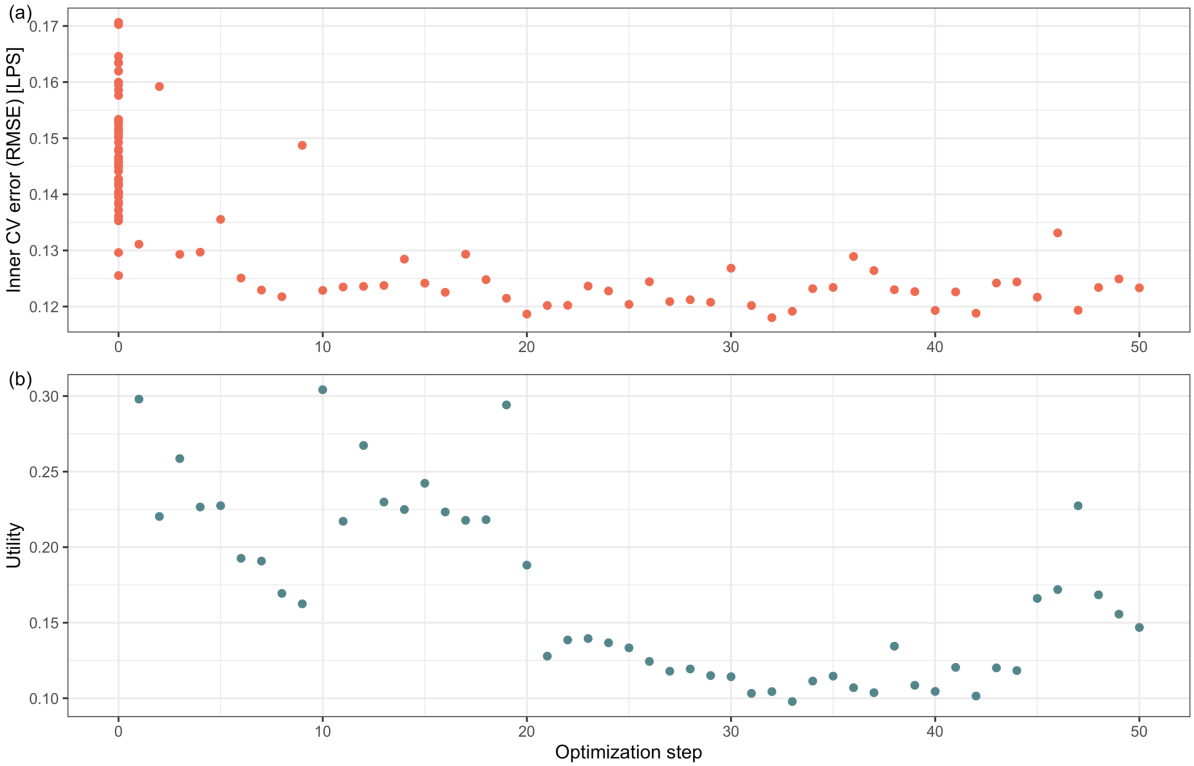

Figure 2. (a) The model's prediction accuracy associated with the hyperparameters evaluated during each optimization step. The prediction accuracy is measured by the root-mean-square error (RMSE) of the predictions obtained during the inner cross-validation (CV) iterations. (b) The expected utility of the candidate hyperparameters evaluated during each optimization step.

We also investigated whether the optimization method resulted in overfitting the model selection criterion. As indicated by the positive correlation between the inner and outer CV errors in Figure 3 (i.e., the good models found during the optimization process also had good performances during testing), the optimization method did not overfit the model selection criterion. Readers do not need to understand the technical details regarding Bayesian optimization, and the method is described briefly in the revised manuscript. In this way, the descriptions on the resampling scheme, the choice of feature engineering and XGBoost hyperparameters, and the model selection method can all be removed or substantially shortened. The models derived using the updated and the old methods were found to have comparable prediction accuracies.

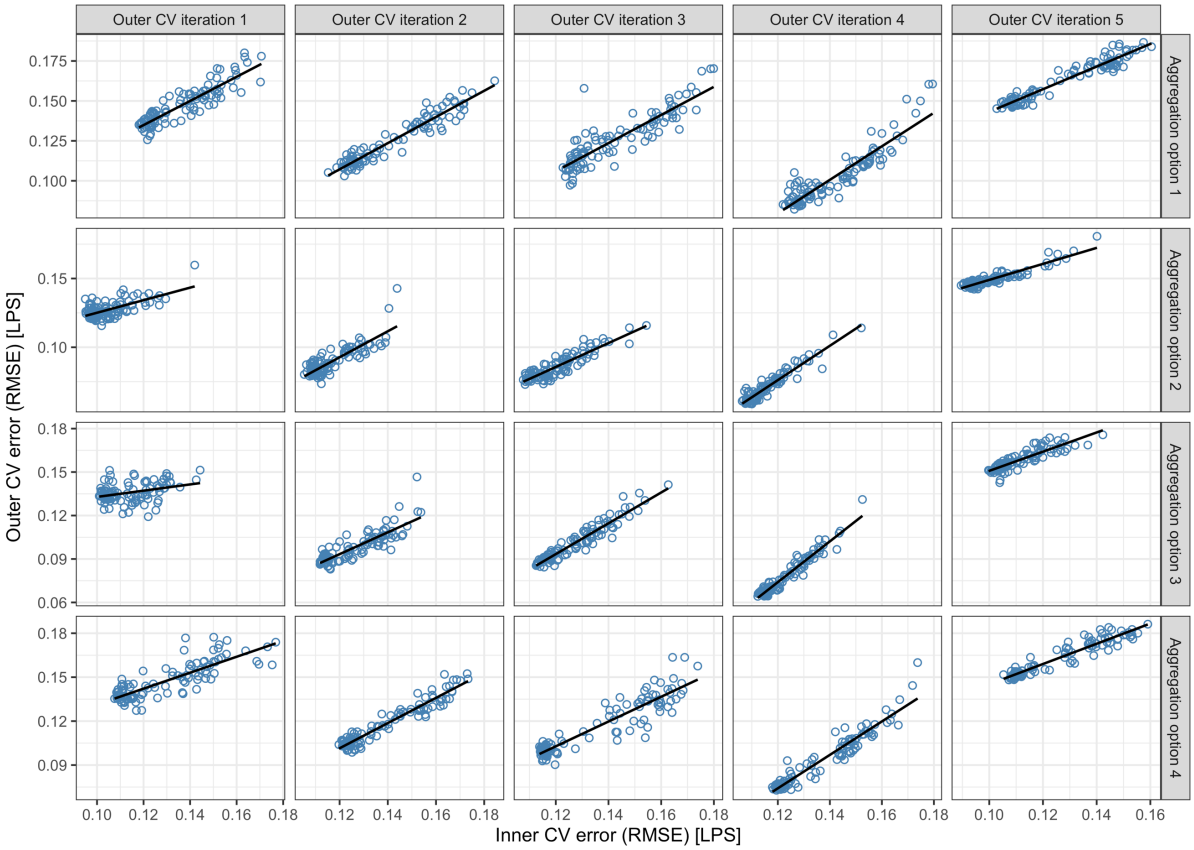

Figure 3. The model's prediction errors estimated during the inner and the outer cross-validation (CV) iterations for each set of candidate hyperparameters evaluated during the optimization. The prediction error is measured by root-mean-square error (RMSE). Each subfigure shows the result obtained for a rainfall depth feature aggregation method and during an outer CV iteration.

(b) Removal of non-essential findings. In the updated manuscript, the results of XGBoost hyperparameters optimization (Section 3.1) were removed due to the updated method and the new research objective. The results on interpreting the feature engineering hyperparameters (Section 3.3.1) were also removed due to their indirect connections to the hydrological processes. The descriptions of goodness-of-fit of the trained models (Section 3.2) were shortened.

(c) Re-organizing the paper according to the new research objective. We removed the following research objective from the updated manuscript, "develop and present tools and methods for building higher quality machine learning models for SuDS-related studies and demonstrate the applications". And the overall objective became "modelling the hydrological responses of SuDS to rainfalls and examining the basis of predictions using interpretation methods". Therefore, we removed the "demonstration" element from the original submission, and focused on developing methods for modelling SuDS and interpreting the model predictions. New

findings applying the proposed methods were also reported. The following changes were made in terms of the content of the paper.

In the introduction section of the updated manuscript, an introduction to the interpretation methods and their applications in hydrology is presented. This will help readers understand the research objectives.

In the methods section, we extended the introduction to the interpretation methods. In particular, the differences between the local and the global interpretation methods (such as the commonly used gain and cover metrics for XGBoost) were discussed, and the two methods (the observational and the conditional methods) to compute the SHAP values and their implications were introduced. We also substantially shortened the description of machine learning model training methods. More focus was given to the interpretation methods, as literature on this topic is currently lacking in the field of hydrology, and readers can get a better understanding of the results if more information on the interpretation methods is provided.

In the results section, we removed some of the results on machine learning model training and the results on interpreting feature engineering hyperparameters. The comparison between SHAP and other global interpretation methods was added, as suggested by Referee #2. We also examined the differences in explanations when using different interpretation methods for different models. In addition, we explained the potential applications of the interpretation methods in greater detail.

2. The logic of this paper is not clear that I cannot figure out what information explained by SHAP model and what relationship of hydrological response and selected hyperparameters.

The SHAP method is a feature attribution method, i.e., it aims to explain how much each feature contributed to the output of a model for a particular sample (Janzing et al., 2019). For this particular study, the rainfall's contribution to the subsequent runoff predictions at each time step is computed. In the updated manuscript, we added a formal introduction to feature attribution problem and their methods. We hope this can help the readers understand the results.

In response to this particular comment, we removed the content on explaining the hyperparameters in the updated manuscript, as we found their connections to the hydrological processes are indirect. The updated manuscript only explains the contribution of rainfall to runoff predictions at different time steps.

3. I think the main question is limited input variables (only Rainfall depth). I cannot agree that the design rainfall depth features (Section 2.1.1) reflect SuDS hydrological process. Thus, the hyperparameters of m, l, q, account_CumRain and account_season have little meaning for interpreting hydrological process in SuDS. Originally, SHAP is a game theoretic approach to explain the output of machine learning model. So maybe more physical observation variables are needed to selected as input variables. Therefore, I suggest this manuscript for Major Revision and Resubmission.

We agree with this comment that more variables can be included in the machine learning models. However, for the study site, only the rainfall and the runoff time series were available. The lack of data (such as the physical properties of the catchment) for setting up process-based models was also a motivation for using machine learning methods. In addition, this study focuses on modelling stormwater runoffs of small-scale urban drainage infrastructures during the wet period (i.e., within 24 hours of rainfall events), thus rainfall is the main driver of the system being modelled. Studying how a runoff prediction is affected by the rainfall of each time step is meaningful. Moreover, we also considered additional input features to the machine learning models to account for potential seasonality of the performance of the SuDS. Finally, this study recommends future studies to include more variables in machine learning models in the conclusion section.

In response to this comment, we removed the content on explaining the hyperparameters, as they are indirectly connected to the hydrological processes of SuDS.

In the updated manuscript, we presented SHAP as a method to explain the basis of a prediction for checking whether that prediction can be trusted. That is, we do not think the hydrological processes inferred by machine learning models are necessarily true. We further clarified this point by adding studies comparing the inferred hydrological processes of different machine learning models derived using different methods for computing SHAP values. As an example, Figure 4 shows that different machine learning models considered the rainfall's contributions to runoff predictions differently. Thus, there were considerable uncertainties in interpreting machine learning model predictions. The existence of various possible explanations was referred to as equifinality in Schmidt et al. (2020), which is an important concept in hydrological modelling (Beven and Freer, 2001). We reported this issue in the updated manuscript.

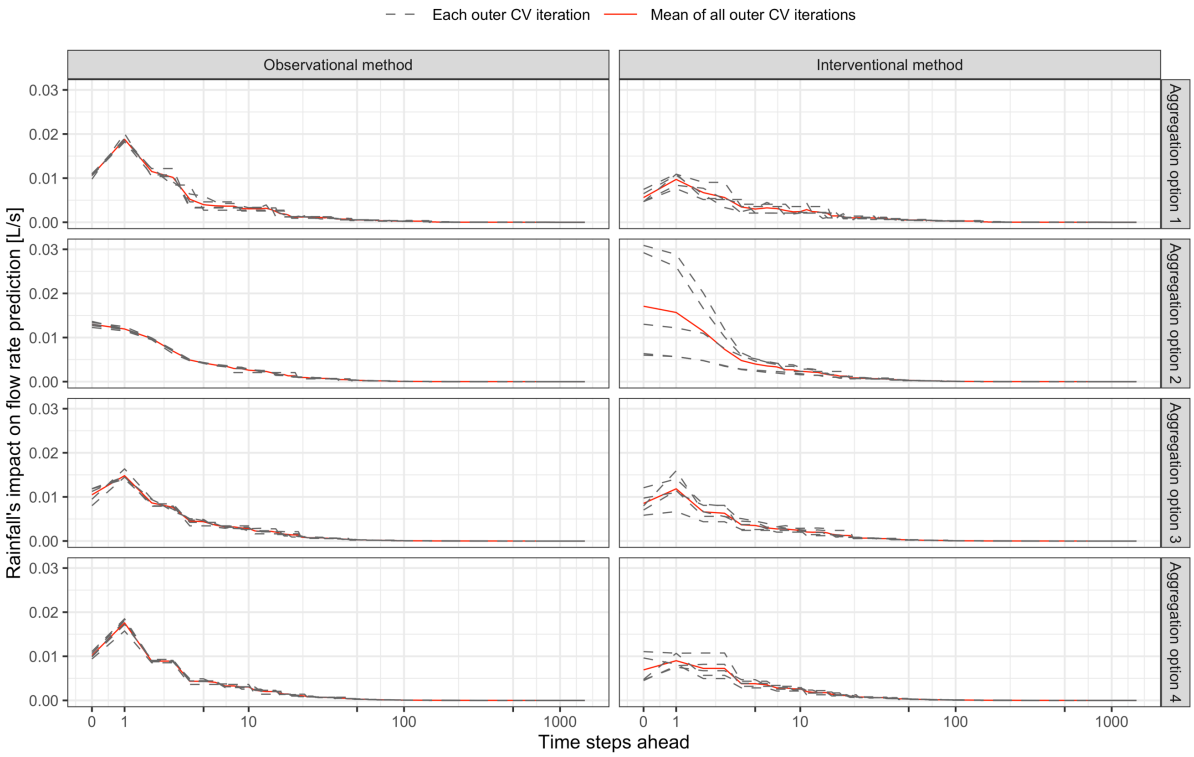

Figure 4. Rainfall's contribution to runoff prediction at different time steps ahead. Each subfigure shows the results obtained in different outer cross-validation (CV) iterations and the mean values derived using a feature engineering method and a SHAP computation method.

4. Point 1:Whether the constructed data feature mining algorithm corresponds to the reference standard in the folded data part?

In the updated manuscript, we no longer discuss the specific values of feature engineering hyperparameters due to their indirect connections to the hydrological processes. We used a Bayesian optimization method to optimize the hyperparameters automatically.

5. Point 2: "The framework is particularly useful for urban catchments where the information for setting up process-based models is insufficient." Is this statement reasonable? Do similar expressions still exist in the full text?

Thank you for raising this question. We think further clarification is needed in the updated manuscript. We claimed that "the framework is particularly useful when information for setting up process-based models is insufficient". Here, the information refers to the knowledge about the physical properties and the physical processes of the system being modelled. This point is demonstrated using two case studies. The physical properties of the drainage systems in the first case study were unknown, and it was also difficult to represent the unknown leakage from

SuDS using process-based models. The second case study site contains a large-scale and complex drainage network, requiring many parameters to characterize their physical properties in process-based models. Machine learning models were built relatively easily in the two case studies and showed relatively high prediction accuracies. However, we argue that process-based hydrological models are useful for examining the involved hydrological processes.

Therefore, machine learning methods are useful for modelling the statistical correlations between interested random variables of the catchment when observation data of the variables are available, and the trained machine learning models can serve as a baseline for evaluating process-based models. We modified the original statement to match this conclusion in the updated manuscript.

6. Point 3:Adding quantitative analysis to the conclusion section should be more convincing.

In the updated manuscript, we added the performance metrics of the trained machine learning models in the abstract and the conclusion section.

7. Point 4:Compared with the commonly used urban rainfall runoff models, what are the obvious advantages of this model?

The advantage of using machine learning methods is that they only require observation data of the interested random variables and do not require the involved physical processes to be characterized. Machine learning models can generally be set up easily and can potentially provide high-quality predictions. Machine learning models may be used as baseline models for evaluation of process-based models. More explanations are provided in our response to comment #5.

8. Line 620-780: It is difficult for finding the references because of improperly format.

Thank you for catching the errors. We updated the citation styles throughout the manuscript to meet HESS requirement.

9. Line 9: How do you define the "fine temporal scales"? It is an important concept in your forecasting, but it is not clear.

"Fine temporal scales" refers to sub-hourly scales. In the updated manuscript, the term "sub-hourly scales" is used as it is more specific.

10. Line 131: Why you use Dt-a,t-b for aggregating rainfall depth?

The lower-dimensional rainfall depth features $D_{t-a,t-b}$ were used because the dimension of the original rainfall time series can be very high (e.g., 1,000 time steps), and some machine learning methods have difficulties to learn the high-dimensional correlation between the input and the output random variables. Thus, we designed a feature engineering method to lower the dimension of the input variables of the machine learning models, and the number of features used is controlled by three hyperparameters. In fact, the feature engineering method allows the rainfall depth features to be very similar to the original time series. And the values of the hyperparameters are chosen according to the prediction accuracy of the resulted machine learning models. This point is explained in the updated manuscript.

10. In Line 84 said many observation data became available, but why only the rainfall data? Do you have other data?

We only have rainfall and runoff data for both study sites, as reported in our response to comment #3. The sentence "more observation data became available" was referring to the fact that the rainfall and runoff are being monitored in more SuDS sites globally. However, as pointed out by Schaffitel et al. (2020), monitoring data of other variables concerning urban hydrology are still currently lacking. Therefore, it can be useful to present a study that focuses only on the correlation between rainfall and runoff time series. More information on this issue is also presented in our response to comment #3. We commented on this issue in the updated manuscript.

11. Line 6-14 and Line 560-595: In the section of abstract and conclusion, the quantitative results are absent and the qualitative descriptions are not enough.

Quantitative results are presented in the updated manuscript.

Reference

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, J. Hydrol., 249(1–4), 11–29, https://doi.org/10.1016/S0022-1694(01)00421-8, 2001.

Janzing, D., Minorics, L. and Blöbaum, P.: Feature relevance quantification in explainable AI: A causal problem, https://arxiv.org/abs/1910.13413, 2019.

Schaffitel, A., Schuetz, T. and Weiler, M.: A distributed soil moisture, temperature and infiltrometer dataset for permeable pavements and green spaces, Earth Syst. Sci. Data, 12(1), 501–517, https://doi.org/10.5194/essd-12-501-2020, 2020.

Schmidt, L., Heße, F., Attinger, S. and Kumar, R.: Challenges in Applying Machine Learning Models for Hydrological Inference: A Case Study for Flooding Events Across Germany, Water Resour. Res., 56(5), https://doi.org/10.1029/2019WR025924, 2020.

Snoek, J., Larochelle, H. and Adams, R. P.: Practical Bayesian optimization of machine learning algorithms, in Advances in Neural Information Processing Systems, vol. 4, pp. 2951–2959, https://arxiv.org/abs/1206.2944v2, 2012.