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

The opinion piece draws the attention towards the size and diversity of data used to train LSTM models for streamflow. I'm really happy that the authors put this material together to emphasize the good practice of using large and diverse datasets for models that will reliably predict streamflow at ungauged locations and time periods that are beyond the time periods the models were trained on.

I think the manuscript is well structured and nicely written. The arguments are easy to follow and the figures are appropriate to support the statements. I actually do not have any major comments; the list of minor (easy to address) comments are attached below. I would like to finish with congratulating the authors to this nice manuscript. It was a very enjoyable read and hope it will be published soon.

Best regards,

Juliane Mai

Response: Thank you very much.

Minor:

L 37-38: "It is important to recognize that there is usually no reason in practice to train LSTM streamflow models using data from only a small number of watersheds." —> Couldn't it be that one could run into storage limitations when, for example, wanting to use 10,000s of basins? I know this is not the target audience here, but maybe it should be mentioned that there is an upper limit of how many basins one can use and still be able to train a model with a standard machine.

Response: Even if your dataset becomes too large to be kept in memory (RAM), there are existing alternatives to stream data from disk into memory during model training, which allows training on any dataset that fits on disk.

L 45-52: Section 2: I think it might be helpful to re-iterate that the behaviour the LSTMs show here is the exact opposite compared to the physically-based models (left panel). I'd probably also add panel IDs to the figure (e.g., A and B) and then refer to them in the text. Up to the authors.

Response: Thanks, we will add panel IDs to Figure 2 and we also adapt the first paragraph of Section 2, as suggested by the reviewer.

Figure 2: caption: "for models trained on individual basins (basin) vs. on multiple basins (regional)" —> "for models trained on individual basins (basin; orange coloured lines) vs. on multiple basins (regional; blue coloured lines)"

Response: Thanks, we will apply those suggestions.

L 63: “a vector that is bounded bounded to 1 (-1)” —> “a vector that is bounded to 1 (-1)”. Also, what does the “-1” in parenthesis mean? Is the vector bound to the interval [-1,1]? I am guessing the “(-1)” refers to “(limiting)” in the next sentence?! But it is confusing. I’d probably not mention the lower bound because you already made the point clear with the upper limit that can be reached.

Response: Yes, this was sloppy notation. Thank you for pointing this out, we will fix it in the revision. And yes, the LSTM output is bound to the interval (-1, 1) (non-inclusive of the bounds).

L 65: Missing closing parenthesis after “Appendix B”.

Response: Thanks.

Figure 3: Wow! This is such an impressive difference!

Response: We agree 🙂

Figure 4 (caption and text): You illustrate the simulation of the single-basin model (13011900). I am assuming that the time period you show is the testing period for the model that was trained with this basin only (no spatial transfer), right? It would be helpful to mention this here as the reader doe not know which 10-year period you use for training vs testing (it’s in the appendix I am sure but it would help the flow of reading).

Response: Indeed this is the test period (i.e. neither the train nor validation period), we will make sure to clarify this. We are not really sure if we understand the second part of your comment though. The plot shows simulations from two different models. The top plot shows results from the regional model (i.e. a model trained on the training period of all basins) and the bottom plot shows the results from the single basin model (i.e. a model trained only with the training data of basin 13011900).

Section 3: Caption is missing a question mark at the end.

Response: Thanks, we will fix this.

Section 3: I would finish that section with a clear statement like: “When you train your LSTM model on a single basin, you will likely not be able to predict any peak event in the future.” This is what I would expect to read as a response to the question you state in the section title.

Response: Thank you for the suggestion. In general, we appreciate (and like) the idea about adding a summary sentence to the end of each section. We will consider the best way to approach this in the revision.

Section 4: I think it might be confusing when you talk about splits here given that hydrologists usually use "splits" to use one set for calibration and then the remaining for validation. When you say N splits (I think!) you mean that you repeat training/testing N times, right? Like N independent experiments. E.g., "The 531 basins are, for example, split into 5 groups (each around 107 basins). Then the first group of basins (first split) is trained on the training period (Oct 1999-Sep 2000) and then all basins of the first split are tested during validation period (Oct 1980-Sep 1989). After training, the model is evaluated for the ~107 basins of the first split using the testing period (Oct 1989 to Sep 1999). This experiment is then repeated for each of the remaining 4 splits." Not sure if that is entirely correct but it took me quite some reading of the appendix to get to this.

Response: We will try to rephrase parts of this section to make this part easier to understand.

Section 4: Do you think that 531 basins might not be enough because they are not diverse enough? Do you think 531 would be enough when the training period is longer than 10 years? I don't think any additional experiments are required here but it might be nice to have a bit of discussion here. Otherwise it might be really demotivating to know that even >500 basins are not enough.

Response: This is a very hard question that we unfortunately don't have an answer for (yet). First of all, what is "enough"? Enough in the sense of getting good predictions in these 531 basins on a different time split or enough for learning a robust and very general understanding of the different types of catchments included in CAMELS? The CAMELS dataset consists of a diverse set of catchments however the number of basins per type of catchment (e.g. catchment with glacier, catchments in the desert, high mountain catchments) do vary a lot. From our intuition, adding more data (basins and/or longer records) should help learning a better understanding of the underlying processes in these environments. Also thinking about the global context, there are certainly types of catchments that are not included in CAMELS.

L 97 (and caption figure 6): "the curve in Fig. A2" —> There is no curve in figure A2. Do you mean Figure 5?

Response: Indeed, thank you for pointing out this mistake.

Figure 6 caption: For completeness, describe what the dashed orange/green line mean.

Response: We will.

Section 5: I love the description of the experiment and its discussion here. Also the take-home message at the end is great. This is something that Section 3 (take-home) and Section 4 (description experiment) are slightly lacking.

Response: Thanks

Line 128: "good" —> maybe "reliable"?

Response: The problem with the word “reliable” is that we didn’t calculate any reliability metrics. We could say “accurate”, but the scope of the preceding discussion is slightly larger than pure accuracy (e.g., Figure 3). We actually thought fairly carefully about this particular wording in the original writing.

Line 129: “We’ve” —> “We have”

Response: Thanks, will be changed.

Line 131: “Of course, it is trivial (but most likely uninteresting) to beat improperly trained models.” Love that statement!

Response: Thanks.

Figure C1: I think it might be helpful to indicate (on x-axis) which the optimal value is for each metric. Table C1 might also be a good place for it. Up to the authors.

Response: Will be added.