

Exploring the ability of LSTM-based hydrological models to simulate streamflow time series for flood frequency analysis

We would like to thank the reviewers for their valuable and constructive feedback. We appreciate the time and effort that was put into the review. All concerns have been carefully addressed. Detailed responses to each of the reviewer's comments are presented below. For clarity, the reviewer's comments are presented in black font, with our responses in blue.

Sincerely,

Jean-Luc Martel, on behalf of all authors.

Reviewer 2: <https://doi.org/10.5194/egusphere-2024-2134-RC2>

This paper evaluates the performance of different LSTM-based frameworks for simulating streamflow time series, focusing on their ability to characterize extreme events and, consequently, enhance flood frequency analysis (FFA). To achieve this, the authors applied a set of 7 different LSTM configurations to simulate streamflow from 88 catchments in Quebec, Canada. These configurations included 1 baseline model (LSTM-base) and 6 alternative schemes, which incorporated observed meteorological inputs, a multihead attention structure, and/or hydrological model-based simulations as inputs in addition to the original ERA5-based data, among other aspects.

In my perspective, this work holds relevance for the hydrology field and is suitable for publication in HESS.

The use of ML-methods for hydrological simulations is rapidly gaining traction in the hydrological community and hence, improving our understanding of their benefits and drawbacks is paramount. As mentioned by the authors, there is still no consensus on how LSTM-based simulations perform on representing extreme flood events and how this can influence FFA.

In my opinion, the methods are sound, and their results are well presented. Overall, it was a pleasant read. I have, however, some questions and suggestions pinpointed below which I believe will help the authors to improve the overall quality of the paper. Once addressed, I believe the paper will be a good addition to HESS.

Thank you very much for your positive comments and suggestions. Please refer to the point-by-point responses to your comments below.

General Comments:

The authors state in the Introduction and Discussion sections that one of the primary goals of the paper is to assess the “potential of LSTM to extend streamflow records.” However, this aspect is not clearly demonstrated in the manuscript. I did not see any analysis or experiment specifically designed to evaluate this claim. So my question is: how are the authors addressing this in their work? Extending streamflow records is indeed a promising application of LSTM techniques with

potential benefits for FFA. For example, to address this gap, the authors could consider an experiment using catchments with longer datasets (e.g., 40 years of data), training the LSTM on subsets (e.g., 20–30%) and assessing how effectively it extends the records and how this lengthening improves FFA. Alternatively, they could revise the manuscript to remove this objective and avoid any misunderstanding.

You are correct. While investigating the potential of LSTM to extend streamflow records was one of our initial secondary objectives, the methodology used in this study did not allow us to achieve this goal. Instead, we focused our analyses on the primary objective: determining whether LSTM-based hydrological models can generate peak streamflow for flood frequency analysis. A methodology similar to the one you proposed could indeed be used to explore the extension of streamflow time series, which we will consider investigating in a future study. We will revise the paper to remove references to this secondary objective and instead introduce it as a potential avenue for future research in the conclusion.

Some of their methodological choices require further clarification. For example, it is not clear why they opt to use the Gumbel and GEV distributions for different stations; why the Cunnane plotting position was chosen; and which parameter estimation method was used. I suggest the authors to reevaluate their manuscript seeking to better detail these aspects to improve its reproducibility.

Thank you for bringing this up. There are valid reasons behind these choices, which we should not have omitted from the methodology section. Note that it is indicated that the GEV was used in Figure 7 panel c, but that was a typo from an earlier version. To ensure the study is reproducible, we propose clarifying these elements and incorporating the following information where appropriate:

In extreme value theory, the Generalized Extreme Value (GEV) distribution is derived from the block maxima method, which corresponding to the annual maximum series and is one of the most suitable approaches for flood frequency analyses. The Gumbel distribution is a special case of the GEV (also known as the Extreme Value Type I distribution or EV-I), when the shape parameter is set to 0 reducing to a two-parameter distribution instead of a three-parameter one. While the GEV generally provides a better fit, annual maximum time series are often too short to allow for a reliable estimation of the shape parameter.

Both the Gumbel and GEV distribution parameters were estimated using the maximum likelihood method. To determine whether the inclusion of the shape parameter in the GEV leads to a statistically significant improvement, we applied the likelihood-ratio test, a hypothesis test used to compare nested distributions within a same family. Based on this test, the Gumbel distribution was selected for all four catchments presented in Figure 7 instead of the GEV.

The Cunnane plotting position is defined as:

$$P = \frac{m - 0.4}{n + 0.2}$$

where P is the non-exceedance probability, m is the rank (with 1 being the smallest value), and n is the total number of observations. This empirical plotting position is an approximately unbiased estimator for extreme quantiles and is widely used when fitting data to the Gumbel and GEV distributions. For example, Environment and Climate Change Canada employs the Cunnane formula in their rainfall frequency analyses.

However, the Gringorten formula was specifically designed for the Gumbel distribution and could provide more precise results in that context. Given that the Gumbel distribution was not explicitly chosen as the preferred model in this study, we opted to use the Cunnane formula for all cases to maintain consistency.

Cunnane, C., 1978: Unbiased plotting positions—A review. *J. Hydrol.*, 37, 205–222, [https://doi.org/10.1016/0022-1694\(78\)90017-3](https://doi.org/10.1016/0022-1694(78)90017-3).

Although it provides valuable insights into LSTMs performance to characterize extreme events, I missed a more in-depth analysis and discussion about the different LSTM configurations and how they perform in FFA. For instance, I believe the manuscript would benefit from an expansion of the results section, including not only the FFA-based assessment for 4 catchments, but for all evaluated catchments, discussing their spatial distribution, general performance and differences between LSTM and HYDROTEL FFA for catchments with different data availability, and uncertainties, which were not included in the original manuscript.

This is a great idea, thank you for proposing it. The reason we initially selected only four diverse catchments with nearly complete observation records for this part of the analysis was the challenge of effectively presenting the results for the entire study site. However, upon further reflection, we believe an approach to achieve this would be to present three sets of maps:

1. NRMSE Qx1day for HYDROTEL;
2. NRMSE Qx1day for LSTM-combined model; and
3. A comparison map highlighting which model provides the lowest NRMSE for each catchment.

This approach would reveal spatial patterns and allow for comparisons with Figure 1 to assess if whether a correlation exists between the number of available years and the NRSME obtained for either hydrological model. Analyzing the annual maximum series (AMS) in this manner would also provide insights into which hydrological model would offer better performance for flood frequency analysis, assuming the observations represent the true values.

We propose conducting this analysis and expanding the result and discussion section accordingly. Additionally, as you have proposed, we will evaluate the performance of different LSTM configurations in simulating Qx1day series to determine which configuration components have the greatest impact on performance.

Regarding the multihead and oversampling approaches, were the lower performances somewhat expected? Given the inherent scarcity of extreme flood data, exploring alternative data lengthening approaches—such as using synthetic series (e.g., Papalexou 2022)—could enrich the discussion and provide directions for future research.

This is a good point that deserves more attention. Indeed, the inherent lack of data for rare events is a well-known problem, and there are approaches to generate synthetic series (CoSMoS-2s is one, and there are a plethora of weather generators and other software to do just that). However, in the case of streamflow in Nordic (i.e. snow-dominated) catchments, peak flows are often caused by snowmelt and more complex processes than precipitation. The text will nonetheless be modified to add alternative ideas to extend timeseries such as those mentioned above in the discussion.

Minor Comments

L78 – (Shen and Lawson, 2021) – Review the reference format

We are unsure about the specific issue with the reference format. If any corrections are needed, we would be happy to make them. The reference in question is:

Shen, C. and Lawson, K.: Applications of deep learning in hydrology, in: Deep Learning for the Earth Sciences, 283-297, 770 10.1002/9781119646181.ch19, 2021.

Figure 1 - Is it possible to improve this figure by adding some additional information in subpanels, such as a histograms of available years of observed streamflow (besides its spatial distribution) and climatic variables for each catchment (such as P, PET, ...)?

Thank you for this suggestion. We propose adding a pie chart to illustrate the fraction of catchments based on the number of available years of observed streamflow (similar to the one you suggested in your comment below). This will convey the same information without requiring an additional panel in the figure.

Regarding the other climatic variables, the 1979-2017 period is fully available since we are using gridded observations and ERA5 reanalysis. Therefore, a histogram would not provide additional relevant insights in this case.

L170 – PETas – Typo

The typo will be fixed as follows: “*PET as*”.

L185 - I believe the text would benefit here from 1 or 2 short sentences explaining the two different configurations of the HYDROTEL (2.3.1 and 2.3.2). It is not clear here if the authors will use both configurations as different inputs for the LSTM or whether the regional model was used only as an initial step (for example, by recalibrating only 11 out of the 27 parameters)

We agree that this section of the methodology may have been unclear. The regional HYDROTEL model (Section 2.3.1) served as the starting point for the local recalibration (Section 2.3.2) of all catchments. The results from the local recalibration were then used both the result comparisons (e.g., HYDROTEL box plot in Figure 2) and as inputs to the LSTM-based hydrological models.

To clarify this, we propose adding the following sentence at the end of L186 (before Section 2.3.1):

“A regional HYDROTEL model, pre-calibrated by the DPEH, served at the baseline for local recalibration on each of the selected 88 catchments. These locally calibrated models were then used in this study for comparison purposes and as an input for the LSTM-based hydrological model structures.”

L323 - Briefly explaining what is the standard scaler will help readers.

We propose to add the following explanation of the standard scaler in the text at L323:

“Prior to training, all variables were normalized using a standard scaler. The standard scaler standardizes values by subtracting the sample mean and dividing by the sample standard deviation, a process commonly known as Z-score normalization or standardization.”

L401 - is Figure 3a, b the same of Figure 2 but displaying all 7 approaches in the same panel?

This is correct – Figure 3a and 3b present the testing period for all 7 approaches already shown in Figure 2. Including these results in Figure 3 facilitates comparison with the map and makes it easier to directly compare the models over the testing period.

Figure 3 - Is it possible to include additional details on model performance in panels c and d? It is challenging to distinguish model performance. A bar or pie chart summarizing the percentage of catchments where each model performed better could enhance clarity and complement the existing text.

This is a great suggestion. We propose adding a pie chart in the white area (upper middle part of the figure), where each color represents the proportion of catchments where a given model performs best overall.

L420 – Suggestion: while Figures 4 and 5 are interesting, they contribute less to the main text. Consider moving them to the Supplementary Material (suggestion only).

Although these two figures show no correlation between KGE and drainage area, this finding is significant in itself. Our initial hypothesis was that the drainage area would be a key explanatory variable for the difference in KGE and NRMSE Qx1day results, particularly for the latter. However, these figures demonstrate the opposite; the LSTM models perform better across all catchment size categories rather than just a specific subset. Therefore, we believe these figures should remain in the main text.

Figure 6 – I believe using distinct background colors (e.g., grayscale) - instead of a line - for different periods (training, validation, test) would improve visualization and readability. Also, is it possible include the daily streamflow KGE for all periods (Hydrotel and LSTM-Combined)? It would help readers to assess and compare performances.

This is a great idea, thank you for the suggestion. We will add background shading in different shades of gray to distinguish the training, validation and testing periods. Additionally, we will display the KGE values for all three periods in the white space below the graph.

L453 - For plot d this is not valid. I suggest the authors include some metrics such as NRMSE here to support what they are claiming. Also try to avoid hyperbolic language, such as "much more similar", "much better",..

Thank you for this comment. We will revise this sentence to remove hyperbolic language and clarify its reference to Figure 6-d. Additionally, we will carefully review the paper to ensure hyperbolic language is avoided throughout.

References

Papalexiou S M 2022 Rainfall Generation Revisited: Introducing CoSMoS-2s and Advancing Copula-Based Intermittent Time Series Modeling Water Resources Research 58 1-33

The proposed reference will be added to the reference list as previously mentioned.