Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., author comment AC6
https://doi.org/10.5194/gmd-2022-213-AC6, 2022

Reply on RC6

Fang Wang et al.

Author comment on "Customized deep learning for precipitation bias correction and downscaling" by Fang Wang et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-213-AC6, 2022

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 (50km2) to the fine grained Stage IV (4km2) 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.

*Response: Thank you for your comments. We agree with you that stochasticity can play an important role. To address your concern, we have run each scenario three more times (4 in total for each scenario) and evaluated the stochasticity with the mean and standard deviation of the four runs. The computational time for each scenario is 20 to 22 hours and running four times for each scenario is feasible for us to consider stochasticity for this study. The results are presented in Table S2 and S3 in the Supplement document, which is discussed in the discussion section (Section 5) of the updated document. Table S1 indicates that Scenario2 to Scenario6 are significantly different (with p-value 0.05 and confidence interval mean±2*standard deviation) for hourly precipitation evaluation in terms of KGE metric, which indicates the differences among the six scenarios are not caused by stochasticity. For daily aggregation, Scenario5 and Scenario6 are significantly better than Scenario2 to Scenario4, which is consistent with the results and stated in Section 4.3. For monthly aggregation, Scenario4 is significantly worse than Scenario2, Scenario3, Scenario5 and Scenario6 and no significant differences are found among Scenario2, Scenario3, Scenario5 and Scenario6, which is consistent with the findings in Section 4.3. For extreme indices (see Table S3), Scenario4 is significantly worse than Scenario3, Sceanrio5 and Scenario6 at 99th percentile, which is consistent with the findings in Section 4.4. For the annual maximum wet spell index, Scenario2 and Scenario3 are significantly better than other scenarios. We have modified the statements in the result section (Section 4) to make sure the findings are consistent with the significance evaluation from the total four runs in the updated manuscript. We have included a paragraph in the discussion section on the stochasticity evaluation as follows "Due to the stochastic nature of DL models, we ran each DL scenario for additional three times (four times in total) to evaluate the effects of stochasticity comparing with the added value of each customized component of DL models (see Table S2 and Table S3 in the Supplement). The results show that KGE values for each scenario are significantly different at p-value of 0.05 at hourly time scale, which indicates that the added value of each customized component is not caused by model stochasticity. Scenario1 is significantly worse than other scenarios including QDM_BI at hourly and aggregated time scales as well as extreme indices, emphasizing the added value of weighted loss function. Scenario5 and Scenario 6 are significantly better than other scenarios including QDM_BI in terms of KGE values at hourly and aggregated time scales, and Scenario4 is significantly worse at monthly time scale. For the 99th percentile extreme index, Scenario4 is significantly worse than Scenario3, Sceanrio5 and Scenario6. For the annual maximum wet spell index, Scenario2 and Scenario3 are significantly better than other scenarios. All these stochastic significance evaluation results are consistent with the findings in Section 4. Due to computational demand (20 to 22 hours for running each scenario once) and resource limits, we ran limited times for each scenario to consider the stochasticity of DL models and incorporating DL models with Bayesian inference is a potential way to quantify systematic uncertainty caused by model itself as indicated by Vandal et al. (2018a)."*

Furthermore, we have noted the total number of iterations for each scenario in Section 3.1.5 of the updated manuscript.

- 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.

Response: Thank you for your comments. Since we have shown that standard MAE has difficulties handling highly unbalanced hourly precipitation data and tends to highly underestimate precipitation (see Table 3, Table 4, and Figures 3 to 8), it is not necessary to test the combination of the standard MAE with other settings.

- 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]

Response: Thank you for your comments. Wang et al. (2021) used regular MSE as loss function, which works well for downscaling daily precipitation through synthetic experiments with no bias, since the precipitation data was first coarsened and then downscaled into the original fine scale. However, in this study we considered two different datasets at hourly scale with large discrepancies (or biases). Particularly, the Stage IV radar observations, as the training target, include outliers (extremely large values). The MSE loss (the square operation) makes the algorithm very sensitive to these outliers (see Ravuri et al., 2021). Ravuri et al. (2021) applied a UNet based architecture (as a baseline) for precipitation nowcasting with radar data and they stated that "We also found that including a mean squared error loss made predictions more sensitive to radar artefacts; as a result, the model is only trained with precipitation weighted mean average error loss.". Here "mean average error" is MAE. Based on the findings from Ravuri et al. (2021), we decided to use MAE instead of MSE as a loss function. We have added the following statement in Section 3.1.4 of the updated manuscript: "Wang et al. (2021) used regular mean squared error (MSE) as loss function, which works well for downscaling daily precipitation through synthetic experiments with no bias, since the precipitation data was first coarsened and then downscaled into the original fine scale. However, in this study, the coarse resolution MERRA2 has substantial biases compared to Stage IV radar data and Stage IV radar data also includes artefacts (e.g., spurious large values) (Nelson et al., 2016). The previous study has shown that the MSE loss function is more sensitive to radar artefacts than the mean absolute error (MAE) loss function (Ravuri et al., 2021). Therefore, we chose MAE as a regular loss function in this study."

- 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).

Response: Thank you for your comments. We have explained why we chose MAE instead of MSE as a regular loss function in the previous response. Your comment that "MAE is not a suitable loss function for downsampling precipitation" is true for downscaling hour precipitation. Regular MAE may work for downscaling daily precipitation with limited biases (Sha et al., 2020a), but to our knowledge, there are no successful cases using regular MAE for downscaling hourly precipitation due to much higher unbalance issues (more no rains for hourly than daily). We have added the following explanations in the discussion

Section 5 of the updated manuscript: "Regular MAE has been used for downscaling daily precipitation data with limited biases in previous studies (e.g., Sha et al., 2020a), but to our knowledge, there are no successful cases using regular MAE for downscaling hourly precipitation data with large biases."

- 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.

Response: Thank you for your comments. The reason that we include Scenario4 is to test whether only using covariates is sufficient for estimating hourly P as stated in Section 3.1.5. In addition, the importance of including the bias-corrected (achieved in MERRA2) coarse-grained precipitation is more clear by comparing Scenario4. Furthermore, considering multitask learning is also positive on certain aspects (e.g., the extreme event in Figure 8 and classification result in Figure 9), even though the improvement is not consistently significant due to including more training parameters. With more and more datasets available in the future, including multitask learning concept may work better someday and gives readers more options. In summary, we think including Scenario 4, Scenario3 and Scenario6 can provide useful information for future research.

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".

Response: Thank you for your comments. We have made significant modifications to the manuscript based on the stochastic evaluation to make sure our findings are consistent and do not over-interpret the results.

- 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.

Response: Thank you for your comments. We have included the reference of Wang et al. (2021) in the caption of Figure 1. We have added units for the colorbars in the caption of each figure. While it is hard to distinguish 8 lines in one plot, Figures 5 and 6 clearly show the performance of different DL scenarios in comparison with QDM_BI, which provides useful information. We have added more information about the IOU metric in Section 3.3 and included examples to interpret IOU metric in the results section.

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

Response: Thank you for your comments. We have added the following explanations in the introduction section of the updated manuscript: "Traditional DL loss functions have difficulties handling hourly precipitation data that are highly unbalanced with many zeros and highly positively skewed for nonzero components, therefore, customized DL with weighted loss function to better balance nonzero components has the potential to improve the DL model performance. Besides the primary task of downscaling and bias correction

task, adding a highly relevant classification task has the potential to improve DL model performance on the primary task. Incorporating covariates selected based on precipitation formation theory (cloud mass movement and thermodynamics) into the DL model also have the potential to improve precipitation downscaling and bias correction. ”

- 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.

Response: Thank you for your comments. We have separated Data and methodology into two sections (Data and Study Area in Section 2, and Methodology in Section 3). The following section numbers are changed accordingly.

- The notation of different models as “Scenarios” is confusing.

Response: Thank you for your comments. We have explicitly described each scenario in Section 3.1.4 of the experiment design and in Table 1. We also added scenario settings for Table 3 and Table 4 and made modifications to make them more clear in the revised manuscript.

- 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.

Response: Thank you for your comments. Li et al. (2021) used intersection over union metric, but they used the short name IOU instead of IoU. So we have changed IoU to IOU in the updated manuscript. We have checked other areas about references inconsistence and made modifications as necessary.

- 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 σs/σo = 0.37, indicating, that the variance is actually under estimated.

Response: Thank you for your comments. The metric γ in the modified KGE is defined as a ratio of estimated and observed coefficients of variation (see Eqn.6,) instead of . Using instead of ensures that the bias and variability ratios are not cross-correlated as stated by Kling et al. (2012). We claimed that Scenario1 highly overestimated the variability (i.e., higher) and “variability” means coefficient of variation (i.e.,) instead of . We have made a note in Section 3.3 of the updated manuscript to emphasize the differences and made necessary modifications.

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.

References

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 424, 264-277, 2012.

Li, Z., Wen, Y., Schreier, M., Behrangi, A., Hong, Y., and Lambrigtsen, B.: Advancing satellite precipitation retrievals with data driven approaches: Is black box model explainable?, Earth and Space Science, 8, e2020EA001423, 2021.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., and Madge, S.: Skilful precipitation nowcasting using deep generative models of radar, Nature, 597, 672-677, 2021.

Sha, Y., Gagne II, D. J., West, G., and Stull, R.: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part II: Daily precipitation, Journal of Applied Meteorology and Climatology, 59, 2075-2092, 2020a.

Vandal, T., Kodra, E., Dy, J., Ganguly, S., Nemani, R., and Ganguly, A. R.: Quantifying uncertainty in discrete-continuous and skewed data with Bayesian deep learning, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2377-2386. 2018.

Wang, F., Tian, D., Lowe, L., Kalin, L., and Lehrter, J.: Deep learning for daily precipitation and temperature downscaling, Water Resources Research, 57, e2020WR029308, 2021.