RESPONSE TO RC2 COMMENTS:

Thanks for the time and effort you invested in reviewing the manuscript, we appreciate the detailed comments you provided. In response to the six questions you raised (*repeated here in red italic text*), please find our detailed responses below in regular black text:

RC2, C1: [There is a main paragraph explaining the importance of large-scale training of LSTMs and how individual catchment LSTMs are not suitable for comparison. However, it seems you're still limiting your data to only regional data from the Great Lakes region. While this is better than training an LSTM on one basin, it is best to give the model all the data you can, including basins outside of the GL region. Have you trained your SR, or lumped, LSTM on all CONUS basins, or done any work regarding that?]

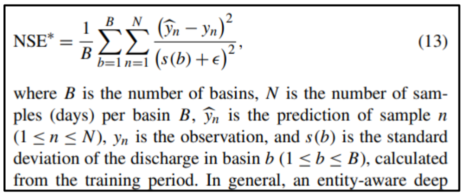
No, we have not done that. We agree with you given the findings of Kratzert et al. (2023) showing that training on more basins is always better. The primary reason we did not train a new model on CONUS basins in this study is because we wanted to show that our SR model built with an **existing** LSTM regional streamflow model (i.e., the GRIP-GL lumped LSTM), can easily be augmented with an uncalibrated hydrological routing approach to enhance streamflow prediction in larger watersheds where the lumped LSTM (with no hydrological routing) predictions are not as good. In this way, no need for researchers to train a new lumped LSTM on new dataset.

RC2, C2: [Figure 1 is confusing as the numbering doesn't align with how the graph is read. It reads like it's circular, everything is somehow mapping to itself. I suggest redoing this figure as it would be easier for the reader to understand the SR workflow.]

We will rearrange the figure in the revised manuscript so the numbering appears more natural with how one would read the graph and we will augment the caption to be more descriptive and help guide readers through this. We note that the arrows in the workflow are definitely not circular. However, data and components are used multiple times and hence multiple arrows.

RC2, C3: [What loss function was used? While it makes sense to have a lot of your information in a supplemental paper, adding this information would be helpful to readers as they would know what information is used in training your lumped LSTM.]

The loss function is the same as the loss function introduced in Kratzert et al. (2019) (see the equation in the screenshot below), which is effectively the averaged Nash–Sutcliffe efficiency (NSE) value across the calibration/training basins.



$$\mathrm{NSE}^* = \frac{1}{B}\sum_{b=1}^{B}\sum_{n=1}^{N}\frac{\left(\widehat{y}_n - y_n\right)^2}{\left(s(b)+\epsilon\right)^2}, \qquad (13)$$

where B is the number of basins, N is the number of samples (days) per basin B, $\widehat{y}_n$ is the prediction of sample n $(1 \leq n \leq N)$, y_n is the observation, and $s(b)$ is the standard deviation of the discharge in basin b $(1 \leq b \leq B)$, calculated from the training period. In general, an entity-aware deep

In response to your suggestion, and details of the LSTM configuration, such as the hyperparameters and training variables, will be added to the appendix in the revised manuscript.

RC2, C4: [I'm unsure of the novelty here. After reading the paper, and the novelty statement, multiple times. I'm not convinced that what you're doing hasn't already been done. It appears you're only routing streamflow from a model created in Mai et al. (2022). I hate to share/recommend reading my own research as it can come off as scummy, but I have an accepted manuscript that has a very similar workflow what you're doing here (Training a LSTM that was created based on another paper's work, applying the LSTM to smaller catchments, routing those outputs, and comparing the routed flows against the lumped predictions. My paper's novelty is that our Muskingum-Cunge routing is a differentiable model). The reason I'm sharing my work is because you may not be aware that I had used a very similar set up in my research. The paper has been under review for ~15 months and only available in preprint. See below for the Title and DOI.

Improving large-basin river routing using a differentiable Muskingum-Cunge model and physics-informed machine learning 10.22541/essoar.168500246.67971832/v1

To be fair to your work, I do think that your paper presents an angle that has not been done yet but is not expressed in the novelty statement: "The distinctive aspect of our research lies in the combination of components to form the SR model. The following subsections explain the specifics of the components we selected in this study to demonstrate the spatially recursive modelling approach."

I believe your work's novelty is both the scale (as my accepted work is done at the catchment scale, not at the regional scale), and my work does not have a reservoir module included in the routing. This will ensure your paper correctly identifies what is an expansion of domain knowledge.]

Below we will detail the four key differences (two of which are noted by the reviewer above) between our SR model approach and the approach described in the Bindas et al. (2023) unpublished preprint.

Our SR model applies an LSTM to directly generate routing model inputs at the spatial scale of the routing model. Specifically, our LSTM directly predicts local subbasin streamflow appearing at the subbasin outlet and these are directly the inputs to our routing model. We have demonstrated that the SR model is robust and works without user intervention when the routing model scale, and hence the scale at which the LSTM predicts local subbasin streamflow, is varied.

In contrast, the approach in Bindas et al (2023) is fundamentally different as their LSTM streamflow predictions do not match the spatial scale of their routing model and hence require an intermediate, and somewhat unclear, scale-specific parameterization to translate in inputs for their 2 km reach length routing model. Specifically, they generate HUC10 subbasin level streamflow predictions with their LSTM and then rescale each of these within the subbasin to lateral inflows of each 2 km reach. How to adjust their rescaling approach for either a new shorter or longer reach and/or a larger or smaller subbasin scale (not HUC10 subbasins) is not reported on.

Beyond the above difference, the Bindas et al. (2023) preprint only demonstrates they can enhance lumped LSTM predictions of larger watersheds (>2000 km^2) *in a single catchment* when their routing model is trained in that catchment. Our method is *at the regional*, not catchment. scale, and demonstrates our untrained SR approach (uncalibrated routing model) enhances LSTM predictions in 27 out of 44 larger watersheds (>2000 km^2) spread across the GRIP-GL study region.

There is also a third key difference. Given an existing applicable lumped LSTM model for streamflow, the Bindas et al. (2023) preprint uses a physics-informed machine learning approach to improve upon the lumped LSTM, but that improvement requires additional ML-model training. Our approach requires no such complex retraining, out of reach to hydrological modellers not trained in machine learning and is thus simpler and directly applicable for all hydrologic modellers. It is not clear the approach in Bindas et al. (2023) would work as successfully at the regional scale, with more spatially variable routing conditions. Particularly, given the scale-specific rescaling approach noted above which may only work in their 5260 km^2 catchment case study. Clearly, if we calibrated our routing model to each of our 200+ catchments, our SR model results would improve even further. Even without calibrating our routing model in each catchment like Bindas et al. (2023) did, we note the magnitude of our streamflow prediction improvements in larger watersheds (>2000 km^2, see our Figure 6 in our manuscript showing KGE increases of roughly 0.05 to 0.3 KGE units in 7 of 8 basins) seem to be equal or larger than the Bindas et al. (2023) improvements which are reported as an NSE metric increase to 0.857 from 0.801. We acknowledge the different units here but believe our point stands.

Another less important novel aspect of our manuscript is that we utilize a routing model that explicitly includes lakes, unlike the BIndas et al. (2023) preprint whose routing model does not simulate lakes explicitly.

We will add a streamlined summary of these 4 differences into our revised manuscript more clearly indicating the relative novelty of our work.

RC2, C5: [Lines 107-108: This sentence reads a little weird. I suggest changing to: "Han et al. (2020) and Mizukami et al. (2016) include examples of hydrologic routing models."]

Thanks for the suggestion, we will rephrase that sentence in the revised manuscript.

RC2, C6: [Can you make the font bigger for figure 5 axis?]

Thanks for the suggestion, we will remake the figure in the revised manuscript.

We look forward to hearing your thoughts on the revised manuscript and hope for a positive outcome. Should you require any further information or clarification, please do not hesitate to contact us.

Thank you once again for your time and expertise.

Qiutong and co-authors

References

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol Earth Syst Sci, 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019.

Kratzert, F., Gauch, M., Klotz, D., and Nearing G.: Never train an LSTM on a single basin, EartharXiv [preprint], https://doi.org/10.31223/X57090, 2023.