General remark about benchmark models and comparability of our results to other model applications

In our original publication, before running any extrapolation experiments, we benchmarked our model implementation against the study from Feng et al, (2022), which was the publicly available model at the time. We can see in Fig A1 of our original manuscript (which we repeat below for clarity), that the models' performance under the original conditions is almost identical.

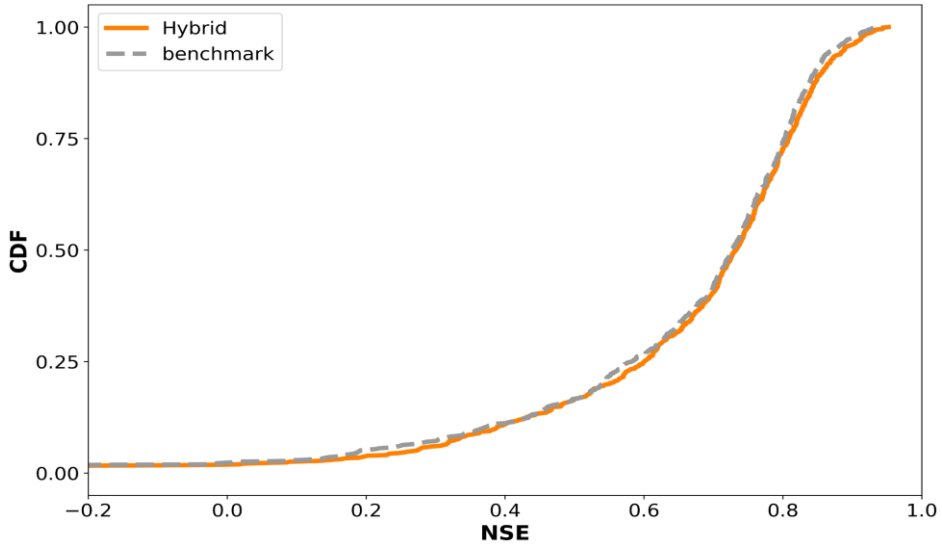

Figure A1. Comparison of model performance against benchmark model – Feng et at, (2022).

Only after our model implementation was verified, we ran the extrapolation experiments, with exactly the same model characteristics we used to run our benchmarks. To our knowledge, we followed the best practices to ensure a fair and reproducible model comparison based on benchmark models reported in the scientific literature.

We would also like to stress that while we appreciate the efforts to compare our results to new model setups (dHBV1.1), the architecture we will refer to in our manuscript is the one proposed by Feng et al. (2022), due to the reasons stated above.

About biased interests

We want to clarify that we do not have any interest in favoring one model over the other, nor we want to "bring everything we got on the LSTM side while not doing much on the dHBV side" as suggested by Chaopeng Shen. For the LSTM we used different initializations, acting as an ensemble, because multiple publications ([2], [3], [4]) have shown that this technique produces more robust results. For the Hybrid model, we used a single initialization, because

that was the implementation done by Feng et at (2022), which is the one we were using as a reference.

About the differences between the models (First comment by Chaopeng Shen in CC3)

We appreciate that Chaopeng Shen ran additional experiments to test the influence on random seeds on dHBV1.0 performance. We also ran our hybrid model using multiple seeds, results we show in Fig B1. For higher return periods, in some cases the hybrid model performs slightly better, for other cases the LSTM performs slightly better. This is in accordance with what we expressed in the original text (Line 160-162), so there is no need for changing our original conclusions. We will include the updated figures and the updated results in a revised version of the manuscript.

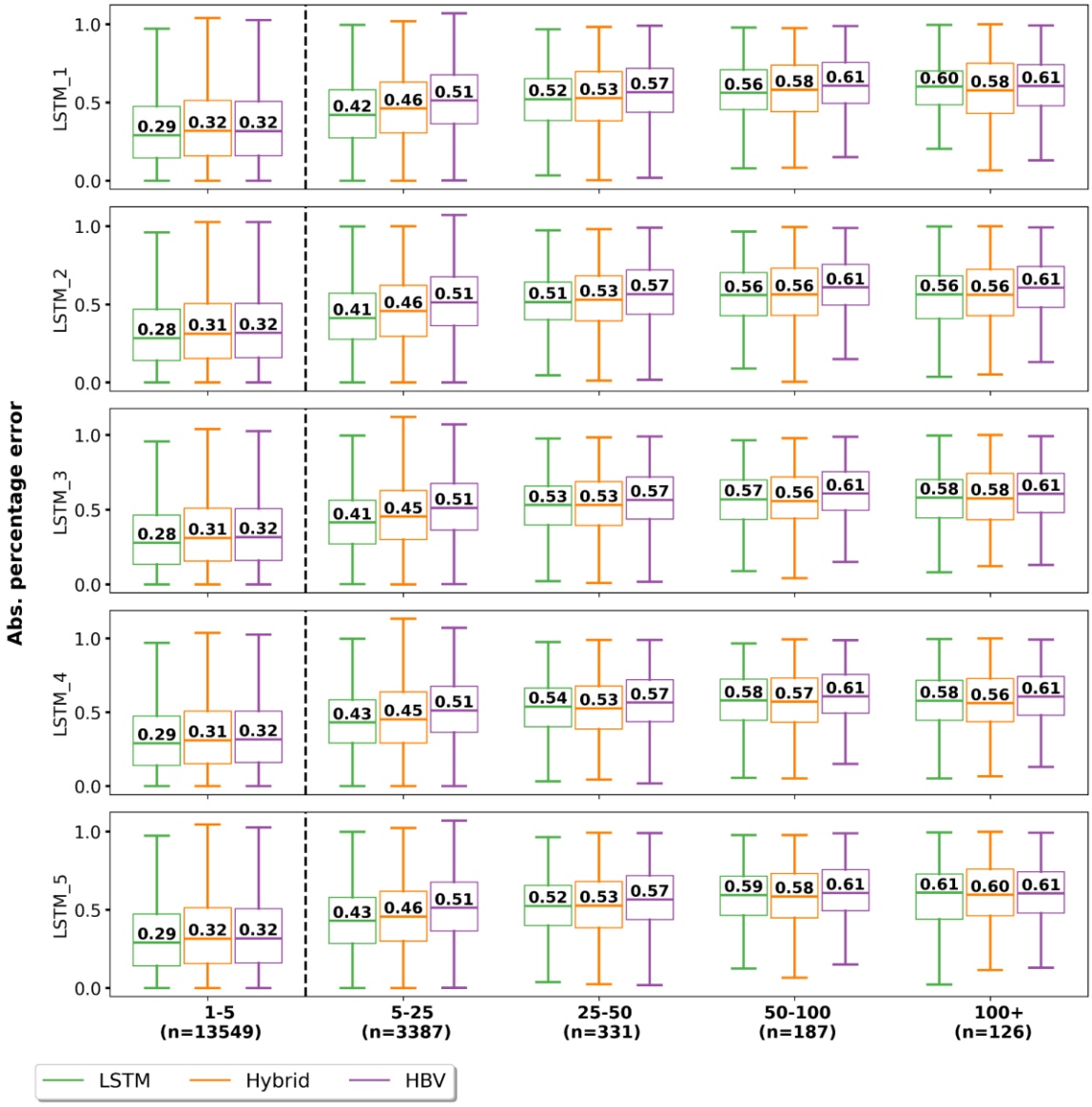

Figure B1 (updated). Variation in absolute percentage error due to random initialization of the models (LSTM and Hybrid)

Chaopeng Shen indicates in his community comment, egusphere-2024-2147 CC3, that there are differences in the results he reported in his community comment, and the results we are reporting in our manuscript. We further analyzed these differences.

We compared the median Absolute Percentage Error from our results (Figure B1, shown above) with those from Chaopeng Shen's (Figure CC1, referenced in comment egusphere-2024-2147). The differences between the two sets of values were calculated and summarized in the table below

Median Abs-Percentage Error			
Return period	**NH-Implementation**	**Chaopeng Shen´s Implementation**	**Error**
5-25	**0.46**	**0.45**	**2.2%**
25-50	**0.53**	**0.51**	**3.8%**
50-100	**0.57**	**0.54**	**5.3%**
100+	**0.58**	**0.55**	**5.2%**
Average			**4.1%**

We can see that the average difference is 4.1%, with a maximum difference of 5.3 %. We argue that these differences are small, especially if we consider the variation of the metric with the different random seeds. Reasons for the possible differences were discussed in our previous response. Therefore, we do not consider it necessary to modify the conclusions stated in our manuscript.

<u>About the different experiment designs (comment "the point about input scaler" in CC3 by Chaopeng Shen)</u>

In the community comment Chaopeng Shen shows an alternative experimental design, in which he does an additional temporal split of the data. He indicates that this experimental design cleanly separates the training and testing data.

We argue that our experimental design also does a clear and clean separation of the training and testing data. As explained in the manuscript, the training and testing conditions are clearly different, because we separate the water years based on the return period, and separate the two regimes by a buffer period as long as the model's sequence length, so absolutely no data

leakage from testing to training is possible. Additionally, by not doing an extra temporal separation, we are able to select training and testing years from the whole data record (1980-2014), which increases the size of our training and testing sets, and help us produce robust results.

Chaopeng Shen also indicates, regarding our training/test split, that "we don't encounter scenarios where we know both the historical and future time series and test in the middle of the time series". However, this never was the idea behind the setup from Frame et al. (2022). Indeed, as we have emphasized multiple times in our manuscript, the experimental design was intended as a stress-test to gain insight into the model's behavior during extreme events. For an operational model, one would not hold out any years, especially those containing extreme events.

About the comment "the final point about ensemble" in CC3 by Shaopeng Shen

About different interpretations of what defines an ensemble prediction, please see our reply 2 in AC2. About the suggestion by Chaopeng Shen that we "bring everything we've got on the LSTM side while not doing much on the dHBV side.", please see our above comment "about biased interests".

Sincerely,

Eduardo Acuna, on behalf of all co-authors

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. Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall-runoff prediction at multiple timescales with a single Long Short-Term Memory network. *Hydrology and Earth System Sciences, 25*(4), 2045–2062. https://doi.org/10.5194/hess-25-2045-2021

3. Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (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

4. Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., & Dadson, S. J. (2021). Benchmarking data-driven rainfall-runoff models in Great Britain: A comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. Hydrology and Earth System Sciences, 25(10), 5517–5534. https://doi.org/10.5194/hess-25-5517-2021