Responses to Comments on “A statistical-dynamical approach for probabilistic prediction of sub-seasonal precipitation anomalies over 17 hydroclimatic regions in China” (Referee #1)

Anonymous Referee #1 Received and published on 23 August 2023.

Our responses are in blue, with the reviewer’s comments shown as normal text.

General comment:

Skillful and reliable sub-seasonal precipitation forecasts are complicated as the sources of predictability are much fewer than short-medium range and seasonal predictions. This study proposes a Spatial-Temporal Projection based Calibration, Bridging, and Merging (STP-CBaM) method to improve sub-seasonal precipitation forecast skills. The manuscript presents many results demonstrating that the STP-CBaM method can provide skillful and reliable sub-seasonal precipitation forecasts using dynamic and statistical models. The strategy appears novel and therefore merits publication after minor revisions.

Thanks for your comprehensive review and recognition of this study. Your constructive comments will help us improve our manuscript after revision.

Major comment:

1. This manuscript uses the intraseasonal oscillation signals forecasted by the ECMWF model as predictors. However, the forecast skill of intraseasonal oscillation is also limited at long lead times. Thus, it is essential to know the potential prediction skill of the STP-CBaM method when observing intraseasonal signals used as predictors.

Thanks for this comment. We agree that it is important to know the potential skill of the STP-CBaM method for sub-seasonal precipitation forecasts. Here, we use the p^{th} 10-60-day signal of atmospheric field derived from ERA5 reanalysis dataset as predictor for the bridging model, instead of the atmospheric field derived from the ECMWF model. Figure S1 presents the potential CRPS skill score of merged forecasts over China. The CRPS skill scores are mostly over 20% even when the lead time is beyond 15 days. Figure S2 shows the differences between potential CRPS skill score and practical CRPS skill score of merged forecasts. The potential CRPS skill scores are slightly lower than the practical CRPS skill scores as the precipitation forecasts derived from the ECMWF model are of high accuracy at short lead times. The potential CRPS skill scores are much higher than the practical CRPS skill scores at longer lead times. This indicates that the forecast skill will be greatly improved when the atmospheric field is well predicted in the GCMs. We will add the above analysis in the revised manuscript.

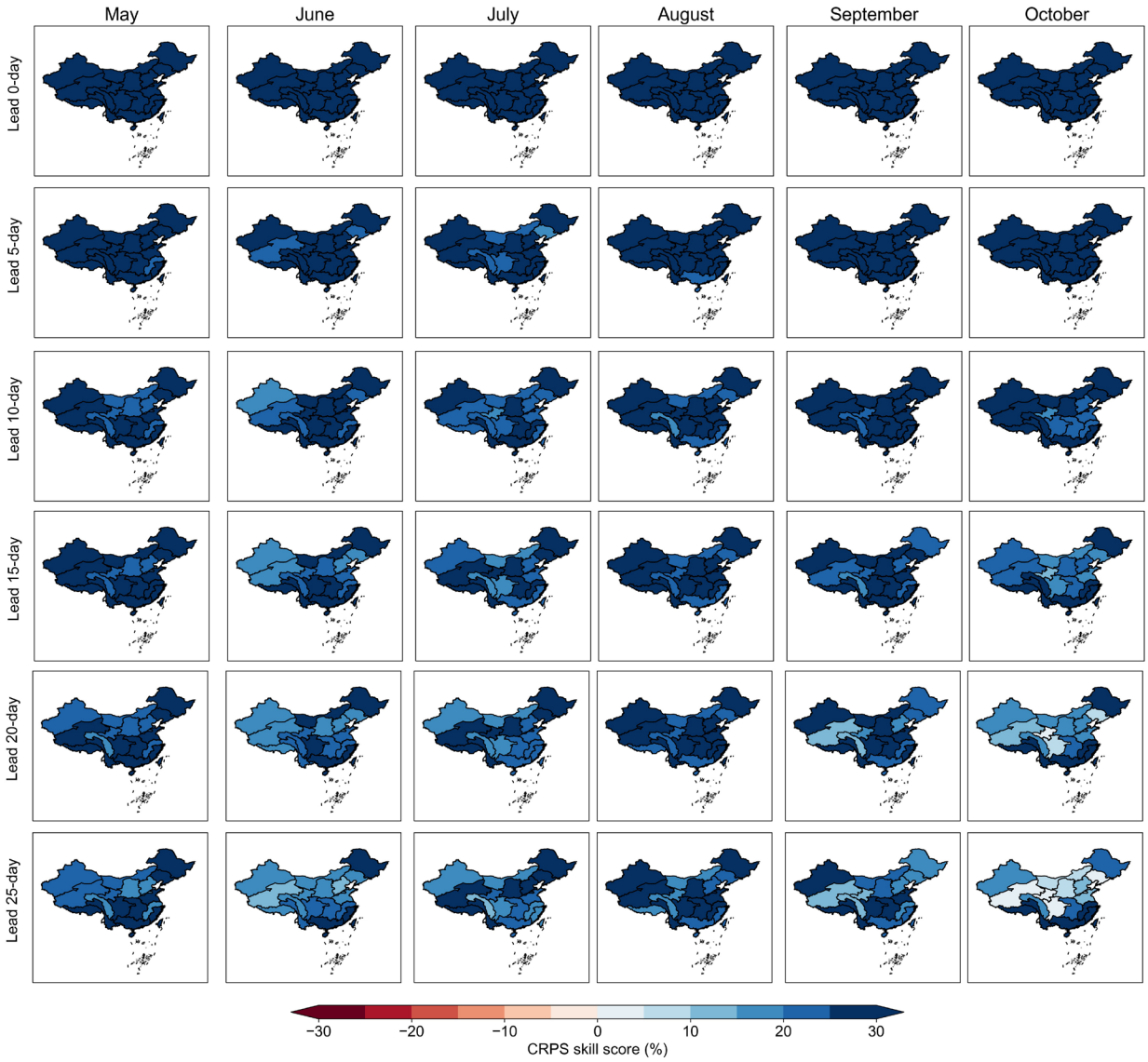

Figure S1. The potential CRPS skill score of merged forecasts over China

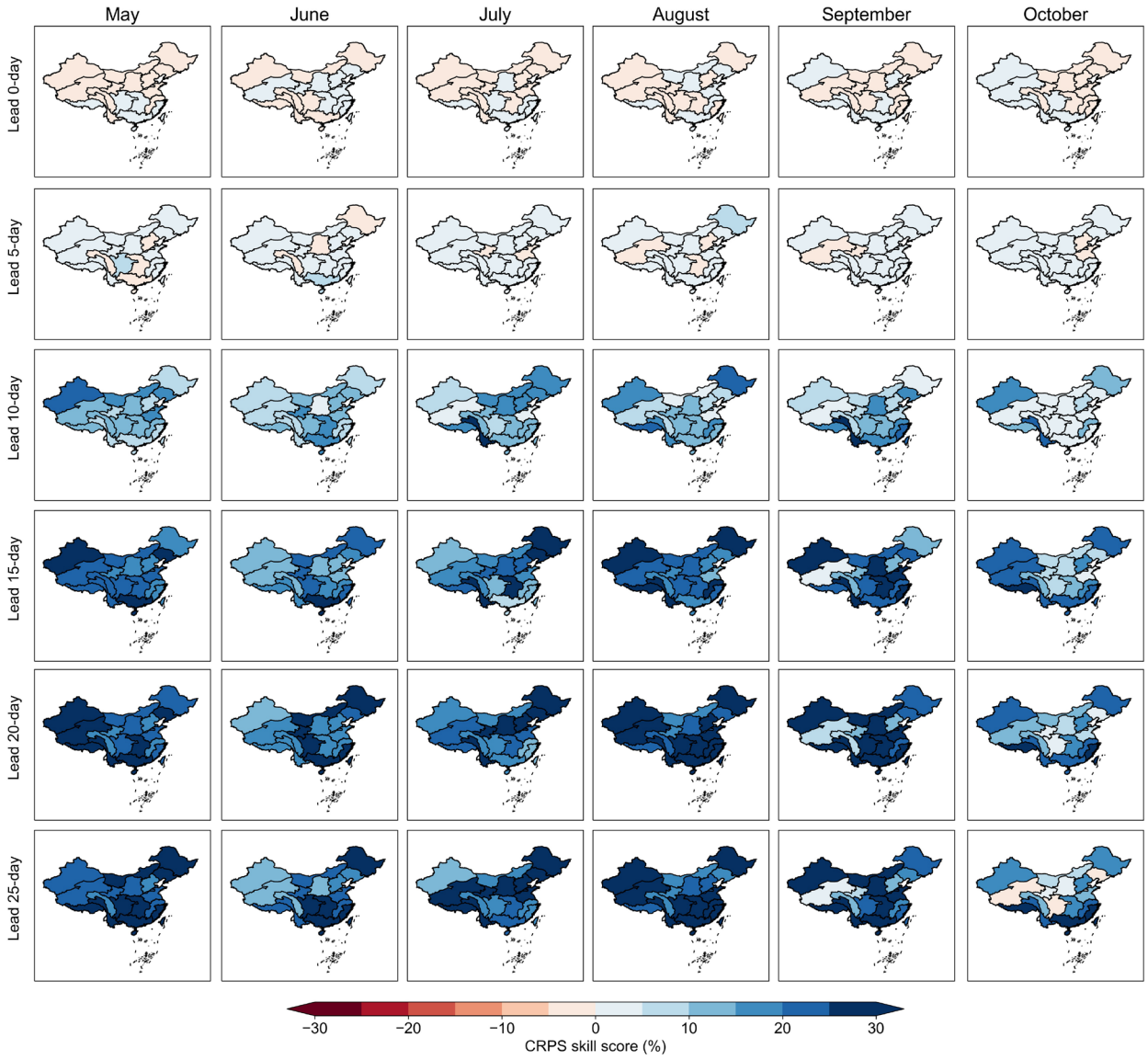

Figure S2. The differences between potential CRPS skill score and practical CRPS skill score of merged forecasts over China

Minor comments:

1. Page 1, Line 9,".." is. ".

Thanks for this comment. We will incorporate this suggestion in the revised manuscript.

2. Page 13, The graphical aspect of Figure 4 and Figure 5 could be improved (e.g., colored bars, etc.).

Thanks for this comment. We have revised the color scheme with discrete color for each level. This will make it easier to identify grid points where the correlation coefficients are statistically significant at the 5 % level. Figure 4 and Figure 5 have been revised to improve the visualization as follows:

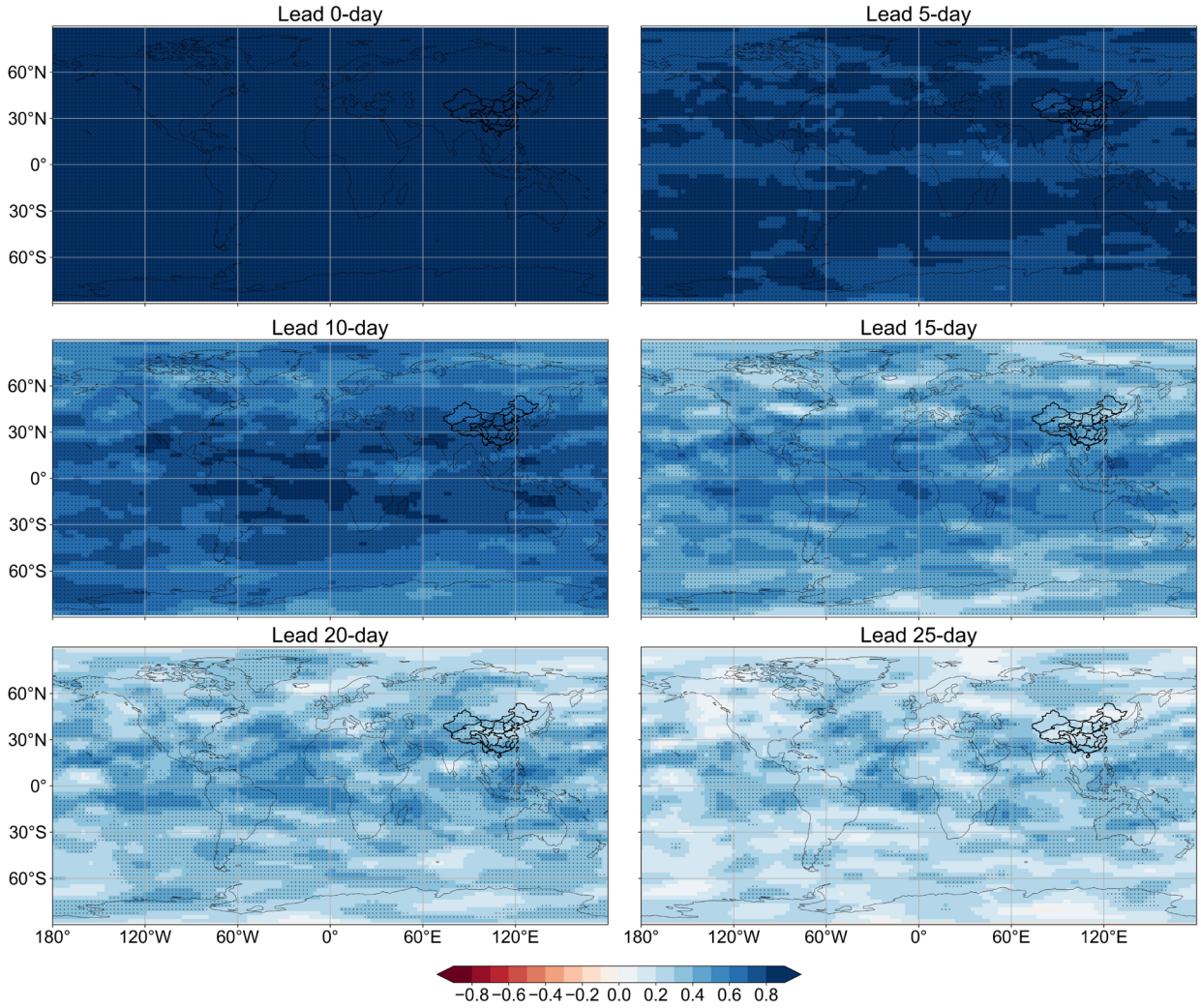

Figure 4. Temporal correlation coefficient (TCC) of the ensemble mean of U200 intraseasonal signals derived from the ECMWF model compared to the ERA5 reanalysis data in May. Correlation coefficients that are statistically significant at the 5 % level are shaded.

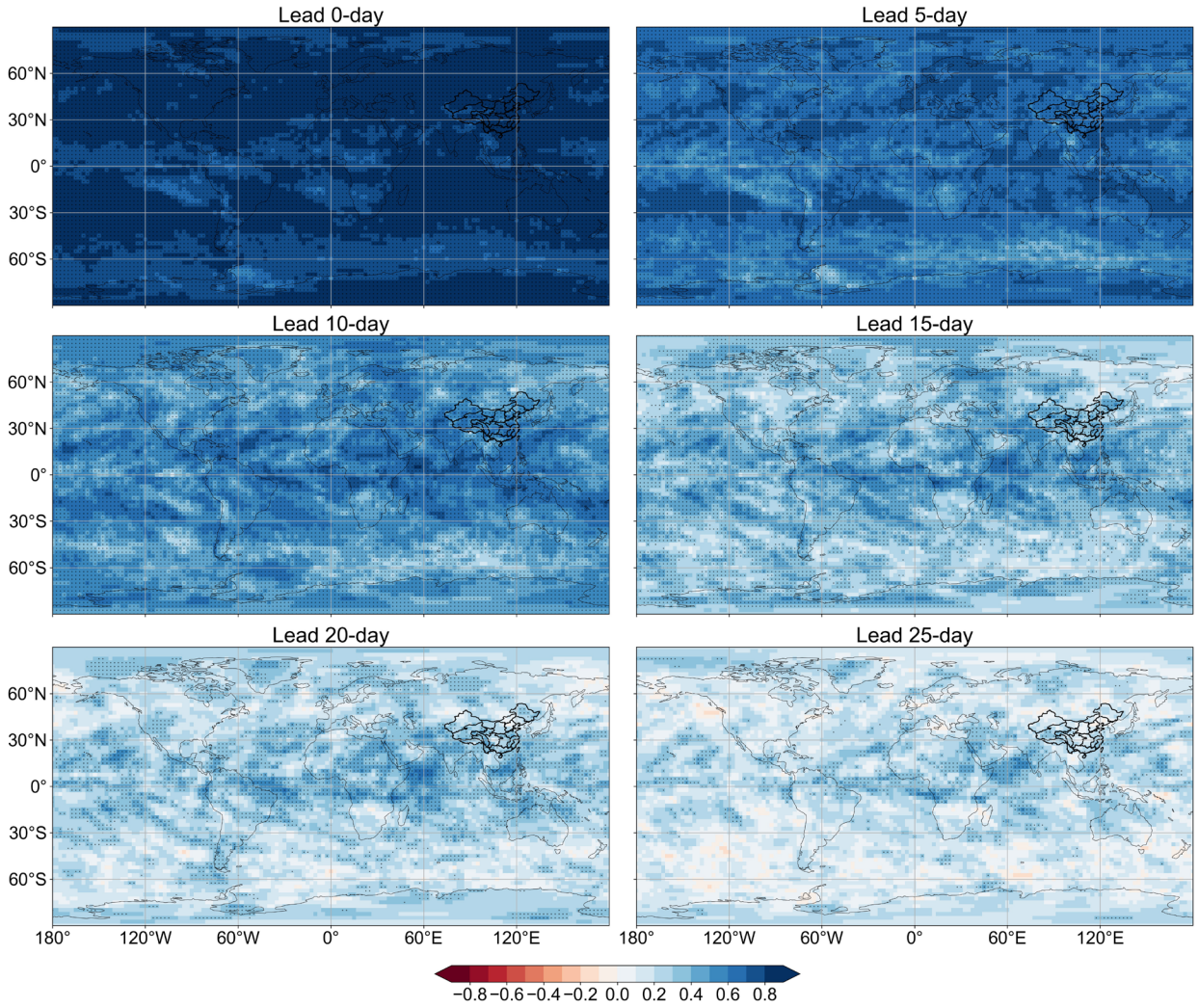

Figure 5. Same as Figure 4, but for OLRA.

3. Page 14, The CRPS skill scores shown in Figure 6 indicate that the STP-CBaM method can provide more skillful forecasts than the calibration and bridging models alone. However, I noticed that the STP-CBaM forecasts are of lower prediction skill than the calibration model in several regions. The authors should provide some explanations.

Thanks for this comment. We agree that the calibration model outperforms the STP-CBaM model in several regions. To have a better explanation of the results, we compare the CRPS skill of merged forecasts to the CRPS skill score of calibrated forecasts, maximum, mean, and minimum CRPS skill score of bridging forecasts as shown in Figure S3. The CRPS skill scores of merged forecasts lie around the calibrated forecasts at short lead times. Although the CRPS skill scores of merged forecasts are slightly lower than the calibrated forecasts, the CRPS skill scores of merged forecasts are always higher than the minimum CRPS skill scores of bridging forecasts. This indicates that the merged forecasts at least appear to

moderate the worst forecast errors. We will further revise the conclusions to have a more accurate description.

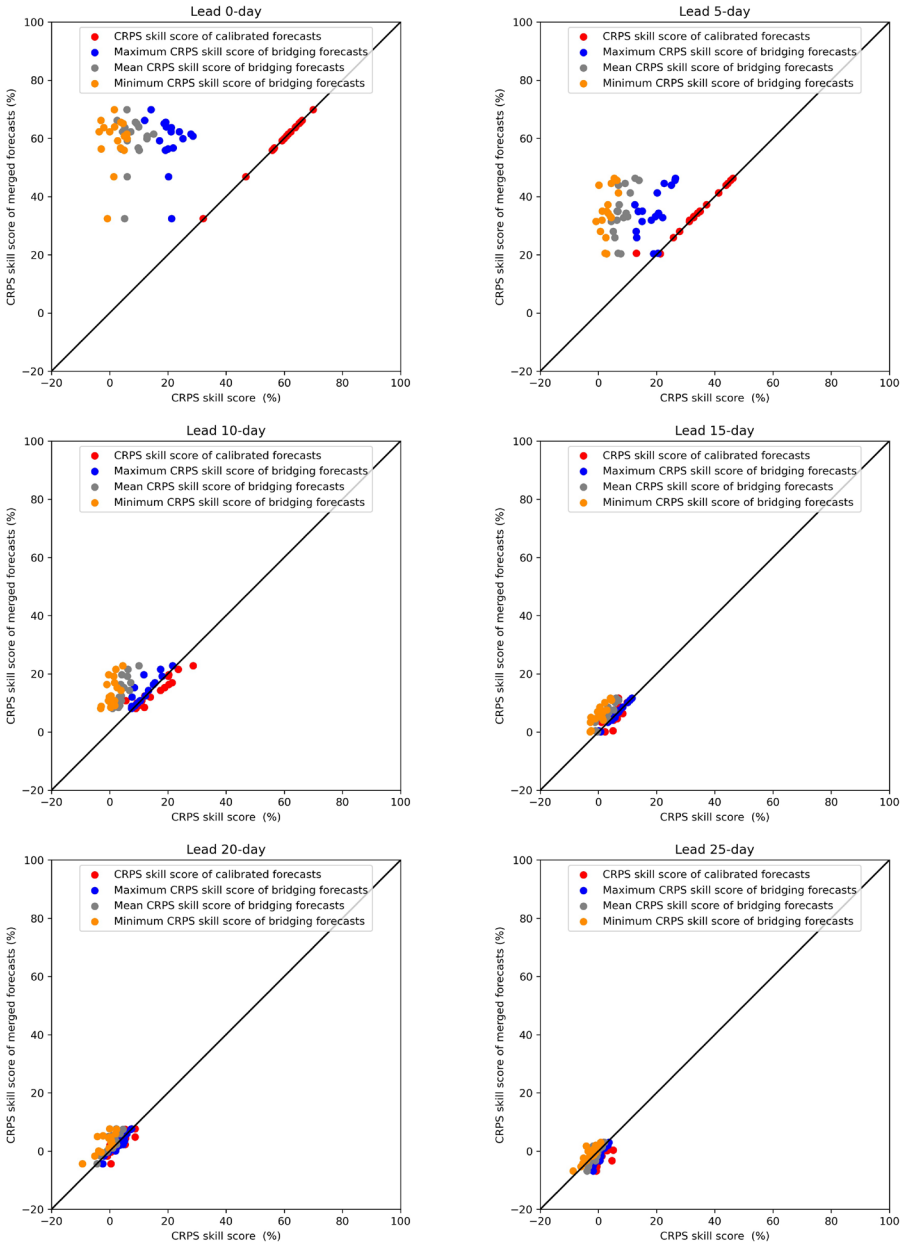

Figure S3. Comparison of the CRPS skill of merged forecasts to the CRPS skill score of calibrated forecasts, maximum, mean, and minimum CRPS skill score of bridging forecasts in May.

4. Page 15, The authors should explain the weights' results in Figure 7.

Thanks for this comment. We noticed that the skill patterns were not always match the weights as suggested by Anonymous Referee #2. At 15-day lead time the weights of OLRA are higher than calibration and U200 in Region 1, but the CRPS of the OLRA is lower than U200 and calibration. In this study, the posterior distributions of model weights are given as

$$p(w_k, k = 1, \cdots, K | x_k^T, y^T, f_k(y|x_k), k = 1, \cdots, K) \propto \prod_{k=1}^{K} (w_k)^{\alpha-1} \prod_{t=1}^{T} \sum_{k=1}^{K} w_k {f_k}^{(t)}(y^t|x_k^t) \quad (14)$$

where ${f_k}^{(t)}(y^t|x_k^t)$ is the cross-validated predictive density.

This indicates that the weights are assigned by the model predictive ability rather than fitting ability. Indeed, there is much literature in support of using predictive performance measures for model choice and combination based on the idea that a model is only as good as its predictions (Eklund and Karlsson, 2007; Stock and Watson, 2006). Thus, the CRPS skill score is not used when inferring model weights. This may lead to the discrepancy between model weights and forecast skill score, especially when none of the models show high predictive skill.

We will have a more detailed discussion on the discrepancy between the model weights and forecast skill in the revised manuscript.

5. L. 359~361: "The values of α-index are mostly over 0.7 for all hydroclimatic regions and lead times, suggesting that the merged forecasts are of high reliability."

I am unsure if the value of the α-index exceeds 0.7, indicating high reliability. I suggest providing some figures to prove such conclusions.

Thanks for this comment. To figure out the differences in reliability between 0.7 and 0.9, we analyze the merged forecasts over Region 3 (Inland Rivers in Inner Mongolia) in May at a lead time of 0-day. The α-index of merged forecasts is around 0.6, suggesting that the merged forecasts are of low reliability. We also investigate the model weights of calibrated forecasts and bridging forecasts. The results suggest that the calibrated forecasts are more important than bridging forecasts, which the cross-validated model weights are over 0.95. This suggests that the low reliability of merged forecasts is mostly caused by the low reliability of calibrated forecasts. Figure S4 presents the quantile ranges of calibrated forecasts and merged forecasts against time. The quantile ranges of both calibrated forecasts and merged forecasts are small, suggesting the forecasts are too narrow (too confident). However, we also note that the forecast accuracy of calibrated forecasts is high, which the CRPS skill score is over 60%. We would like to focus on improving the forecast reliability in the future.

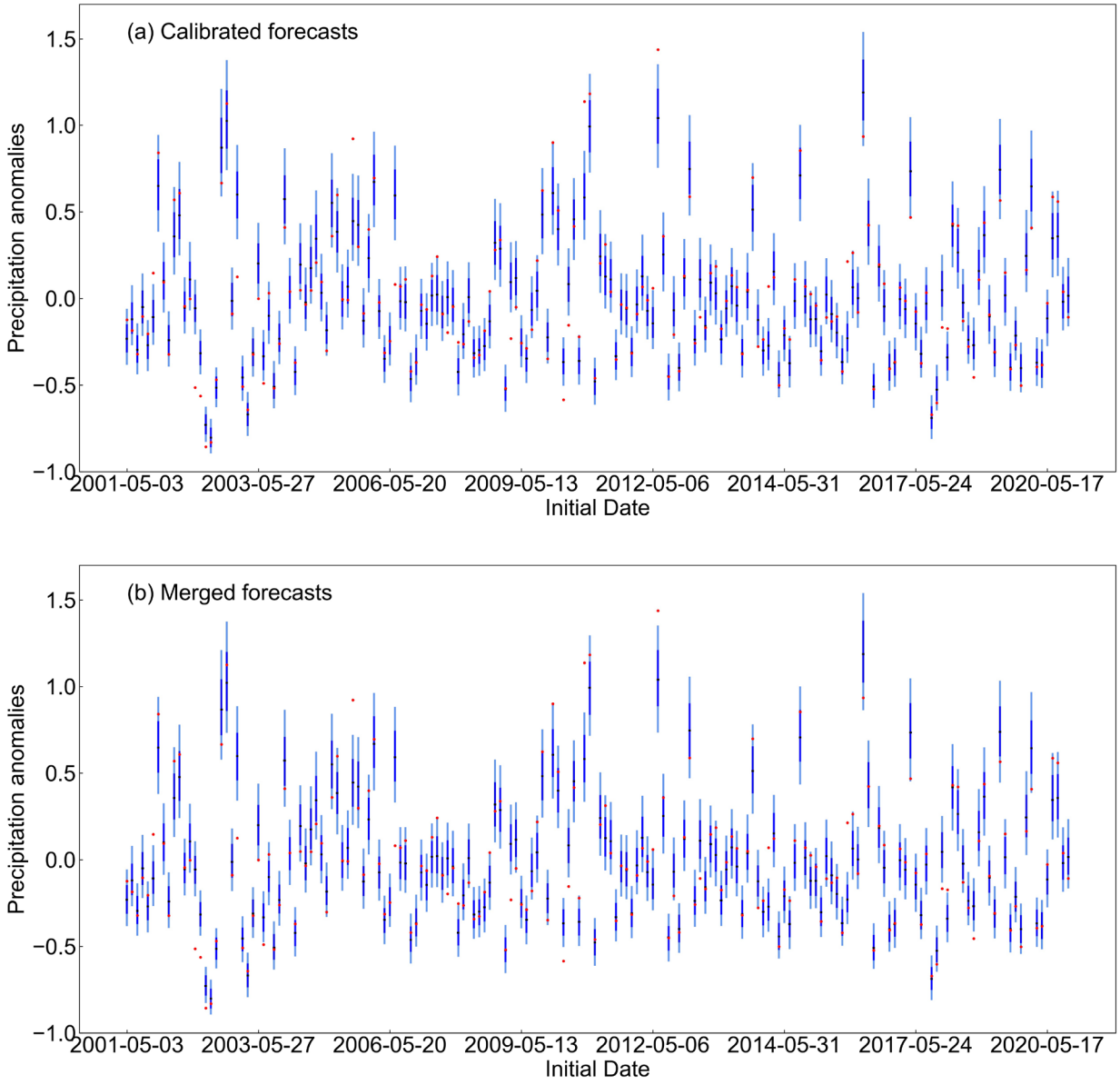

Figure S2. Forecast median, quantiles ranges and observed value against time for sub-seasonal forecasts over Region 3 (Inland Rivers in Inner Mongolia) at a lead time of 0-day. The black dots, forecast median; dark blue vertical line, forecast [0.25, 0.75] quantile range; light and dark blue vertical line, forecast [0.10, 0.90] quantile range; red dot, observed precipitation anomalies.

6. L. 365: The authors should also discuss recent progress on the prediction of the East Asian Monsoon. The prediction skill of extreme events should also be addressed.

Liu, B., Zhu, C., Ma, S., Yan, Y., & Jiang, N. (2023). Subseasonal processes of triple extreme heatwaves over the Yangtze River Valley in 2022. Weather and Climate Extremes, 40, 100572.

Yan, Y., Zhu, C., & Liu, B. (2023). Subseasonal predictability of the July 2021 extreme rainfall event over Henan, China, in S2S operational models. Journal of Geophysical Research: Atmospheres, 128(4), e2022JD037879.

Zhu, C., Liu, B., Li, L., Ma, S., Jiang, N., & Yan, Y. (2022). Progress and Prospects of Research on Subseasonal to Seasonal Variability and Prediction of the East Asian Monsoon. Journal of Meteorological Research, 36(5), 677-690.

Thanks for this comment. We will incorporate this suggestion in the introduction section.

Eklund, J. and Karlsson, S.: Forecast Combination and Model Averaging Using Predictive Measures, Econometric Reviews, 26, 329-363, 10.1080/07474930701220550, 2007.

Stock, J. H. and Watson, M. W.: Chapter 10 Forecasting with Many Predictors, in: Handbook of Economic Forecasting, edited by: Elliott, G., Granger, C. W. J., and Timmermann, A., Elsevier, 515-554, https://doi.org/10.1016/S1574-0706(05)01010-4, 2006.