Review: Deep learning rainfall-runoff predictions of extreme events

This paper seeks to answer an extremely pertinent question, to what extent do LSTMs (the current state of the art for rainfall-runoff modelling) continue to perform adequately when predicting out of sample extreme events? This idea that LSTMs will fail whenever there are extreme events that we do not observe in the training period has been a key criticism of data-driven approaches and yet there has been very little empirical validation of this. This paper is well timed and addresses this research gap.

The main results are drawn from the experiment which separates years of data by the peak flow in that year, thus simulating unseen extremes, and therefore testing model adequacy in a future with more extreme flows.

Ultimately, this is an extremely valuable contribution to HESS and I recommend that it is published.

Comments:

The paper is incredibly clear, with a clear hypothesis: "*data-driven streamflow models are likely to become unreliable in extreme or out- of-sample events*" and a clear conclusion: "reject the hypothesis". I really enjoyed the separation of text and appendices. It makes the material much more easily digestible and allows extra information to be kept for the interested reader.

This is potentially out of scope and I do not want to add clutter to this paper since it is incredibly concise and a valuable contribution. That being said, I am left wondering how these patterns vary spatially over the 498 basins used for training. Is there anything that can be said about the model conditions in which the LSTM/MC-LSTM/SAC-SMA difference is large? Is it the same as found by Nearing et al (2020) where the post-processed SAC-SMA model was most improved in snowy catchments? Please feel free to ignore this recommendation!

Is there any way that we can see a hydrograph showing the difference for an extreme high-flow in the two experiments: 1) showing how well the models perform when the training period is split the same way as the original papers cited 2) showing how the performance differs when the training period is split by the max annual return period. This may also not be necessary if the authors feel that a single example will focus too much attention on a sample of 1. I understand that choosing an archetypal example can be difficult and potentially lead to cherry-picking of results. Completely up to the authors!

L112-129: This paragraph describes the main experimental setup for the paper. Why do the authors use a threshold of a 5-year return period (20% chance of annual exceedance) split for

train-test? Was it because it allowed a sufficient number of samples in the train/test split? Would it be possible to make this choice explicit?

Figure C2: Looking at the peak-timing metric below, it seems as though the MC LSTM (**orange**) outperforms the LSTM on high return period flows. I recognise that the performance seems at odds with the other results, but is one interpretation of this finding that Mass-Conservation helps with peak-timing for the really high return period flows?

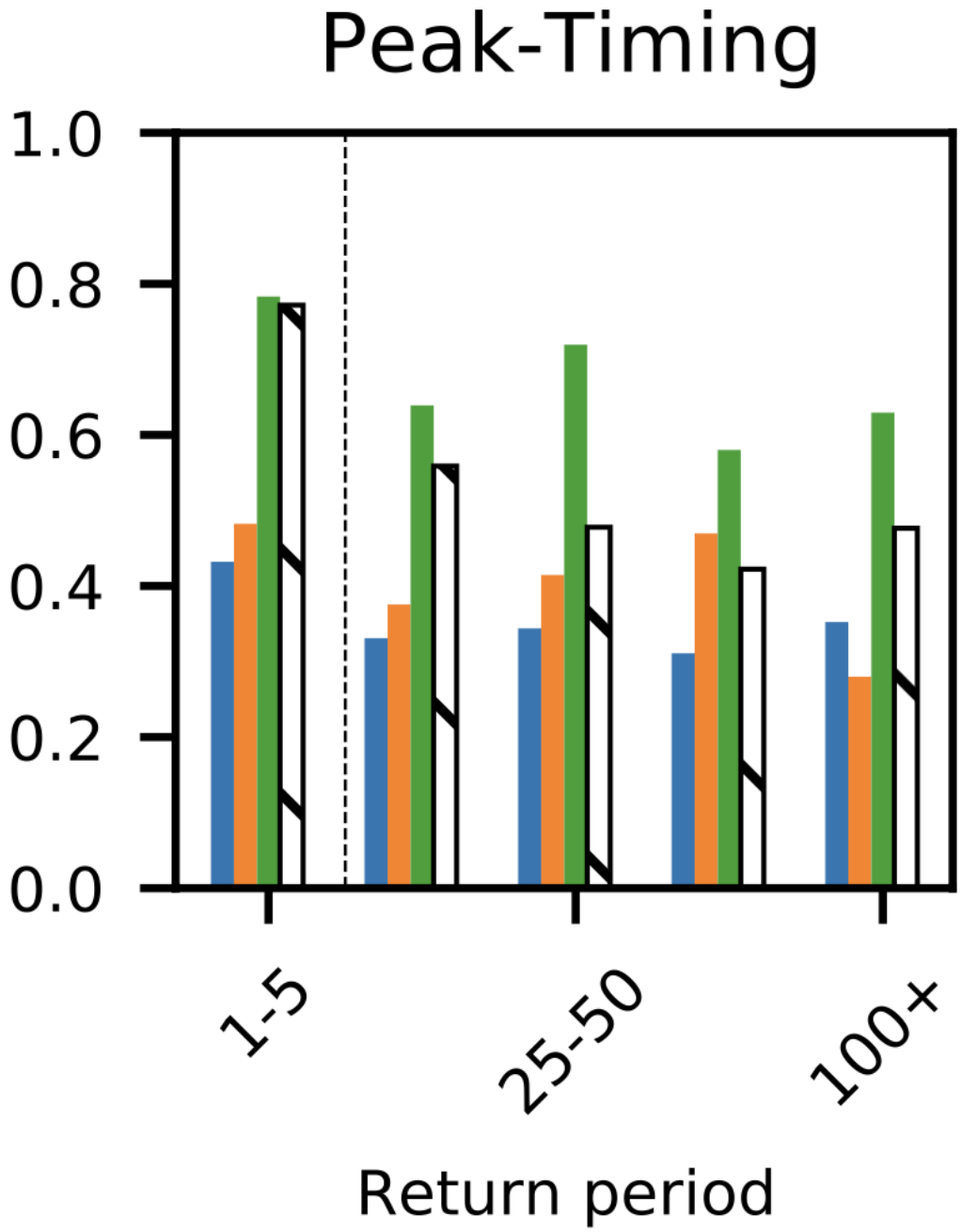
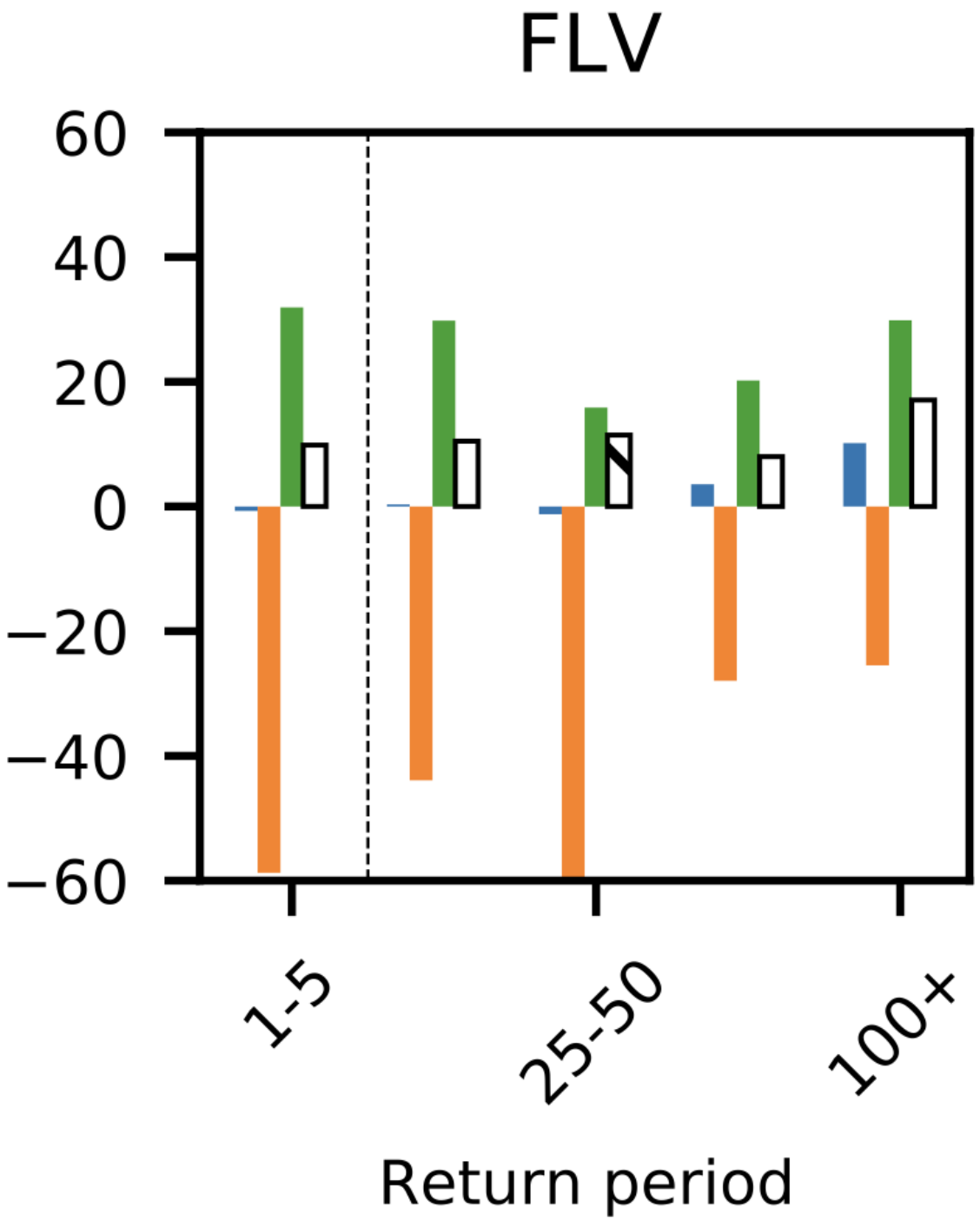


Figure C2: I find the following graph interesting, and perhaps worthy of discussion (at least in the appendix C). The results for the MC-LSTM / LSTM seem to be most different for the FLV metric (looking at the bias of the low-flows). What is it about the physical constraints that could cause such large differences for the low flows? Is it in these low-flow conditions that errors in the underlying data (and therefore the mass-balance) have the largest impact?



Minor Comments:

L44-45: "*the basins that were used for model benchmarking by Newman et al. (2017), who removed basins with (i) large discrepancies between different methods of calculating catchment area, and (ii) areas larger than 2,000 km2.*" What is the justification for removing these larger catchment areas?

L47: "because we benchmarked against the National Water Model retrospective" → "because we benchmarked against the National Water Model retrospective**ly**" ? I'm not sure what is trying to be said here?

L298-299: "*In our case, the mass input is precipitation and the auxiliary inputs are everything else (e.g. temperature, radiation, catchment attributes)*" Is there any way that energy can be conserved using this model too? In terms of the energy available for evaporation of water? This is definitely out of scope. I am just interested in whether this is possible given the approach.

L320-322 "*we accounted for unobserved sinks in the modeled systems by allowing the model to use one cell state as a trash cell. The output of this cell is ignored when we derive the final model prediction as the sum of the outgoing mass*" This is incredibly interesting! Is there any information about the times/locations when this "trash cell" contains a lot of water? Not for this paper but if you had any sense of whether these learned outflows correspond with subsurface transfers of water then that would be super interesting!

L323-325: "The NWM calibration does not correspond to the training/calibration period of SAC-SMA, LSTM or the MC-LSTM." Can we possibly have a more immediate answer to what time period was the NWM calibration? It's written elsewhere () but since this is an appendix it would help the reader to know what the NWM calibration was here too.
"*was calibrated by NOAA personnel on about 1400 basins with NLDAS forcing data on water years 2009--2013*"

L333: "SAC-SMA model according every" → "SAC-SMA model according **to** every"

References

Nearing, G., Sampson, A. K., Kratzert, F., and Frame, J.: Post-processing a Conceptual Rainfall-runoff Model with an LSTM, 2020a.