Response to Referee 1. Comments are in black, responses are in blue.

Introduction – The various concepts of the context are explained well and are interesting. However, the review of existing literature related to these concepts is almost completely missing. A good introduction should include key concepts of the problem at hand, i.e. water routing (which is addressed in the current version*). It should then review what has been done so far (both classical and AI-based methods) related to these concepts, thereby revealing the current gap(s) to which your paper would contribute (this is poorly addressed in the current version).
* Although this section could be more engaging by concisely explaining water routing ideas in physics-based models.
We have added further description of physics-based models and highlighted advantages and disadvantages of both classical and AI approaches. We hope that the immediately following "Our Contribution" makes it clearer how our approach addresses certain gaps and what is left to be done in future work.

Methods – Important LSTM training details are missing. For example, the loss function, with its full definition, is a crucial element of the optimization algorithm and should be presented in the Methods section, not in the Metrics section (and it is not sufficient to refer the reader to another work for its definition). The optimization algorithm and the LSTM architecture are also completely missing here and throughout the paper.
In the Methods section (2.2), we added the definition of the loss function and described the optimizer and hyperparameters. We agree that the LSTM architecture should be described. However, because the design of the LSTM architecture was taken from previous work as cited, we chose not to highlight it in the main text, but instead modified Appendix B now to include a detailed description. We reference Appendix B in the main text. Additionally, we modified some notation in Section 2.2 and Figure 3 to make the paper consistent end-to-end.

Benchmark – In its current form, the comparison with LISFLOOD is not fully justified in my opinion, as the existing LISFLOOD simulations were conducted under a different setup that seems to be unknown to the authors. Such simulations involve several subtleties that need to be carefully managed; otherwise, any conclusions drawn would be biased. Why don't the authors conduct these simulations themselves under controlled conditions corresponding to their LSTM experiments?
In our opinion, conducting their simulations is beyond the scope of this paper. For example, it is documented in Alfieri et al. (2020) that a full calibration exercise took 2 months on a HPC Cluster. Moreover, Nearing et al. (2024) conducted similar comparisons as to the ones we have, also without recalibration. That said, we agree that the comparison can be improved, and we have attempted to do so.

In our comparison, we require knowledge of the GloFAS training protocol and predictions at specific gauges. However, we originally only had access to the GloFAS predictions on a grid, and hence we had to associate these grid points with nearby gauges. This left room for uncertainty as to whether we picked the correct gauge. To improve the comparison, we have emailed their team and received the set of gauges from GRDC used for the calibration of the model, which we used to adapt Section 3.2.1.

With this data we were able to separate two experiments in GLOFAS, which we refer to as basin-split* and time-split*. We made clear that both experiments are slightly different from our own basin- and time-split experiments, but they are similar and likely biased towards GloFase, as we document. We think the comparison has a good place in the article as it demonstrates the generalizability skills of the LSTM.

Results (and Discussion) – The analysis of the results does not appear to be sufficiently in-depth, particularly in relation to the few previous regionalization studies using LSTMs over the US continent. There should be a thorough discussion comparing the findings of this study with those from previous research to highlight the contributions and significance of your work, or, to explain any potential divergence from their results (this is missing in the current version).
We have gone through the literature again regarding application of LSTMs to the streamflow prediction problem. We found that LSTMs in previous studies use a different set of gauges and different input datasets (observed precipitation and observed atmospheric conditions, instead of reanalysis data), and this makes the direct comparison between them challenging to interpret. In Section 3.2.2, we discuss this and compare our results with other LSTM models.

The Use of the Term "Forecast" – Based on the content, this paper is not about forecasting but rather about prediction (simulation). This error should be corrected (this mistake does not appear in the Conclusion, where it correctly states: “We have successfully trained and validated an LSTM for the task of predicting streamflow from runoff worldwide”).
We have adjusted the text throughout.

MINOR COMMENTS

- PDF Version – The PDF version of the paper did not include line numbers, which made it very impractical for review.
We agree with the referee and apologize for the inconvenience. This happened because we posted our preprint both via EGU and via arXiv. EGU recommends having a single preprint, but arXiv doesn’t allow papers with line numbers.

- P.2, Introduction – The phrase “common ungauged basins” in the sentence “This indicates that information in large-scale hydrological datasets is sufficient for generalization tasks, especially to the common ungauged basins (Nearing et al., 2021)” needs clarification. What does “common ungauged basins” mean in this context?
We rephrased this part to make the text more clear.

- Table 1 – Please specify the range for each level presented in the table.
We interpreted this comment as providing the range (minimum and maximum) catchment area for each level. That information is now included in Table 1. Additionally, we note the median catchment size in Table 1, and contrast with the median size of each basin as noted in the text (Section 2.1.4).

- Page 9, Line 6 – Remove "mean" in “mean squared error,” as the term does not include any averaging.
We altered the text for clarity. The numerator in the NSE metric per basin is the mean squared error, and the denominator the variance, but the 1/n factor cancels out between the two. Our loss function is slightly different, as explained.

- Appendix B – What are the tested values for each of the three hyperparameters? This information is important and concise enough to be included in the main text.
We have changed Appendix B to provide a detailed explanation of the LSTM architecture. The values of the hyper-parameters were moved to the main text. Note that we have not run extensive tuning experiments with them, but rather used the values used in similar studies, which we have now explicitly referenced.

- Figure 5 – Place the legend above the subplots as it applies to both of them.
We have added “The legend in (a) equally applied in (b).” at the end of the figure’s caption.

- Figure 6 – (Maybe) Place the labels (a), (b), (c), (d) inside the respective subplots to save space.
We decided to keep the labels out of the plots as it gives more liberty to the editor.

- P. 16 – It would be interesting if you could present the top 4-5 attributes that, according to your results, show a relationship with model performance. For instance, in which regions are variables like “karst percent cover” and “groundwater table depth” explanatory to some degree in terms of model performance?
We agree that this is an interesting question. In our study, we used many of the same static features that were used in the LSTM model of Kratzert et al (2019b), who did carry out a detailed feature importance study. However, as noted, their model represents the entire land hydrology system, while our model does not. Because of this, there are likely some differences in relative feature importance between the two.

- Figure 8 (caption) – Provide the definition of the aridity index both in the caption and in the text. For instance, you mention "Drier regions" (i.e., regions with lower aridity index),” but it is natural to expect that the higher the index of a region, the more the climate lacks effective moisture. Additionally, the following sentence in the caption should be stated more carefully: “There is a tendency for worse scores for smaller aridity indexes (i.e., drier basins),” since at the same range of aridity indices, there are basins with good NSE values. Also, state what each point represents in the figure.
We have defined the aridity index and referenced the original and most recent work from the authors of the dataset containing it. We have also rephrased “There is a tendency for worse scores for smaller aridity indexes (i.e., drier basins),” to “We have observed that lower NSE scores preferentially are found in more arid basins.”. We also provided further details about the meaning of each small square (“points”) in the figure.

- P.18 – “However, it is not clear if this increase in performance is due to a change in the LSTM model.” How is Nearing et al.’s LSTM model different from yours? This is an example of studies that should be included in the literature review. The provided context and highlighted differences can then be used as an element of result analysis in your discussion section.

We provide a comparison of our model results to other LSTM models (including Nearing et al., 2024) in Section 3.2.2.

- P.18 – The equivalency between gauged and time-split configuration, as well as between ungauged and basin-split, should be mentioned at their first introduction. Additionally, consider using the terms "gauged" and "ungauged" instead of "time-split" and "basin-split," as these are far more intuitive.
In our understanding, the terms `gauged' and `ungauged' denote a physical feature of a basin, while `time-split` and `basin-split` indicate an experimental design, and there is not a direct correspondence between these concepts. For example, only gauged basins - with measured streamflow - are used across all of our tests, while a model trained in either a time-split or basin-split configuration can be used to predict streamflow in a gauge or ungauged basin.

The terms time-split and basin-split correspond to the notions of temporal and basin generalizability introduced in the introduction. We consider the terms time/basin-split to be technical enough to only include them in the Results section. We have checked the manuscript to make sure that our notation is clear and consistent throughout, and added a small section on Notation in the introduction to make sure the reader is aware of our convention.

- P.18 – The conclusion "suggesting that drier regions pose unique challenges for the LSTM model" is incorrect. As mentioned above, many of your basins with good NSEs fall within the same aridity interval. Please revise this conclusion to reflect the actual results.
We rephrased this conclusion as follows: "Additionally, our analysis revealed a correlation between the model's performance and the aridity index. While streamflow in arid basins can be modeled well by the LSTM, it is also true that all basins with a poor NSE have a lower aridity index. This suggests that drier regions pose challenges for the LSTM model, but that other basin features may affect performance as well."

- P.19 – Remove the parentheses around "Hoedt et al., 2021."
We fixed this.

- I find the mass balance analysis interesting. You may consider placing it inside the main body of the paper.
We kept the mass balance analysis in the appendix as we felt it was more secondary, but we appreciate the referee's comment.