RESPONSE TO RC1 COMMENTS:

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

Overall, this is a very interesting study considering how to integrate deep learning models with process-based models. The paper introduces a Spatially Recursive (SR) model based on GRIP-GL data. The model first trains a basin LSTM and then uses the trained LSTM to simulate the flow of subbasins. Finally, it obtains the ultimate basin flow through a Routing-only mode. While the aspects mentioned in the paper are not individually new approach, this combination demonstrates a certain level of innovation.

We are pleased you recognize the innovation here.

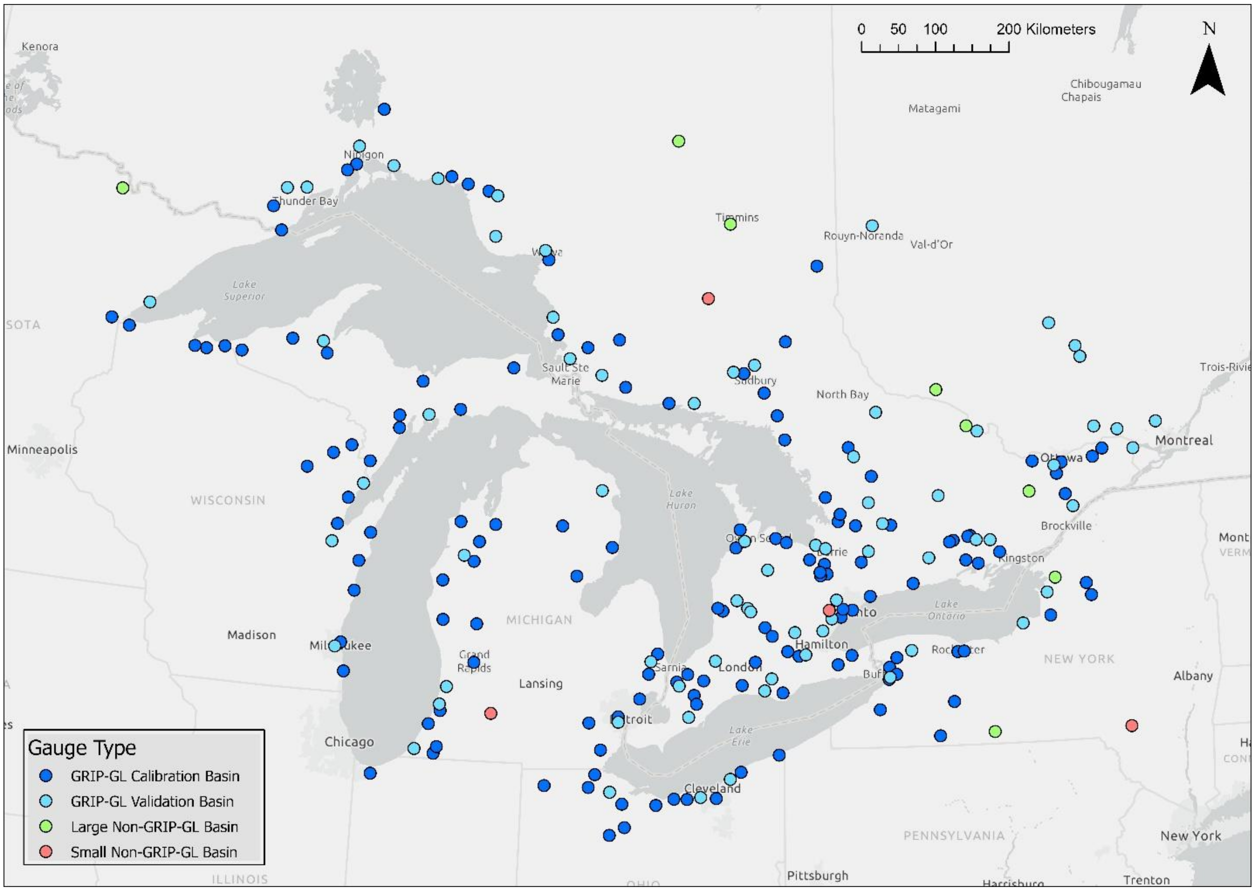
RC1, C1: [The "Routing-only mode" is an important part of the document, and perhaps a flowchart would help readers better understand its workflow. Additionally, the document mentions that the input for the Routing-only mode is hourly data. The question is how the authors transform data from the LSTM into hourly data.]

The workflow of the routing mode was briefly explained in Section 2.1.3. We don't see how a flow chart can enhance the workflow understanding given that for "Routing-only mode", the workflow is really only input file conversion and then a Raven model configured only for routing. Instead, in response to your suggestion, precise details of the Raven routing-only mode configuration, such as routing method algorithms, will be added to the appendix in the revised manuscript ensuring readers can precisely replicate our Raven Routing-only configuration.

The daily LSTM-predicted outputs are in units of millimeters per day (mm/d). We transformed the daily LSTM predictions to hourly routing simulation inputs by assuming the precipitation intensity is constant over the day. That is, for a given date, the LSTM-predicted streamflow is assigned to all 24 hourly time steps. We will ensure this transformation is noted in the revised Section 2.1.3.

RC1, C2: [Including a map of the study area would indeed be beneficial for readers unfamiliar with the GRIP-GL project.]

Thanks for your valuable suggestion, we will add a detailed map of the study region in the revision, featuring the classification of calibration and validation basins. The draft new map is shown below:



RC1, C3: *[The paper mentions that initially, basin data is used to train the LSTM, which is then applied to predict streamflow in subbasins using the subbasins' input data. Since there is sufficient subbasin data available, the question is why not directly train a subbasin-specific LSTM for predicting subbasin streamflow, which could then enable the prediction of basin streamflow through Routing-only mode?]*

We assume the reviewer is asking us why we did not train specifically on subbasin-scale (i.e., very small local drainage basins) watersheds. That would be ideal since we assume that finer resolution (i.e., smaller-scale data) would result in higher quality of predictions at that scale. The primary reason we did not do that is because we wanted to show that our SR model built with an existing LSTM regional streamflow model, trained with watersheds beyond the subbasin scale delineated in the routing product (The GRIP-GL calibration basins are ranging from 200 km^2 to 16000 km^2, while the average subbasin scale is approximately 131 km^2 in the GRIP-GL routing product) 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, our SR model as implemented is a lower bound estimate of the optimal performance our SR modelling approach could achieve. For example, if we instead trained our LSTM by adding additional subbasin scale watersheds (< 200 km^2) to the training set, we can assume performance would improve, or at the very least, not degrade. Overall, our work shows hydrological modellers a reliable way to improve LSTM-based streamflow predictions without having to do any more training or calibration.

One of the practical reasons we did not train with subbasin-scale watersheds is that the baseline lumped LSTM we duplicated from our previous work in Mai et al. (2022), specifically targeted gauged watersheds greater than 200 km^2. Given the findings of Kratzert et al. (2023) showing that training on more basins is always better, we would caution against the assumption that developing any subbasin-scale specific LSTM (targeting and training on only watersheds smaller than 200 km^2 for example) would be optimal within our SR modelling approach. Another practical reason we did not do this is that the discharge observations are unavailable for most of the subbasins delineated in the GRIP-GL routing product and the North American Lake-River Routing product. Even though we have sufficient training features (i.e., dynamic forcings and static attributes) for the subbasins, we do not have the target variable (i.e., subbasin streamflow).

We will integrate some of the above rationale into our revised manuscript to highlight this question of utilizing subbasin-scale watersheds in the subbasin level LSTM development.

RC1, C4: [In the discussion in section 3.1, for smaller basins, the spatial segmentation might still represent a sub-basin. In this scenario, is the structure of the Spatially Recursive (SR) model still the same as it would be for multiple sub-basins, or is the Routing-only mode not used under these conditions?]

Yes, for smaller basins, the delineated routing network would only contain one subbasin which is geometrically identical to the basin outline. The structure of the SR model is still the same as it would be for multiple sub-basins, that is, the routing simulation will still be applied on these smaller basins. In such case, the routing model only functions to take the subbasin streamflow (LSTM-predicted) as an input and it would be directly flushed without delay to the basin outlet, making it equivalent to a lumped prediction. In other words, routing model application does not change LSTM-predicted inputs for such a case.

We will ensure our revision in section 3.1 makes the above answer clear.

RC1, C5: [What is the role of Figure 8? Providing different delineations of the routing network within a single basin might better help in understanding the impact of routing network delineation.]

As explained in the paragraph from line 400 – 417, Figure 8 is to demonstrate how our findings might relate to different lake densities in the different watersheds. On the other hand, Figure 2 is used to demonstrate the impact of different delineations of the routing network in a single basin. In response to this comment, we will revise the caption of Figure 8 to make its purpose clear to the readers.

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

Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., Fortin, V., Fry, L. M., Gauch, M., Klotz, D., Kratzert, F., O'Brien, N., Princz, D. G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N. K., Temgoua, A. G. T., Vionnet, V., and Waddell, J. W.: The Great Lakes Runoff Intercomparison Project Phase 4: The Great Lakes (GRIP-GL), Hydrol Earth Syst Sci, 26, 3537–3572, https://doi.org/10.5194/hess-26-3537-2022, 2022.