Response to RC1: 'Comment on egusphere-2024-2147', Basil Kraft

We want to thank the referee for the detailed evaluation of our paper. In this document we answer the questions, comments and suggestions given. We will address those comments individually. For clarity, the original comments posted by the referee are written in blue.

Short summary and highlights

This study examines how well an established hybrid hydrological model generalizes under extreme conditions, comparing its performance to a data-driven model (LSTM) and a conceptual hydrological model. They split the discharge time series into training and test sets based on the magnitude of events within a hydrological year. Normal years were used for training, while extreme ones were used for testing.

The results indicate that the LSTM model performs best across all test data. For extreme events, the performance gap between the LSTM and hybrid models narrows, and eventually disappears for events with a 50-year return period or above. All models face challenges with extremely large discharge events, and the LSTM shows limitations due to saturation, where it cannot produce outputs beyond a certain threshold. While the hybrid and conceptual models slightly better reproduce extreme event patterns, they all exhibit substantial errors under these conditions.

Overall, the study suggests that while hybrid models may offer a modest advantage for modeling extreme discharge events, the LSTM performs better across all conditions. Thus, the choice of the best model should be guided by the specific objectives, such as whether the focus is on overall discharge or extreme conditions, as well as considerations for control and interpretability.

The study offers a critical perspective on hybrid models, which have gained popularity in recent years. While the benefits of hybrid models are often promoted, a thorough evaluation is still needed. Therefore, I find the study highly relevant. The study design is clear, and the use of established models lends credibility to the findings.

We thank the referee for the well-structured summary of our paper.

Major remarks

1. Could you explain why a fixed number of epochs was chosen for training instead of early stopping?

Response: The choice of using a fixed number of epochs instead of early stopping was done to keep consistency with the two studies we used as a base for our experiments ([1] and [2]). Moreover, early stopping based on validation scores requires a third split, a validation split, which means that we lose additional test data.

2. The study evaluates a specific hybrid model. While this is valuable, it is important to note that these findings may not apply to all hybrid models. I suggest updating the title to “Analyzing the

generalization capabilities of a hybrid hydrological model for extrapolation to extreme events", and to discuss this limitation more in-depth.

Response: Thank you for the suggestion. We will modify the title accordingly. We did discuss the limitations in the original manuscript. In section 3.5, between lines 223-226, we stated that:

> "The comparison that we presented here was done using the hybrid architecture proposed by Feng et al. (2022). This architecture was chosen because it gave a competitive performance with LSTM in their original experiment and because the code was open source. Other hybrid model architectures might give different results, and we encourage the hydrological community to expand the test cases presented here."

However, we do admit that this is perhaps a bit spartan and will expand upon it in the revised manuscript.

3. If feasible, a comparison regarding low flow conditions would be appreciated.

Response: Thank you for the suggestion. Even though a comparison regarding low flow conditions would be interesting, our whole study and the proposed methodology focus on high flow conditions. Consequently, evaluating low flow conditions are outside of the scope of the current paper.

4. The overall structure could be improved. Currently, the Results section includes some Method descriptions, and within the Results and Discussion section, results and discussion are not clearly separated into paragraphs.

Response: Thank you for the suggestion. In a revised version of the manuscript, we will clearly separate the Results from the Method. However, with regard to the separation of Results and Discussion, we would like to keep these sections together, as we believe this creates a concise and fluent manuscript, and allows the reader to better understand the narrative of the study.

5. Although it may be beyond the scope of this study, examining the robustness of the (interpretable) hybrid model parameters would be very interesting.

Response: Thank you for the suggestion. In the revised version of the manuscript, we will add an Appendix discussing the variation of the dynamic parameters of the hybrid model.

6. The discussion falls a bit short in general. I would appreciate a more in-depth discussion of the findings.

Response: Thank you for the suggestion. In a revised version of the manuscript, we will provide a brief discussion of other advantages provided by the hybrid models, and some hypotheses about how our findings might generalize to other models.

Minor remarks

L11 Clarify “the latter study”

Response: We will modify this.

L9-L13 Consider revising the first paragraph for improved clarity.

Response: We will do that.

L35/L37 Choose either “large events” or “extreme events” for consistency.

Response: Thank you for pointing this out, we will use only the term “extreme”.

L36 Perhaps rephrase to “How does a hybrid model compare to a process-based model”? See Major remarks.

Response: We will modify this in a revised version of the manuscript.

L38 Specify the type of advantage being discussed.

Response: We will modify this in a revised version of the manuscript.

L41 Rephrase to: “In Section 3, we compare the results of various tests that assess generalization capabilities.”

Response: We will modify this in a revised version of the manuscript.

L50 A different name for the hybrid model might be clearer; “$\delta(\gamma t, \beta t)$” is somewhat cumbersome.

Response: This was done to maintain consistency with Feng et al., (2022), so we prefer keeping this name.

L53 Consider using “Experimental setup” as the section title.

Response: “Experimental setup” is a more general term. We would like to keep the title as “Data handling: training/test split” because it is more specific of what the subsection is about.

L60 Move “(a water year is defined as the period from October 1 to September 30)” to where you first mention“ water years” on line 59.

Response: We will modify this in a revised version of the manuscript.

L71 Provide a reference for USGS NWIS.

Response: We will add the reference in a revised version of the manuscript.

L74ff Please rephrase for clarity.

Response: We will modify this in a revised version of the manuscript.

L89 Add a comma before “respectively”.

Response: We will modify this in a revised version of the manuscript.

L89 Define NSE and provide a reference.

Response: We will modify this in a revised version of the manuscript.

L90 Specify which study is referred to as “the original study”.

Response: We will modify this in a revised version of the manuscript.

L91 Provide full names and add a space between variable names and units.

Response: We will modify this in a revised version of the manuscript.

L92 Include categories of static variables, such as “27 static variables describing topography, soil properties, and land surface cover . . . ”.

Response: We will modify this in a revised version of the manuscript.

L96 Briefly mention the benefits of using ensemble methods.

Response: We will add a short sentence, about the benefits of using ensembles, in a revised version of the manuscript.

L98 Replace “As mentioned in the introduction,” with “For the hybrid model architecture,”

Response: We will modify this in a revised version of the manuscript.

L103 Rephrase to: “210 parameters (16 ensemble members, each with 13 HBV parameters plus 2 routing parameters)”.

Response: We will modify this in a revised version of the manuscript.

L105 Did you use the same warm-up period for the LSTM, considering it also needs to initialize its states?

Response: The warmup period of the stand-alone LSTM is done during the 365 timesteps considered in the sequence length, so the warmup periods are equivalent. The main difference is that, with the LSTM, we only retrieved one value after the warmup period (seq-one) while in the hybrid we retrieved a whole year (seq-seq).

L107 Change “365 elements” to “365 time steps”.

Response: We will modify this in a revised version of the manuscript.

L108 Provide the full name of the YAML file.

Response: We will modify this in a revised version of the manuscript.

L113 Rephrase: “To ensure a comprehensive comparison of the model spectrum,”

Response: We will modify this in a revised version of the manuscript.

L114 Rephrase to “14 parameters (12 HBV plus 2 routing)”.

Response: We will modify this in a revised version of the manuscript.

L115 Note that this HBV instance has 12 parameters, while the hybrid model has 13.

Response: Yes. We explain the reason for this difference between lines 115-117.

L122 Rephrase to: “their generalization capabilities in the time domain to extreme events.”

Response: We will modify this in a revised version of the manuscript.

L124 Consider “Model comparison for the whole test period” as a section title.

Response: We would like to keep the word “performance” because it is more specific of what the subsection is about.

L125 Introduce NSE before line 89.

Response: Thank you for pointing this out. We will modify this in a revised version of the manuscript.

L125 Change “reported for each model”.

Response: We will modify this in a revised version of the manuscript.

L126 Rephrase: “The LSTM outperforms the hybrid model, with a median NSE of 0.75 and 0.71, respectively.”

Response: We will modify this in a revised version of the manuscript.

L127 Rephrase: “The hybrid model has a median NSE of 0.64. This indicates that even with a different training-test split than the usual temporally contiguous subsets, our results align with those reported by Feng et al. (2022) and Acuña Espinoza et al. (2024), where the same model ranking was observed.”

Response: Thank you for the suggestion. We will modify this in a revised version of the manuscript.

Fig. 2 Caption: “Cumulative Density Functions (CDF)”

Response: We will add the (CDF) abbreviation to the figure´s caption.

L131 Consider “Model Comparison for peak flows” as a section title.

Response: We would like to keep the word “performance” because it is more specific of what the subsection is about.

L135ff This section sounds more like methods; consider restructuring.

Response: Thank you for the suggestion. We will move this part to the method´s section.

L148 Omit: “This trend will also be observed in other experiments discussed in the following sections.”

Response: We will remove this sentence from a revised version of the manuscript.

L150 The sentence is quite generic. Consider removing it and discussing the point in detail later.

Response: We would like to keep this sentence because it is supported by the results we showed in that section.

Fig. 3 Add ‘)’: “The results of subplot b) show the error distribution.”

Response: We will make the respective change.

L168 Use “Eq. 1” instead of “equation 1”.

Response: We will modify this in a revised version of the manuscript.

L176 Replace “has to go out”. Maybe “has to leave the system”? Also, the water could just stay in the system and accumulate over time, right?

Response: We will modify this in a revised version of the manuscript. About the second comment, your are correct, in the last time step there can still be water stored in the system. We will make the respective changes.

L181 Omit: “This phenomenon as such is not necessarily an indication of model deficiencies.”

Response: We would like to keep this sentence because it helps us connect with the next idea.

L228 Replace “For the reasons stated in previous sections, . . . ” with “This stress-testing . . . ”

Response: We would like to keep the sentence structure as it is, because it helps us connect with the next idea.

L229 “In a practical case, one should use all the data during model training, to increase the performance of the models.” Could you mention that you mean using also extreme events, and not literally all data (because we want training/test split)?

Response: We will modify this in a revised version of the manuscript.

L259 Replace “As explained in the manuscript,” with “For our hybrid model,”

Response: We will modify this in a revised version of the manuscript.

L260 Please rephrase: “Because our experiment pipeline was executed in the NeuralHydrology package, we did first had to benchmark our model . . . ”.

Response: We will modify this in a revised version of the manuscript.

Final remarks

We would like to thank the referee for the overall positive evaluation of our manuscript and hope we could address the questions raised in a satisfactory manner.

References

1. Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. Water Resources Research, 58, e2022WR032404. https://doi.org/10.1029/2022WR032404
2. Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., & Nearing, G. S. (2022). Deep learning rainfall-runoff predictions of extreme events. *Hydrology and Earth System Sciences, 26*(13), 3377–3392. https://doi.org/10.5194/hess-26-3377-2022