NOT FOR PEER REVIEW.

Supplemental analysis from the revision of:

On Strictly Enforced Mass Conservation Constraints for Modeling the Rainfall-Runoff Process

Jonathan M. Frame

January 31st 2023

This document is for discussion and contextualizing some of the results in (and left out of) our paper submitted to Hydrologic Processes entitled “On Strictly Enforced Mass Conservation Constraints for Modeling the Rainfall-Runoff Process”. In the process of revising the manuscript, we did many analyses to shed light on the nature of precipitation events across the Contiguous United States with runoff coefficients greater than 1, the distribution of model performance, the nearest neighbors analysis and more. In general, these plots are kind of interesting, and lead to some interesting ideas about what is going on with the models, but there isn’t enough to draw good conclusions from, and do not help us answer the hypothesis posed by Beven (2020), so it was best to leave them out of the paper.

1.0 Precipitation events with runoff coefficients greater than one

Beven (2019) defines the runoff coefficient as “the proportion of the rainfall appearing as discharge from the catchment”. First off, this is fundamentally unobservable in most catchments, as there is no way to track the timing and distribution of all water that ends up as streamflow. This is because there are often many complex interacting water storages within a watershed, such as snowpack, soil moisture, groundwater, and surface ponding/storage (natural and man-made). There may be accurate measurements of a watershed runoff coefficient in a basin with little-to-no surface water-groundwater interaction and no snowpack, but these basins are rare. We do, however, have a prior estimate of these watershed characteristics in the CAMELS dataset, and we can separate our analyses based on these characteristics.

Given that we are dealing with a large and diverse sample of watersheds, it is not possible to define a single event-splitting scheme, such as the recession curve method that Beven (2019) used. We described in our first paper revision, when including the event-based analysis of runoff coefficients as a proxy for short-term mass balance, how we defined a single precipitation event. “Since we included a diverse range of basins in our analysis, we needed to define a precipitation event that could be representative across many different types of hydrologic regimes.” Our event separation method is described in Section 4, and this was done to accommodate basins with influence from baseflow, snowpack, glaciers, etc.: “an event starts when precipitation is greater than the 25th percentile and stops when less than the 5th percentile.”

1.1 Distribution and frequency

Based on the above defined precipitation events, there are many such events that end up with a runoff coefficient greater than 1, and it was suggested by a reviewer that this is unreasonable. we do not see this as unreasonable, and we would actually expect this watershed behavior due to sequential storms, snowpack, interflow, etc. We have done several analyses, and we can see a clear signal that the basins with the most events greater than 1 are located in Northwest U.S (Figure 1).

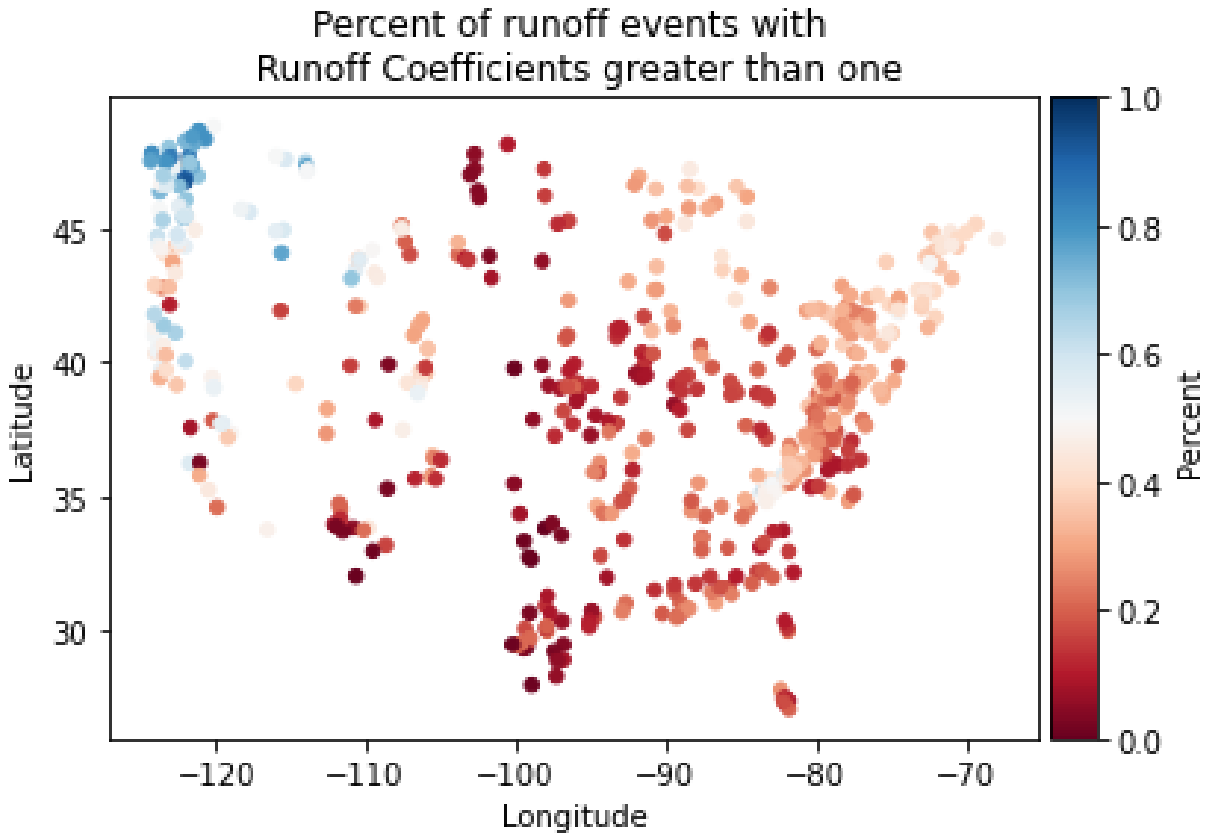

Figure 1. The spatial distribution of precipitation events resulting in runoff coefficients greater than 1, meaning that more water leaves a watershed from observed streamflow than enters the watershed from precipitation.

Below is the same plot, but broken up by month. It is fairly obvious that the runoff coefficients greater than one are largely a result of snow melt. The main clue is that runoff coefficients in the late summer and early fall (August-November) are relatively rare. Keep in mind that we are performing these analyses on ~500, all with much different hydrologic characteristics. So when we define the “events” we do so in a way that probably does not completely isolate each such event. Many events have runoff coefficients greater than 1, it is likely that many event windows are too small for some catchments, and there are fat tails to the hydrograph, but if we extend the event window, we would reduce the total number of events that we can analyze. I was trying to find an event window that gave us 100+ events on every catchment, so we can do the whole “100 similar events” type of analysis.

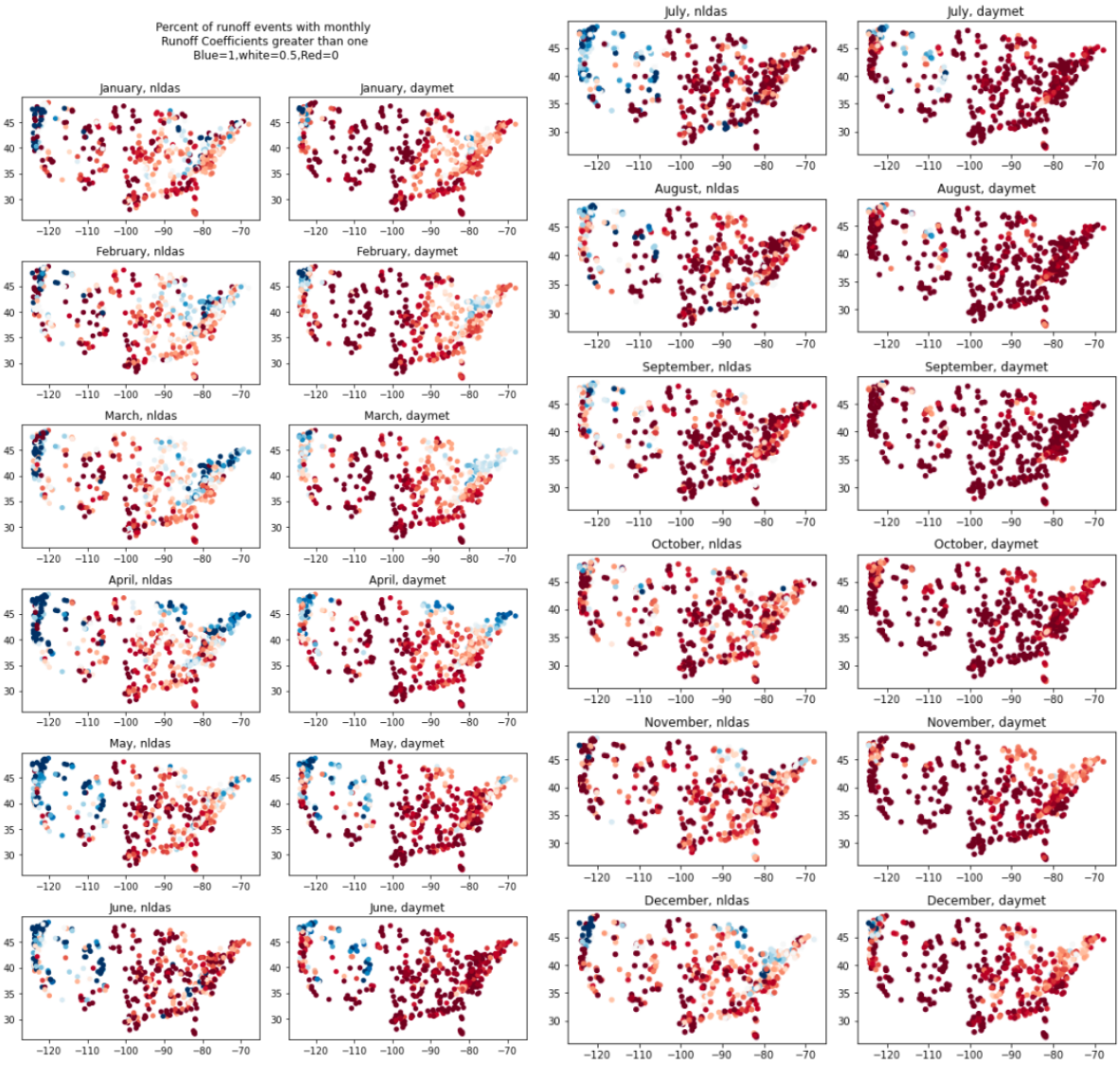

Figure 2. The temporal and spatial distribution of basins with runoff coefficients greater than 1.

1.2 MC Models can simulate runoff coefficients greater than 1

Please notice that it is not true that MC-LSTM or SAC-SMA are unable to simulate event-based runoff coefficients (as defined by event Q/P) greater than 1 - both models are able to (and demonstrably do) produce runoff coefficients greater than 1. For instance, the plots below shows that all the models in this study make some event runoff coefficient predictions greater than 1.

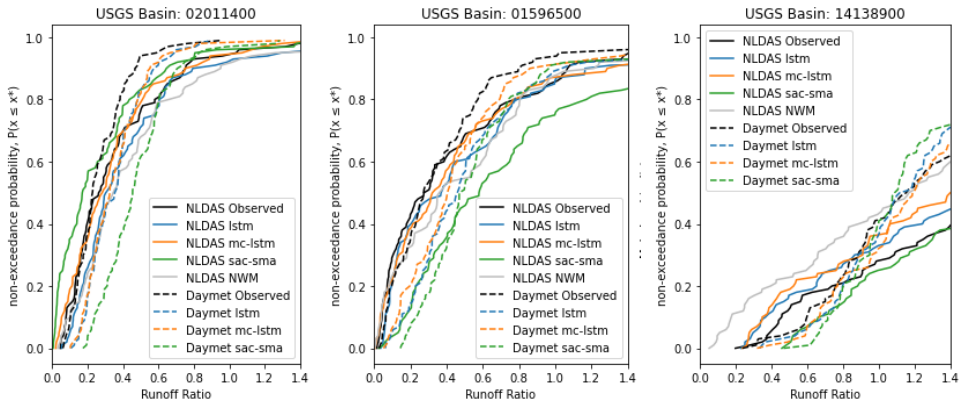

Figure 3. Examples of simulated runoff from precipitation events generating runoff ratios greater than one, for LSTM, MC-LSTM, SAC-SMA and NWM.

It was suggested by a reviewer that perhaps the authors were deliberately clipping the runoff coefficients to 1 in our plots, and it is insinuated that we did this because we do not expect runoff ratios greater than 1. Yes, they were clipped to one, more as a means of having a consistent plot across all examples. The code that was published along with the paper back in July 2022 included an option to run the plots with the x-axis (runoff ratio as rr_bound) to any extent, but the default was one. This was in no way an attempt to cover up any results with runoff coefficients >1.

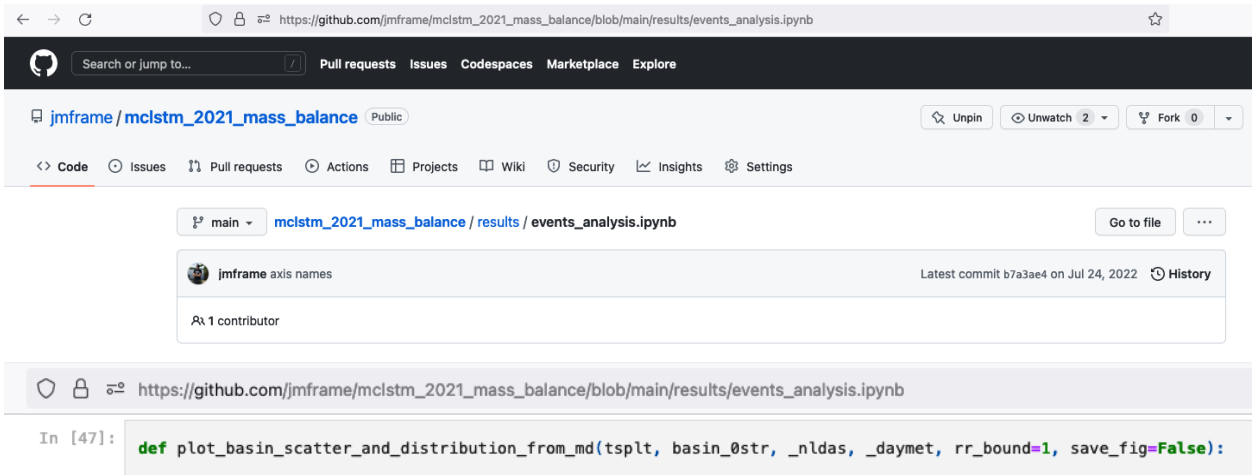

Figure 4. Screenshot of the plotting function, publicly available on Github, taking a variable for the runoff coefficient bounds (rr_bounds). Default is set to one, but can be modified to show the runoff coefficient bounds on the x-axis extending out to any chosen value.

It was further suggested in the review of our paper that it is unreasonable for mass conserving models to make streamflow predictions resulting in simulated runoff coefficients greater than 1. This is not the case. There is nothing conceptually limiting SAC-SMA, NWM or MC-LSTM from making streamflow predictions that are calculated as runoff coefficients greater than 1, because all of these models have states that represent the water stored in a watershed before the precipitation event. Beven (2019) had suggested that the runoff coefficient analysis maybe inappropriate for a watershed with a strong baseflow influence. Since we are doing a large sample study, it is not a standard approach for us to select sub-sets of basins to run particular tests, instead, we can bin each of the large sample of basins into categories to get a more wholistic picture of what is happening.

1.3 Model performance

Below is a figure that shows the event-based results from Table 5 of the original manuscript, but broken down by the baseflow index, where those basins with baseflow indices of between 0-0.1 are calculated separately from those with baseflow indices between 0.1-0.2, etc. These results shows that the basins with lower baseflow indices are better at matching the event runoff coefficient than those with higher baseflow indices. The LSTM runoff coefficients are better in basins with baseflow indices between 0-0.1, 0.3-0.4, and above 0.5. The MC-LSTM is better at matching runoff coefficients from basins with baseflow indices between 0.1-0.2.

The y-axes on these plots is the mutual information and r2 coefficient between the predicted and observed event-based runoff coefficients. The x-axes are the basins split by one of the CAMELS attributes.

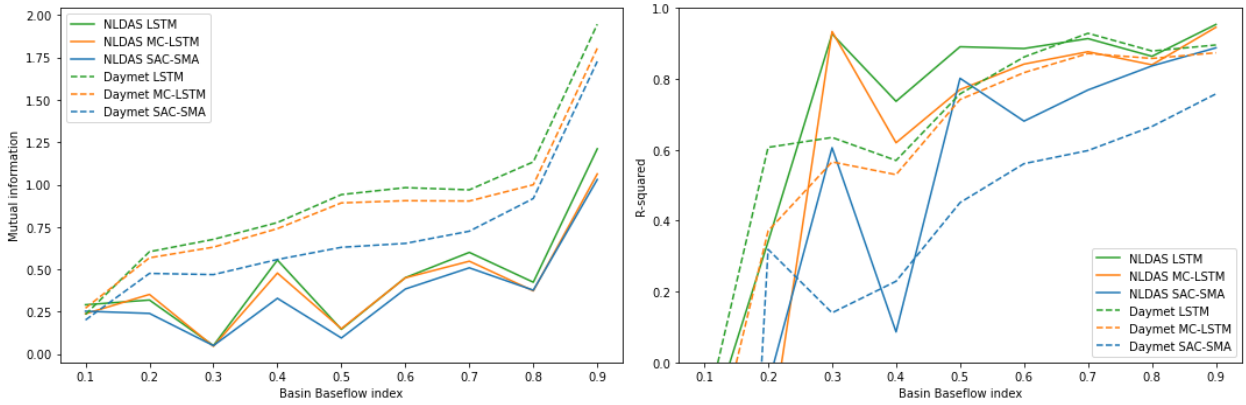

Figure 5. Mutual information and R-squared split by basin runoff ratio index

Melting snow is a common hydrologic phenomenon that results in much more runoff than was introduced to a basin from a single precipitation event. This plot can also show how different basin attributes affect the ability of a model to match runoff coefficients, which is useful to know for those basins that have attributes most likely to obscure the runoff coefficients, like snowpack, shown below.

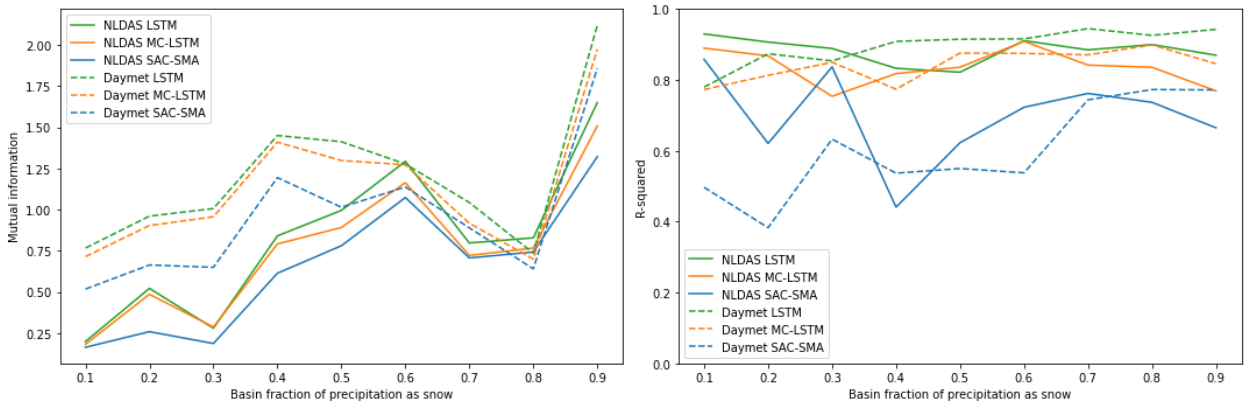

Figure 6. Mutual information and R-squared of precipitation events split by the fraction of precipitation that falls on a basin as snow.

The LSTM, MC-LSTM and SAC-SMA can all consistently capture events with runoff coefficients > 1. Below is a figure that shows the mean NSE values for the three models split by the basin's percentage of events with runoff coefficients greater than 1.

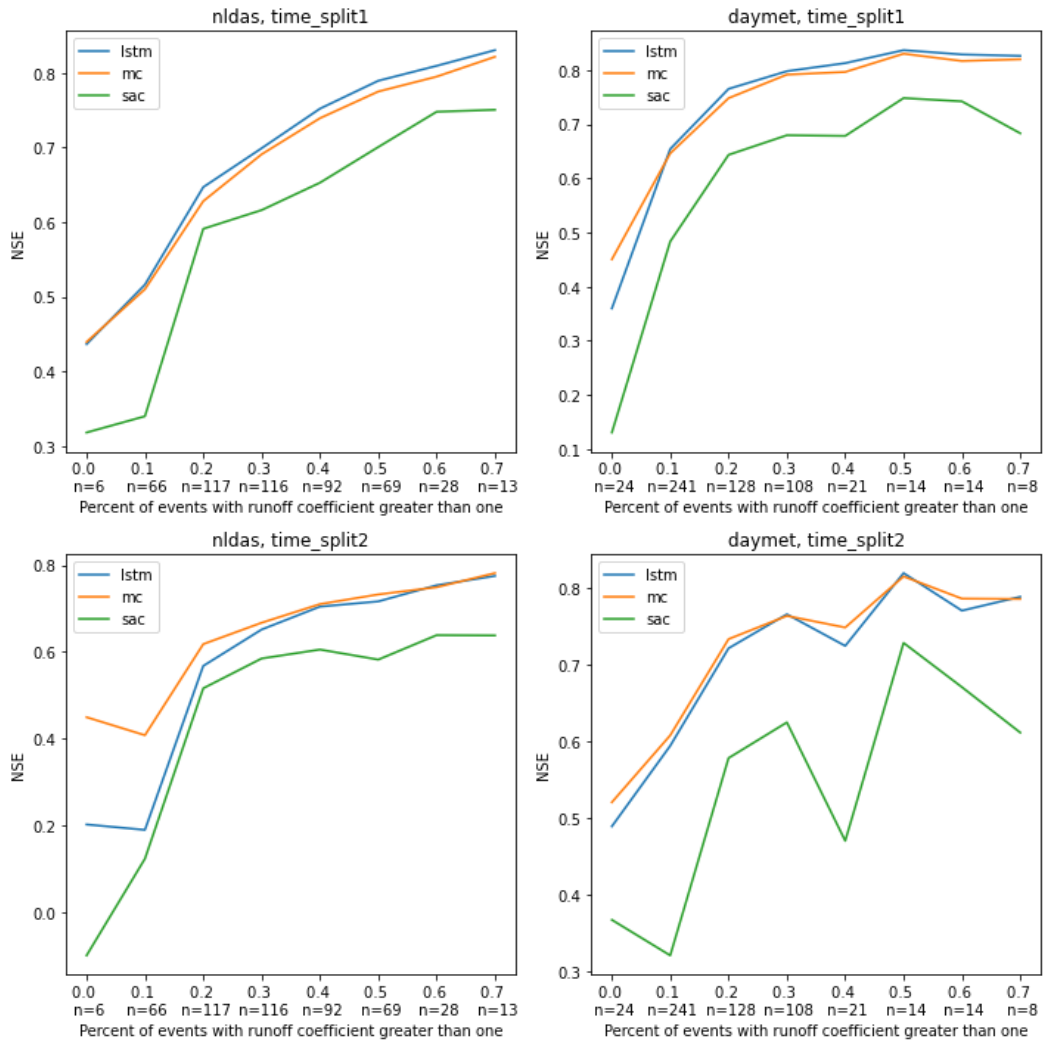

Figure 7. Nash-Sutcliffe Efficiency grouped by the percent of runoff coefficients greater than 1.

Please note that the SAC-SMA model includes Snow-17, a snow melt model that allows for snowmelt to influence discharge at times when the event precipitation does not account for the total runoff (rain-on-snow for instance). The states of the LSTM models have been shown to correlate to snowpack/snow melt.

Below is a similar plot to those above, but splitting up the basins in terms of their long term runoff ratio from the CAMELS catchment attributes (which is the same as the runoff coefficient Q/P). LSTM, MC-LSTM and SAC-SMA all have increasing mutual information with the observed runoff ratio as it exceeds 1.

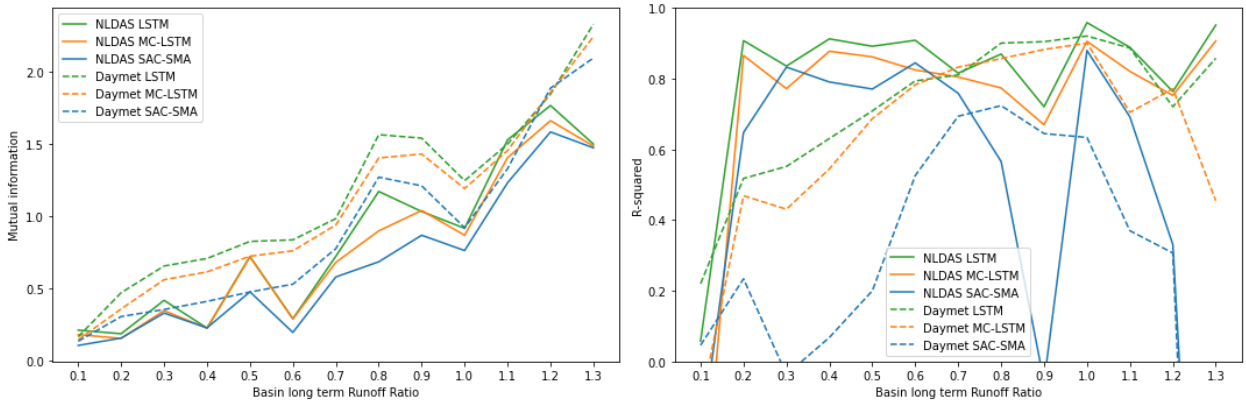

Figure 8. Mutual information and R-squared split by the long term, basin average, runoff ratio (runoff coefficient).

1.4 Summary

There is no reason that these events would be disinformation. There is nothing physically restricting runoff coefficients to be less than one. As shown above, there is a clear signal that events with runoff coefficients greater than 1 are clustered around snowy areas, and are far less common in arid areas. We would expect any hydrological model to capture this.

2.0 Nearest neighbors of event runoff

We used the nearest neighbors for events in an attempt to use the Beven (2019) method for evaluating precipitation events, because that is how we interpreted a comment from the last revision round, but clearly we misunderstood. We were hesitant to include these results, because we were not sure how best to show and interpret them. We included them because it seems that there is something interesting going on when we plot the results on a quantile-quantile (q-q-) plot, particularly the difference between the NLDAS and Daymet forcing, which is how we are attempting to show that deep learning models can learn to make accurate predictions, event when there is a bias in the data.

The q-q plot shows that all three models deviate very little from the q-q plot for NLDAS, but for Daymet they deviate a little bit more. The LSTM deviates most, followed by MC-LSTM (constrained only by mass conservation), then by SAC-SMA (constrained by its hydrologic conceptualization, including mass conservation). This can be attributed to the model's ability to respond uniquely to individual events, deviating from the events with most similar runoff coefficients (the ratio of rainfall to runoff), and I believe this may be a mechanism to compensate for the systematic error in the Daymet data in the eastern united states. The diversity likely

comes from other atmospheric variables that aren't captured by runoff ratio, i.e., not precipitation. This would indicate that the LSTM makes more use from the non-mass inputs than the MC-LSTM. As a response to your comment below, we will remove this section.

2.1 Regional Q-Q plots

The Q-Q plots generated on a region by region basis show that the regions with the Daymet positive mass bias in precipitation totals have a higher cumulative divergence from the 1:1 line from LSTM, followed by MC-LSTM and then SAC-SMA has relatively little. This is a clue for me that this divergence in the nearest neighbors distributions is some sort of compensation for the biased precipitation. Perhaps the LSTM learns a more complex relationship from the forcings that can make up for the erroneously large precipitation values.

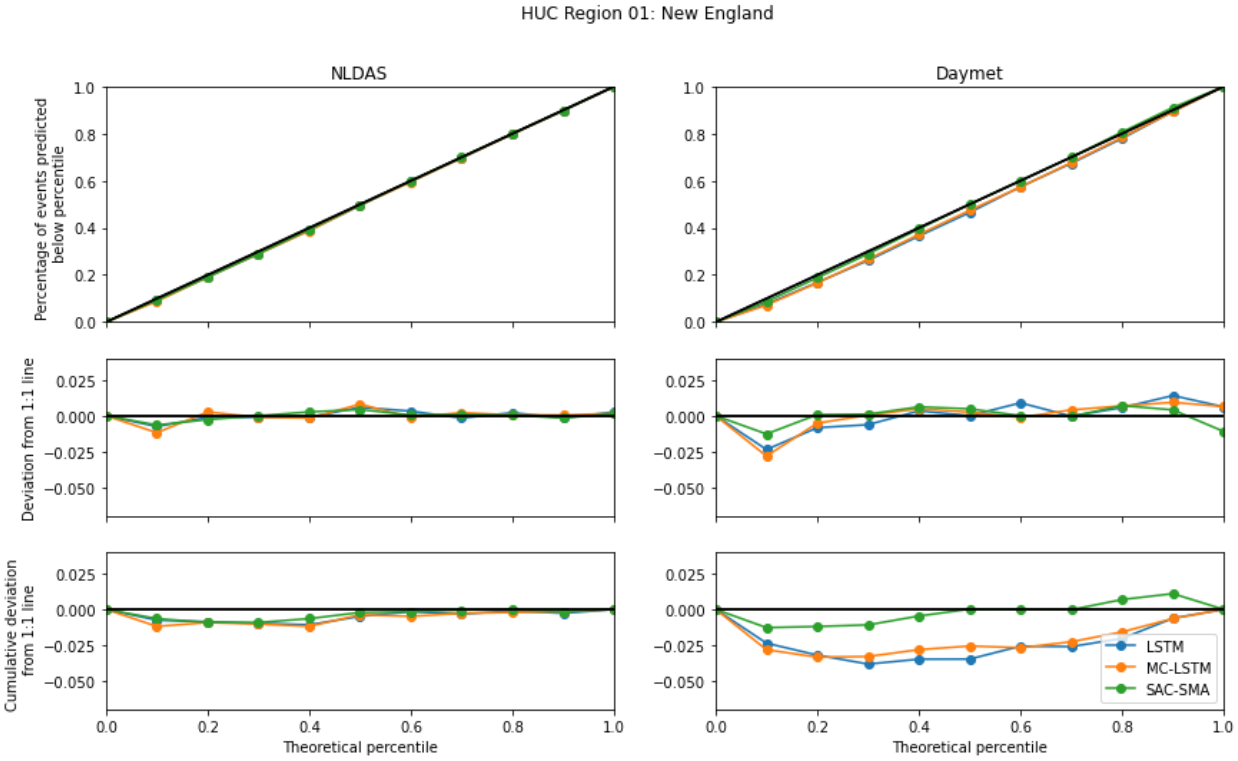

Figure 9. The Q-Q plot, including the cumulative divergence from the 1:1 line in the bottom right corner, on an example East Coast region, which shows that the LSTM has the greatest divergence with Daymet data, followed by MC-LSTM, and SAC-SMA diverging relatively little.

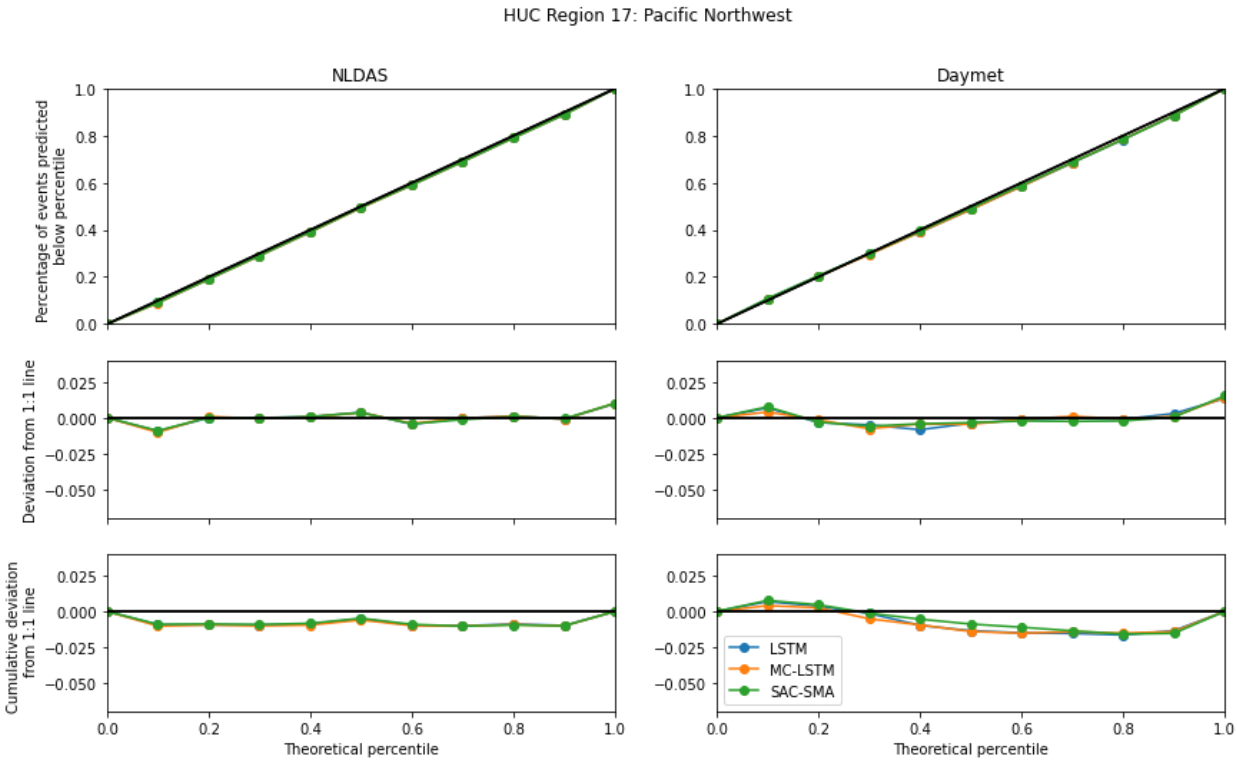

Figure 10. The Q-Q plot, including the cumulative divergence from the 1:1 line in the bottom right corner, on an example West Coast region, which shows little divergence on Daymet data from either model.

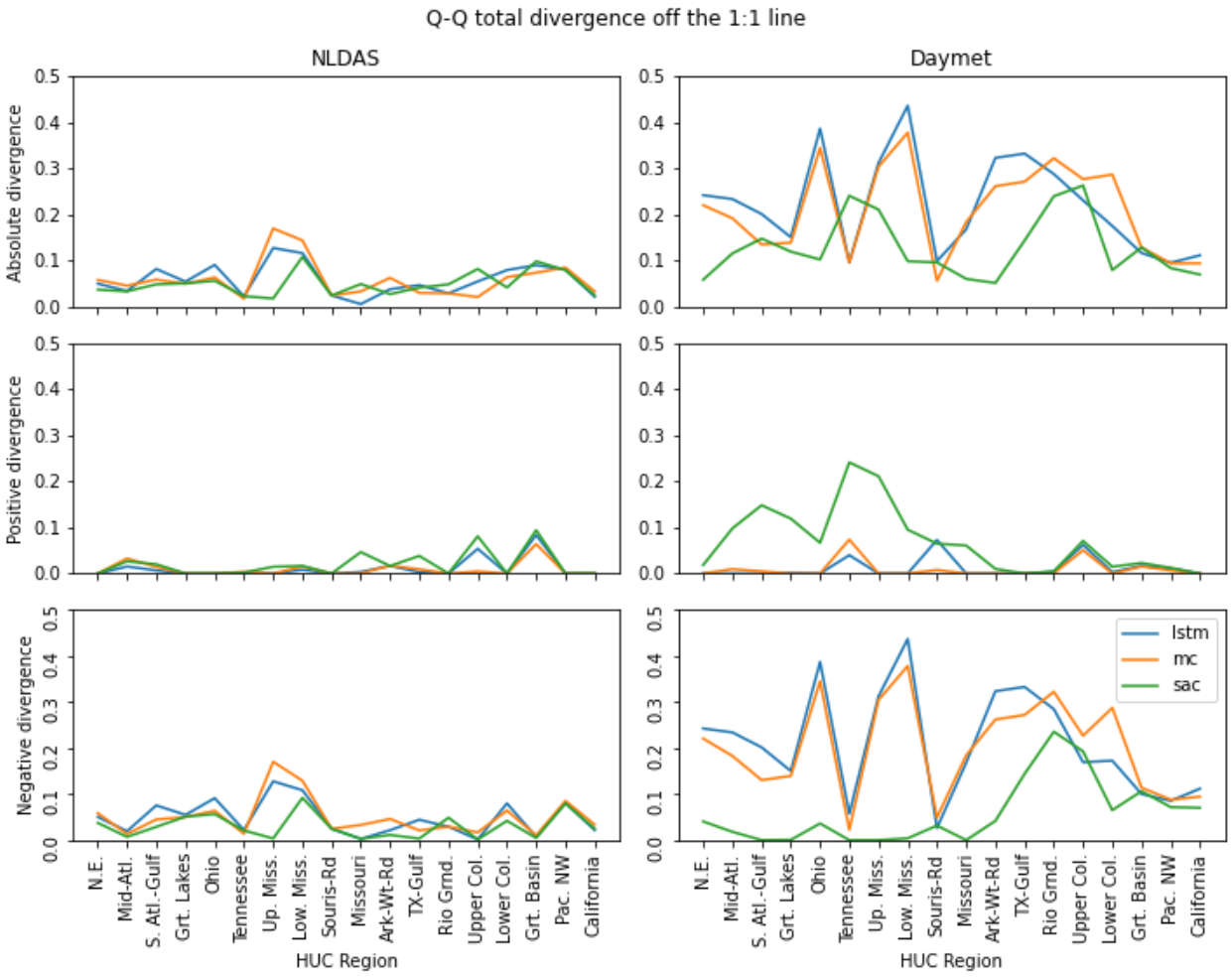

Figure 11. The divergence from the Q-Q plot 1:1 line for all the regions, with both NLDAS and Daymet forcings.

2.2 Theoretical percentiles and extreme events

There was some confusion as to what the Q-Q plots were showing, particularly with the theoretical percentiles. Theoretically, 10 percent of the events should fall in the 10th percentile, 20 percent of events should fall in the 20th percentile, etc. What I believe this is showing is that the LSTM is better at simulating the correct volume of water because it has more freedom to treat each uniquely. That enforcing constraints on the model architecture, starting with mass conservation, causes runoff responses to look similar, event if the event in general is not supposed to look like other such events.

The extreme events, which would lie on the tails of the distributions are in the 10th and 90th percentiles. This is because if you select the 100 nearest neighbor events to the largest precipitation event, all the events in the distribution would be to the left of the largest event.

3.0 MC-LSTM sometimes does better than LSTM

There were several review comments regarding the fact that sometimes the MC-LSTM does better than the LSTM. The MC-LSTM and the LSTM are very, very similar models. The only difference being that one is designed to strictly enforce mass conservation. Because they are so similar, and because it does not appear that enforcing mass conservation is detrimental to the model, we should expect one or another to do better just based on randomness in the training. It could be the case that the enforced mass conservation aids the search for a set of model weights that minimizes the cost function on just some of those instances, but since there is no discernible pattern to where/when this happens, it is not likely. We have searched extensively to find any evidence of this. Below is a table and some plots that might help the reviewer understand this point.

Table 1. The number of basins that the LSTM performs better than mass conserving models. A percentage over 50% means that the LSTM does better in more basins, and a percentage lower than 50% means that the referenced mass conserving model.

	Time split 1				
	NLDAS			Daymet	
	MC-LSTM	SAC-SMA		MC-LSTM	SAC-SMA
KGE	46%	71%		60%	84%
Absolute mass balance	43%	52%		52%	64%
	Time split 2				
	NLDAS			Daymet	
	MC-LSTM	SAC-SMA	NWM	MC-LSTM	SAC-SMA
KGE	62%	81%	74%	58%	90%
Absolute mass balance	50%	58%	56%	52%	82%

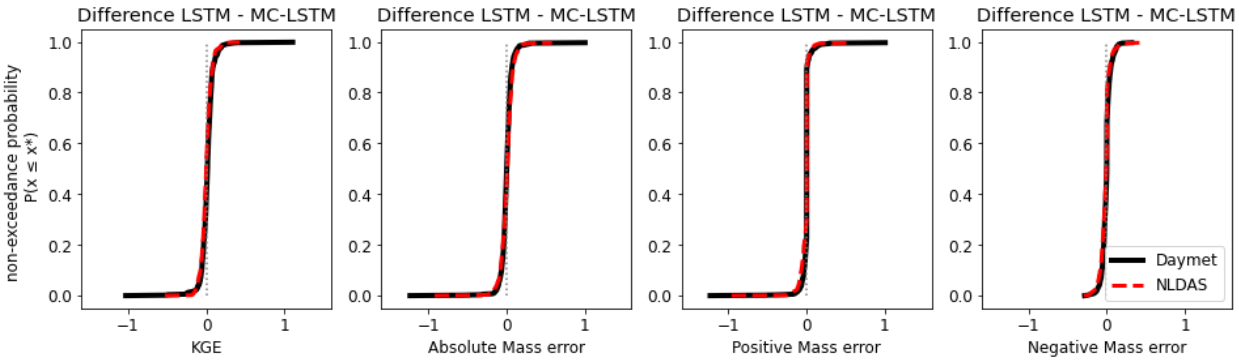

Figure 12. Shows that the LSTM and MC-LSTM do better on about half the basins in both efficiency and mass bias.

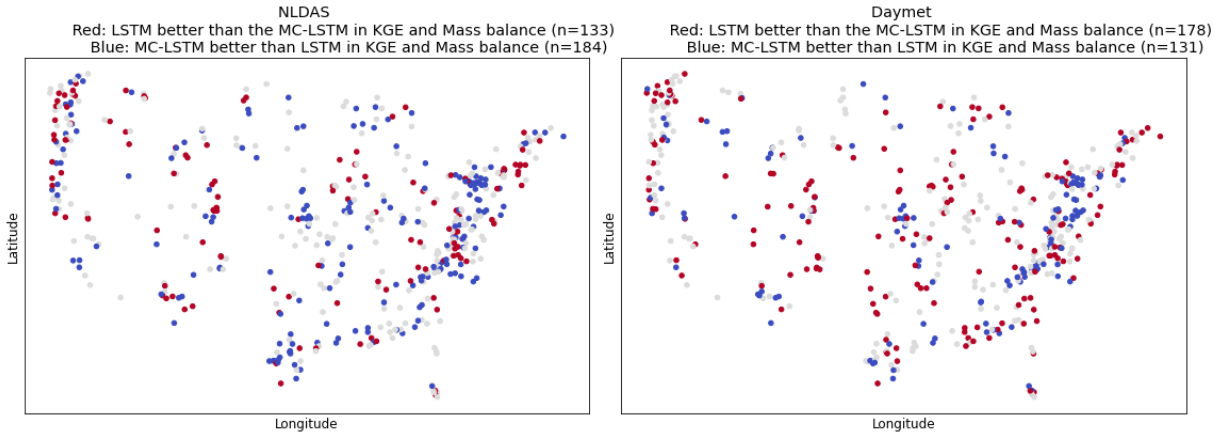

Figure 13. Shows the basins in which LSTM or MC-LSTM do better in both efficiency and in mass balance. The LSTM does do better in both efficiency and mass balance in slightly more basins with the Daymet forcing, which could be a result of the biased precipitation totals in Daymet, but there just isn't enough to determine if that is true.

Just for reference, here are the sample plots with SAC-SMA, insead of MC-LSTM

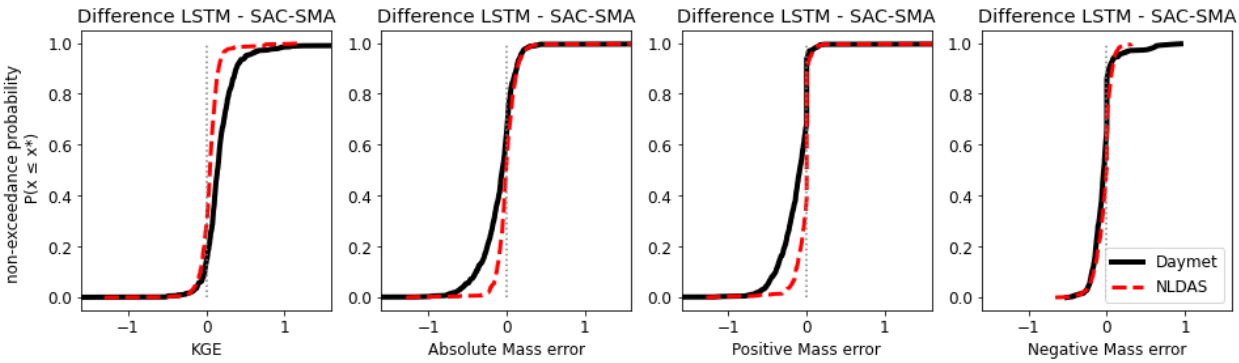

Figure 14. shows the cumulative distribution of basin which LSTM or SAC-SMA perform better in both efficiency and in mass balance.

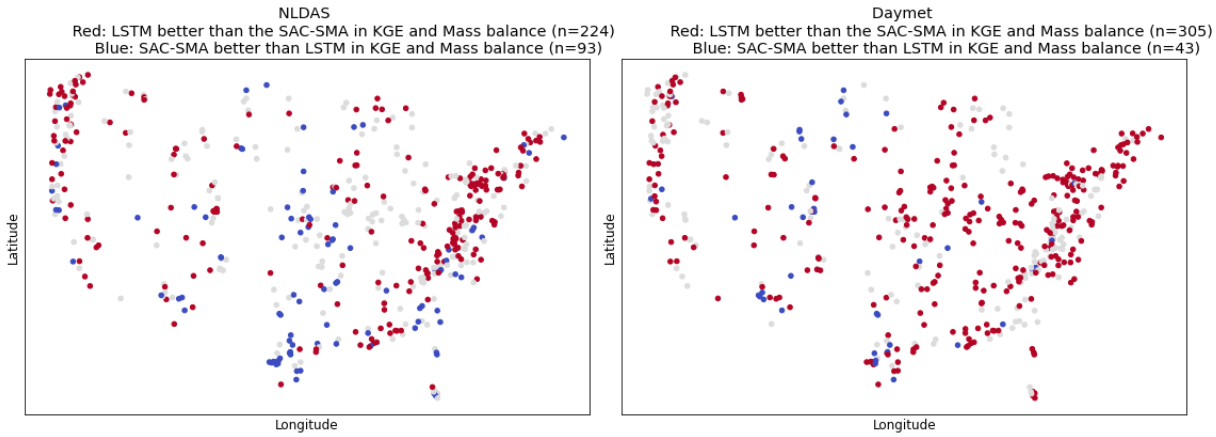

Figure 15. Shows that spatial distribution of basin which LSTM or SAC-SMA perform better in both efficiency and in mass balance.