Responses to Comments on “A statistical-dynamical approach for probabilistic prediction of sub-seasonal precipitation anomalies over 17 hydroclimatic regions in China” (Referee #2)

Anonymous Referee #2 Received and published on 12 September 2023.

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

General comment:

This paper investigates the application of calibration, bridging and merging to forecast subseasonal precipitation anomalies in China based on ECMWF model output of precipitation and atmospheric circulation patterns (zonal winds, geopotential height, OLR). Observations are taken from ERA5, MSWEP and NOAA. Individual models are constructed using BJP and then merged with BMA. Forecast performance is assessed using leave-one-year-out cross-validation and in terms of CRPS, reliability and model weights. It is found that calibration is dominant at short lead times (5-10 days) and U200 and OLRA have increasing relevance at longer lead times. It is concluded that the BMA forecasts have best overall skill and the forecasts are reliable.

Overall, the paper is structured well and the presentation of figures is appropriate. The methods can be followed sufficiently well. My concerns with the paper lie around the marginal performance differences and the strength of the conclusions, particularly around outperformance and reliability. I therefore have some moderate comments for the authors to address ahead of publication, mostly minor, but perhaps requiring some further analysis.

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

Major comment:

Abstract: First line, add a supporting statement about where subseasonal forecasts are of value or delete.

Thanks for this comment. We will remove the first line in the revised manuscript.

Abstract: I suggest adding details about the study area

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

L37: is it better to say the variability is too slow (rather than too short)?

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

Section 2.1: Suggest some commentary on the quality of the MSWEP rainfall dataset and underlying sources over the 17 regions.

Thanks for this comment. We will include some commentary on the quality of the MSWEP rainfall dataset as follows:

This dataset is developed by optimally merging precipitation data derived from gauge, satellite, and reanalysis datasets. It covers the period from 1979 to near recent with a spatial resolution of 0.1° × 0.1°. Many studies have found that the MSWEP dataset is of high quality over China (Li et al., 2023b; Liu et al., 2019; Guo et al., 2023).

Li, Y., Pang, B., Zheng, Z., Chen, H., Peng, D., Zhu, Z., and Zuo, D.: Evaluation of Four Satellite Precipitation Products over Mainland China Using Spatial Correlation Analysis, Remote Sensing, 15, 1823, 2023b.

Liu, J., Shangguan, D., Liu, S., Ding, Y., Wang, S., and Wang, X.: Evaluation and comparison of CHIRPS and MSWEP daily-precipitation products in the Qinghai-Tibet Plateau during the period of 1981–2015, Atmospheric Research, 230, 104634, https://doi.org/10.1016/j.atmosres.2019.104634, 2019.

Guo, B., Xu, T., Yang, Q., Zhang, J., Dai, Z., Deng, Y., and Zou, J.: Multiple Spatial and Temporal Scales Evaluation of Eight Satellite Precipitation Products in a Mountainous Catchment of South China, 10.3390/rs15051373, 2023.

Figures 4 & 5: suggest using different colors for the outline

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:

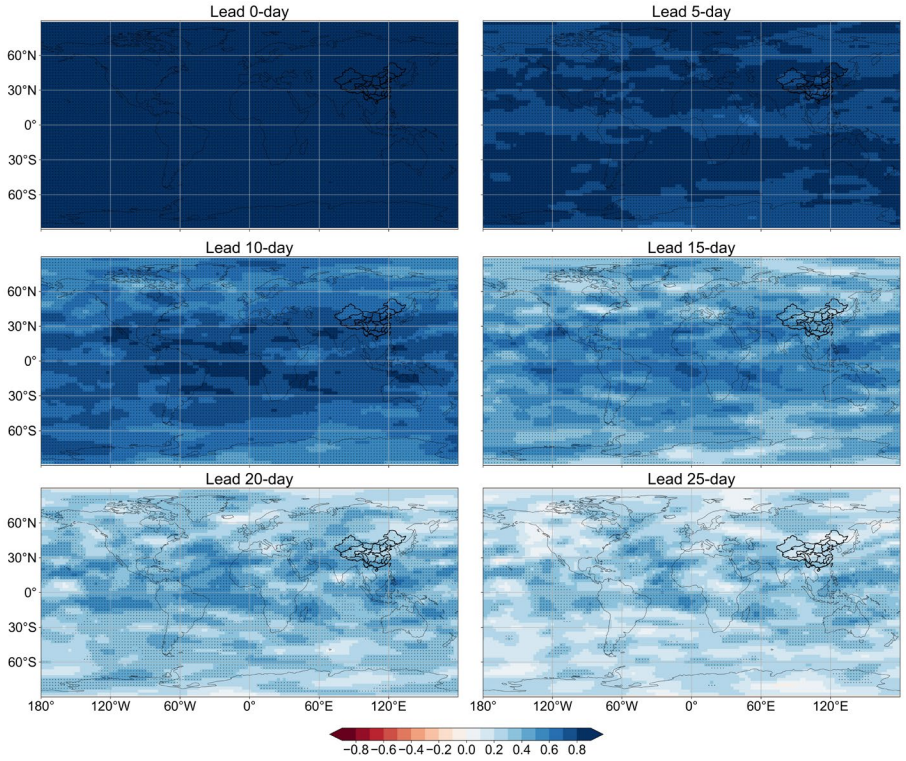

Figure 4. Temporal correlation coefficient (TCC) of the ensemble mean of U200 intraseasonal signals derived from the ECMWF model compared to the ERA5 reanalysis data in May. Correlation coefficients that are statistically significant at the 5 % level are shaded.

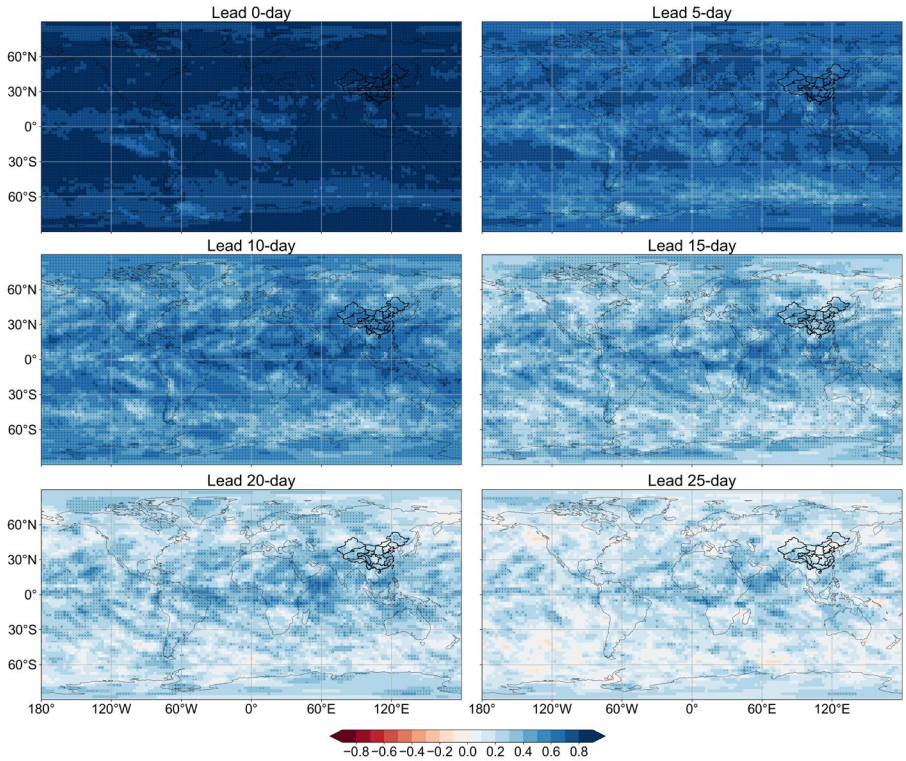

Figure 5. Same as Figure 4, but for OLRA.

L328-330: It is difficult to interpret the differences visually from blue shading on the maps. To my eye the difference between BMA skill and calibration skill is marginal and sometimes the merged skill is even lower. I suggest the authors find some way to highlight that indeed “the BMA CRPS skill scores are higher compared to both calibration and bridging”. Perhaps include some statistics.

Thanks for this comment. We will compare the CRPS skill scores of merged forecasts to the calibrated forecasts, maximum, mean, and minimum CRPS skill score of bridging forecasts as shown in Figure S1. We agree that the CRPS skill scores of merged forecasts are slightly lower than the calibrated forecasts in several regions. Nevertheless, 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.

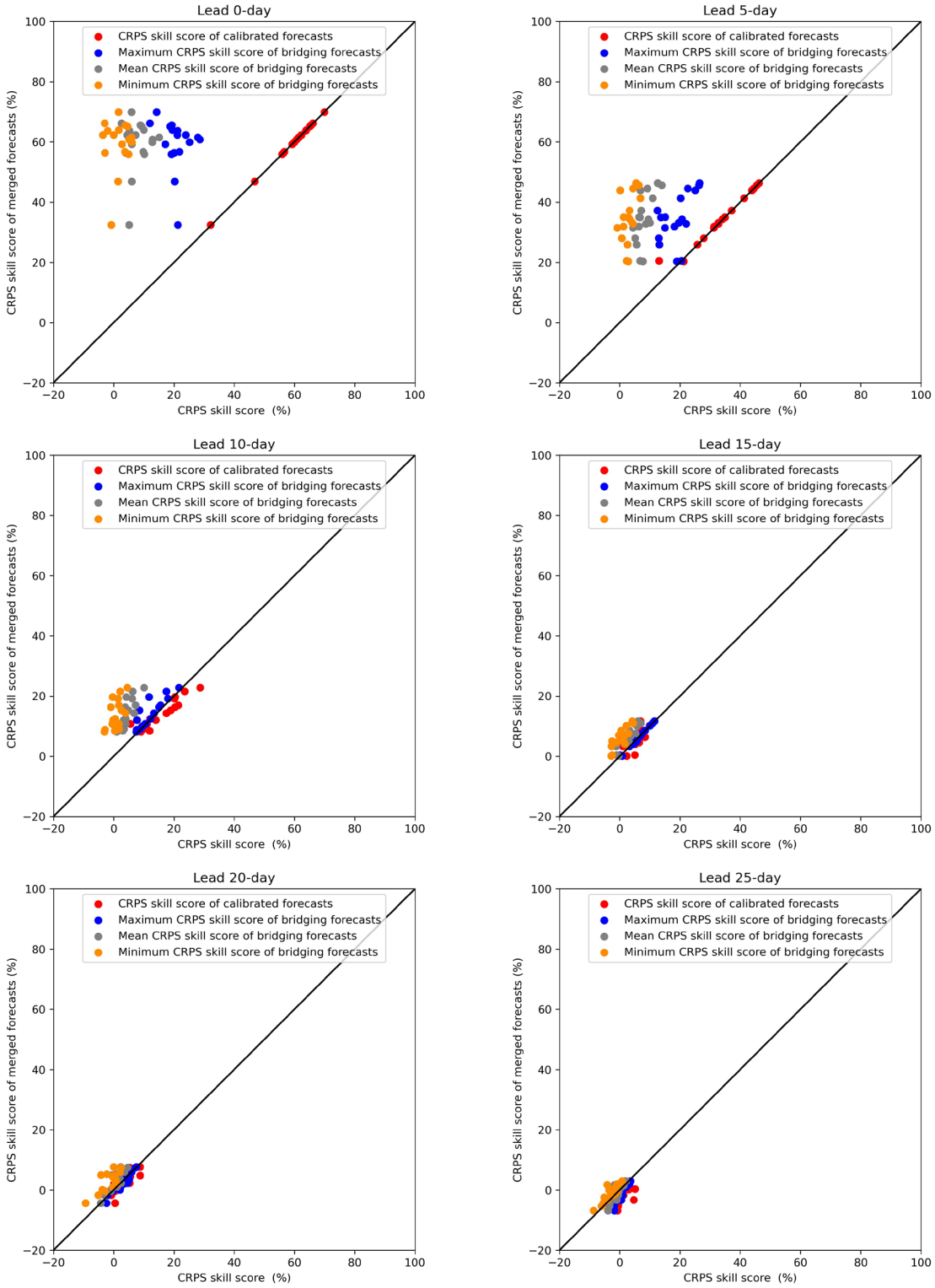

Figure S1. Comparison of the CRPS skill of merged forecasts to the CRPS skill score of calibrated forecasts, maximum, mean, and minimum CRPS skill score of bridging forecasts in May.

Figure 7: The weights don't seem to match the skill patterns. 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. I suggest the authors discuss the discrepancy.

Thanks for this comment. 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 of the discrepancy in the revised manuscript.

L339-341: Related to the above, it is stated that the OLRA and U200 models are more useful at longer lead times based on the higher weights in Figure 7. However, it is difficult to discern from Figure 6 that the bridging models are skilful beyond about 15 days. I suggest revising this sentence discuss the value in terms of skill rather than the weights alone.

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

L350-355: I wouldn't give too much credit to weakly positive skill scores at longer lead times, they may not be significantly different from zero. I suggest the paragraph be rewritten to focus on the stronger patterns of skill and not worry too much, e.g., about the differences between May and June at longer lead times.

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

Figure 9: Some of the reliability index values are quite low, around 0.7, which would indicate some problems with the reliability. I suggest further investigation is required to unpack what is the difference in reliability between 0.7 and 0.9. Perhaps the merging is causing some problems with uncertainty representation.

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 S2 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.

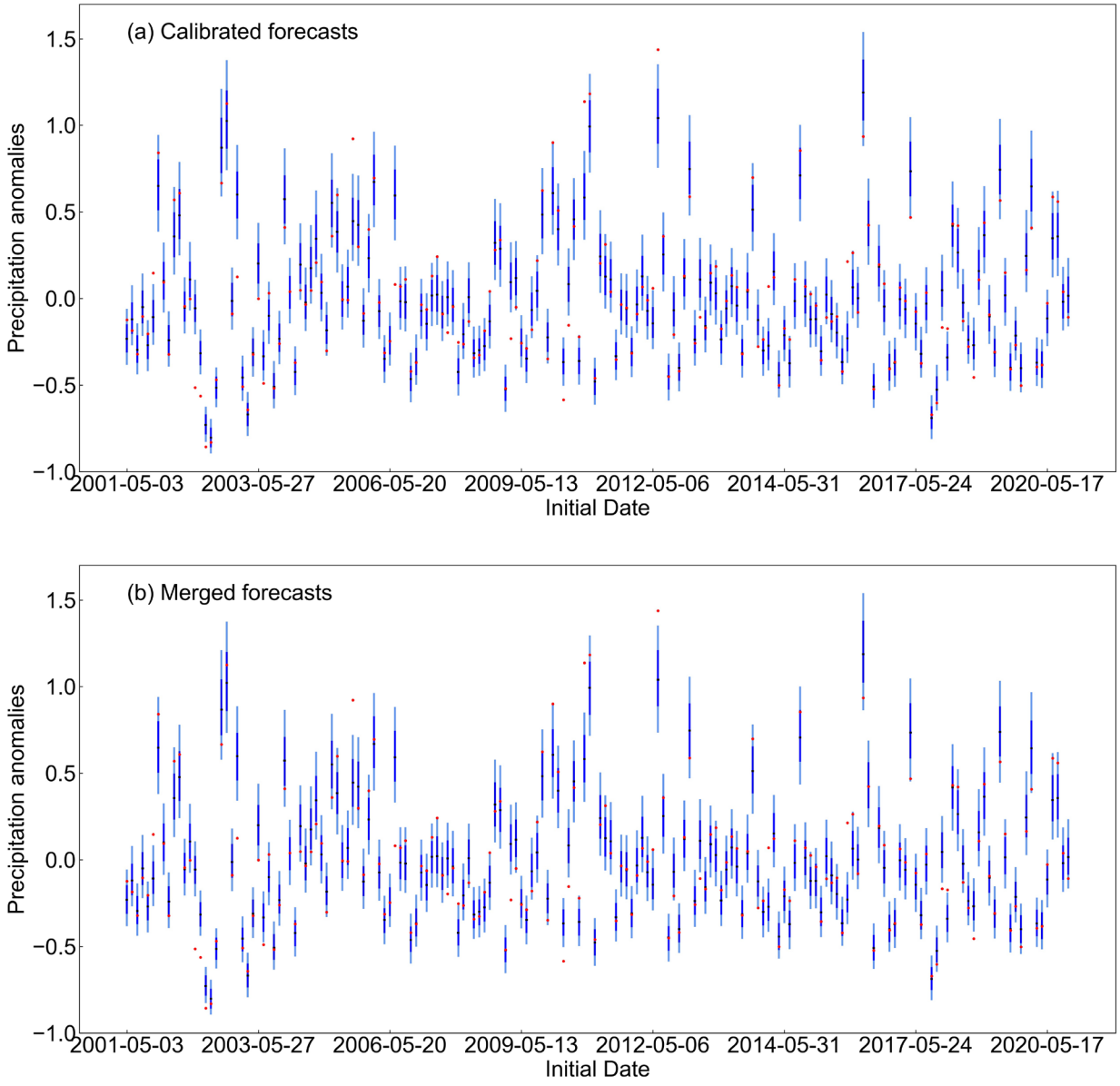

Figure S2. Forecast median, quantiles ranges and observed value against time for sub-seasonal forecasts over Region 3 (Inland Rivers in Inner Mongolia) at a lead time of 0-day. The black dots, forecast median; dark blue vertical line, forecast [0.25, 0.75] quantile range; light and dark blue vertical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

Conclusions: I suggest the first and last paragraphs are not really necessary.

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

L396: It’s not certain the skill will be improved, could just say it will be investigated.

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

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.