"Analyzing the generalization capabilities of hybrid hydrological models for extrapolation to extreme events"

Manuscript #*2024-2147*

September 5, 2024

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.

Major remarks

1. Could you explain why a fixed number of epochs was chosen for training instead of early stopping?
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.
3. If feasible, a comparison regarding low flow conditions would be appreciated.
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.
5. Although it may be beyond the scope of this study, examining the robustness of the (interpretable) hybrid model parameters would be very interesting.
6. The discussion falls a bit short in general. I would appreciate a more in-depth discussion of the findings.
   a) A comparison with existing studies would be beneficial. How do your findings relate to other research in this area, for example the study mentioned by another reviewer, Song at al. (2024), 10.22541/essoar.172304428.82707157/v1?

b) There are other advantages of hybrid (and conceptual) models over neural networks, such as interpretability. I would appreciate a brief discussion of this aspect.

c) Could these findings apply to other hybrid models and domains, why (not)?

Minor remarks

Here I list some typos and suggestions for improving clarity:

L11 Clarify “the latter study”.

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

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

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

L38 Specify the type of advantage being discussed.

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

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

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

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.

L71 Provide a reference for USGS NWIS.

L74ff Please rephrase for clarity.

L89 Add a comma before “respectively”.

L89 Define NSE and provide a reference.

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

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

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

L96 Briefly mention the benefits of using ensemble methods.

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

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

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

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

L108 Provide the full name of the YAML file.

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

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

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

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

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

L125 Introduce NSE before line 89.

L125 Change “reported **for** each model”.

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

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.”

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

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

L135ff This section sounds more like methods; consider restructuring.

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

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

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

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

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?

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

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

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)?

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

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