Letter to Editor egusphere-2025-425

We would like to thank the editor for managing the review process and the referees for their constructive comments, which significantly improved our manuscript.

Following the suggestions from Referee #2 (please refer to the discussion 'Reply on RC2'), we implemented a grid-search-based hyperparameter tuning procedure for the stand-alone LSTM used in our study. Specifically, we optimized three hyperparameters: the number of hidden states, the initial learning rate, and the dropout rate (frequently identified as critical in DL models). Additionally, besides the stand-alone LSTM and the hybrid model, we calibrated 196 individual HBV models locally, one for each catchment analyzed in this study. The calibration approach and results for the HBV models are detailed and discussed in the newly added Appendix B. In response to Reviewer #1's recommendation, we explored two further strategies to address the saturation behavior of the LSTM under extreme scenarios. These strategies included modifying the loss function and developing a specialized LSTM variant (sLSTM). The methodological description and results from these tests are provided in the newly included Appendix D. Furthermore, we refined the methods section and expanded our discussion of results to integrate these modifications effectively. We also incorporated all minor suggestions provided by both reviewers.

Below, we summarize the reviewers' major comments and outline the specific manuscript revisions made in response. Reviewer comments are highlighted in blue, with our responses in black. Key revisions in the manuscript are marked in **bold**. For a detail response to the reviewer comments, we refer to the discussion in HESSD.

Referee Comments: RC1, Basil Kraft:

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.

Response: In the first submission, we discuss the results from larger LSTM networks trained on more diverse datasets. However, following the referee's suggestion, we tested two additional strategies to mitigate the saturation effect, namely, a modified loss function and the sLSTM architecture. **Appendix D** gives the description of methods and results from these models. We also adjusted our discussion.

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.

Response: **Section 2.4 (Line 147-151)** of the revised manuscript includes an explanation of the chosen training approach (seq-to-one for the LSTM and seq-to-seq for the Hybrid) and the respective sequence lengths used for both the 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.

Response: We modified **Section 2.4 (Line 152-157)** of the manuscript for a clearer explanation of the use of pet_sim (mm/d) as an explicit input to the HBV components of the Hybrid model.

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.

Response: We revised **Section 2.5 (Line 169-174)** of the manuscript to include a clarification with respect to the said source of error.

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.

Response: We modified **sections 3.1 (Line 227-230)** to include the RMSE for the predictions of the stand-alone LSTM for extreme events within the test data. **In section 3.3 (Line 295-297)** we talk about the LSTM saturation (%) for these events without the input of synthetic precipitation data.

RC2: Anonymous:

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.

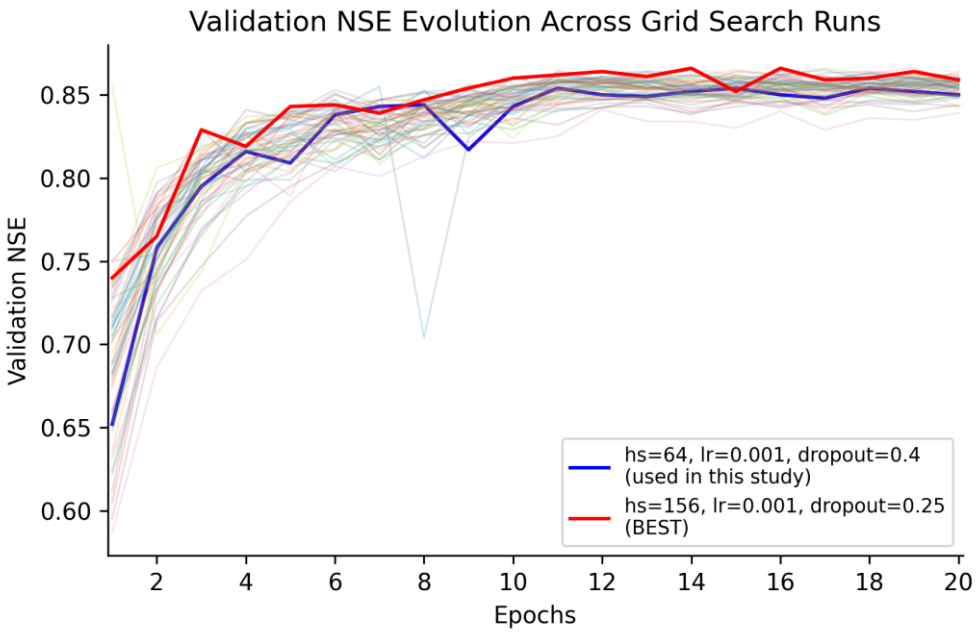

Fig 1: Validation NSE for grid search based hyperparameter tuning. The best model (in red) does not affect results of this study in comparison to the model originally used in the study (in blue).

Response: On the referee's suggestion, we implemented a grid search based hyperparameter tuning for our LSTM model. The following hyperparameters were searched for from among the given values, respectively:

1. Number of Hidden states: 64, 96, 128, 156, 196, 224, 256
2. Initial learning rate 0.001, 0.0005
3. Dropout rate: 0.0, 0.25, 0.4, 0.5

We trained a single model for every combination of the hyperparameters for the training period (01.10.1995 - 30.09.2005) and tested them for the validation period (01.10.2005 - 30.09.2010). The evolution of the validation performance of these models in terms of the CDF of the catchment-wise NSE is given in Fig. 1. The best model hyperparameter combination is given in red, which is a hidden size of 156, an initial learning rate of 0.001 and a dropout rate of 0.25. The validation performance of the hyperparameters used in this study is shown in blue. Though the slight difference between the two, might be significant for model benchmarking, it is not critical for the experimental set-up in our study. Training the LSTM with the best hyperparameter setting does not change the 'theoretical prediction limit' or the 'design limits' significantly, and hence does not change the nature of results in this study. Also, when we test ensemble models with differing hidden states (ranging from 8 to 2048, while keeping all other hyperparameters unchanged), it does not affect the nature of our results. Thus, we believe that, the results from the hyperparameter tuning or any inferences drawn from it can be left out from the revised manuscript. Instead, we stick to the adopted hyperparameters based on previous studies by Acuna Espinoza et al., (2024) and Kratzert et al., (2019). We also believe that it is justified to adopt the same hyperparameters of the LSTM for the Hybrid model, for a fair comparison.

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 locally calibrated a variant of the HBV (Seibert, J. (2005)) model for all the catchments in our study. We include the description of the calibration process and discuss the results **in Appendix B** in the revised manuscript.

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: We added the metrics High Flow Bias, fraction of Missed Peaks and Peak Mean Absolute Percentage Error in **Fig. 1** of the revised manuscript and also modified the **section 3.1 (Line 220-227).**

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 modified **section 3.2 (268-269, 277-280, 281-285)** to include a clarification of why we show only three flood events in Fig. 2. We added additional results in **Appendix B.** 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 a research line on its own.

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 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). For instance, Froidevaux et al. (2015) found that 0–3 days of accumulated precipitation is the main driver of floods, while longer-term (4 days–1 month) antecedent precipitation has only weak, region-specific effects—especially relevant in gentler plateau areas, but negligible in Alpine catchments. While Staudinger et al. (2025) found that only 18–44% of extreme annual floods coincided with maximum precipitation, highlighting the crucial role of antecedent soil moisture and snow storage. Several attempts from our side showed that given these complexities, it is challenging to identify a consistent, clear signal across a large-sample dataset covering Switzerland, with its diverse hydrological regimes. We have improved our manuscript in this regard. We modified **section 4 (line 407-422)** of 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.

References:

Acuña Espinoza, Eduardo, Ralf Loritz, Manuel Álvarez Chaves, Nicole Bäuerle, and Uwe Ehret. 2024. "To Bucket or Not to Bucket? Analyzing the Performance and Interpretability of Hybrid Hydrological Models with Dynamic Parameterization." *Hydrology and Earth System Sciences* 28 (12): 2705–19. https://doi.org/10.5194/hess-28-2705-2024.

Froidevaux, P., J. Schwanbeck, R. Weingartner, C. Chevalier, and O. Martius. 2015. "Flood Triggering in Switzerland: The Role of Daily to Monthly Preceding Precipitation." *Hydrology and Earth System Sciences* 19 (9): 3903–24. https://doi.org/10.5194/hess-19-3903-2015.

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.

Seibert, J. (2005) HBV Light Version 2. User's Manual. Department of Physical Geography and Quaternary Geology, Stockholm University, Stockholm.

Staudinger, Maria, Martina Kauzlaric, Alexandre Mas, Guillaume Evin, Benoit Hingray, and Daniel Viviroli. 2025. "The Role of Antecedent Conditions in Translating Precipitation Events into Extreme Floods at the Catchment Scale and in a Large-Basin Context." *Natural Hazards and Earth System Sciences* 25 (1): 247–65. https://doi.org/10.5194/nhess-25-247-2025.