An argument for parsimony in differentiable hydrologic models

Sandeep Poudel and Scott Steinschneider

Corresponding author: Sandeep Poudel, sp2596@cornell.edu

Dear Editor,

Thank you for reviewing our work and collecting the reviewers' comments. We have prepared a revised manuscript addressing yours and both reviewers' comments. Please see the points below, where we outline the summary of changes made in the revised manuscript, and we refer to our point-by-point response to both reviewers comment in the interactive discussion section.

Major changes:

1. Following the comment from Reviewer #1, we re-trained our model on the same time period as the reference benchmark of Feng et al., 2022. Our results are consistent with those before, and we discuss this in the main manuscript and provide an additional result figure in the Supplementary.
2. Reviewer #1 also commented on the challenges in interpretability of the differentiable models, even with the MLP-based architecture. We believe this is a valid point and have substantially expanded the discussion section of the main manuscript to incorporate these challenges.
3. Both reviewers had similar comments on our alternative input experiments. We realized our original manuscript was very brief in its explanation and have therefore substantially revised it to provide a clearer and more expanded presentation of our alternative experiment results. The Editor also raised concerns about the possibility of alternative interpretations of these results. Although our ablation-based results suggest otherwise, we agree with the Editor that alternative interpretations cannot be ruled out. We therefore revise wording in our entire manuscript and also provide a discussion of this in the revised manuscript, along with a recommendation for more mechanistic analysis of how neural networks use attribute inputs to produce hydrological responses, as a direction for future work.
4. Reviewer #1 also commented that our LSTM-based model parameters were changing only at 60-day intervals and questioned whether different results might be obtained if they changed at 1-day intervals. We ran simulations with a 1-day stride so that LSTM-based parameters change at each time step, and our results are consistent. We provide figures from this experiment in the Supplementary and discuss them in the main manuscript.
5. Reviewer #2 commented that MLP-based parameterization has been used in the past, which we believe is true. However, in the revised manuscript we clarify that this was only in the case of routing models, not rainfall-runoff models. When used for rainfall-runoff models, it was not for static parameterization but for hybrid parameterization, with the motivation of reducing the computational burden of LSTM in high-resolution modeling applications.

6. Reviewer #2 also raises a point that ensemble-based models outperforming single models is not new, which is true. We therefore shift our emphasis toward the finding that the simpler MLP-based architecture matches the performance of the LSTM-based architecture as our primary contribution, rather than the ensemble performance gains. Along with that, we now also provide additional discussion on how ensemble-based models, while offering better performance, can introduce challenges to interpretability.

Minor changes:

1. Previously, we only showed model performance metrics including high flow bias, low flow bias, KGE, etc. for out-of-sample testing in time. As per Reviewer #1's comment, we now show results for both out-of-sample in time and out-of-sample in space in the main manuscript.
2. As per the suggestion of Reviewer #2, we have moved the detailed description of the integrated gradient method from the methods section to Appendix D.

Overall, these changes do not alter our conclusions, rather we believe they have meaningfully improved the quality of the manuscript. We are grateful to the Editor and the reviewers for their valuable input.

Best,

Sandeep Poudel and Scott Steinschneider