**Review of "Beyond Observed Extremes: Can Hybrid Deep Learning Models Improve Flood Prediction?" by Guan et al. (2025)**

https://egusphere.copernicus.org/preprints/2025/egusphere-2025-1509/

Summary:

This study assesses the extrapolation skill of traditional conceptual hydrological models (HBV, GR4J and SIMHYD), a purely data-driven model (LSTM), and three hybrid versions in predicting unprecedented floods on a large-sample data set from Austria. The findings support current issues identified by the community: LSTM-based models outperform traditional models in in-sample interpolation, improve somewhat in extrapolation, but show extrapolation limits. Many questions and hypotheses about extrapolation skill are raised, but left uninvestigated and unanswered.

Overall evaluation:

The research questions raised in this paper definitely are within the scope of HESS. Unfortunately, these questions are mainly left unanswered. The study brings little additional insight to what has been proposed in the literature so far. Given some shortcomings in the methodology and a too shallow discussion, I recommend a rejection at this point and encourage the authors to follow up on the questions they have raised in this manuscript – they are relevant and point in the right direction; answers will turn this effort into a meaningful contribution.

Specific comments:

- Abstract: Large parts of the abstract read like a wishlist for the future; I would rather like to understand from the abstract what is achieved within the paper, and read about future work in the Conclusions/Discussion/Outlook section.
- L. 36: It comes as a bit of a surprise to me that data-driven models should be able to solve the extrapolation problem, since they purely rely on provided data and are said to lack extrapolation/generalization skill (although they have shown great skill in specific settings) – one or two additional sentences might help explaining why specifically data-driven models are expected to suffer less from "training bias" than physics-based models (I assume the point here is to allow for more flexibility in modeling as compared to rigid process-based structures that might be too much of a constraint?).
- L. 65: The paragraph on the actual contribution of this paper seems too short. First, it needs to be outlined which broad types of hybrid modeling approaches exist (post-processing of model outputs, time-dependent parameters, data-driven sub-components, …) and which ones the authors assume here (how are the physics-based models and the LSTM "integrated"?). Second, how do the authors intend to analyze unprecedented flood events? Will this be a synthetic analysis with synthetic forcing data? Or will certain events be excluded from

training to be used for testing? The paper lacks a clear formulation of the specific research question to be addressed.

- L. 80: “The historical observation period spans from 1981 to 2017 for most catchments, providing a comprehensive dataset for model evaluation.” I assume that not the whole period was actually used for evaluation, but there was a split into training, validation, and testing?
- L. 125: “Given that we focus on simulating extremes, where input data are highly skewed, Min-Max scaling was used for feature preprocessing instead of standardization to better preserve extreme values.” – This needs more attention, please explain in more detail what the min-max scaling does and whether, in principle, it would allow for predicting something outside of the training data range (concerning theoretical limits of extrapolation skill of LSTMs). Current research on these aspects exists (e.g., Baste et al., 2025), such that the literature review needs to be extended in that regard.
- Section 3.2: Does the LSTM predict sequence-to-1, or sequence-to-sequence?
- L. 136: “…thus generating intermediate state variables that reflect key physical processes that drive floods (see Table A1 for a summary of the intermediate state variables)” – since this is the only information provided about what makes the hybrid models hybrid, these variables need to be listed in the main body of the manuscript. Also, it needs to be discussed that the three conceptual models provide different intermediate variables, and hence the number and types of inputs differ between the three hybrid versions.
- L. 167: Why did the authors choose different loss functions for the conceptual hydrological models and the LSTM?
- Section 4.1: The NSE values achieved by the three conceptual models and also the pure LSTM are relatively low – can the authors provide an explanation? What is hard about modeling these catchments?
- L. 195: The pure LSTM does not outperform GR4J in validation as measured by NSE; only very slightly when looking at RMSE.
- L. 200: It seems all three hybrid models perform practically equally well, which triggers the question whether the provided inputs are actually informative for the LSTM, or whether it simply benefits from additional input channels that offer more flexibility. A test with randomized/nonsense input could provide some insight (idea presented by Bárdossy et al. at EGU25).
- L. 210: “RMSE values for out-of-sample floods were consistently higher across all models, likely due to the higher discharge magnitudes in these events, which lead to larger absolute simulation errors despite similar NSE values.” – the absolute value of simulation errors enters RMSE and NSE in the same way (squared errors); if only the larger discrepancies for out-of-sample flood events were causing the deviations, the RMSE should actually show more similar results between in- and out-of-sample, because the root is reducing this effect. What about the normalization in the NSE? Did the authors normalize with the variance of the observations in-sample vs. out-of-sample? Then the in-sample-

variance would be much smaller than the out-of-sample variance, which would explain that NSE stays relatively constant during both periods.

- L. 221: “The underestimation of in-sample flood peaks may be partly due to issues with the quality of rainfall and discharge data or catchment-specific characteristics not fully captured by the models.” – the pure LSTM and also the hybrid versions do not care about the quality of rainfall so much, they can learn to bias-correct. I would rather argue that conceptual models are limited by volume balance closure (so yes, they struggle if rainfall or discharge data is imperfect), while data-driven models are limited by theoretical extrapolation limits (the way the activation functions are defined is designed for interpolation, so generally good skill in interpolation, but not suitable for extrapolation; again, see e.g. Baste et al., 2025).
- Section 4.4: The discussion raises many questions, but fails to answer at least some of them. I agree that, apparently, the different conceptual hydrological models leave different amounts of room for improvement (this is an almost trivial conclusion given that they show differences in performance); however, the fact that all three hybrid versions practically show the same performance in my view clearly triggers the questions whether (1) it is not the modeled information that helps the LSTM, but the added degrees of freedom, and (2) whether this shared performance shows us the performance limit given those degrees of freedom.
- L. 272: “We suggest synthetic design storms (Macdonald et al., 2025) can be generated carefully to demonstrate how hidden states in LSTMs saturate near the activation function bounds, capping the model’s capacity to represent unprecedented events. Specifically, we can simulate a series of increasing storm intensities beyond the training range, evaluate the model’s response to these inputs by tracking activation values in hidden states, and examine whether LSTM predictions plateau due to activation function saturation” – Yes, this is the type of analysis I would suggest to answer these questions (and which is currently pursued in the community, see, e.g., Baste et al., 2025, or Martel et al, 2024, both in the same journal). Why did the authors not try something in this direction in this case study? To me, the findings presented here are not new, surprising or enlightening; rather, they raise questions that are already being discussed by the hydrological community.
- L. 281: “Our current hybridization strategy, which integrates hydrological states, lays a foundation for such advancements…” – This statement is too strong. This type of “integration” has been proposed before (Frame et al., 2021) and used many times (e.g., Martel et al., 2024), and it is a very limited type of integration, given that model-predicted intermediate state variables are passed to the LSTM as inputs, which it can basically choose to use or ignore, whatever the data say. Several authors have pointed out that this post-processing type of hybridization is a very weak constraint for an LSTM (e.g., Frame et al., 2021; Álvarez Chaves et al., 2025). So neither is the proposed strategy new, nor will it be a basis for actual integration of physics into machine learning.
- L. 286: “In this study, models were implemented independently for each test catchment, without considering the potential benefits of a regional approach.” – only now I learn that the LSTM models weren’t trained over all catchments. It

has been discussed broadly in the community that LSTMs perform much worse when trained on individual basins (Kratzert et al., 2024), and the open question is whether this is only due to the vast amount of data available, or because the LSTM needs the diversity in the value ranges for extrapolation skill (e.g., Frame et al., 2022; Acuna et al., 2025). Training the LSTM on individual basins without any additional tweak to improve extrapolation skill doesn't provide an advance for the community.

References:


- *Acuña Espinoza, E., Loritz, R., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., and Ehret, U. (2025): Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events, Hydrol. Earth Syst. Sci., 29, 1277–1294, https://doi.org/10.5194/hess-29-1277-2025.*
- *Álvarez Chaves, M., Acuña Espinoza, E., Ehret, U., and Guthke, A. (2025): When physics gets in the way: an entropy-based evaluation of conceptual constraints in hybrid hydrological models, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2025-1699.*
- *Bardossy, A., Seidel, J., and Acuna, E. (2025): Is Artificial Intelligence the Ultimate Solution for Hydrological Modelling?, EGU General Assembly 2025, Vienna, Austria, 27 Apr–2 May 2025, EGU25-12083, https://doi.org/10.5194/egusphere-egu25-12083.*
- *Baste, S., Klotz, D., Espinoza, E. A., Bardossy, A., and Loritz, R. (2025): Unveiling the Limits of Deep Learning Models in Hydrological Extrapolation Tasks, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2025-425.*
- *Frame, J.M., F. Kratzert, A. Raney II, M. Rahman, F.R. Salas, and G.S. Nearing (2021): "Post-Post-Processing the National Water Model with Long Short-Term Memory Networks for Streamflow Predictions and Model Diagnostics." Journal of the American Water Resources Association 57(6): 885–905. https://doi.org/10.1111/1752-1688.12964.*
- *Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S. (2022): Deep learning rainfall–runoff predictions of extreme events, Hydrol. Earth Syst. Sci., 26, 3377–3392, https://doi.org/10.5194/hess-26-3377-2022.*
- *Kratzert, F., Gauch, M., Klotz, D., and Nearing, G. (2024): HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin, Hydrol. Earth Syst. Sci., 28, 4187–4201, https://doi.org/10.5194/hess-28-4187-2024.*
- *Martel, J.-L., Arsenault, R., Turcotte, R., Castañeda-Gonzalez, M., Brissette, F., Armstrong, W., Mailhot, E., Pelletier-Dumont, J., Lachance-Cloutier, S., Rondeau-Genesse, G., and Caron, L.-P. (2024): Exploring the ability of LSTM-based hydrological models to simulate streamflow time series for flood frequency analysis, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2024-2134.*