Review

This article demonstrates a data-driven approach to simulate extreme stream flow temperatures. The authors conducted an extensive ablation of input features and loss terms in training LSTMs for extreme streamflow temperature. They develop one catchment LSTM across multiple stations in the Garonne catchment and compare this with local LSTMs, trained on stations separately. In their ablation the authors focus especially on 1) using multi-scale data that is catchment scale forcing and auxiliary information, and 2) modifications to the loss term that aim at penalizing extreme values

While the approach to optimising simulation of extreme temperature through targeted loss functions would be a valuable experiment to show and learn, one major thing I am sceptical about is the evaluation and choices regarding their ablation. In order to make this a more robust part of their study, I would kindly ask the authors to add the full picture to their methods description on the evaluation and consider the comments below to revise or discuss their comparative evaluation.

Because statistical testing was used for quantifying significant performance differences, the choice on the scoring function gets blurred in the evaluation framework (section 3.3). My reading is that all models were evaluated on the test set with the MAE computed on the upper 10% - which is a limitation, and this choice needs to be much more visible in the methods. E.g., its not given that this is a robust choice for example for figure 5 where MAEs are computed over the whole test period.

From this point of view, issues I see that need to be address are:

a) MAE and MSE as loss functions optimize on different properties of the predictive distribution: The MAE optimizes toward median and the MSE toward the mean, and which one is appropriate is, in principle, non-arbitrary (e.g. 10.5194/gmd-15-5481-2022). What for this work matters is that mean and median are only equivalent if the residual distributions are symmetric - which I assume is not the case for streamflow temperature, hence the focus on capturing extreme values. Using only the MAE as an evaluation score on models that are trained with MSE is not a fair comparison if the target distributions are asymmetric. This may particulary be relevant for evaluation on whole test periods, i.e. for figure 5.

b) An evaluation with MAE for the median may be inconsistent with the training objective – hence the evaluation only on the 10% highest values, i.e. an evaluation of the median of that specific subset. But training with high values of mu or lambda may force the model to optimize on higher quantiles than 10% percent. This may be a possible explanation for what the authors state in l. 306 (section 4.1).

c) It should be made more obvious if the same lookbacks were chosen for all the models, or if the comparative evaluation that we see in Figures 2 and 3 allows varying lookbacks, as they were seemingly part of the ablation too (explained in section 3.3).

Fig. 3: So the middle row left and right columns use the same loss function, while the bottom row different ones? Clarify.

Minor comments

l. 190: This builds quite a tension arc towards section 3.3., but in this section you don't state a best lookback length either. Could you be more clear about your choice, and at best already state here which lookback length you chose?

Section 3.3.: Here the choice of the lookback is stated as part of an original ablation, so do lookback lengths vary across the results we see? (see also c) in major comment above)

Section 3.2.1: nice description of the loss term.

Tab. 2: Listing input variables and models of those variables feels a bit redundant. I'd say a table of just input variable combinations is enough. Regarding CatAttrs – are reach attributes are explicitly named in Tab.1 or do they fully intersect with catchment attributes?

l. 470: The punctuation in this sentence is weird.