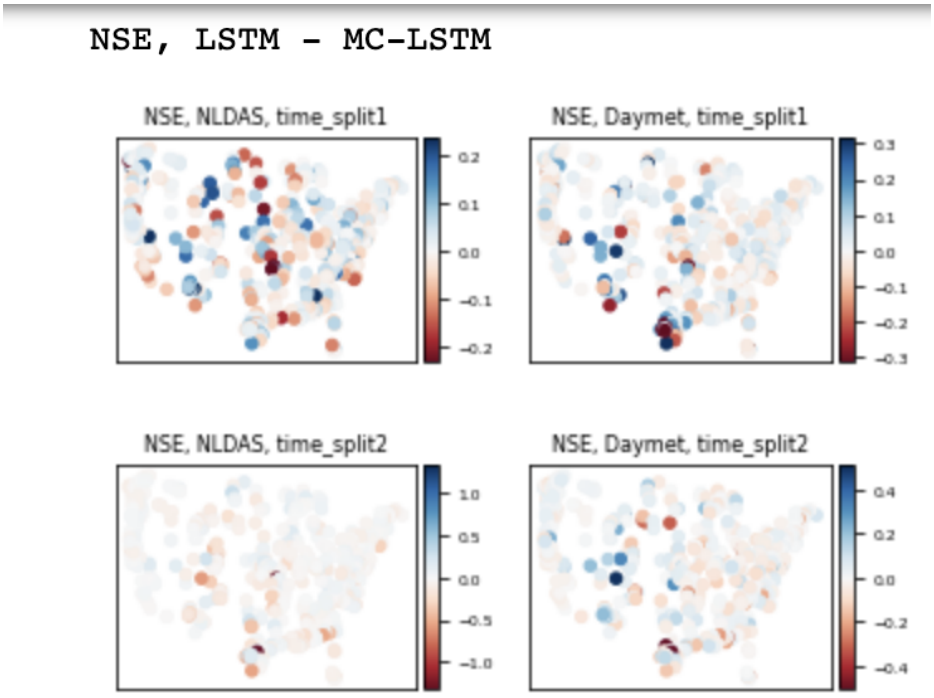
Thank you for your comments. I am very pleased you enjoyed the paper.

RC3: ["I am left wondering how these patterns vary spatially over the 498 basins used for training."]

We spend considerable time trying to describe the results spatially. No obvious discernible patterns come out of the spatial distribution of the results. Some small clusters exist where the difference of LSTM and MC-LSTM is positive or negative, but nothing particularly worth mentioning.




RC3: ["Is there anything that can be said about the model conditions in which the LSTM/MC-LSTM/SAC-SMA difference is large?"]

The differences aren't really large enough to make that distinction.

RC3: ["Is it the same as found by Nearing et al (2020) where the post-processed SAC-SMA model was most improved in snowy catchments?"]

I'm not sure that Nearing et al (2020) made that claim? We also do not see that in this data here.

RC3: ["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. “]

Just as you caution, I am very reluctant to cherry pick some hydrographs. Perhaps we can host the complete series of hydrographs on Hydroshare.

RC3: [“L112-129: This paragraph describes the main experimental setup for the paper. “]

RC3: [“Why do the authors use a threshold of a 5-year return period (20% chance of annual exceedance) split for train-test? “]

This was an arbitrary choice made before the experiment was run. Perhaps we could have done lower to get more “extreme” high flow events, as there are plenty of lower flow years for training, but I would be worried about potential “p-hacking” type of activity.

RC3: [“Was it because it allowed a sufficient number of samples in the train/test split? Would it be possible to make this choice explicit? “]

We did consider the number of training/testing splits, but just in the sense that once we picked 5-yr as the threshold, we checked to make sure we had enough data for both training and testing. Again, we picked this threshold BEFORE looking at the segregated data or conducting any experiments, in order to avoid experimental bias.

RC3: [“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? “]

That is a possibility, and perhaps worth further discussion.

RC3: [“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? “]

Our response to Reviewer 2 provides figures that show the FLV metric is wildly unstable when flows are very low, or zero. We will add the following discussion on this metric:

“The results indicate that the MC-LSTM performs much worse according to the FLV metric, but we caution that the FLV metric is fragile, particularly when flows approach zero (due to dry or frozen conditions). The large discrepancy comes from several outlier basins that are regionally clustered, mostly, around the south-west. The FLV equation includes a log value of the simulation and observed flows. This causes a very large instability in the calculation. Flow duration curves (and flow duration curve of the minimum 30% of flows) of the LSTM and the MC-LSTM are qualitatively similar, but they diverge on the low flow in terms of log values.“

RC3: [“Minor Comments: “]

RC3: [“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? “]

Critically, the reason that the current study removes these basins is to be consistent with previous benchmarking experiments, specifically the now rather large set of community experiments that inherit from Newman 2017. Andy Newman’s justification for doing this initially back in 2017 was that the basin area is used to convert between streamflow [L3/T] and surface runoff depth [L]. Since the two basin area calculations disagree by a large magnitude, we cannot be sure which is closer to the truth.

RC3: [“L47: “because we benchmarked against the National Water Model retrospective” → “because we benchmarked against the National Water Model retrospectively” ? I’m not sure what is trying to be said here? “]

The National Water Model Retrospective is a specific dataset. The word “retrospective” in the name of this dataset refers to the fact that these data are model hindcasts. Will make this more clear by changing the phrasing to include the new documentation “NOAA National Water Model CONUS Retrospective Dataset” with a link to NOAA’s page: https://github.com/NOAA-Big-Data-Program/bdp-data-docs/blob/main/nwm/README.md

RC3: [“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. “]

The basic structure of the MC-LSTM is conservative – any quantity or quantities can be conserved (e.g., mass, energy, momentum, counts, etc.). In the current study, we aren't using surface flux data, so it would not be possible to use the CAMELS data to (reasonably) constrain energy. But we have done this with FluxNet data in other studies.

RC3: ["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! "]

That is planned for another experiment, possibly even in review already.

RC3: ["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?"]

Yes. Good idea.

RC3: ["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""]

RC3: ["L333: "SAC-SMA model according every" → "SAC-SMA model according to every""]

Thank you, will be revised.