Reply to RC2: 'Comment on egusphere-2025-425', Anonymous

We thank the referee for the detailed evaluation of our manuscript and the insightful comments, which will help us improve the paper. In this document we address the comments, questions and suggestions posted by the referee. Please find the referee remarks in blue and our response in black.

General comment

This paper compares a stand-alone LSTM model with a hybrid HBV-LSTM model on the CAMEL-CH dataset. It also examines the impact of training the stand-alone LSTM on the CAMEL-US dataset, alongside CAMEL-CH, and also using 256 nodes instead of 64. The main focus of the discussion is on the ability of both models (stand-alone LSTM and hybrid) to show a linear pattern between simulated peak flows and rainfall when applying "synthetic" rainfall far higher the observed ones. The results clearly show the impossibility for the stand-alone LSTM model to exibit this linear pattern due to simulated discharges tending to a limit values as predicted by a previous study. But more interestingly, it clearly shows that this observed limit is far lower than the theoretical limit expected by the authors.

In my opinion, this paper is very interesting as it is the first to clearly and honestly address the limitations of LSTM models in hydrology.

I recommend publication after revision.

Response: We thank the reviewer for this interesting and positive assessment of our work.

Main comments

My main comments are:

1) Even though it is outside the scope of the paper, I would have appreciated a "deeper" and "fairer" comparison between the stand-alone LSTM model and the hybrid HBV-LSTM model. The paper is short (only 3 figures of results in the main text), there is room for that. My main criticism

is that no hyperparameter (HP) tuning is done for either model. The HP values are simply taken from previous studies. I think that the results could be different if a proper fine tuning was done for each model.

Response: We agree with the reviewer that hyperparameter tuning is an important part of any deep-learning model training. However, for our setting, the hyperparameters are not as critical as for benchmarking settings. Hence, we used the ones that have been identified in previous studies (Acuna Espinoza et al., 2024 and Kratzert et al., 2019). In some sense, this might be suboptimal. Nevertheless, we tried several hyperparameter combinations. In that regard, we tested the LSTM for a various number of hidden states, ranging from 8 to 2048 and obtained the same results. As the primary goal of the study is to highlight the saturation behavior in the LSTM, the nature of our results will not change even with a better set of hyperparameters. Having said that, we will implement a hyperparameter tuning for the stand-alone LSTM. If it changes the nature of our results significantly, we shall revise the manuscript accordingly. However, we believe that it is justified to adopt the same hyperparameters from the LSTM for the Hybrid model, for a fair comparison. Moreover, as it wouldn't change the nature of our results by large, hyperparameter tuning for the hybrid model might not be worth the computational demand for this study.

2) For analysis, it would also be very interesting to see the results for a single HBV model as a benchmark, which is very "cheap" to calibrate locally. Is there an improvement and is it "worth" the huge amount of data and GPU time required to process it? For example, the authors added US CAMEL data to their CH CAMEL learning dataset and moved from 64 to 256 nodes, which would have required a considerable amount of additional resources, but they don't show the corresponding improvement.

Response: We would first like to address the reviewer's comments regarding increased computational demand for the different LSTM models in our study. Technically, with four times the original number of hidden states (256 instead of 64), the increased computational demand is 16 times the original. In terms of time required: training a single LSTM using CAMELS-CH takes about ~900s (64 hidden states) and ~1700s (256 hidden states) on a GPU type V100. Including

CAMELS-US in the training data requires increased time of ~6000s and ~13500s respectively. This is not necessarily a concern for us, since we specifically wanted to test the effect of increased LSTM size and more training data. The fact that the models don’t show corresponding improvement was only known as a result of these experiments. We agree that it will be interesting to see the results from locally calibrated conceptual models for our design experiments. We will implement a single HBV model calibrated locally for catchments in our study. The manuscript will be revised accordingly to include these results, either in the main text or in an Appendix.

3) As the paper focuses on extremes, I also think that the evaluation against the observed runoff should not be limited to the NSE criteria as in Fig. 1 (which is the only figure presenting models performances), but should include a deeper analysis, including for example signatures calculated on flood events.

Response: Thank you for your suggestion. It is a good idea, and we will consider including additional metrics in the revised manuscript.

4) The same comment applies to the second part of results (Figs. 2 and 3, using synthetic rainfall): only 1 flood for 3 catchments (a little more in the appendix), whereas the authors have thousands of examples. A synthetic metric should be found that "summarises" the different observed behaviours (between catchments, but also for the same catchment but under different conditions). A "visual" analysis on a few examples, as in this paper, is a first step to draw first hypotheses. But then these hypotheses should be tested in depth.

Response: We only focus on the three flood events shown in Figures 2 and 3 because these events highlight the saturation behavior of the LSTM most prominently. We summarize the overall trend in our results in lines 249 to 251. From the reviewer's comment, we assume that this is not discussed clearly enough, and we shall rephrase this for better clarity in the next revision. Regarding the reviewer’s suggestion of developing a synthetic metric: we believe that at this stage, developing such a metric is best left as a part of future work, since it can be an intensive task. We will however address this discussion in the revised manuscript and speak to the potential of developing such a metric, as suggested here by the reviewer.

5) This last point (the need for a synthetic metrics that allows a "deep" analysis) leads me to my main comment. The authors don't clearly explain why, from a hydrological point of view, peak discharge should increase linearly with extreme rainfall. I fully agree with this, and even if it seems obvious, I think it would be valuable to anchor the paper with more basic hydrological references. In terms of synthetic metrics, I would, for example, calculate a regression coefficient between peak discharge and synthetic rainfall and see how it changes as a function of rainfall, as in the paper, but also as a function of the initial moisture content before a flood and/or the runoff coefficient. I would also not look at flood by flood, but try to find a graphical representation of all floods and catchments together.

Response: The sensitivity of flood peaks to an increase in maximum precipitation likely varies significantly across catchments, depending on multiple factors such as topography, soil characteristics, land use, and antecedent moisture conditions (as correctly highlighted by the reviewer). Given these complexities, we believe it is challenging to identify a consistent, clear signal across a large-sample dataset covering Switzerland, with its diverse hydrological regimes. Nonetheless, at a fundamental level, one would generally expect runoff to increase with increasing rainfall, particularly under extreme precipitation scenarios. For instance, in a simple linear reservoir model, the runoff response is inherently linear, meaning the total runoff volume (the integral of Q(t) over time) remains proportional to the total rainfall input, assuming negligible losses or constraints. Thus, the runoff coefficient in such a system is constant irrespective of rainfall magnitude. In contrast, conceptual models such as the TOPMODEL (Beven et al., 2021) demonstrate clear nonlinearities due to the exponential relationship between subsurface flow and water-table depth. This nonlinearity implies a substantial increase in runoff generation as saturation thresholds within the catchment are approached, resulting in runoff coefficients that vary with antecedent moisture conditions and rainfall magnitude. Interestingly, our analysis revealed that the LSTM model exhibited an unexpected and physically counterintuitive trend: runoff coefficients start decreasing with increasing precipitation magnitudes, especially for extreme precipitation values. This is particularly true for catchments with higher runoff generation. Motivated by the reviewer's suggestion, we will explore graphical representations of this phenomenon and include them if a consistent spatial pattern emerges across Switzerland. Additionally, we will refine the manuscript by clearly articulating, from a hydrological perspective, what physically reasonable runoff responses should look like, and explicitly discuss the limitations observed in the LSTM predictions for extreme events.

Minor comments

L100 : why did not you do a hyperparameter tuning? (a LSTM expert told me one day that hyperparameter training is absolutely required in any case, and that, if "hydrologists" don't have the necessary GPU resource, they should not use LSTM)

Response: Kindly refer to our response to the reviewer's first major comment. While hyperparameter tuning can improve the overall model performance, the behavior of the LSTM under our test conditions comes from its inductive bias and will hence not change fundamentally.

L200: You should give more details on the models performances, for instance using flood signatures

Response: Additionally to the answer above we would also like to highlight that such research has been presented extensively in the studies by Acuna Espinoza et al. (2024) and Frame et al. (2022). The hybrid model is acting mainly as a benchmark here.

L224: The results for the 256 node LSTM and/or the training using US-CAMEL should be presented in Fig .1 and discussed. Does this huge amount of additional data improve models performances?

Response: We believe it is better for a reader's comprehension to only include results from the stand-alone LSTM with 64 hidden states trained on the CAMELS-CH and the hybrid model, in Figure 1.

L235: You should do more clearly the link with basic hydrological processes, such as soil saturation and the effect of initial humidity condition.

Response: We will add a stronger hydrological perspective to the discussion in the revised manuscript.

Figure 2 and 3 : the terme "observation" is misleading. There is no observed discharges in this figure.

Response: Thank you for highlighting this disparity. We will change the legend for better clarity in the next revision.

L260: this affirmation is supported only by 1 flood over 3 catchments. You should try to exhibit that using much more discharge simulation (...that you have)

Response: We will consider including more results from the 3-day and 5-day events, either in the main text or as an appendix.

L299 : "Extreme hydrological events often coincide with distinct regime shifts": I fully agree but could you explain what do you mean to a "non-hydrogist", in term of involved processes.

Response: We will rephrase this in the revised manuscript.

References:

Acuna Espinoza, Eduardo, Ralf Loritz, Frederik Kratzert, Daniel Klotz, Martin Gauch, Manuel Álvarez Chaves, Nicole Bäuerle, and Uwe Ehret. 2024. "Analyzing the Generalization Capabilities of Hybrid Hydrological Models for Extrapolation to Extreme Events." https://doi.org/10.5194/egusphere-2024-2147.

Beven, Keith J., Mike J. Kirkby, Jim E. Freer, and Rob Lamb. 2021. "A History of TOPMODEL." *Hydrology and Earth System Sciences* 25 (2): 527–49. https://doi.org/10.5194/hess-25-527-2021.

Frame, Jonathan M., Frederik Kratzert, Daniel Klotz, Martin Gauch, Guy Shalev, Oren Gilon, Logan M. Qualls, Hoshin V. Gupta, and Grey S. Nearing. 2022. "Deep Learning Rainfall–Runoff Predictions of Extreme Events." *Hydrology and Earth System Sciences* 26 (13): 3377–92. https://doi.org/10.5194/hess-26-3377-2022.

Feng, D., Liu, J., Lawson, K., and Shen, C.: Differentiable, Learnable, Regionalized Process-Based Models With Multiphysical Outputs can Approach State-Of-The-Art Hydrologic

Prediction Accuracy, Water Resources Research, 58, e2022WR032 404, https://doi.org/https://doi.org/10.1029/2022WR032404, e2022WR032404 2022WR032404, 2022.

Houska, Tobias, Philipp Kraft, Alejandro Chamorro-Chavez, and Lutz Breuer. 2015. “SPOTting Model Parameters Using a Ready-Made Python Package.” Edited by Dafeng Hui. *PLOS ONE* 10 (12): e0145180. https://doi.org/10.1371/journal.pone.0145180.

Kratzert, Frederik, Daniel Klotz, Guy Shalev, Günter Klambauer, Sepp Hochreiter, and Grey Nearing. 2019. “Towards Learning Universal, Regional, and Local Hydrological Behaviors via Machine Learning Applied to Large-Sample Datasets.” *Hydrology and Earth System Sciences* 23 (12): 5089–5110. https://doi.org/10.5194/hess-23-5089-2019.

Vrugt, Jasper A. 2016. “Markov Chain Monte Carlo Simulation Using the DREAM Software Package: Theory, Concepts, and MATLAB Implementation.” *Environmental Modelling & Software* 75 (January):273–316. https://doi.org/10.1016/j.envsoft.2015.08.013.