Comments by MH in *blue*, our responses in black.

Summary

The Opinion paper addresses a common issue and misconception in the application of LSTM models for hydrologic streamflow prediction: Often, LSTM models are trained and evaluated on only a few basins or even a single one. This leads to sub-optimal performance as LSTM models benefit from being trained on a large variety of data - as they are increasingly available in hydrology. Therefore, the authors suggest to conduct LSTM training with such large-sample datasets as best-practice (outlining that additional fine-tuning might be an option for single or small set basin applications). Further, the paper focus on training data diversity and optimal setup-up of training sets.

Evaluation and Recommendation

The Opinion paper covers a timely topic since LSTM models are state-of-the-art tools for streamflow prediction and a variety of other tasks in the broader geosciences. With LSTMs being increasingly used (which is also shown in the paper), the presented topic is important and meets the community's interest.

The manuscript is well written and referenced. The codes that were used are freely available and data sources are referenced. The figures are of good quality. Yet, the current manuscript requires some specifics to be addressed (see below). In particular, section "5 Is hydrological diversity always an asset?" addresses a very important topic – at the same time, it would benefit from some iteration as is also specified in the comments below.

I recommend publication after minor revisions.

Specific comments

> *l.18-19: "We do not mean top-down vs. bottom-up in the sense discussed by Hrachowitz and Clark (2017)." Please specify briefly their definition of top-down and bottom-up.*

Response: Thank you for this suggestion. We will do exactly what the reviewer suggests in the revision.

> *l.28-29: "We see no reason why…" Please briefly elaborate on these reasons.*

Response: Thank you again for helping us to add clarity to the writing (sincerely). We will do exactly what the reviewer suggests in the revision.

> *l. 32-34: Please rewrite for clarity, e.g. split up in two sentences.*

Response: We will split the run-on sentence and do what we can to make the passage easier to understand.

> *l.50 ff & Figure 2: It is stated that 400+ catchments from CAMELS were modelled with mHM and VIC for comparison. How many did overlap with your 531 basins? There has to be a large overlap anyway with 671 basins in CAMELS in total, but I think it would be an interesting information to know. Or better: why not showing the cumulative plots between only the basins that are in both the VIC+mHM set and the LSTM set? This would make the comparison stricter.*

Response: This is an artifact of some legacy benchmarking experiments where some of the physically-based models only ran over a subset of the CAMELS basins. The 400 is a strict subset of the 531. Our reasoning in not making the basin groups match exactly is that we wanted all of our ML-based CAMELS results to be identical throughout the paper. The “fair” comparison in this figure is Figure 4 and Table 3 in Kratzert et al. (2019b). Changing the basin group in this figure does not change the message, but we agree with the reviewer that it is probably cleaner to make these subplots match. We will do that in the revision.

> *l.63: “to 1 (-1)” -> unclear, is this [1 -1]? Please specify.*

Response: Yes, it should be (-1, 1). We will fix this notation in the revision.

> *l.64-65: “size equal to the number of cell states.” -> Please add one or two explanatory sentences. This refers to the model architecture and might not be clear to all readers.*

Response: Thank you again for offering suggestions about how we can make the manuscript more accessible to more readers. We will take the suggestion in the revisions.

> *l.73: “no model captures all of the extremes” -> Agreed. Yet, I wonder whether the chosen basin in Fig. 4 represents "extreme events" well. The hydrograph shows a rather regular pattern with a peak flow ranging roughly between 5 and 15 mm/d every year. I think a more irregular hydrograph with, e.g., one or two intense peaks (or missing peaks in some years) would better illustrate extreme events.*

Response: It is a good suggestion to add more examples to this figure and discussion, to help illustrate diversity. We will do that in the revision.

> *l.83-84: “In other words, even these 531 basins are most likely not enough to train optimal LSTM models for streamflow.” Why do you think that is? Please elaborate. Are there indications for an upper limit of the objective value and a corresponding sufficient training size from research with CARAVAN that could be shown here? (This last point could also be part of conclusions and outlook)*

Response: The reason is because we have not found the saturation point in terms of improving skill scores from training on more basins. We could use Caravan, however there is currently no “official” LSTM benchmark, and we want to make sure that the message in this paper is as clear as possible. We also do not want to run the risk of readers thinking that the results might be different because we used a new dataset. We don’t think this is the right venue for showing LSTM results with Caravan.

> *l.85ff: This section covers quite a range of aspects. I suggest to restructure it a little, e.g. into subsections: One that contains Figure 6 and the corresponding text, and one that covers conjectures about effects of larger datasets and data split approaches. Nonetheless, these are interesting points to be discussed since they also might pave the way for further research.*

Response: We will look for a way to subdivide this section to make the message more concise and focused.

> *l.87ff: “more is always better, as far as we have seen), and variety refers to the (hydrologic) diversity of data.” -> Both points, volume and variety, are important to be pointed out. There is an issue that I think could be mentioned and discussed in this context as well: class imbalance (even if this is not a classification problem). This might also be part of the explanation behind the things discussed in the rest of section 5. In the CARAVAN-paper (Kratzert et al., 2023), there is a histogram showing the distribution of catchments over the different climatic zones of the earth and the corresponding distribution in the dataset. There, a class imbalance is visible which indicates why predictions of certain climatic basin classes are better or worse since they are overrepresented or underrepresented.*

Response: First, there are no model results reported in the 2023 Caravan paper (and we have not published any modeling results on Caravan anywhere else) so we are a little unclear on the connection that the review is making between model performance and class imbalance in the Caravan dataset. That being said, we are not convinced that the reviewer is correct that class imbalance is an issue here. We have not seen any empirical evidence that suggests that this is an issue.

> *l.148: “training (1 October 1999 through 30 September 2000), validation” -> mix-up of dates? This is a very short training period.*

Response: Thank you, this is a typo.

> *l.210-211: “without carbonate rocks fraction and the seasonality of precipitation” -> I agree on dropping the seasonality of precipitation. But I would assume that carbonate rock fraction and related karst flow properties might be an important feature to be included in the clusters. Did you investigate in which clusters those*

basins fall, that have a carbonate rock fraction value larger 0 or above a certain threshold? Maybe having a dedicated "karst" cluster might be an option?

Response: Thank you, this is a typo.

l.215: "into 5 groups" -> should be "into 6 groups", right?

Response: Thank you, this is a typo.

Fig A4: Interesting to see how geographically aligned the different clusters are. (Apart from separated small accumulations apart from the bulk of a certain cluster – like the small groups of cluster 2 (orange) and 6 (brown) north from cluster 1 (blue)). With 6 being the detected optimum number of clusters, did you look into the neighboring 5 and 7 clusters with respect to the spatial distribution on the map and performance gain/loss of the model?

Response: Yes, the attribute clusters follow fairly well-known geologic and climate patterns. We did not train on 5 or 7 clusters. We aren't sure that is really within the scope of this paper, which has a fairly focused point about not training on a single basin.

l.249: "Fig ??" -> compilation error, please check reference

Response: Thank you.

Tables and Figures

Fig. 3.: y-axis label on right hand plot not necessary if figure outline kept like this

Response: Thank you.

Fig. 5 could be dropped since its content is also shown in Figure 6 and I think the value of showing the blue line alone is not as significant.

Response: That's true. We were really focused on readability, and though the two plots helped isolate the main message from the "extra" experiments. But we see the reviewer's point and will take the suggestion.

Language

Very good and clear language, only small remarks as follow:

1. *63: "bounded bounded" -> bounded*
2. *69-70: "… in total 10 timesteps of streamflow observations…" -> … in total 10 streamflow observations…*

Response: Thank you.