Comments by Anonymous Reviewer in *blue*, our responses in black.

In this study, rainfall-runoff data from 531 watersheds in the camel dataset were simulated utilizing the Long Short Term Memory (LSTM) model. The LSTM model, when trained with multiple basins, exhibited superior performance compared to those trained with a single basin. The authors assert that LSTM streamflow models yield optimal results when trained with extensive and hydrologically diverse datasets, encompassing as many watersheds as feasible. Regardless of the specific objectives a researcher may have for training a machine learning (ML)-based rainfall-runoff model, there appears to be no compelling reason not to employ a large-sample dataset for training purposes. Segregating basins into groups based on hydrologically-relevant characteristics appears to enhance the performance of models trained with limited data, surpassing models trained on randomly grouped basins of comparable size.

I concur with the authors' viewpoints. The research outlined in the manuscript is methodical and enhances our comprehension of the necessity to utilize data from numerous basins for effectively training LSTM models for rainfall-runoff modeling. The statistical analyses conducted by the authors are robust. Nevertheless, significant revisions are imperative prior to considering this manuscript for acceptance.

My suggestions and critiques are enumerated as follows:

Abstract: The abstract contains excessive information regarding the popularity of ML and LSTM in rainfall-runoff modeling. To align with the traditional structure of scientific papers, I propose including more details regarding the methodology implementation, delineating how the LSTM model's performance improves with a larger dataset, and specifically identifying the most crucial influencing factors for LSTM training in rainfall-runoff modeling.

> **Response:** This is a fair point, but we largely do not agree. Scientific papers are typically motivated by an impact statement, and since this paper is a paper about modeling itself (instead of, for example, about how modeling allows us to learn about the real world), the impact statement in this case is about why ML models are important. The analogy would be if we were writing a paper on flood forecasting, we would motivate the paper with a few statements about floods being the most common type of natural disaster, impacting XX people, etc.

Section 1: While it is true that ML algorithms vary, it would be beneficial for the authors to elaborate on why LSTM has gained prominence, particularly due to its development for time series prediction, and elucidate why other ANN systems are seldom employed in hydrological studies.

> **Response:** This is a good point. We will add a sentence or two about this in the revision.

Section 2: Figure 2 presents results that may be surprising to readers. The authors should provide more information on why VIC basin and mHM basin outperform VIC regional and mHM regional. While these findings support the authors' assertions, additional context is required to evaluate their accuracy. Consider transferring relevant information from the appendix to Section 2 for clarification.

> **Response:** To our understanding, it is common knowledge that regionalization strategies under-perform compared to local calibration. Given that this is important to the point of the paper, we will certainly provide some insight into why this is the case.

Section 3: The utilization of 531 CAMEL basins, encompassing more extreme runoff data, to enhance LSTM prediction is comprehensible. However, clarity is lacking regarding Figure 3 and the content between Line 61 and 67. Additionally, it remains unclear why there is a limitation on regional models in Figure 2, and how the bounded vector in LSTM influences runoff prediction as mentioned in Lines 63-64.

> **Response:** Thank you for pointing out sections of the writing that could be improved. This level of detail is sincerely appreciated. We will revise these sections for clarity.

Section 4: Would presenting Figure 5 on a logarithmic scale for the x-axis be more appropriate? This aspect requires further consideration.

> **Response:** We did try a log scale (and a log-log scale) for figures 4 and 5. We felt like the message is clear using the figures in their natural scales, and it is just *slightly* more complicated to glance at a log-scaled figure. Also, this way there is no risk of a reviewer worrying that we were using a figure that exaggerates the real effect.

Section 5: Enhancements to Figure 6 could involve using more contrasting colors to ensure clarity in black-and-white printed versions.

> **Response:** Thank you for noticing this. We will use different line styles in Figure 6 to accommodate printed versions.

Code and Availability: While the code and data are accessible for download, they remain challenging to execute, especially for those unfamiliar with the NeuralHydrology Python package. I recommend providing supplementary materials to furnish more detailed information on these packages.

> **Response:** Unfortunately, this comment is not specific enough to provide us with the information that we would require to add such supplementary material. NeuralHydrology is already as “plug-and-play” as Python packages get. Reproducing our experiments requires only installing the software and running a single command. These steps are well documented in the documentation of NeuralHydrology.

Additionally, NeuralHydrology comes with full documentation and tutorials. You can run the experiments on personal computers (with a single GPU) so that anyone should be able to use it with almost no effort. Additionally, we have previously published a peer-reviewed software paper about this code package (cited in the current manuscript), and we do not see the need to reproduce the contents of that software paper in each publication that uses this codebase.

Considering the significance of the content, it is advisable to integrate essential information from the appendix into Sections 1-5, reserving detailed model settings for supplementary materials.

Response: We believe that the current organization makes the main message of the manuscript more accessible (after all, it is an opinion paper and not a research article). The organization of this paper is designed so that readers can quickly get the main message, and the appendix then also allows people to reproduce some of the experiments that we use to highlight the opinion. To re-emphasize, the purpose of this opinion paper is not to present new research but instead to address a systemic problem in the hydrology literature.