Authors' response to Anonymous Referee #1

Italicized text: comments made by Referee #1.
Blue text: Authors' responses. The line numbers mentioned below correspond to those in the revised version of the manuscript.

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. Almost the whole paper has been rewritten. The details are as follows.

(a) Simplification of the methods. We updated the feature engineering method and the hyperparameter optimization methods 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 a rainfall depth feature. 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 (Eq. 5 to Eq. 7) in the original submission were removed. More details of the updated feature engineering method are provided in Section 2.2.2, lines 214 to 236.

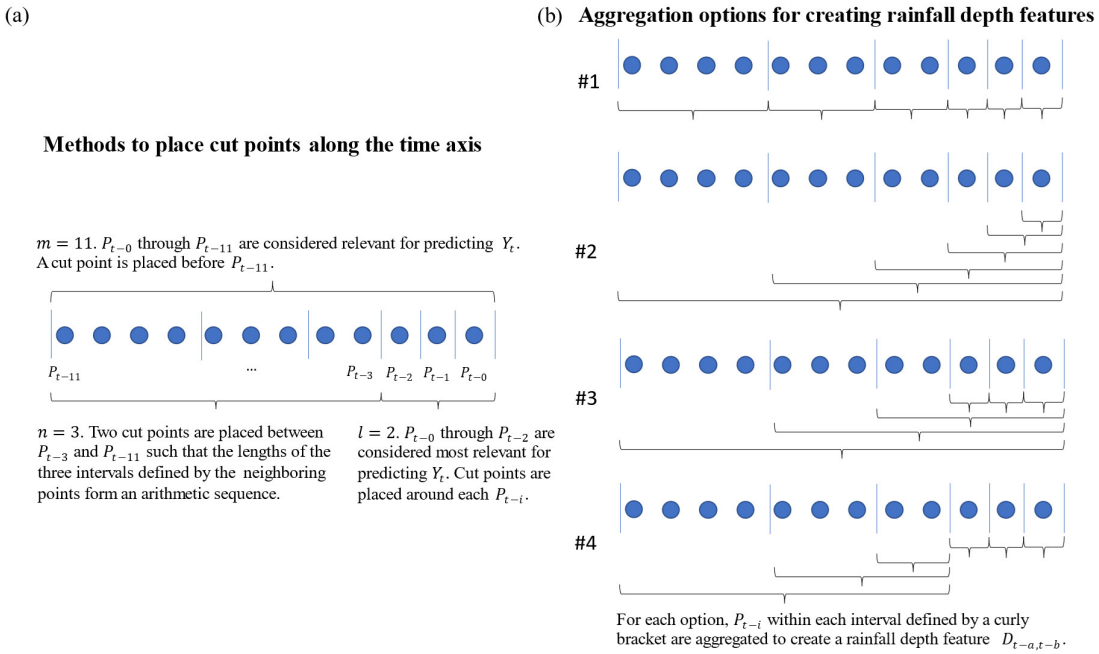

Figure 1. (a) Illustration of the methods to place cut points along the time axis. (b) Illustration of the four aggregation options for creating rainfall depth features after the cut points are selected.

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.

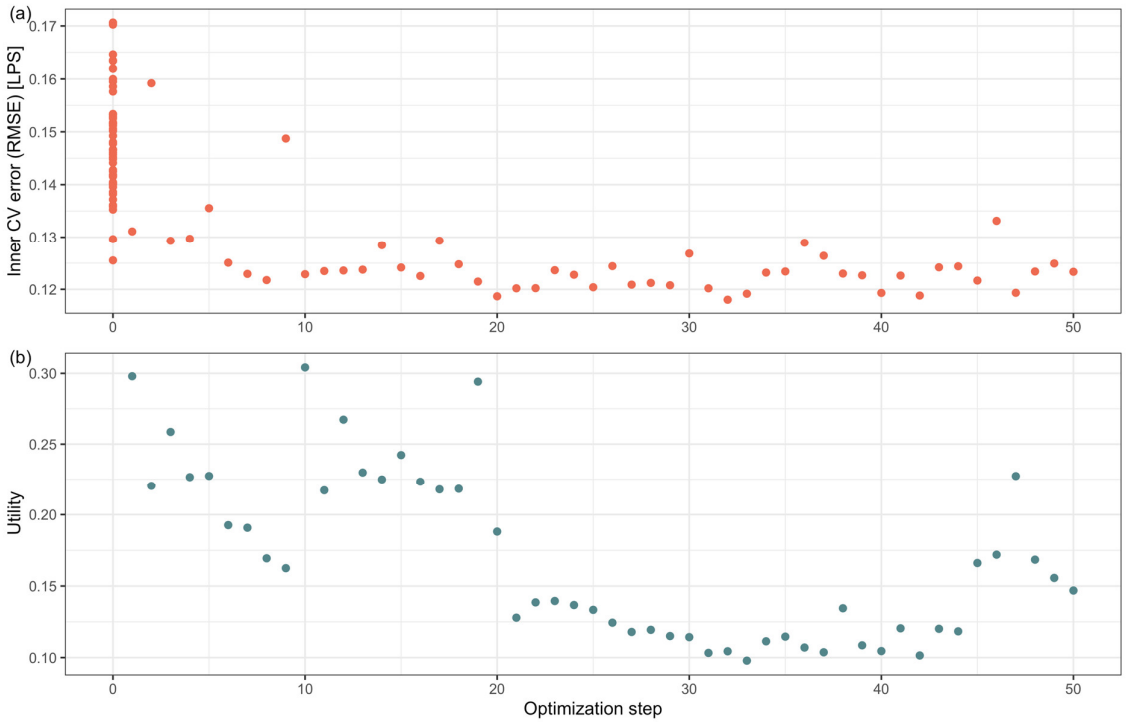

Figure 2. (a) The model's prediction accuracy associated with the hyperparameters evaluated during each optimization step. The prediction accuracy is measured by the root-mean-square error (RMSE) of the predictions obtained during the inner cross-validation (CV) iterations. (b) The expected utility of 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 (this point is mentioned in lines 334 to 344 in the updated manuscript). Due to the adoption of the updated methods, the descriptions on the resampling scheme, the choice of feature engineering and XGBoost hyperparameters, and the model selection methods were shortened. The models derived using the updated and the old methods were found to have comparable prediction accuracies to those found in the original submission.

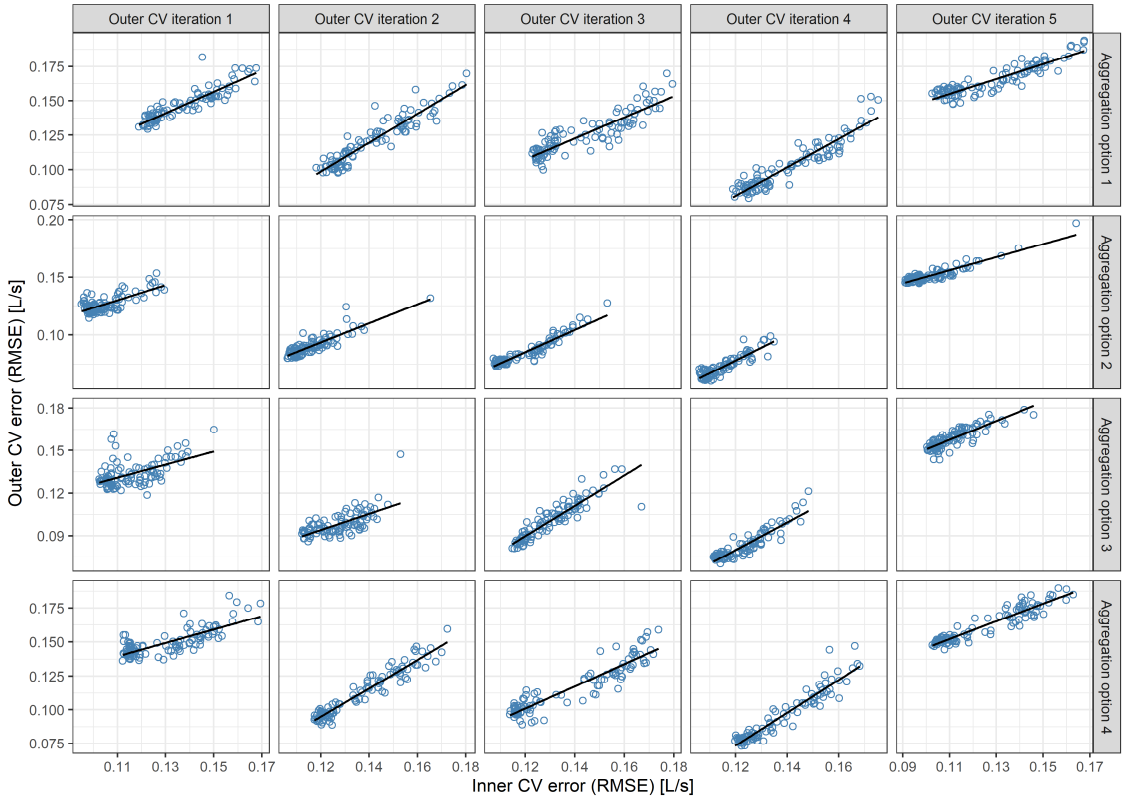

Figure 3. The model's prediction errors estimated during the inner and the outer cross-validation (CV) iterations for each set of candidate hyperparameters evaluated during the optimization. The prediction errors are measured by root-mean-square error (RMSE). Each subfigure shows the result obtained 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 of the original submission) were removed due to the updated methods and the new research objectives. The results on interpreting the feature engineering hyperparameters (Section 3.3.1 of the original submission) 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 of the original submission) 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 new objective is "to evaluate the usefulness of explainable machine learning methods in modeling and interpreting the hydrological responses of SuDS to rainfall at sub-hourly time scales". Therefore, we removed the "demonstration" element from the original submission and focused on developing methods for modelling SuDS and interpreting the model predictions and presenting findings regarding the accuracy and the interpretability

of the machine learning models. 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 applications of interpretation methods in hydrology is presented (lines 94 to 109).

In the methods section, we extended the introduction to methods for interpreting machine learning models. The differences between the local and the global interpretation methods (such as the commonly used gain and cover metrics for XGBoost) were first discussed, and the two methods (the observational and the conditional expectation methods) to compute the SHAP values and their implications were introduced. The methods to assign the importance and the contribution scores of the rainfall of each time step to runoff prediction were then introduced. A method to combine the explanations of multiple input samples was subsequently introduced. The other commonly used global interpretation methods were then presented. The descriptions of machine learning model training methods, including nested-cross validation procedure and hyperparameter optimization methods, were substantially shortened. 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. Results on the prediction accuracies of the models were first presented (Section 3.1). Regarding the application of the interpretation methods, the results on interpreting the global model structures were first presented, where a comparison between SHAP and other global interpretation methods was also added, as suggested by Referee #2 (Section 3.2). This results on examining the models' structures for predicting discharges of different magnitudes is then presented (Section 3.3). Finally, a few applications using the local explanations (i.e., SHAP values for individual input samples) were presented, including hydrograph decomposition, and determination of the source of predicted discharge (Section 3.4). Section 3.2 through Section 3.4 were organized according to the aggregation levels of the explanations, from global levels (all samples) to multiple samples levels, then to region levels (each sample).

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 study, the contribution of rainfall of each time step to each discharge prediction is

computed. In the updated manuscript, we added a formal introduction to the feature attribution problem and their methods (Section 2.1.2). In Section 3.2, when presenting the comparison between the results produced by the observational and the interventional SHAP, we commented that,

"The contributions resulting from the observational SHAP values can be interpreted as the expected difference in the predicted discharge when a particular rainfall is observed/not observed, accounting for the presence/absence of the other rainfall measurements." (lines 547 to 549)

"The contributions that result from the interventional SHAP values are the expected prediction changes when rainfall is set to a specific value, accounting for the presence/absence of other rainfall measurements." (lines 550 to 552)

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 rainfalls of each time step to each runoff prediction.

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 the 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 in lines 542 to 545.

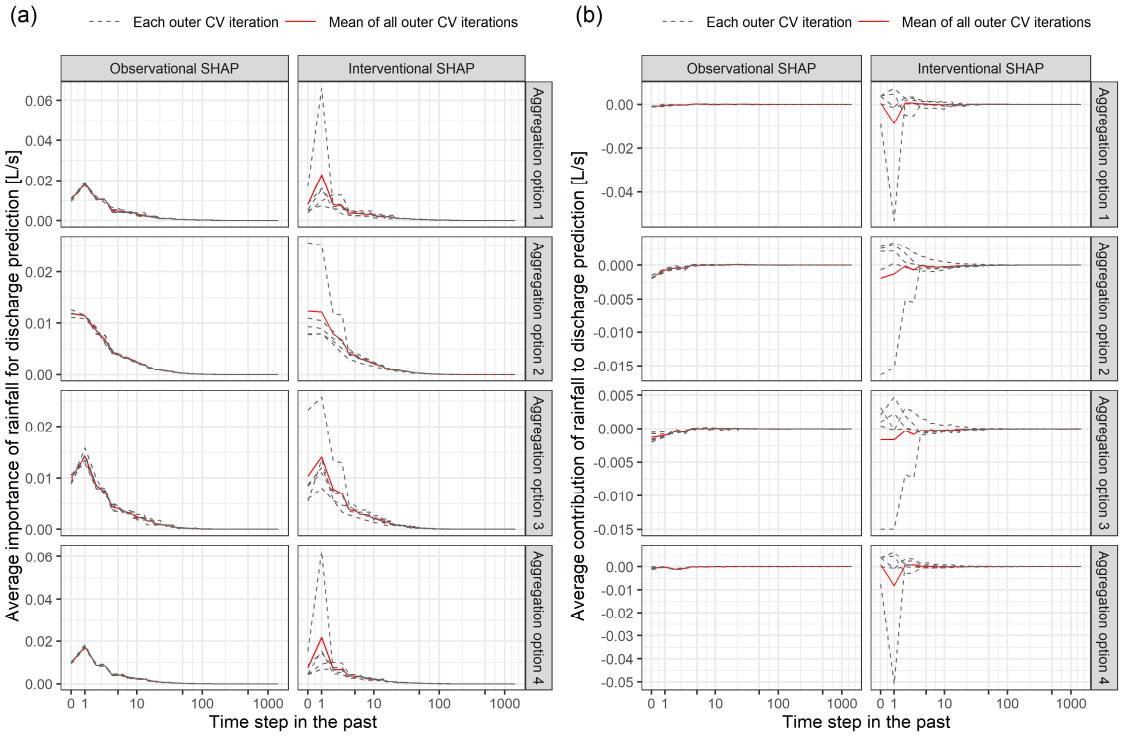

Figure 4. The average (a) importance and (b) contribution of the rainfall of different time step in the past for discharge predictions in different models. Each subfigure shows the results obtained in different outer cross-validation (CV) iterations and the mean values derived using a feature engineering method and a SHAP computation method.

In this study, we showed that the interpretation method can be useful tools for model diagnosis, for instance, the models built for SHC were found to be inadequate, as the weights they assigned to the rainfall depths at each time step during prediction were not realistic (lines 562

to 565). The interpretation methods were also used to solve some hydrological tasks, such as hydrograph decomposition. We do not consider the hydrological processes inferred using the interpretation methods are true, we commented on their shortcomings,

"However, this hydrograph decomposition method has some limitations. First, negative contributions are assigned to the rainfalls, which are difficult to interpret. Second, there is a constant bias term in the hydrograph, which does not correspond to any element in the commonly used conceptual rainfall–runoff models. Third, a model might use features that are not derived from rainfall (e.g., temperature) as a predictor, which will also be assigned with contributions to the runoff predictions when SHAP methods are used. It is unclear how to represent the contributions of factors other than rainfall when decomposing a hydrograph. Nevertheless, the hydrograph decomposition results shown in Fig. 9 are still useful for understanding why a given prediction is made by the model and to what extent the rainfall of each time step contributes to the model runoff prediction." (lines 638 to 645)

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. The updated optimization methods are described in lines 334 to 344.

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,

"It was difficult to build process-based hydrological models for the two SuDS catchments examined in this study due to insufficient information regarding the physical properties and drainage processes. This study designs a simple feature engineering method to facilitate the application of the commonly used machine learning algorithms to model rainfall–runoff correlations at sub-hourly time scales". (lines 688 to 691)

As for the "insufficient information", we mean that the physical properties of the drainage systems in the first case study were unknown for the first site, 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. We mention these points in the updated manuscript in lines 389 to 401. A table was added to the updated manuscript to compare the characteristics of the selected study sites (Table 1). 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. We pointed out the shortcomings and the uncertainties related to the data-driven approach (see our responses to comment #4).

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 removed the original statement 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.

"The resulting models have high prediction accuracies (the Nash–Sutcliffe model efficiency coefficient (NSE) > 0.70 and the coefficient of determination (R^2) > 0.70 for all models)." (lines 17 to 18)

"The proposed methods are applied to two SuDS catchments of different sizes, SuDS practice types, and data availabilities to predict discharge and produce models with good prediction accuracies (NSE $>$ 0.7 and $R^2 >$ 0.7 for all models)." (lines 670 to 672)

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 the evaluation of process-based models. More explanations are provided in our response to comment #5. In the Section 2.3.1, we discussed why machine learning models are used,

"Process-based models can be difficult to develop for both sites. In WS, the physical properties and exact design of the different drainage system elements are not precisely known (Darner et al., 2015). The rain garden is also not isolated from the gravel storage layer of the porous pavements, which permits an unknown amount of stormwater from the rain garden into the underdrain system of the porous pavement. However, commonly used process-based models are mostly designed to model SuDS with standard designs and may not be directly applicable to WS. In the SHC, the main challenge lies in the heavy workload and uncertainties in estimating the model parameters that characterize the complex drainage system. The SHC can be divided into multiple subcatchments connected by the drainage network, and a number of parameters must be determined for each catchment. In particular, the portions of impervious area that are directly and indirectly connected to the drainage network must be specified to accurately represent the flow paths of each subcatchment." (lines 389 to 397)

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 time scales". In the updated manuscript, the term "sub-hourly time scales" was 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.

"The hydrological responses of SuDS can be affected by relatively long-term hydrometeorological conditions that occurred in the past. $\boldsymbol{X_t}$ can thus be a long time series of hydrometeorological condition measurements. As pointed out by Nielsen (2019), many machine learning algorithms are not designed for modeling time series data. Therefore, long time series are often converted into lower-dimensional features that are then used as the input variables of machine learning models. The input variable transformation process is known as feature engineering and is expected to produce higher-quality models (Kuhn and Johnson, 2019)." (lines 197 to 202)

"Representing a rainfall time series using a set of $D_{t-a,t-b}$ can reduce the data dimension at the cost of losing information regarding the temporal rainfall distribution. Note that fewer cut points are selected for rainfalls in the long-term past (e.g., a few days), implying that they play less important roles in predicting Y_t. This is reasonable considering the relatively fast response time of SuDS (DeBusk et al., 2011). Gauch et al. (2020) also showed that the hydrometeorological time series recorded in the long-term past can be represented using a coarser temporal resolution in machine learning models built for rainfall–runoff modeling without deteriorating their prediction accuracy. In the proposed method, the three hyperparameters and aggregation options control the aggregation level and approach by which rainfall data recorded at different time steps are aggregated." (lines 229 to 236)

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.

"There is a limited amount of publicly available data concerning the hydrological processes of SuDS because they represent relatively new technologies and monitoring is often conducted by the local authorities and other interested parties. A lack of data is also common for other data types that are useful in urban hydrological studies, such as the soil moisture content and soil temperature (Schaffitel et al., 2020). It may be therefore useful to demonstrate the applications of machine learning methods to predict SuDS discharge based on preceding rainfalls because rainfall and discharge are the most commonly monitored features in SuDS

sites and several rainfall–discharge datasets are available online (e.g., the United States Geological Survey Water Data for the Nation, https://waterdata.usgs.gov).” (lines 110 to 116)

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. See our response to comment #6.

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.

Authors' response to Referee #2–Dr. Georgia A Papacharalampous

Italicized text: comments made by Dr. Georgia A Papacharalampous (Referee #2).
Blue text: Authors' responses. The line numbers mentioned below correspond to those in the revised version of the manuscript.

***Summary**: The paper focuses on the predictive modeling of sustainable drainage systems (SuDS) at fine temporal scales using boosting (Friedman 2001). Several boosting variants are formed and exploited in two case studies, while comparisons with the linear regression algorithm and the Storm Water Management Model (SWMM; Rossman 2015) are also provided. Furthermore, the SHapley Additive exPlanations (SHAP) method (Lundberg and Lee 2017) is used to explain the contribution of each variable (else referred to as "feature") to the issued predictions, thereby facilitating interpretability to some extent.*

Thank you for providing a nice summary of our research.

***General comments:** In general, I find that the manuscript is well-formulated and -written, and I think that the work done so far (including the release of the R codes at GitHub) should be appreciated. Nonetheless, I also think that there is some room for improvement before publication.*

I recommend minor revisions. My comments are given right below.

Thank you very much for the positive assessment. We have revised the manuscripts according to your comments. In particular, we added an introduction to machine learning model interpretation methods in the introduction sections (lines 94 to 109), which can help the readers understand the findings of this study better. We also added a comparison between the proposed SHAP-based interpretation method and the commonly used feature importance assessment methods (lines 279 to 295, and lines 575 to 594). The source code provided on GitHub has also been updated, where more explanations were provided. The proposed methods have been tested on several additional case studies, the source code and the results can be found on GitHub.

In addition, in response to the comments raised by Anonymous Referee #1, we have updated the methods and substantially changed the paper structure. (a) The feature engineering and hyperparameter optimization methods have been simplified. (b) The non-essential findings on the model training processes have been removed. (c) The structure of the paper has been reorganized, focusing more on explaining machine learning models. These modifications did not change the overall content and the conclusion of the paper. Detailed information can be found in Authors' response to Referee #1.

Comments:

(1) To my view, the following clarification is required: Which are the similarities and differences between basic variable importance measures (available in the xgboost R package) and the SHAP methodology (available in the SHAPforxgboost R package)?

In the updated manuscript, we added an introduction to the methods for interpreting machine learning models in Section 2.1. We showed that SHAP is a local feature attribution method, and the other commonly used feature importance measures (such as gain, cover, and frequency) are global interpretation methods. In addition, we showed that local explanations derived using SHAP can be combined to understand the global structures of the model in Section 2.2.3 to Section 2.2.4. And in Section 2.2.5, the other global feature importance measures are introduced.

The main differences between SHAP and the other feature importance measures are as follow:

(a) SHAP is a model-agnostic interpretation method, and the other importance measures provided by the "xgboost::xgb.importance" function in R package are model-specific (Chen et al., 2020). The advantage of model-agnostic methods is that they can be applied to various machine learning models and thereby allow comparisons between different types of models in terms of the derived interpretations (Ribeiro et al., 2016).

(b) SHAP is a local interpretation method while the other methods provided by the "xgboost::xgb.importance" function are global interpretation methods. The local interpretation methods are designed for interpreting the prediction made for individual input samples, and the global methods are independent of the input samples and often explain the structure of the model (Lundberg and Lee, 2016). Therefore, in this study, SHAP can be used to analyse a specific runoff prediction, and the other methods cannot be used for this task.

(c) Theoretically, SHAP is the only method that provides interpretations that satisfy a series of desired properties, such as local accuracy, missingness, and consistency (Lundberg et al., 2020).

(2) Since interpretability is one of the main themes of the present work, I feel that a comparison (direct or indirect, depending on the answer to comment #1) between basic variable importance measures and the SHAP methodology is currently missing from the manuscript and should be necessarily made for both case studies. New computations are needed for this comment to be fully addressed (independently of the answer to comment #1); however, these computations will only require the xgboost R package (which is already used in the paper).

Thank you for your suggestion. In the updated manuscript, we added a comparison between the feature importance derived using different interpretation methods (lines 575 to 589 and Figure 7).

The original submission, however, did not present the computation steps for deriving global feature importance measures from local explanations computed using SHAP. In the updated manuscript, we formally defined the methods to compute the contribution of rainfall at each time step to runoff predictions (Section 2.2.3, lines 238 to 254) and the method to combine local explanations to understand model structures and rainfall-runoff correlations (Section 2.2.4, lines 262 to 278). We believe the addition of the formal definitions can improve the readability of the paper.

In Figure 1, the importance of the rainfall of each time to discharge prediction obtained using various feature importance measures are compared. The results obtained using various methods are generally similar, confirming the validity of the opposed SHAP-based method. More discussions on this result can be found in lines 575 to 589.

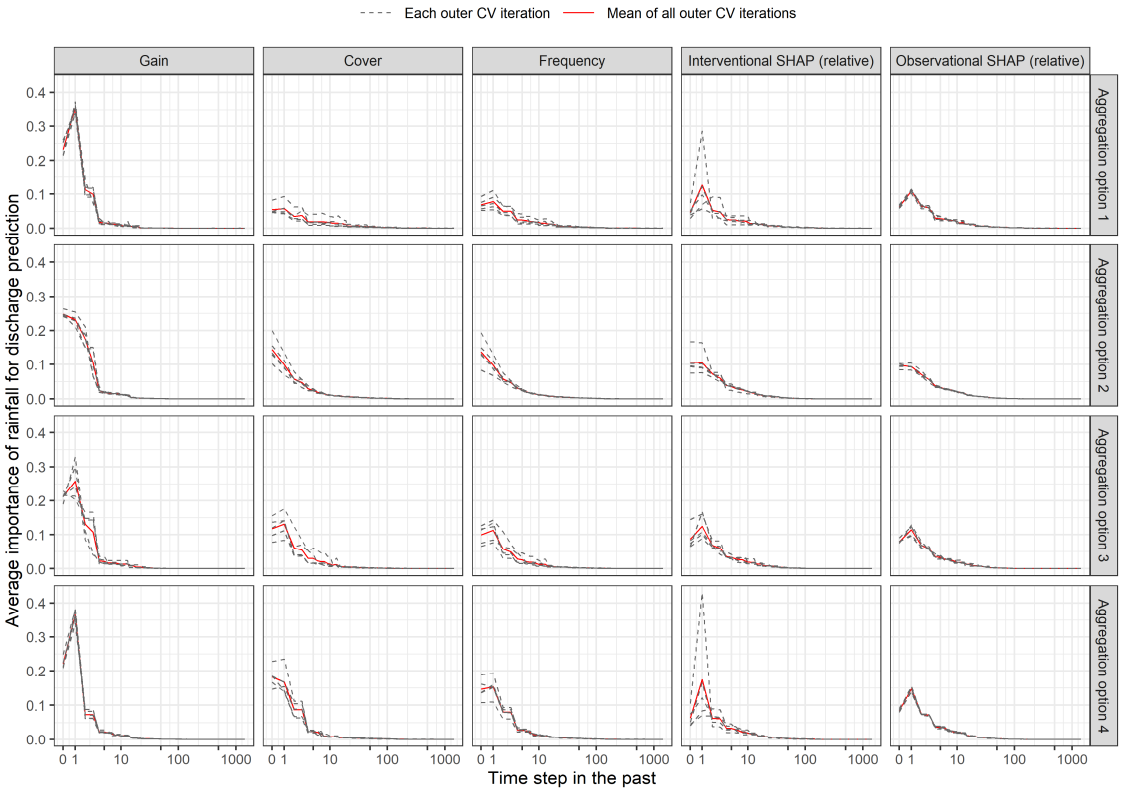

Figure 1. The importance of rainfall of each time to runoff prediction. Each subfigure shows the results obtained in different outer cross-validation (CV) iterations and the mean values derived using a feature engineering method and an interpretation method.

Following the recent discussions on the "correct" methods to compute SHAP values in Chen et al. (2020) and Janzing et al. (2019), we used both the observational and the interventional methods to compute the SHAP values. The results obtained using the two methods, as shown in Figure 2, are overall similar. The implications and reasons to use each method are explained in the updated manuscript. The Python package "shap" was used in the computation, as it offers both methods (Lundberg et al., 2017). The two methods are introduced in lines 150 to 158, and more descriptions of the results can be found in lines 546 to 556 of the updated manuscript.

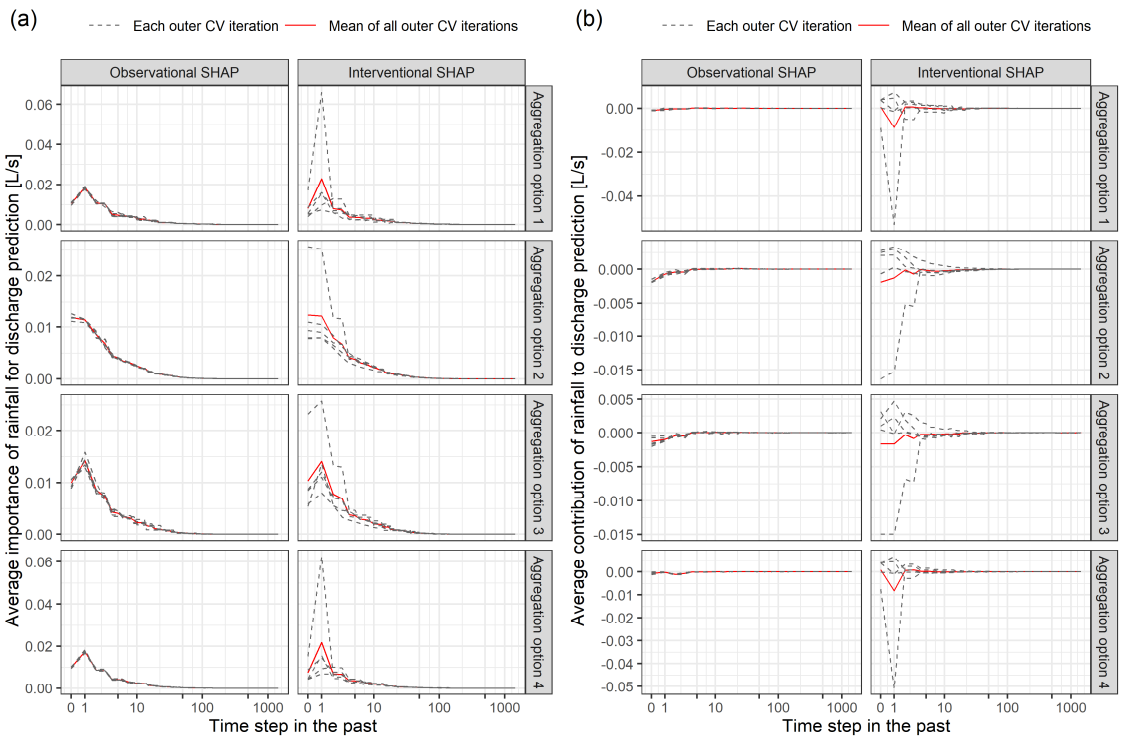

Figure 2. The average (a) importance and (b) contribution of the rainfall of different time step in the past for discharge predictions in different model. Each subfigure shows the results obtained in different outer cross-validation (CV) iterations and the mean values derived using a feature engineering method and a SHAP computation method.

(3) In light of comments #1 and #2, other hydrological studies using boosting or random forests while also emphasizing on interpretability (by using variable importance measures) could be discussed (in comparison to the present study) somewhere in the manuscript. What is the added value of the present work with respect to such existing works?

The main contribution of this study is as follows. (a) It presents the applications of model-agnostic local interpretation methods for interpreting individual predictions of rainfall-runoff models, whereas existing studies mostly use model-dependent global interpretation methods. (b) This study proposes a feature engineering and model training method to automatically find

the optimal lower-dimensional representations of high-dimensional input time series for machine learning models. (c) This study shows that machine learning methods can be effective for modelling the rainfall-runoff correlations of SuDS. (d) It defines various methods to use the local explanations derived by the SHAP method for model diagnosis and analysing the hydrological processes being modelled. We discussed the contributions of this paper in the conclusion section, lines 688 to 707.

(4) In the "Introduction" section, it is written that "only a few studies adopted machine learning methods to investigate the hydrological processes of SuDS", with the studies by Eric et al. (2015), Khan et al. (2013), Li et al. (2019), and Yang and Chui (2019) being discussed as examples of such studies. Since such studies are quite close to the present work, more of them could be reported (provided that they exist).

Currently, there are very few studies applying machine learning methods to study the hydrological performances of sustainable urban drainage systems (SuDS). We explained the reasons behind the lack of popularity (lines 72 to 116), and listed all the literature we can find in the updated manuscript, e.g., Hopkins et al. (2020) was added. More details of these studies are also presented in the updated manuscript. Additionally, we clarified our contributions more clearly (see our response to comment #3).

(5) In the same section, it is also written that "modeling the responses of SuDS at fine temporal scales requires high-dimensional hydrometeorological time series to be used as input, which is difficult in machine learning". Could this sentence be further elaborated? I would say that the opposite holds, i.e., that machine learning methods are ideal for handling high-dimensional hydrometeorological time series.

Our original statement was inaccurate. There are some machine learning methods that are very efficient in handling high-dimensional data, such as deep learning methods. However, as discussed in Nielsen (2019), high-dimensional time series data are usually converted to lower-dimensional features before feeding to a machine learning model, unless the model is specifically designed to model sequence data. We were also not sure if the XGBoost method works well with high-dimensional time series. Therefore, we designed a very flexible feature engineering method, that certain hyperparameter values would generate rainfall depth features that are very close to the original rainfall time series. The final features chosen were those that corresponded to the highest prediction accuracy, and they were generally in low dimensions. In the updated manuscript, we updated this statement by saying that modelling high dimensional data can be challenging for some machine learning methods.

"Thus, modeling the responses of SuDS at fine temporal scales requires a high-dimensional hydrometeorological time series to be used as input, which is difficult for machine learning algorithms that are not specifically designed for modeling data sequences (Nielsen, 2019)." (lines 88 to 91)

And in the conclusion section, we suggested future studies to explore the usefulness of machine learning methods that are specifically designed for high dimensional sequence data, such as LSTM networks in deep learning.

"However, feature engineering methods can be designed arbitrarily and it can be computationally expensive to identify the optimal methods. Future studies can thus explore the application of machine learning methods that are specifically designed for modeling high-dimensional time series data, such as the long short-term memory (LSTM) networks in deep learning." (lines 709 to 712)

(6) The reader could also be referred to several specialized books (e.g., Hastie et al. 2009; James et al. 2013; Witten et al. 2007), for further information on the machine learning (or statistical learning) methods used in the paper.

Thank you for your suggestion. In the updated manuscript, we provided a list of suggested references for the machine learning methods and the model interpretation methods. For instance,

"More information on the various interpretation methods can be found in Molnar (2021)." (line 132)

"More information on the Shapley value can be found in Osborne and Rubinstein (1994)." (lines 145 to 146)

"More information on the resampling techniques for testing machine learning models can be found in Kuhn and Johnson (2013) and Hastie et al. (2009)." (lines 354 to 355)

(7) Another concern of mine is related to the small number of real-world cases examined in the paper. I think that the application of the proposed procedures to large real-word datasets (comprising hundreds of cases) should be addressed at least with extensive relevant discussions in the manuscript (e.g., future research recommendations). (Currently, it is only suggested using "the SHAP method in more case studies"). To my view, these extensive discussions are important, especially given that (i) there are studies in the hydro-meteorological literature validating their models using big datasets, and (ii) the first aim of

the paper is to "evaluate the usefulness of machine learning methods in predicting the hydrological responses of SuDS at fine temporal scales". The necessity of evaluating machine learning methods using big datasets is extensively discussed by Boulesteix et al. (2018).

We agree that the proposed method should be thoroughly tested on different datasets to prove its usefulness. However, to the knowledge of the authors, there are no publicly available regional or global datasets on the rainfall and runoff time series of SuDS.

We have the rainfall-runoff data of SuDS for a few sites in the U.S., and the proposed methods were found to be effective for these sites. The two sites, WS and SHC, were chosen to be reported in the manuscript for the following reasons. (a) The two sites are in very different scales: WS is about 1,000 m^2, and SHC is about 1 km^2. We intended to show our methods are useful for catchments of various scales. (b) A few years of data were available for WS, and only two months of runoff data were available for SHC. We aimed to show our methods can be useful even when the data is not abundant. (c) The two sites faced different difficulties in setting up process-based models: the physical properties of the SuDS were unknown in WS, and the drainage system of SHC was very complex and the characterization of which requires thousands of parameters. We believe these difficulties are common in practice, and thus we presented the proposed methods as potential solutions to the common problems. The reasons and implications for choosing the two sites were listed in Table 1 in the updated manuscript.

To address this comment, we suggest the proposed methods be tested on more SuDS sites in the conclusion section of the paper.

"Second, the methods proposed in this study are only applied to model the correlations between rainfall time series and discharge in a few U.S. sites. The proposed methods should therefore be tested in more sites worldwide to model the correlations between other variables, although this may require the development of new feature engineering methods." (lines 712 to 715)

Finally, we added results of more SuDS cases studies as the demonstration applications in the documentation of the source code on GitHub.

(8) In the "Conclusions" section it is written that "the proposed model training methods are semi-automatic, requiring minimal user input". It would be useful to discuss (somewhere in the paper) which parts of the proposed methods are not (fully) automatic, and how one could overcome this limitation to allow large-scale (even global-scale) investigations (see also comment #7).

In response to this comment and the comments made by Anonymous Referee #1, we used Bayesian optimization algorithms to automatically find the optimal features and hyperparameters for training machine learning models (Snoek et al., 2012). This eliminates the need to select a predefined set of candidate feature engineering and XGBoost hyperparameters. The updated methods thus only require the lower and upper bounds for each hyperparameter. The user can also use the default values, if she/he so desires (the method then becomes fully automatic). This change allows the method to use regional scale data, where multiple sites were analyse,

"The model training process is automatic and only requires that the range of possible hyperparameter values be defined." (line 696)

The quality of the models derived using the updated method was found to be similar or better when comparing to that derived using the old methods (some of the results are presented in our responses to comment #1 made by Referee #1).

(9) Currently, the use of the xgboost and SHAPforxgboost R packages is reported in the manuscript. To my view, all utilized software packages (which, of course, at the moment can be found online at https://github.com/stsfk/explainable_ml_hydro, since the R code has been made available) should necessarily be reported and cited in the paper.

In response to this comment, we listed the packages used in this paper in the code availability section. In addition, we updated the source code using the "tidymodels" R packages (Kuhn and Silge, 2020), the source code is now easier to understand.

(10) Finally, the manuscript is not typo-free at the moment. Particular attention should be placed on the mathematical notations. For instance, the transpose operator should not be written in italics (therefore, T should be replaced with T*) and the vectors should be bolded (therefore, Xt−m,t should be replaced with* $\boldsymbol{X}t-m,t$*).*

Thank you for catching the errors, we updated the math notations accordingly. The manuscript has been thoroughly checked. We also hired a professional English editor to correct grammar mistakes.

Reference

Chen, T., and Guestrin, C.: Xgboost: A scalable tree boosting system, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794, 2016

Chen, H., Janizek, J. D., Lundberg, S. and Lee, S. I.: True to the model or true to the data?, http://arxiv.org/abs/1805.11783, 2020.

Hopkins, K. G., Bhaskar, A. S., Woznicki, S. A. and Fanelli, R. M.: Changes in event-based streamflow magnitude and timing after suburban development with infiltration-based stormwater management, Hydrol. Process., 34(2), 387–403, https://doi.org/10.1002/hyp.13593, 2020.

Janzing, D., Minorics, L. and Blöbaum, P.: Feature relevance quantification in explainable ai: A causal problem, arXiv, 2019.

Kuhn, M. and Silge, J.: Tidy Modeling with R, Version 0.0.1.9007, https://www.tmwr.org, 2020.

Lundberg, S. and Lee, S.-I.: An unexpected unity among methods for interpreting model predictions, http://arxiv.org/abs/1611.07478, 2016.

Lundberg, S. M., Allen, P. G. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, https://github.com/slundberg/shap, 2017.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.-I.: From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell., 2(1), 56–67, https://doi.org/10.1038/s42256-019-0138-9, 2020.

Nielsen, A.: Practical Time Series Analysis, O'Reilly Media, Inc, https://www.oreilly.com/library/view/practical-time-series/9781492041641, 2019.

Ribeiro, M. T., Singh, S. and Guestrin, C.: Model-Agnostic Interpretability of Machine Learning, http://arxiv.org/abs/1606.05386, 2016.

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, 2951–2959, https://arxiv.org/abs/1206.2944v2, 2012.