Response to Reviewer #1

Thank you for your comments. Below are our replies to each of your comment.

We believe that the planned changes will improve the clarity and significance of our manuscript.

Major Comments:

Comment: In introduction, L45-70 introduce the advantages and disadvantages of dynamic downscaling and statistical downscaling, which are well known to us, and just to illustrate a few deep learning methods. Failure to highlight the research focus of this paper, i.e., the progress of research on traditional statistical downscaling and statistical downscaling combined with deep learning.

Reply: We appreciate the reviewer's comment. We will revise the introduction to emphasize the progress made in statistical downscaling methods, particularly when combined with deep learning techniques. This will better align the introduction with the specific research focus of our paper.

Comment: I do not understand the role of L70-75, it seems to describe weather prediction models like Pangu and GraphCast which can achieve entirely through deep learning have demonstrated the potential to achieve forecast skills comparable to state-of-the-art numerical weather prediction systems. Is this relevant to the research in this paper?

Reply: The mention of Pangu and GraphCast was intended to introduce the state-of-the-art deep learning models for weather prediction in a broader context. To bridge the gap in deep learning models from the broader context to the focus of our paper, we will revise these lines to ensure that this literature review remains directly relevant to our study.

Comment: About result, some of the conclusions seem too brief. L255-260 have no results about the relative error stands, the mean absolute error and relative error of simulated maximum daily flow.

Reply: We acknowledge the need for more detailed descriptions of results. We will expand this section to include the results of relative error and mean absolute error of the simulated maximum daily flow against the observations.

Comment: The author writes "while the RMSE of EC-QM forecasts sees a relatively steady reduction over all lead times"(L280-285), but from the Fig.4, we can clearly find that EC-QM showed more RMSE compared to EC for the lead times of 23-24 days, 27 days. The author can explain the reason? And please draw the spatial distribution of the bias of EC-CNN, EC-QM and EC because it mentioned in Line 310.

Reply: We agree with the Reviewer that the QM method does not reduce the RMSE in an exactly steady manner with lead time. In particular, for lead times such as 23, 24 and 27 days, QM-based forecasts show a slightly larger RMSE than that of raw forecasts. Our finding is in line with these of recent studies using QM for statistical downscaling precipitation. For example, Li et al. (2023), Huang et al. (2022), and Mao et al. (2015) show that, while QM is generally effective in adjusting model bias towards observations, it does not always lead to improvements of the forecast accuracy.

One plausible reason could be the limited applicability of the relatively simple QM method. Specifically, QM primarily adjusts the distribution of forecasted values towards the distribution of historical observations, with no account of the atmospheric conditions associated with those forecasts. However, physics-based numerical weather predictions involve complex and nonlinear errors that QM may not be able to fully correct. For example, for lead time of 23 days, the QM consistently predicts lower precipitation than raw forecasts for light rain events across all initial dates, while the actual biases are positive for some initial dates and negative for other dates. Another reason could be that QM does not explicitly account for errors in the spatial dependencies of precipitation field, which can also contribute to the increased RMSE in specific lead times.

We will revise the manuscript to include the above-mentioned explanations and to add relevant references accordingly. We will provide comprehensive discussions about the limitations of QM in the context of sub-seasonal forecasting. For visualization, a spatial plot of forecasting error will be added.

References

Li, X., Wu, H., Nanding, N., Chen, S., Hu, Y., & Li, L. (2023). Statistical Bias Correction of Precipitation Forecasts Based on Quantile Mapping on the Sub-Seasonal to Seasonal Scale. *Remote Sensing, 15*(7), 1743. https://doi.org/10.3390/rs15071743

Huang, Z., Zhao, T., Xu, W., Cai, H., Wang, J., Zhang, Y., Liu, Z., Tian, Y., Yan, D., & Chen, X. (2022). A Seven-Parameter Bernoulli-Gamma-Gaussian Model to Calibrate Subseasonal to Seasonal Precipitation Forecasts. *Journal of Hydrology, 610*, 127896. https://doi.org/10.1016/j.jhydrol.2022.127896

Mao, G., Vogl, S., Laux, P., Wagner, S., & Kunstmann, H. (2015). Stochastic Bias Correction of Dynamically Downscaled Precipitation Fields for Germany Through Copula-Based Integration of Gridded Observation Data. *Hydrology and Earth System Sciences, 19*(4), 1787-1806. https://doi.org/10.5194/hess-19-1787-2015

Comment: In Chapter 4.3, "the EC-QM (EC-CNN) forecasts reduce the relative error of the raw forecasts by 12% (20%), 16% (24%) and 9% (21%), respectively, and reduces the relative error of maximum daily flow by 27% (29%), 16% (32%) and 11% (18%)", whether EC-QM or EC-CNN showed the RMSE decreased more or the lead times of 11-20 days than 1-10 days, but we think that the smaller the lead time, the better the results. Can the author give a reasonable explanation?

Reply: Thank you for your comment. To clarify, Chapter 4.3 in the original manuscript discusses the relationship between the *improvements* due to statistical downscaling techniques like QM or CNN and the lead time. While it is true that the forecast accuracy generally decreases with longer lead times (as shown in Figures 4 and 7), this relationship may not be the case when examining the *improvements* brought about by QM or CNN for different lead times. For example, many studies have shown that, in specific cases, the improvement from statistical downscaling could be larger for longer lead times and smaller for shorter lead times (Zhang et al., 2023; Lyu et al., 2023; Li et al. 2024).

We will revise the manuscript to clarify this distinction, ensuring that the discussion on the varying impact of QM and CNN with lead time is clear and well-supported by the literature.

References

Zhang, T., Liang, Z., Li, W., Wang, J., Hu, Y., & Li, B. (2023). Statistical Post-Processing of Precipitation Forecasts Using Circulation Classifications and Spatiotemporal Deep Neural

Networks. Hydrology and Earth System Sciences, 27(10), 1945-1962. https://doi.org/10.5194/hess-27-1945-2023

Lyu, Y., Zhu, S., Zhi, X., Ji, Y., Fan, Y., & Dong, F. (2023). Improving Subseasonal-To-Seasonal Prediction of Summer Extreme Precipitation Over Southern China Based on a Deep Learning Method. Geophysical Research Letters, 50(24), e2023GL106245. https://doi.org/10.1029/2023GL106245

Li, L., Yun, Z., Liu, Y., Wang, Y., Zhao, W., Kang, Y., & Gao, R. (2024). Improving Categorical and Continuous Accuracy of Precipitation Forecasts by Integrating Empirical Quantile Mapping and Bernoulli-Gamma-Gaussian Distribution. Atmospheric Research, 298, 107133. https://doi.org/10.1016/j.atmosres.2024.107133

Comment: Since the article analyses individual cases, in conjunction with Figure 8, please give quantitative indicators to analyse the description.

Reply: We agree that quantitative analysis will strengthen the discussion of individual cases. We will include quantitative indicators alongside Figure 8 to provide a more robust analysis of the streamflow forecasts, comparing them across different scenarios.

Minor Comments:

Comment: L38: "S. Zhu et al., 2020" should be modified.

Reply: We will correct the citation formatting in L38 as recommended.

Comment: L60: "Wilby, R. L., et al., 2004; Vrac, M., & Friederichs, P., 2015" should be modified.

Reply: We will adjust the citation formatting in L60 for consistency and correctness.

Comment: L87-88: "For example, Humphrey et al. (2016) achieved improved streamflow forecast skills by combining Bayesian artificial neural networks with traditional models of GR4J." What's the conclusion?

Reply: We will clarify the conclusion drawn from Humphrey et al. (2016) in the revised manuscript.

Comment: L142-143: "streamflow and flood peak forecasts of XAJ-LSTM and standalone XAJ driven by EC-CNN forecasts are then quantitatively evaluated against those driven by raw and QM-based forecasts using a series of metrics. " Please write the full name for the first occurrence.

Reply: We will ensure that the full names of all abbreviations are provided upon their first occurrence in the manuscript.

Comment: L262-263: "The results indicate that the daily Nash-Sutcliffe Efficiency (NSE)", abbreviations have already been mentioned.

Reply: We will correct this by removing the repeated explanation of the abbreviation.