We thank the reviewers for their insightful comments and constructive suggestions that have led to the improvement of our paper. Our responses to all the comments and suggestions are detailed as below.

Reviewer #1:

Hu et al. present a methodological framework to improve the efficiency and effectiveness of data-driven models using a two-part approach: (1) using Sobol global sensitivity analysis for hyperparameter selection (HS) and (2) using Bayesian optimization for hyperparameter tuning (HT). This generalizable framework was demonstrated using a case study of daily EOF runoff predictions in the Maumee domain, demonstrating that a combination of HS-HT as developed in the framework is most effective in improving performance of the data-driven model in addition to reducing overfitting. The manuscript is well written and suitable for publication in GMD. I recommend publication after addressing my minor comments below.

Specific comments:
1. L25: The authors point in the introduction several barriers to the wide use of physics-based numerical models, which are accurate. For a balanced discussion, I also recommend a brief sentence discussing potential barriers to usage of data-driven machine learning models. e.g., the availability (and storage requirements) of data, and need for computational resources (i.e., GPUs).

Thanks. Following the reviewer's suggestion, we add a brief sentence to discuss the limitations of data-driven modeling.

Despite the fact that data-driven models are data-intensive and not generalizable, they do not require an explicit mathematical formulation of all underlying complex processes to perform predictive analysis.

2. L81: A set of initial values is assigned to the influential hyperparameters during HS. Is the HT phase sensitive to the choice of initial parameter values (would this affect the outcome of the HT process)? How are these initial values chosen?

Thanks. The initial values of the influential hyperparameters can impact the outcomes of hyperparameter tuning in terms of the convergence time and optimal values identified for the hyperparameters. To minimize such impacts in the study, we choose a powerful, automated optimization approach, namely Bayesian hyperparameter optimization. This approach can help avoid being trapped in the local optima by using a probabilistic model to approximate the objective function and using this model to guide the search for the optimal hyperparameters. We made some modifications to improve the clarity (See Ln 122 – 124).

3. L207-209: Following up on Reviewer #1's comments regarding imbalanced data, I would suggest elaborating the added paragraph with the general advice offered in the authors' response - namely, the choice of an effective ML algorithm, a good CV strategy, and weight the minority class in the class weights. This will be useful for future readers of the paper as guidance beyond the specific problem presented.

Thanks. We added some descriptions and references to show what steps we have taken to improve model performance against imbalanced data for the case study.

To further mitigate the impact of the imbalanced runoff data, besides the effective XGBoost algorithm, we used the Stratified K-Fold cross-validation across different scenarios to ensure the training and test datasets follow a similar distribution and defined a loss function (e.g., R-Squared) that penalizes more the missing predictions of large runoff events, that is, the minority class in this study.

4. The results section for the test case is well presented. I would only suggest, if possible, to add a comparison of the HS-HT approach's test performance with prior work (either using physics-based or data-driven models) to provide better context for the performance of the proposed framework.

Thanks. A comparison of the performance between the physics-based model and data-driven model in runoff prediction was discussed in Hu et al. (2021). As for the comparison of the performance of the data-driven models with/without using the HS-HT approach, we demonstrated the results in two steps, as alluded in Figures 5 and 6, data-driven models without using the HS approach are more likely to be overfitting (Step 1) and without using the HT approach can lead to long convergence time for the underlying ML algorithm to identify the optimum of the hyperparameters (Step 2).

Hu, Y., Fitzpatrick, L., Fry, L. M., Mason, L., Read, L. K., and Goering, D. C.: Edge-of-field runoff prediction by a hybrid modeling approach using causal inference, Environmental Research Communications, 3, 075 003, 2021.

Technical corrections:
5. L57: I suggest keeping "Category I" and "Category II" numbering consistent with the numbering above, now 1) and 2).

Thanks. We change the numbering to I and II to be consistent with “Category I” and “Category II”.

Reviewer #2

I appreciate that the authors have done a huge amount of work to modify the paper deeply. I have no further questions about the manuscript except for one language issue. I do not think it needs to be sent to me for review again, as it is beyond the line. The language issue is about the verb in the tense. For example, "choose" seemed like using the past tense "chose", Please double-check the verbs in the manuscript to keep consistency in the manuscript.

Thanks. For the methodology section, we tend to use the past tense as it is describing what has been done in the study. We went through the manuscript and made some modifications to ensure that the tenses are appropriate for each individual section.

Reviewer #3:

The manuscript presents a novel approach for automatically determining the best hyper-parameters. The writing is technically accurate, and the topic is both current and compelling. I recommend minor revisions before the manuscript publication.

1. It would be beneficial if the author could provide more details about Figure 3a. Specifically, an explanation for why the distribution of the first two hyper-parameters differs from that of the middle four and last three hyper-parameters in the diagonal of the matrix.

Thanks. We intend to sample the hyperparameters uniformly across the range of their values and use the histogram plots on the diagonal to evaluate the actual sample distribution for a given hyperparameter. Each bar in the histogram plot shows the sample size in the corresponding interval. As the differences in bar heights are small, we can thus claim that the selected samples for each hyperparameter follow a uniform distribution. We added some descriptions to clarify this (See modifications in Figure 3a caption).

2. In Figure 5a, the x-axis scale is not consistent, with a spacing of two months in the training set and three months in the test set. It is recommended that the author standardize the scale. Additionally, there are some short lines at the top of the HT result figure, and their meaning should be explained.

Thanks. We modified the time scale to ensure consistency with the training set (See Figure 5a). The short lines at the top of the HT result figure indicate the rain rate (mm/day) as shown on the right y-axis.

3. The author may consider providing a more clearly defined sub-figure in the Great Lake Region figure at Figure 2, or alternatively, removing that sub-figure.

Thanks for the suggestion. To improve the clarity, we changed the position of the subfigure that defines the Great Lakes Region as illustrated in Figure 2.