"Unveiling the Limits of Deep Learning Models in Hydrological Extrapolation Tasks"

Manuscript #*2025-425*

B. Kraft

March 14, 2025

Short summary and highlights

This study evaluates how well LSTMs, both in a standard setup and as part of a hybrid LSTM/process-based model, generalize to extreme conditions. The authors train their models on a subset of catchments from the CAMELS-CH dataset. Some experiments also incorporate CAMELS-US data. They then test the models using extreme precipitation events derived from extreme value analysis. These events cover 1- to 5-day precipitation totals with return intervals ranging from 1 to 300 years. On the hold-out set—a time period excluded from training—the LSTM significantly outperforms the hybrid model. However, the main focus is on how the models respond to extreme precipitation inputs.

The experiments reveal a strong saturation effect in the LSTM when exposed to large precipitation events. In contrast, the hybrid model does not exhibit this saturation and maintains a quasi-linear response. It predicts higher runoff values for larger precipitation events. While the LSTM's saturation effect has been observed and discussed in prior work, it is surprising that it fails to predict output values even within the range seen during training. In practice, this saturation occurs at levels lower than the theoretical maximum.

The study highlights a major limitation of the widely used LSTM architecture in hydrological applications. The findings raise concerns about whether LSTMs are suitable for extreme event prediction.

Major remarks

1. I see potential for making the study much more impactful by testing solutions for the saturation effect. The authors speculate a lot in the discussion, and I was a bit disappointed that these ideas were not tested. However, I respect the authors' decision to not include further analysis. I do not insist on further experiments, but I would find them very interesting.

2. The hybrid model receives 730 days of input, while the LSTM only receives 365 days. The authors argue that the hybrid model needs the additional data for spin-up of the internal states. This makes sense, but the LSTM also needs to initialize its states. If the LSTM is used in a many-to-many setup (where the prediction period corresponds to the input period), the warm-up period should be the same for both models.

3. Why does the hybrid model receive an additional input? I believe the comparison between the models would be fairer if both models received exactly the same input data. Please clarify.

4. The synthetic precipitation data is extracted for a selection of catchments within a distance of 2.5 km. If I understand correctly, the model is fed with point observations, while it was trained on the average precipitation over the catchment area. This could be a source of error. Please discuss or clarify.

5. I would appreciate a brief analysis of the extreme values within the test data. How do the predictions compare to the observations? If you see the same saturation in the test data, this would strengthen your argument. It would also rule out that the experimental setup (changing precipitation but not temperature, etc.) is the cause of the saturation. You mention this in the discussion (L365-370), and I suggest showing it instead of speculating.

Minor remarks

Here I list some typos and suggestions for improving clarity:

L6 In the first reading, I did not understand where the 73 mm d^{-1} came from. I suggest clarifying this, e.g., "...show that this limit (which we have calculated for this study to be 73 mm d^{-1}) is below ...".

L68 "plausible" instead of "possible"?

Sect. 2.1&2.2 The description of the data splitting and how it is used to train the model (L104-107 and L116-117) is a bit misplaced. I suggest moving it to the Methods section.

L199 Could it be that this is because you use different catchments in your study?

Fig. 2a-c Consider adding the prediction limit also to the top row, also in Fig. 3a-b.

Fig. 3 The second panel should be (b), not (a).

L318 "...can lead to ~~more~~ behavior that is more plausible ...."

L323 "Alternatively, one can also think to directly train for the warranted behavior." This is a bit vague. What do you mean exactly?

L351 Since you already speculate: Would using other NN architectures be a solution?

L363 I suggest replacing "ceteris paribus" with an English term.

L377 The hybrid model coefficients (the outputs of the LSTM) can, in theory, also saturate. Consider mentioning this, or, even better, calculating the theoretical limits.

L380 Consider making a sentence out of the parentheses. This would make the sentence easier to read.

Tab. A1 Is mean temperature missing in the table, or did you not use it?

Tab. C1 Could you add the actual prediction limit (the "design limit") to the table?