Response to RC2: 'Comment on egusphere-2024-2147', Shijie Jiang

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.

The manuscript "Analyzing the generalization capabilities of hybrid hydrological models for extrapolation to extreme events" compares the generalization capabilities of hybrid models, LSTM networks, and process-based models for rainfall-runoff simulations, with a particular focus on extreme events. The study examines whether hybrid models provide a meaningful advantage over standalone data-driven or process-based models. The results suggest that hybrid models show marginal improvements in predicting extreme peak flows, but overall perform similarly to LSTM networks. The authors argue that given the comparable performance, the choice of model depends on user needs. Overall, the study does a great job of providing a balanced perspective on the hybrid models. The paper is valuable in stimulating further discussion in the field.

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

Major comments

1) One of the central claims for hybrid models is that they combine the predictive power of data-driven approaches with the interpretability of process-based models. However, the manuscript focuses more on marginal differences in predictive performance than on the added interpretability that might justify hybrid models. I suggest including a discussion of the trade-off between accuracy and interpretability. For example, does the hybrid model help to better understand the causes of extreme flows, such as snowmelt, soil moisture dynamics, or precipitation anomalies? Could the explicit encoding of hydrologic concepts in the hybrid model be more valuable for decision making, even if the predictive gains are minimal?

Response: Thank you for the suggestion. We agree that both performance and interpretability can be important, and that is why in previous studies, as Acuña Espinoza et al. (2024), we ran multiple experiments to evaluate model interpretability. In this study, we are tackling the question from a practitioner's point of view, in which performance is the main interest. However, in the Limitation section, we will add more emphasis on the fact that other criteria, besides model accuracy, play an important role in an integral evaluation of the model.

2) While the paper touches on model errors during extreme events, it does not provide an analysis of where and why each model is better or worse, e.g., under which geophysical, climatic, or soil conditions. This could be helpful to better understand the strengths and limitations of each model type and provide a useful guide to when hybrid / LSTM models are most beneficial.

Response: Thank you for the suggestion. In a revised version of the manuscript we will include a map, indicating for each basin, the difference in performance of the models. This way we will be able to visualize if there are geographic settings in which one model consistently outperforms the other.

3) A related comment is that while the authors conclude that the choice of model depends on user needs, the manuscript does not provide clear guidance on how to make this choice. For example, in data-poor environments where high-quality or long-term observational data may not be available, should hybrid models be preferred because they incorporate process-based knowledge that could compensate for sparse data? Is it possible to make a comparison that assumes limited data? I think it would be helpful for practitioners working in regions with poor monitoring infrastructure.

Response: Thank you for the suggestion. To evaluate if hybrid models can be trained with less data is by itself a full study, with a different set of experiments, which is beyond the scope of our study. A good overview on related studies for process- and data-based hydrological models is given by Jiang et al. (2024). Specifically for the fully integrated hydrological model (ATS), they conclude that about 4 years of data allow for robust parameter learning. From our own recent work comparing the learning ability of single-basin process-based (HBV) and data-based (LSTM) models (publication in preparation), we found that HBV learns all it can from 2-3 years of data, that the LSTM achieves good performance with 2-3 years of training data but keeps learning when more data are available, and that LSTM outperforms HBV beyond ca. 10 years of training data.

Moreover, both data-driven methods and hybrid models have shown to perform better when trained regionally. Therefore, even in data-sparse regions, one could train the models on public databases and then fine-tune it to the specific areas, which would somehow help mitigate the limited data problem. Therefore, even though this is an interesting question, we believe it is considerably outside of the scope of the current study. Nevertheless, we will include a short discussion, similar to the above paragraph, in the revised manuscript.

Specific comments:

L12, the term “out-of-sample conditions” is somewhat ambiguous. Please specify what type of generalization is meant (temporal or spatial domains).

Response: Agreed. We will better specify this term in a revised version of the manuscript.

L16, the phrase "notion of interpretability" could be clearer. What does "notion" mean in this context? It sounds vague. If interpretability is considered to be a key reason for adopting hybrid models over purely data-driven ones, it should be more clearly defined and quantified. Does interpretability mean the ability to interpret the parameters, processes, or outputs in a hydrologically meaningful way? Or are you suggesting that it's a "so-called" interpretability?

Response: The interpretability gained in this type of hybrid model is that we associate the parameters and buckets of the process-based models with interpretable processes, domains and states (baseflow, interflow, snow accumulation…). However, we argue that this type of interpretability is based on association, and the physical principles represented on process-based models, such as the HBV, have major simplifications. We will clarify this in a revised version of the manuscript.

L30, what specific structural deficiencies are you referring to here?

Response: The hybrid structure we present in this study consists of a data-driven part that predicts the parameters (static and time-varying) used to operate a process-based model. This is the same architecture type used in Acuña Espinoza et al. (2024). They show that the data-driven part, through the dynamic parameterization, is able to increase the performance of the model, compared to the stand-alone process-based benchmarks. This was attributed to the fact that process-based models present a relatively simple structure that in a lot of cases oversimplifies the actual physical processes. One example is assuming that all the flows have a linear relationship with the storage and that the storage/discharge rate does not change over time. Or that snow melting is a linear process, proportional to the difference between a threshold temperature and the air temperature.

By giving additional flexibility through the dynamic parameterization, the LSTM is able to compensate for some of these deficiencies. Acuña Espinoza et al. (2024) discussed these aspects in more detail, and that is why we refer to that publication. We believe that including that in our current manuscript would reduce the fluency of the reading, as it is not the main point we are trying to establish.

L35, the focus on "higher predictive accuracy" may overlook the fact that accuracy alone may not be the best criterion for assessing model suitability. Authors should clarify that other criteria (such as robustness, model transparency, applicability) besides accuracy may be equally important in model evaluation.

Response: We agree that other criteria, besides model accuracy, play an important role in an integral evaluation of the models. However, in this study, the main focus of our experiments is model accuracy. In a revised version of the manuscript we will indicate that while we focus on accuracy in this study, future studies can expand the comparison tests in the other points.

L100, the explanation of the hybrid model's parameterization is complex and may not be easily understood by just reading this paper. At least a clearer explanation of the buckets and parameters is needed.

Response: Thank you for the suggestion. In a revised version of the manuscript we will add a figure of the model setup in an Appendix, to better illustrate the idea.

L127 without discussing the potential limitations of the HBV model, this claim seems overly simplistic. It is useful to explain here why the HBV model underperformed, even though it has been studied in previous studies.

Response: In a revised version of the manuscript we will add some explanation of why the HBV model has a lower performance. Something similar to what we indicated in our previous response to the referee's comment on L30.

L150, again, this conclusion of equivalence is overly simplistic and could lead to believing that there are no meaningful differences between the models. Are there certain types of basins or hydrological conditions (e.g., arid basins) where one model clearly outperforms the other?

Response: As discussed in Major Comment 2, we will add a map to indicate if under certain conditions, one model outperforms the other.

L167, it's hard to read from the figure about the "slightly lower errors".

Response: In a revised version of the manuscript, we will update Figure B1 with additional runs for the hybrid model using different seeds, which will give us more information about the differences between the models.

L215, this observation is important but lacks sufficient follow-up. If the dynamic parameterization reaches its limits during extreme events, it indicates a potential flaw in the model design, but the text does not discuss how this issue could be addressed or what its implications are. Could the predefined intervals be adjusted or extended to better handle extreme events?

Response: We defined the parameter intervals according to Feng et al, (2022), which was the model we were using as a benchmark. In a revised version of the manuscript we will expand on strategies to address these implications.

L220, I am very confused here. How does the snowmelt effect indicate the potential bias in the input data? If the snowmelt flux is high, it's not surprising to see a discrepancy between precipitation and runoff. This statement also raises the question of a structural flaw in the HBV model, but it is not elaborated. I'm left wondering what specific deficiencies in the snow module are responsible for the poor performance and how these deficiencies could be addressed in future work. For example, is the snowmelt process not adequately modeled due to insufficient temperature data, or is the parameterization of the snow module too simplistic?

Response: In this paragraph we are looking at possible causes of why the hybrid model underestimates the peak discharges. We mentioned that for the specific events presented for basins 14182500 and 14138800, the cumulative water volume that comes from precipitation is smaller than the cumulative water volume given by the observed discharge. Given the mass conservative structure of the hybrid model, the simulated values will therefore be smaller than the observed discharge, unless most of the simulated discharge comes from snowmelt. However, given the flow underestimation, this is not the case. In a revised version of our manuscript, we will make the explanation more understandable.

We also agree with the referee that two possible deficiencies in the snow module could be insufficient temperature data and an overly simplistic parameterization of the snow module. However, because we are conducting a regional study in 531 basins, looking in detail at model deficiencies in each basin is not feasible, nor is it the main point of our study.

L225, it's vague and doesn't provide enough insight into what types of hybrid architectures might yield different results. In my opinion, the hybrid model used in this paper considers model with a conceptual model as the backbone and neural networks for parameter learning. It would be more actionable to point out some other types of hybrid models, e.g., component replacement or more conceptual frameworks (e.g., https://hess.copernicus.org/articles/26/1579/2022/) that might address some of the limitations identified in the study.

Response: In a revised version of the manuscript we will expand upon this point, and add the reference provided by the referee.

L230, I'm afraid this recommendation is too general and simplistic…

Response: If we understood correctly, the recommendation the referee refers to is: “In a practical case, one should use all the data during model training, to increase the performance of the models.” We would argue that this recommendation is correct, and it was previously stated in Nevo et al (2022).

L241, is it possible to use more precise numbers or statistical analysis to support the claim of “slight” outperformance. If the differences are marginal, do you think they might still matter in practical scenarios?

Response: We did additional runs for the hybrid model using different seeds. In a revised version of the manuscript we will report those results, which will provide more information.

L245, the mention of "possible bias in the input data" is speculative without further analysis. And if that's the case, does it imply that LSTM is insensitive to the bias?

Response: In the analysis accompanying Figure 6 of our manuscript, we point to precipitation bias as a possible cause for the peak underestimation in the hybrid models. Similar discussions have been carried out in the literature, indicating that biases in precipitation measurements can be caused by point uncertainty, interpolation uncertainty, and equipment malfunction (Westerberg & McMillan, 2015; Bárdossy & Anwar, 2023), especially if one is working with catchment-averaged values. Therefore, we believe that our hypothesis is correctly justified. Doing a bias analysis for the whole CAMELS-US dataset is outside of the scope of our current publication.

About the second question, the LSTM does not have a mass conservative structure, and therefore, systematic biases in the inputs can be accounted for. We will add this discussion and references to the revised manuscript.

L249, the statement about dynamic parameterization is not sufficiently elaborated. It doesn't provide enough detail about how this adaptation happens or why it is particularly useful for extreme events. Also, the comparison with LSTM gating is interesting, but lacks further discussion.

In a revised version of the manuscript we will expand upon this point.

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. Acuña Espinoza, E., Loritz, R., Álvarez Chaves, M., Bäuerle, N., & Ehret, U. (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), 2705–2719. https://doi.org/10.5194/hess-28-2705-2024
2. Bárdossy, A., & Anwar, F. (2023). Why do our rainfall–runoff models keep underestimating the peak flows? *Hydrology and Earth System Sciences, 27*(10), 1987–2000. https://doi.org/10.5194/hess-27-1987-2023
3. 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
4. Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., Nearing, G., Shalev, G., Noga, H., Shavitt, I., Yuklea, L., Royz, M., Giladi, N., Peled Levi, N., Reich, O., Gilon, O., Maor, R., Timnat, S., Shechter, T., Anisimov, V., Gigi, Y., Levin, Y., Moshe, Z., Ben-Haim, Z., Hassidim, A., & Matias, Y. (2022). Flood forecasting with machine learning models in an operational framework. *Hydrology and Earth System Sciences, 26*(15), 4013–4032. https://doi.org/10.5194/hess-26-4013-2022
5. Westerberg, I. K., & McMillan, H. K. (2015). Uncertainty in hydrological signatures. *Hydrology and Earth System Sciences, 19*(9), 3951–3968. https://doi.org/10.5194/hess-19-3951-2015
6. Jiang, P., Shuai, P., Sun, A. Y., and Chen, X.: Optimizing parameter learning and calibration in an integrated hydrological model: Impact of observation length and information, Journal of Hydrology, 643, 131889, https://doi.org/10.1016/j.jhydrol.2024.131889, 2024.