Review of "Customized Deep Learning for Precipitation Bias Correction and Downscaling" by Wang *et al.* (gmd-2022-213)

The article proposes some improvements to the authors model for downscaling precipitation presented in Wang *et al*, (2021) where they use the loss function MSE instead of MAE.

The three proposed improvements are:

- using a weighted MAE as the loss instead of the MAE,
- using a second loss function on a quantized version of the upscaled Stage IV data
- and including other coarse grained predictors.

They evaluate these improvements in the task of downscaling hourly precipitation from the coarse grained MERRA2 (50km^2) to the fine grained Stage IV (4km^2) in a rectangle coastal area of the Gulf of Mexico covering the states of Alabama, Mississippi and Louisiana. They evaluate the performance of models by comparing the KGE-score on different aggregations as well as extreme events. The authors conclude that all three of their proposed improvements are helpful. In two marginal notes, the authors evaluate whether coarse-grained predictors make precipitation redundant as an input and whether model performance is related to its complexity. While they state the first to be negative, they state the second to be true.

The problem of downscaling precipitation is relevant and tailoring proven deep learning methods to this problem is a valuable contribution. However, the presented study has severe issues that considerably weaken its interpretation and the possible impact of the study considerably. Further, parts of the manuscripts need major updates.

This article requires **major revisions** before publication.

General comments

Study

Unfortunately, the results presented in Tables 3 and 4 are not enough to estimate the usefulness of the three proposed improvements. The differences between the "Scenarios" 2 to 6 are marginal and the order differs a lot between tasks and metrics. This is especially critical, since the chosen method (a deep neural network) is inherently stochastic and hence, differences between different "Scenarios" might be due to this stochastisity. This stochastisity is further increased by the special training method that the authors use. Instead of presenting each time step once in each epoch, they present random 1897 independent sampled batches of 64 random time-steps (which should be mentioned in the manuscript). To distinguish between these random effects and the effect of the proposed improvements it would be necessary to run the models multiple time and assess the significance of the differences between results.

Further, to evaluate the three different improvements independently it would be interesting to test and compare all eight possible combinations. The paper, unfortunately, only reports results on five of the eight combinations. Having the three missing combinations (standard MAE + categorical Loss, standard MAE + covariates and standard MAE + categorical Loss + covariates) will help immensely in disentangling the effects of the individual improvements.

The only difference that is apparent without the need of a test of significance is the difference between the "Scenarios" that use the weighted MAE and the "Scenario" that use the standard MAE. However, it is unclear, why the authors choose to modify the baseline (Wang *et al.*, 2021) to use a MAE instead of a MSE. In fact, to understand the performance of the proposed model in comparison to the state of the art, a comparison to the original baseline would be very helpful. Especially since

the baseline reached a KGE of 0.951 on a slightly different task, which indicates that it might be very competitive. Especially to support claims like "These results highlight the advantages of the customized DL model compared with regular DL models as well as traditional approaches, which provides a promising tool to fundamentally improve precipitation bias correction and downscaling and better estimate P at high resolution."[lines 502 to 505]

Since the central result of the paper seems to be that MAE is not a suitable loss for downsampling precipitation, the paper should include a discussion on why someone might consider this to be a sensible idea in the first place, which will amplify the impact of the result. However, at this stage, the motivation for this change in the baseline is unclear from the paper and (to the best of my knowledge) it was not suggested to use MAE in the literature (at least not in the related work presented in the paper).

Finally, the two side nodes of whether the coarse grained precipitation can be excluded as a predictor ("Scenario4") and whether larger models are overfitting in this specific example do not fit naturally in this study and distract from the main point of the study. Since both of them cannot be answered significantly from the results, I recommend to exclude them to further the readability.

Manuscript

Many of the above mentioned comments have implications on the manuscript. For example the discussion is quite long and discusses many aspects that are not reflected in the experimental results in any significant way (lines 307-314, 330-340, 349 - 351, 357-362, 367-375, 380-387, 388-392, 405-419, 422-428, 447-479). I recommend to focus the discussion of results mainly on significant results to not "over-interpret" the results and, consequently, "over-claim".

Further, many of the figures require more work. Figure 1 should maybe reference the very similar figure in (Wang *et al.*, 2021). Many figures have incomplete or no colorbars. Figure 5 and 6 are hard to read and I recommend to exclude them. The interpretation of Figure 9 is unclear.

Additionally, it would be helpful if the motivation for each of the three individual contributions is clearly stated in the beginning of the paper.

Further, the structure of the paper is unclear, for example "Data and methodology" is combined into one section, but is immediately split into two parts, which are data and methodology in 2.1, 2.2. It would be helpful to my understanding of the manuscript to restructure the work.

The notation of different models as "Scenarios" is confusing.

Often the choice of references is confusing. For example Li *et al.* (2021) is cited for IoU even though the paper includes no information on IoU that is not also included in this manuscript. This is just an representative example for other cases.

Finally, the interpretation of the KGE, more specifically the interpretation of β and γ is surprising. The authors, for example, state that "Scenario1" "highly overestimated the variability"[line 306] however, if we calculate

$$\frac{\sigma_s}{\sigma_o} = 0.37$$

indicating, that the variance is actually under estimated.

Summary

In summary I believe that the study aims to close a relevant research gap. Further, the proposed method of testing different models with combinations of different improvements is effective. By repeating the experiments to reach significant results, comparing the results to a state-of-the-art baseline and adding more explanation on the motivation of the proposed changes, the paper will be a valuable contribution.