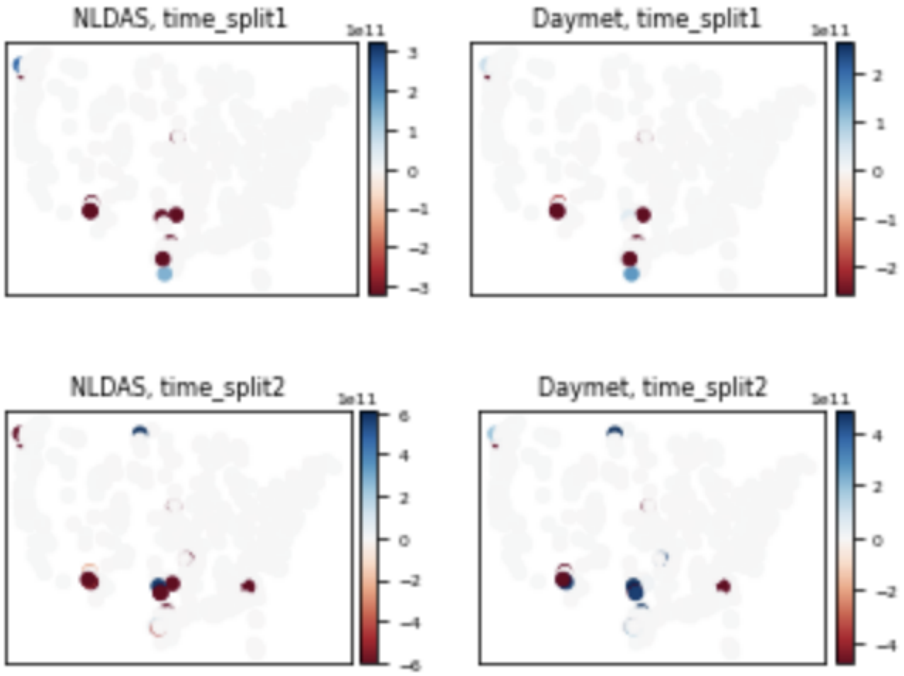
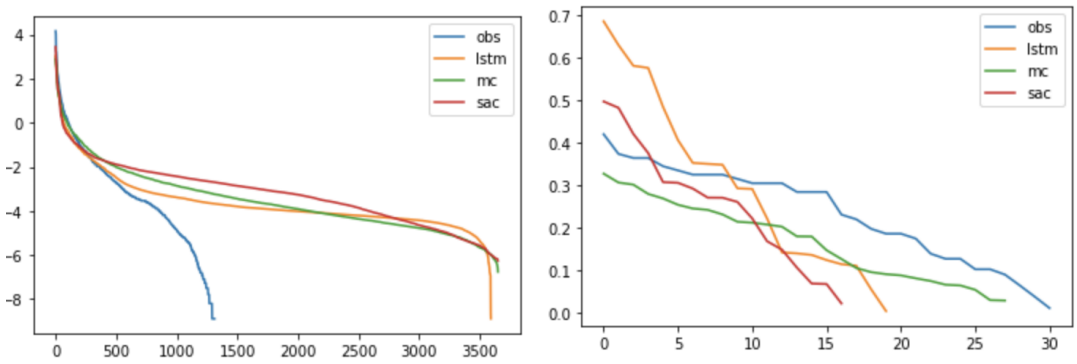
Thank you for your comments.

The only action item in this review is suggesting additional analysis into the large discrepancy of the MC-LSTM for the FLV metric. This is indeed an interesting result. It turns out that outliers play a major role in the FLV results. And that they are regionally clustered, mostly, around the south-west. Basically, because the FLV depends on a log value of the simulation and observed flows, they have to be either removed, or artificially set above zero. This causes a very large instability in the calculation. Basically, this metric is not viable when streamflow approaches zero. Below is a little further analysis. I believe we can convey the following in the discussion, without adding additional figures.

FLV, LSTM - MC-LSTM



Flow duration curve (left) and flow duration curve of the minimum 30% of flows for Basin 09513780, which is in the cluster of Arizona basins shown on the map above. The actual curves of the LSTM and the MC-LSTM are not that far off, but the difference in FLV metric between the two is wildly different (~1e10) : lstm -9.506222e+11, mc-lstm -1.907157e+11, sac -1.517753e+11.



We will add the following discussion of the FLV metric describing the fragile, and potentially misleading results:.

“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.“

RC2: [“Another exciting thing is about the FLV (Bottom 30% low flow bias). Analysing why FLV increases(in magnitude) drastically for MC-LSTM could be an interesting direction to explore. Especially for the low probability years. As theoretically, the machine learning model should have seen such low flow data. The author illustrates that any constraint restricts the space of possible functions that the network can emulate. MC-LSTM is developed primarily to model this type of situation where an entity is conserved. Furthermore, unlike other metrics, which did not deteriorate much, we see a drastic drop (increase in magnitude) in FLV. More analysis in this direction would be interesting for the readers as well.”]

As mentioned above, the FLV has some problems when flows approach zero, so we will add the description above. Thank you for pointing this anomaly out.

RC2: [“The paper highlights the potential of deep learning models to predict extreme events, while the hypothesis is that the data-driven models lose reliability in extreme events more than models based on process-understanding. The notion of reliability can be somewhat vague and should be clarified. The paper is only focusing on the predictive reliability here.”]

Will clarify about focusing on “predictive reliability”.