The authors (referred to as Espinoza24 thereafter) have made a valuable contribution with their analysis. This comparison provides useful insights. Espinoza24 demonstrates that the LSTM model performed moderately better than the hybrid model (NH-hybrid) for return periods of 5-10, 10-25, and 50-100 years, while the NH-hybrid was slightly superior for return periods exceeding 100 years. They also showed their Hybrid model's results are comparable to observed soil moisture which is encouraging.

We have run similar experiments on our end, which show that our version of single hybrid model, dHBV, outperformed LSTM in nearly all return-period categories. These results are documented here: https://t.co/BnWtEy6NEk. The conclusions seemed to be modestly different from Espinoza24.

To understand where discrepancies lie, we performed extensive due diligence by running multiple experiments with the same setups as the authors to understand the observed differences. We appreciate the authors for making their code available, enabling this exploration. Our experiments used the same data split as Espinoza24. The models compared include:

- **LSTM**: NeuralHydrology version of LSTM from Espinoza24.
- **NH-Hybrid**: Differentiable HBV from Espinoza24 based on the NeuralHydrology package.
- **dHBV1.0 hydroDL**: Differentiable HBV by Feng et al., 2022, based on the HydroDL package.
- **dHBV1.1p hydroDL**: An improved version with updates to the loss function and capillary terms from dHBV1.0 hydroDL.
- **HBV**: Traditional HBV.

A notebook reproducing the results of dHBV1.0 hydroDL (blue box in Figures C1 and C2) is available here (https://colab.research.google.com/drive/12xUvTu9NoGVdcRWqJvypy4DGg5jy5q9T#scrollTo=v0JIEuFZjkxq).

Both dHBV1.0 hydroDL and dHBV1.1p hydroDL showed advantages over LSTM and NH-Hybrid, particularly for high peaks (Figure C1) and the 531 largest peaks (Figure C2), with the benefits being more pronounced for larger return-period event. Notably, NH-Hybrid exhibited larger peak errors than LSTM for the 25-50 and 50-100 return periods, while the HydroDL versions showed smaller errors. This discrepancy may influence perceptions of the relative strengths of these models.

Further, we observed that alternative choices (Table C1) in experimental design, which might align more closely with Frame et al., 2022, or represent more realistic tests, could yield greater advantages for dHBV HydroDL over LSTM (Figure C3). Specifically, the experiments in Espinoza24 may allow LSTM to operate more within an "interpolation" regime, whereas true experiments should challenge models in an "extrapolation" context.

Additionally, Espinoza24 interpreted the 16 multicomponents in dHBV as an ensemble, but this setup is more analogous to the "hidden size" concept in LSTM. Due to the process-based

nature of differentiable models, a true ensemble should consist of different model structures (e.g., HBV, SAC-SMA, PRMS, CFE). Preliminary results (not shown here, but to be provided in a subsequent publication) indicate that such structural variations improved NSE metrics.

While Espinoza24 conducted an experiment to verify that NH-hybrid could reproduce earlier results from Feng et al. (2022), it is important to note that this does not imply that other experiments would yield the same outcome. The claimed equivalence, which could influence how some readers interpret the results, is not established here.

In conclusion, while the LSTM model in Espinoza24 represents a substantial effort, it may not reflect the state-of-the-art due to differences in training frameworks and frontend LSTM configurations. The community may benefit from more explicit specification of implementations used, pulling the original dHBV1.0 HydroDL code into the comparison, and that alternative methodologies in the community could produce different results.

We respect the authors' alternative implementations of our idea, which adds to the healthy discussion of pure data-driven vs. interpretable hybrid models. Hopefully more research can go this way to understand the Pros and Cons of each method.

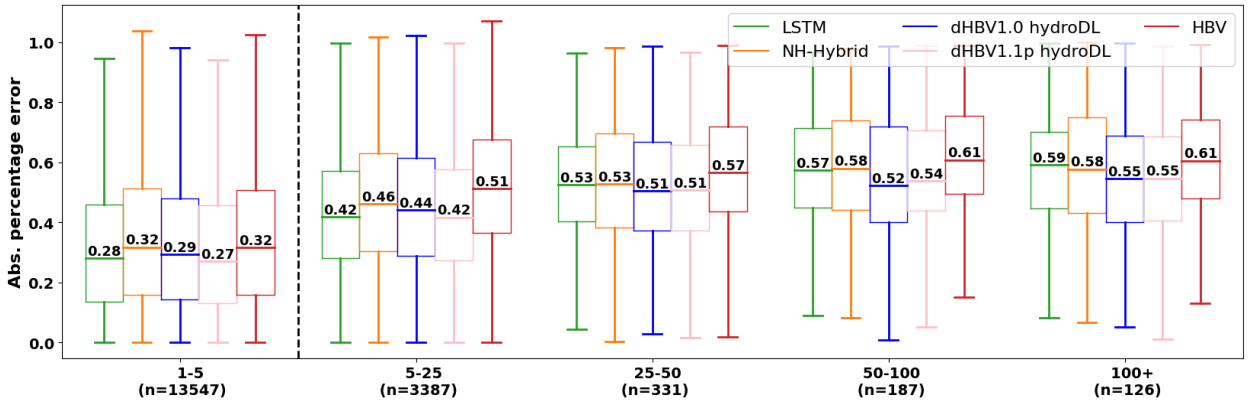

Figure C1 Absolute percentage error between the observed peak discharge and the associated simulation value for the different models, classified by the return period of the observed peaks. The four categories to the right of the dashed vertical line present the errors associated with observed discharge above the 5-year return period threshold, evaluating the out-of-sample capabilities of the models. The n-value below each category indicates the amount of data used to produce the box-plot.

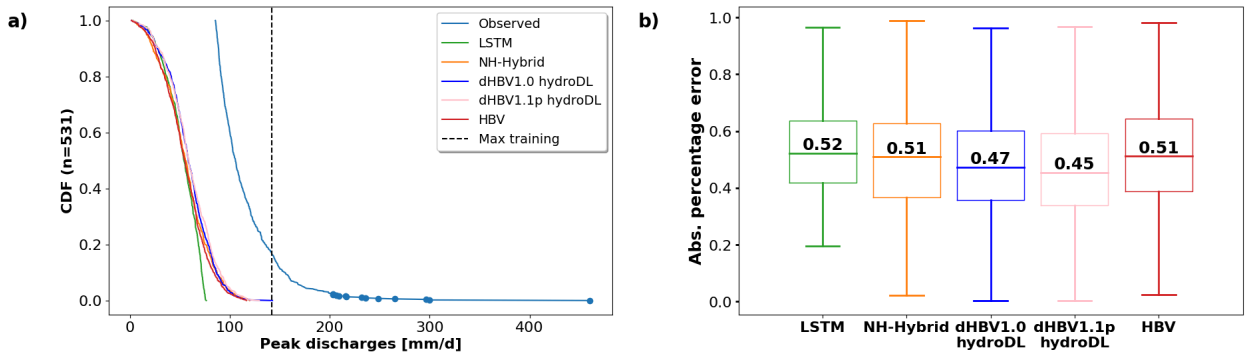

Figure C2. a) CDF of the 531 observed highest discharge values across all basins and their respective simulated values. The blue dots help visualize that under 3% of the events have values between 200 and 400 mm/day. b) Absolute percentage error of the 531 highest discharges for the different models.

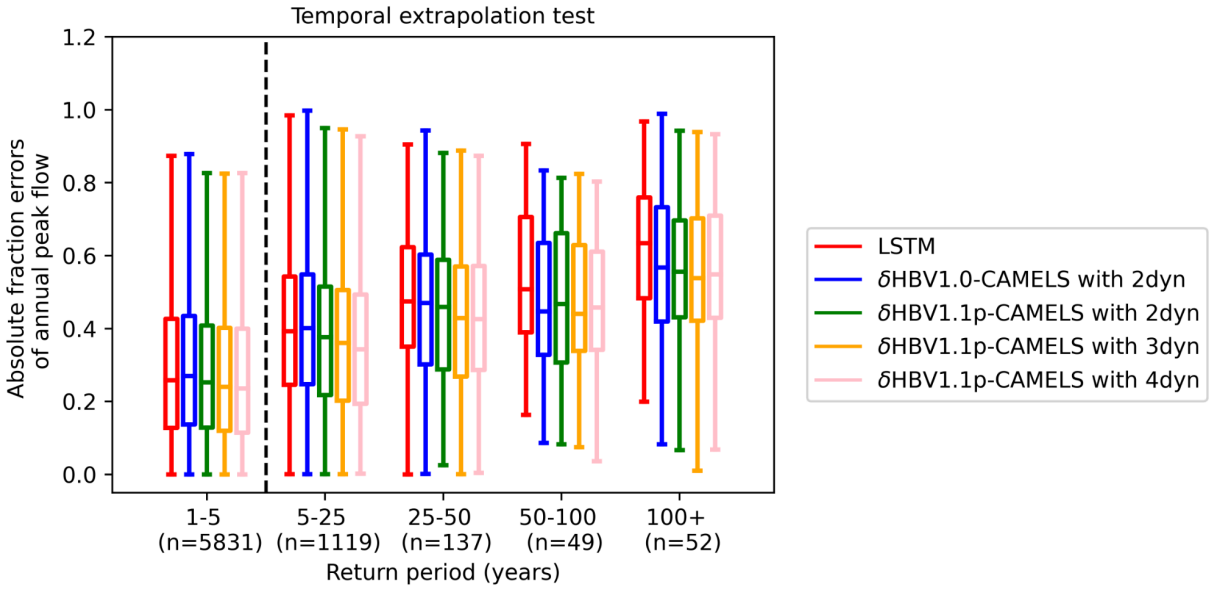

Figure C3. Comparisons in the experiments we have run, presented in Song et al, 2024.

Reasons --- Impact	**Differences**	**Comments**

1. Package difference --- significant favor for LSTM	Espinoza24 used their own training framework built on NeuralHydrology (NH) package including a sequence-to-one LSTM network, while we employed our framework (HydroDL) with a sequence-to-sequence LSTM network with slightly different implementations.	While it seems the NH-hybrid package can give the same performance in the benchmark case used in Feng et al 2022, it gives suboptimal performance in extreme event tests.
2. Ensemble strategy --- large on NSE; maybe minor on extreme events	Espinoza24 used ensemble LSTM to compare with a single dHBV model	Espinoza24 misinterpreted the dHBV hybrid model. The multicomponent in HBV is not an “ensemble” . Rather, it is similar to the hidden size in LSTM. The real ensemble of differentiable models should be composed of different model structures, e.g., HBV, SAC-SMA, PRMS. We have ongoing work that shows the true model ensemble provides better NSE. That being said, the impact on extreme impact needs more time to understand.
3. Experiment design --- moderately favored LSTM impact on extreme events	Espinoza24 et al. validated the model using holdout years within the training period --- they trained the model with CAMELS data from 1980-2014, and tested it using years within 1980-2014 that had extreme events and were held out during training. We trained from 1995/10/01 to 2014/09/30, holding out water years with peak flows greater than a 5-year return period. We then tested models on a separate continuous time period, from 1980/10/01 to 1995/09/30.	We argue a true test should purely exist in continuous history or future years to avoid any kind of information leak. Even though Espinoza24 would say the data in some years is held out for test, during training LSTM still sees what the future looks like after the extreme events. Somehow this makes it a simpler task than purely predicting in untrained years. We have run the tests, a pure extrapolation like what we did represents a harder case, and LSTM shows more disadvantages less favored under such a more real-world scenario.

Reference:

Feng, D., Liu, J., Lawson, K. and 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*(10), p.e2022WR032404. https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2022WR032404

Song, Y., Sawadekar, K., Frame, J.M., Pan, M., Clark, M., Knoben, W.J., Wood, A.W., Patel, T. and Shen, C., 2024. Improving Physics-informed, Differentiable Hydrologic Models for Capturing Unseen Extreme Events. *Authorea Preprints*. https://essopenarchive.org/doi/full/10.22541/essoar.172304428.82707157

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. *Hydrology and Earth System Sciences*, *26*(13), pp.3377-3392. https://hess.copernicus.org/articles/26/3377/2022/hess-26-3377-2022.html