We would like to thank Dr. Georgia A Papacharalampous (Referee #2) for providing insightful comments for improving our paper. Our responses to her comments are as follow.

Italicized text: comments made by Dr. Georgia A Papacharalampous (Referee #2).
Blue text: Authors' responses.

***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 for the positive assessment. We revised the manuscripts according to your comments. In particular, we added more discussions on the machine learning model interpretation methods and a comparison between the explanations derived using different interpretation methods. We also plan to improve the documentation of the source code on GitHub when submitting the revision. Please find our responses and some of the new results below.

In addition, in response to the comments provided by Anonymous Referee #1, we improved the readability of the paper by making the following changes: (a) simplification of the methods, (b) removal of non-essential findings, and (c) re-organizing the paper according to the new research objective. The modifications did not change the overall content and the conclusion of the paper. Detailed information can be found in Author Comment #1 (AC1).

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)?

The main differences between SHAP and the basic importance measures (such as gain and cover) are as follow.

(a) SHAP is a model-agnostic interpretation methods, 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).

In the updated manuscript, we added discussions regarding the differences between various model interpretation methods.

(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 feature importance derived using different interpretation methods. As shown in Figure 1, for each machine learning model, the rainfall's contribution to the runoff prediction at each time step was computed using different methods. As we explained in the response to comment #1, the gain, the cover, and the frequency are all global interpretation methods, it is not possible to

compute the importance of the rainfall for a specific input sample. Thus, each dashed line in Figure 1 corresponds to the results obtained for a model for all input samples. In general, we found that gain and SHAP offered similar explanations regarding the relative importance of rainfall to runoff predictions. We also showed that the explanations are dependent on the machine learning models and the interpretation methods. The implication is that if we use these methods to investigate the involved hydrological processes, then various explanations are plausible. Schmidt et al. (2020) suggested the many possible explanations associated with different machine learning models are similar to the equifinality phenomenon in process-based hydrological modelling. The new results are discussed in greater detail in the updated manuscript.

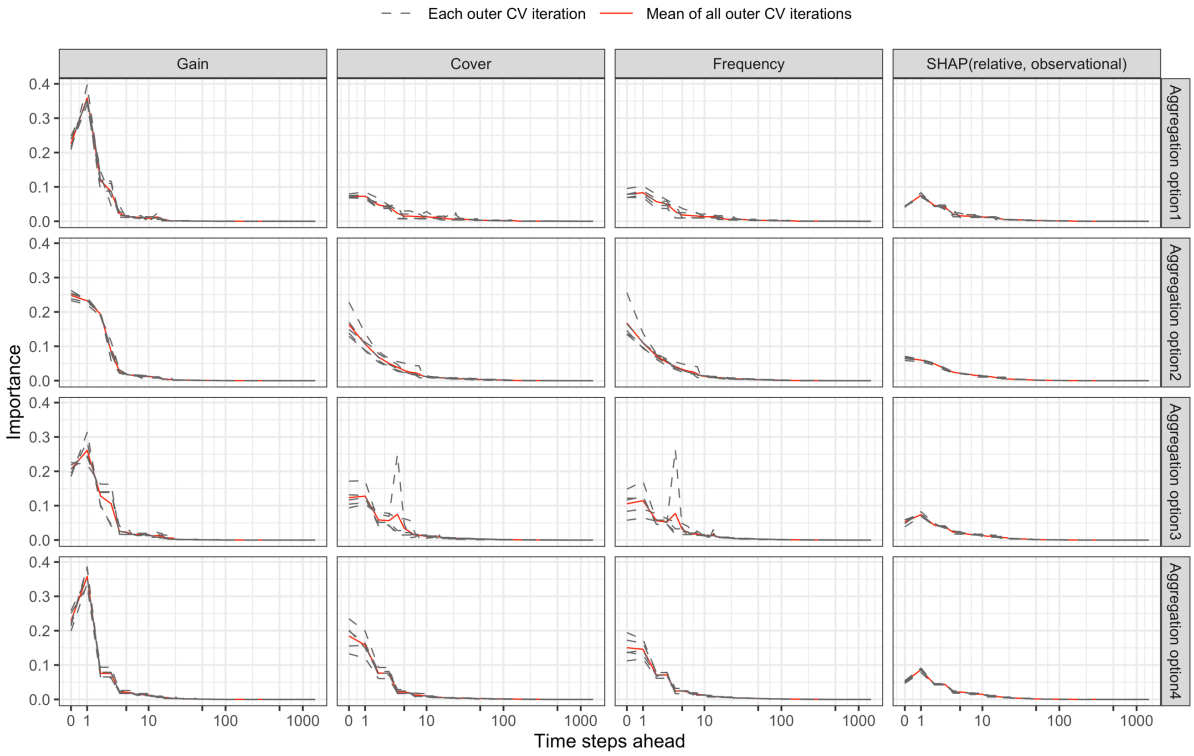

Figure 1. Rainfall's contribution to runoff prediction at different time steps ahead. Each subfigure shows the results obtained in different outer cross-validation (CV) iterations and the mean values derived using a feature engineering method and an interpretation method.

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 source code for computation will be posted on GitHub when submitting the revision.

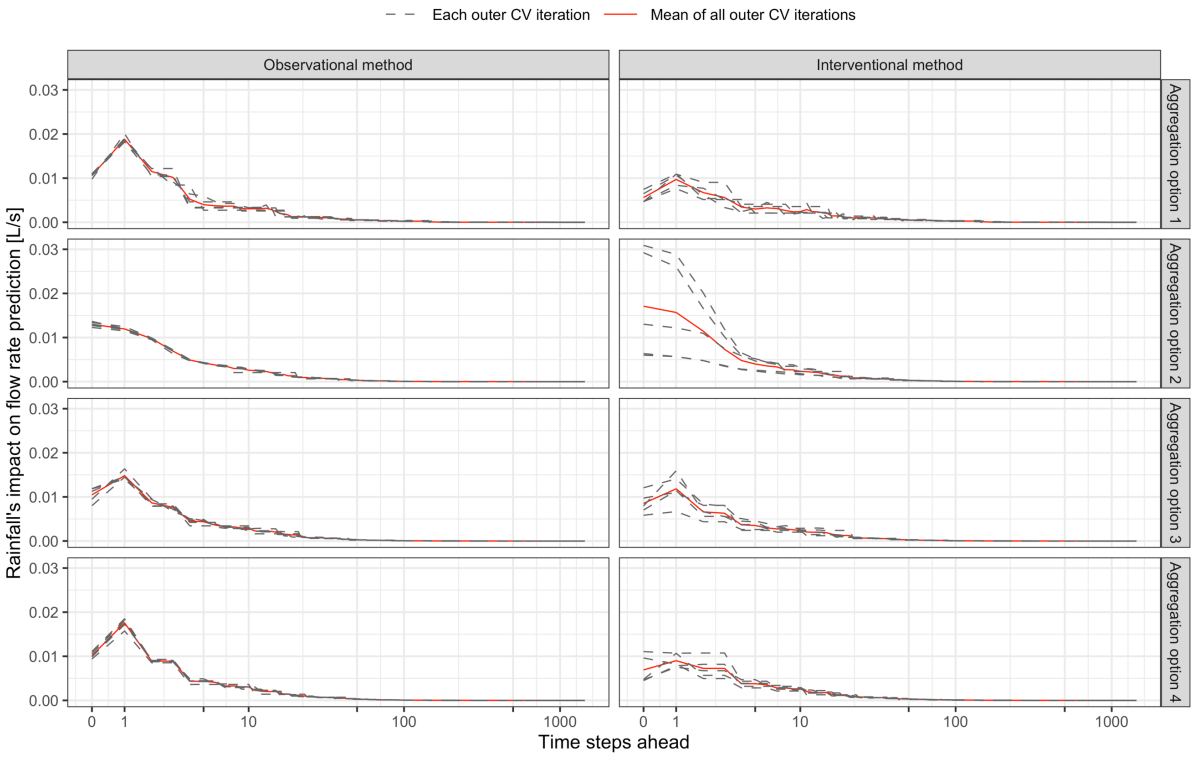

Figure 2. Rainfall's contribution to runoff prediction at different time steps ahead. Each subfigure shows the results obtained in different outer cross-validation (CV) iterations and the mean values derived using a feature engineering method and a SHAP computation method.

(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?

In the updated manuscript we added a short review of the methods applying boosting and random forests methods in hydrology.

The main contribution of this study is as follow. (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. The contribution of the rainfall at each step to a runoff prediction can be easily computed using the proposed method. (c) This study shows that machine learning methods can be effective for modelling the rainfall-runoff correlations of SuDS. It also shows that the hydrological processes inferred by interpretation methods are dependent on the

machine learning models and the interpretation methods, i.e., there are many different but likely explanations to the predictions of the same event. Relevant discussions were added in the updated manuscript.

(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, 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 corresponded to the highest prediction accuracy, and they were generally in low dimensions. In the updated methods, we updated this statement by saying that modelling high dimensional data can be challenging for some machine learning methods. 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.

(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.

(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 clarified in the updated manuscript.

To address this comment, we suggest the proposed methods to be tested on more SuDS sites in the conclusion section of the paper. We also added discussions on usefulness of the proposed method in other fields of hydrology, where regional and global data, such as the CAMELS

dataset (Newman et al., 2015), are available. We also plan to include 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 analysed. Relevant discussions are added in the updated manuscript.

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 all the packages used in modelling in the updated manuscript.

In addition, we updated the source code using the "tidymodels" R packages (Kuhn and Silge, 2020), the source code is now easier to understand. We are also updating the documentation of the source code. The code and the documentation will be posted on GitHub upon submission of the revision.

(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 will thoroughly check the manuscript when submitting the revised version and also hire a professional English editor to correct grammatical 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.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T. and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, Hydrol. Earth Syst. Sci., 19(1), 209–223, https://doi.org/10.5194/hess-19-209-2015, 2015.

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.