Response to Reviewer #2

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.

Comment: Although QM is a widely applied technique and should be familiar to the general audience of HESS, the reviewer believes that some key information or references may need to be incorporated. Precipitation data is known to be highly skewed, making the selection of the distribution function critically important for the effectiveness of QM. However, such details appear to be missing in the current manuscript.

Reply: In our current implementation of QM, a specific distribution function is not required as we use a non-parametric approach. This method directly adjusts the quantiles of the forecasted and observed data without assuming a specific distribution. To be specific, the empirical cumulative distribution functions of observed and forecasted daily precipitation are built respectively, and each percentile of the forecasted data is adjusted to match the corresponding percentile in the observed data. Dry days with a precipitation amount less than 0.1 mm are excluded from the derivation of empirical cumulative distribution functions. Several studies have indicated the effectiveness of this approach in improving the overall precipitation forecasts (Manzanas et al., 2018; Cannon et al., 2015).

We will revise the manuscript to provide details on the implementation of the QM approach and provide the necessary references to support it.

References

Manzanas, R., Lucero, A., Weisheimer, A., & Gutiérrez, J. M. (2018). Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts? Climate dynamics, 50, 1161-1176.

Cannon, A. J., Sobie, S. R., & Murdock, T. Q. (2015). Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? Journal of Climate, 28(17), 6938-6959.

Comment: Additionally, the reviewer is curious about the specific implementation of QM. Was seasonality or the variation in forecast lead times considered in the QM-based bias removal process? More detailed documentation on this aspect should be included in the manuscript.

Reply: In our current implementation, the QM is constructed separately for each lead time to account for forecast bias variations across different lead times. For each lead time, a single model is applied across all months, which is aligned with the structure of the CNN model built in this study. This approach is considered able to more effectively capture biases across different lead times while maintaining a uniform correction across the seasonal cycle.

We will clarify this aspect in the manuscript and provide more details on the implementation of the QM process to ensure a clear understanding of the methodology.

Comment: Furthermore, given that QM is primarily designed for bias removal rather than enhancing the temporal correspondence between forecast time series and observations, the reviewer suggests that the authors also evaluate the resulting forecasts (both precipitation and streamflow) in terms of their bias. Specifically, the overall CDF of precipitation forecasts generated by different statistical downscaling methods should be compared. Given that these downscaled precipitation forecasts eventually run through lumped hydrologic models, CDF of the areal averaged precipitation forecasts is perhaps a good way to demonstrate the bias condition at all percentiles across the study region. While the proposed DL technique improves the predictive skill of S2S precipitation, it would be valuable to see whether it also reduces forecast bias compared to QM.

Reply: We agree that while QM primarily addresses bias removal, it would be beneficial to evaluate both the precipitation and streamflow forecasts in terms of bias. In response to your suggestion, we plan to compare the cumulative distribution functions (CDFs) of the areal-averaged precipitation forecasts generated by QM and the proposed deep learning (DL) framework, which could provide a comprehensive view of the bias at all percentiles.

We believe this additional analysis will allow us to assess not only the predictive skill but also the extent to which the DL technique reduces forecast bias compared to QM. We will incorporate this comparison into the revised manuscript to offer a clearer picture of the effectiveness of our approach in reducing forecast bias.

Comment: The reviewer feels that the description of the employed statistical downscaling techniques is unclear in general. It appears that the proposed CNN-ResNet generates a single precipitation prediction value while using multiple spatially distributed forecast variables as inputs (Figure 2). If this is indeed the case, the proposed framework seems more like an "upscaling " rather than "downscaling" technique. This also raises questions about how the authors produced the spatially distributed precipitation climatology plot (Figure 6). Additionally, given that CNN-based structures typically produce square-shaped outputs, were any masks applied during the training of the proposed CNN-ResNet?

Reply: The CNN incorporating residual blocks, or the CNN-ResNet framework, in our study is indeed a downscaling technique. Specifically, it downscales ECMWF S2S reforecasts from a coarse 1.5-degree resolution to a finer 0.25-degree resolution, which corresponds to the resolution of the CN05.1 observation-based dataset. The CNN uses spatially distributed inputs (e.g., geopotential height, temperature, humidity) from the ECMWF dataset, covering 7×7 1.5-degree coarse grid cells centered on the target 0.25-degree fine grid cell. Due to the square-shaped input structure of the CNN model, some ECMWF data from outside the basin boundary are included in the input. For the outputs, the CNN loops over each fine-resolution grid cell (0.25 degrees) within the basin boundary, generating a high-resolution precipitation forecast for the region of interest. Therefore, no masks are applied during training, as the model is trained to predict each fine-resolution grid cell individually within the basin boundary.

We will clarify this process in the manuscript to avoid any potential confusion about the methodology and ensure that the CNN-ResNet framework is well understood.

Comment: Similarly, is QM conducted at each pixel across the study watershed? If so, does the raw spatial resolution of the S2S precipitation forecast match that of the reference precipitation? These questions are particularly relevant considering the employed hydrologic models are lumped. It is

important to clarify for the audience at which specific technical step(s) the spatially distributed forecast variables are converted into area averages.
In general, it is recommended that the entire methodology section be revised to avoid potential confusion and to ensure clarity on the steps involved in the downscaling process.

Reply: The raw ECMWF S2S precipitation reforecasts are released at a 1.5-degree resolution, while the reference CN05.1 observation dataset is at a 0.25-degree resolution. To match the forecast resolution with reference dataset resolution, QM is performed by looping over 0.25-degree fine grid cells to establish the empirical cumulative distribution function for each fine grid cell based on its corresponding 1.5-degree coarse grid cell from ECMWF reforecasts. This method is widely used in statistical downscaling to account for the resolution mismatch between coarse forecast models and high-resolution observational data (Gudmundsson et al., 2012).

Given that the hydrologic models employed in this study are lumped, the spatially distributed precipitation forecasts (both from QM and CNN) are converted to area averages after the downscaling step. This averaging occurs within the study area before the precipitation data are input into the hydrologic models.

We agree that these technical steps should be clarified to avoid any confusion. We will revise the methodology section to clearly outline the process, including how the downscaling is conducted at each pixel and how the spatially distributed data are subsequently converted into area averages for use in the lumped hydrologic models.

References

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., & Engen-Skaugen, T. (2012). Downscaling RCM precipitation to the station scale using statistical transformations – A comparison of methods. Hydrology and Earth System Sciences, 16(9), 3383-3390. doi:10.5194/hess-16-3383-2012

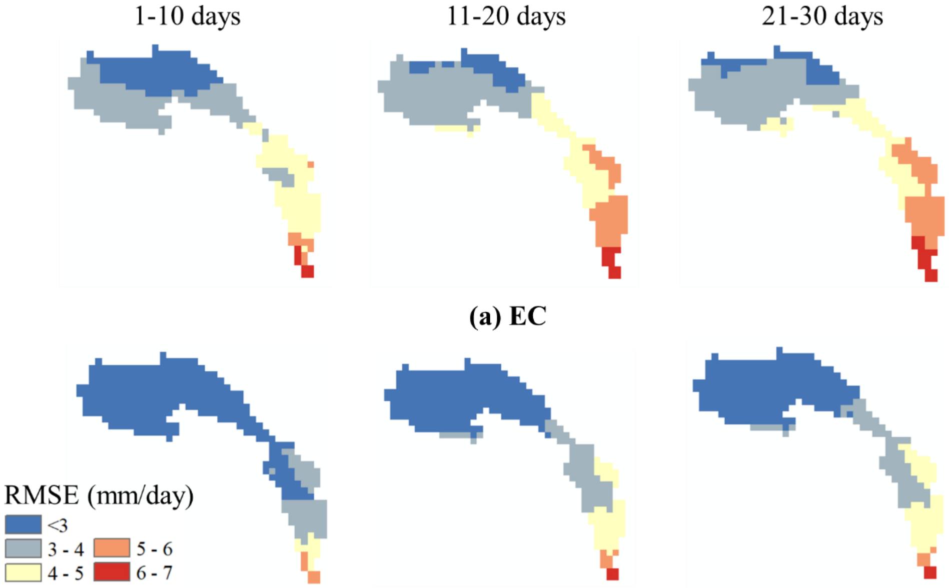
Comment: The reviewer suggests conducting additional seasonal and spatial analysis to better highlight the strengths and weaknesses of the proposed CNN-ResNet downscaling technique. First, the reviewer notes that the study watershed covers a broad area (around 10 degrees, or approximately 1000 km, in both the north-south and east-west directions). This suggests significant spatial and seasonal variability in terms of precipitation generation mechanisms, magnitude, frequency, etc. However, this variability is not discussed in the manuscript, limiting the audience's understanding of the study watershed.
Building on this, it would be interesting to examine whether the proposed framework is equally effective across different seasons and geospatial locations, or if its performance varies. The reviewer believes such an analysis would be crucial in further enhancing the quality of the manuscript. Consequently, it is recommended that the authors evaluate the post-processed precipitation both spatially and seasonally. Since the proposed method is a statistical downscaling technique, it is important to demonstrate its skill over such a large study region. This additional analysis would provide valuable insights into the effectiveness of the method across different conditions.

Reply: Thank you for your suggestion. We fully understand the importance of examining spatial and seasonal variability across such a large and diverse study area.

While we agree that spatial analysis is crucial due to the broad area covered by the basin and its variability in precipitation generation mechanisms, our downscaling and model evaluations were all conducted during the wet season, specifically from May to August. This period is chosen because it accounts for the majority of annual precipitation, which is key to water resources management and flood prevention in the region. Therefore, an evaluation across different seasons may not be the primary focus of this paper. We will revise the manuscript to highlight the focus is on the wet season.

However, we fully agree that conducting a spatial analysis within the wet season, with the geographic variability (e.g., northern plateau versus southern hill regions) taken into account, will provide important insights into the performance of the CNN-ResNet model. This analysis will allow us to assess if the model performs consistently across the different parts of the watershed and help identify any areas where its performance may vary.

A preliminary spatial analysis on the RMSE of EC and EC-CNN forecast for lead times of 1-10, 11-20 and 21-30 days is presented in the Figure R1 below. It can be seen that the EC-CNN improves the forecast skills of the raw ECMWF forecasts over the majority of the basin for all lead times. For example, the RMSE is reduced from 3.4 mm/day to an average of 2.2 mm/day at the northern headwaters of the basin for the lead time of 21-30 days. Similar improvements can also be seen around the southern part of the basin.



(a) EC

(b) EC-CNN

Figure R1. RMSE of EC and EC-CNN forecasted precipitation for lead times of 1-10, 11-20 and 21-30 days.

We will perform a comprehensive spatial analysis and revise the manuscript accordingly.

Comment: If the potential workload is manageable, the reviewer strongly recommends that the authors utilize the entire ensemble of S2S precipitation forecasts from ECMWF in their experiments, rather than focusing on the ensemble means. The primary reason for this suggestion is that neither precipitation forecasts nor the corresponding streamflow predictions can be applied deterministically at a subseasonal timescale due to limited skills at longer forecast lead times.

At this timescale, probabilistic forecasts are typically constructed using multiple predictions (i.e., ensemble forecasts). While the proposed framework appears effective and interesting, the reviewer believes its full potential can be better demonstrated with a revised experimental design that aligns more closely with real-world needs (i.e., ensemble predictions).

Following this suggestion, the reviewer suggests the authors to incorporate additional probabilistic evaluation metrics, such as CRPS or CRPSS, for a more comprehensive assessment of the framework's performance for both post-processed precipitation forecasts and the corresponding streamflow predictions.

Reply: We agree with your suggestion to incorporate ensemble-based predictions, as probabilistic forecasts are more suitable for sub-seasonal timescales due to the inherent uncertainties at longer lead times (Li et al., 2019; Ferranti et al., 2018). To address this, we will use all ensemble members from the ECMWF S2S precipitation reforecasts and build ensemble CNN models to generate probabilistic forecasts. For the probabilistic evaluation, we will apply Continuous Ranked Probability Skill Score (CRPSS) to the ensemble forecasts. This metric is widely used for evaluating ensemble precipitation forecasts and accounting for uncertainties across multiple ensemble members (Bremnes, 2020).

We will revise the manuscript to present probabilistic forecasting and its benefits for improving the predictability of precipitation and streamflow on sub-seasonal timescales. Thank you again for this suggestion.

References

Bremnes, J. B. (2020). Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. Monthly Weather Review, 148(1), 403-414.

Li, W., Pan, B., Xia, J., and Duan, Q. (2021). Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. Journal of hydrology, 605, 127301.

Ferranti, L., Corti, S., & Janousek, M. (2018). Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. Quarterly Journal of the Royal Meteorological Society, 144(712), 317-326.

Comment: Lien 126: What is the naive spatial resolution of the collected S2S precipitation forecasts from ECMWF?

Reply: The S2S precipitation reforecasts from ECMWF collected in this study are with a spatial resolution of 1.5 degrees. We will clarify this point in the revised manuscript to ensure the spatial resolution of the input data is clearly understood.

Comment: Line 142: EC-CNN is referenced here for the first time in the manuscript, but without a clear explanation.

Reply: We will provide a more detailed description when first introducing EC-CNN, which refers to the statistically downscaled ECMWF S2S reforecasts using the proposed CNN framework.

Comment: Line 250: It seems a standardized metric is employed here (i.e., NSE) to evaluate the hydrologic model calibration. The reviewer wonders why switch to RMSE and other metrics for later streamflow predictive skill evaluation? While RMSE is a widely applied metric in many fields, standardized metrics such as NSE and KGE might be more familiar to researchers in the hydrology community.

Reply: It is true that the hydrologic model calibration is evaluated using the Nash-Sutcliffe Efficiency (NSE). For consistency and to align with hydrologic model evaluation metrics, we agree that it would be appropriate to add NSE for the later streamflow predictive skill evaluation as well, and we will revise the manuscript accordingly.

Comment: Line 354: Perhaps “forecast issue date” is more appropriate for the titles of different panels in Figure 8. Also, it would be interesting to see these examples where the proposed framework delivers more accurate streamflow predictions. Overall skill evaluation would still be more informative in general. Perhaps these figures could be included in the supplementary material so that previous suggested additional evaluation and analysis could be included in the main manuscript.

Reply: We agree that using ‘forecast issue date’ would be more appropriate for the titles of different panels in Figure 8. We will update the figure accordingly in the revised manuscript.

Additionally, we will follow your recommendation to prioritize more comprehensive evaluation and analysis in the main manuscript, while moving very detailed figures and additional examples to the supplementary material. We believe this will allow for a clearer focus on the overall skill evaluation in the main text, while still providing valuable examples for interested readers.