Key summary of our responses to the comments raised by referees and community notes:

We sincerely thank Referee #1, Dr. Werner, and the community reviewer (Dr. Nima Zafarmomen) for their thoughtful and constructive comments on our manuscript. We have carefully considered all of the comments and will revise the manuscript accordingly. We greatly appreciate the time and effort devoted to reviewing our work, and we believe that the comments and suggestions will help improve the clarity, rigor, and overall quality of the manuscript. Detailed point-by-point responses are provided below. For ease of review, the proposed revisions and corresponding responses are highlighted using different font styles where appropriate.

Reviewers or editor: Black

Responses: Black and underline

Referee#1's Comments:

This is an exhaustive analysis of the sub seasonal forecasts produced by a number of models that participated in two different model inter comparison campaigns. S2S and SubC. While I think this article should be published, it is more of a time capsule of simulations performed nearly a decade ago and experimental design that is somewhat limited by the computational limitations at that time. The number of inter and intra model ensembles is quite small compared to what would be feasible today at these spatial resolutions. In that sense I am not sure how much this work reflects the current model capabilities that extend to convection resolving scales and much larger ensemble sizes. With that in mind a few additional comments:

<u>Response: We thank the reviewer for the time and effort devoted to reviewing our manuscript and for the thoughtful and constructive comments. We agree that the evaluated hindcast datasets may not fully reflect current state-of-the-art subseasonal forecasting capabilities. We will therefore revise the Introduction and Discussion to better clarify the study's scope, summarize recent advances in operational S2S forecasting, and emphasize that this work provides a benchmark evaluation of widely used publicly available multi-model hindcast archives. Please see our point-by-point responses to the reviewer's specific comments below.</u>

a. Why was the choice made to scale everything to 0.25 degree, this is neither native of PRISM or the models. I suspect much of the "noise" or striations seen on the western mountainous regions in. figure 4 and 7 for example is probably an artifact of the remapping of the model outputs from 100km to 25 km (which seems to be a direct interpolation without consideration for terrain effects of other features).

   <u>Response: We thank the reviewer for this helpful comment. The resampling was carried out to provide a common spatial framework for comparing all hindcast datasets with the PRISM reference dataset. One of the reasons for selecting a 0.25° analysis grid was to preserve the relatively high spatial detail of the observational reference during regional-scale evaluation. In addition, the use of a common 0.25° analysis grid has been widely adopted in large-scale hydrological and hydroclimatic studies evaluating multiple gridded precipitation datasets, including native 0.25° products (e.g., ERA5, PERSIANN-CDR, and CMORPH) together with finer-resolution products (e.g., CHIRPS, IMERG, and MSWEP) that are commonly aggregated to a common analysis grid for consistent intercomparison (e.g., Rajulapati et al., 2021; Aerts et al., 2022). We agree, however, that interpolating relatively coarse forecast products to finer grid resolution (i.e., 0.25° in this work) may introduce artificial fine-scale spatial patterns, particularly over complex terrain. To assess the potential impact of the chosen analysis resolution, we will conduct a targeted sensitivity analysis using a coarser common grid and compare the resulting PBIAS patterns with those presented in Figures 4 and 7. We will also revise the manuscript to clarify the rationale for the selected analysis grid and discuss the limitations associated with spatial remapping.</u>

b. It is remarkable that past two weeks all the model variability of this (I amassing a mix of intra and inter model) variability collapses into approximately the same bias geographically (fig 5 for example). This seems to suggest that the deterministic equation setup in these models make this reach a climate 'state' after the initial conditions and no external perturbations (for example as in a climate models). Do all these models use the same SSTs and how often do they get updated?

Response: We thank the reviewer for this insightful comment. We agree that several broad spatial patterns, such as relatively higher ACC values (Figure 5) along the western coastal regions, are consistently observed across many forecasting systems at longer lead times. But we would also like to note that noticeable inter-model differences are also observed in both the magnitude and spatial distribution of ACC, particularly over the interior CONUS. For example, over the West North Central region at week four, ECMWF retains appreciable ACC, whereas BoM and CMA exhibit only marginal ACC, demonstrating that substantial regional differences remain among forecasting systems (Figure 5; Lines 252–259).

To address the reviewer's question regarding SST forcing, we reviewed the version-specific SST and ocean configurations of the forecast systems evaluated in this study. The available documentation indicates that these systems employ heterogeneous ocean and SST representations (Saha et al., 2014; Richter et al., 2020; Pegion et al., 2019; Hamill et al., 2022; ECMWF S2S Database model descriptions). Most use initialized coupled atmosphere–ocean models in which SST evolves prognostically during the forecast, including ECMWF, UKMO, CMA, CFSv2, and CESM. In contrast, CPTEC, HMCR, and GEFSv12 are atmosphere-only systems that use fixed, relaxed, or otherwise prescribed SST treatments, while CNR-ISAC employs a slab-ocean configuration. The frequency and manner of SST evolution also differ among systems: where documented, coupled models exchange information between the atmosphere and ocean at intervals ranging from sub-hourly or hourly to daily, whereas atmosphere-only systems may retain the initial SST, relax it toward climatology, or impose a lead-dependent SST trajectory. Thus, the evaluated systems do not use identical SST fields or a common SST-update protocol.

In summary, we believe the inter-model differences and consistencies in terms of their predictive skill are the product of multiple factors. Our interpretation is that, as forecast lead time increases and the influence of atmospheric initial conditions gradually diminishes, more slowly varying and larger-scale sources of predictability may exert a relatively greater influence on forecast performance. This is likely to be the major cause of the observed similar large-scale spatial patterns across different forecasting systems. On the other hand, the differences in model formulation, initialization, and boundary-condition treatments may contribute to the remaining inter-model variations. In the revised manuscript, we will summarize the newly collected information in terms of the SST setup in the Supplementary. We will also expand the Discussion to address the potential role of SST in subseasonal precipitation predictability. In particular, we will discuss the potential implications of these heterogeneous SST and ocean configurations for both the common large-scale ACC patterns and the remaining inter-model differences at longer lead times.

c. Have you looked at the inter model variability (for example BOM or CNRM) that have larger ensemble size? Does the bias compared to PRISM for the inter model ensemble remain consistent with each other?

Response: We thank the reviewer for this valuable suggestion. We agree that examining the consistency of forecast bias among ensemble members can provide additional insight into whether the observed biases primarily reflect systematic characteristics of the forecasting system or variability among individual ensemble realizations. To address this comment, we will perform an additional analysis using forecasting systems with relatively large ensemble sizes (e.g., BoM and CNRM) and evaluate the consistency of ensemble-member bias patterns relative to PRISM. The corresponding results and discussion will be incorporated into the revised manuscript.

d. The CRPS results applied across different models with different initializations may be not appropriate? If applied to the same model ensemble it helps understand the model variability and predictability. Across different models it really not very useful as it much impossible to understand what makes these ensembles collapse.

Response: We thank the reviewer for this comment. We agree that CRPS alone cannot explain the physical or methodological reasons underlying performance differences among forecasting systems with different ensemble generation strategies, initialization methods, or model configurations. But we would like to clarify that our intention in this study is not to attribute these differences. Instead, our aim is to provide a systematic intercomparison of publicly available subseasonal precipitation forecast products from the perspective of forecast performance. In this regard, CRPS provides an integrated measure of probabilistic forecast quality by evaluating the full forecast distribution against the corresponding observation.

Nevertheless, we agree that the interpretation of absolute CRPS values can be improved by providing an appropriate reference forecast. Jointly considering your comment and another comment on CRPS raised by Referee #2, Dr. Werner, we will introduce an observation-based climatological forecast as an additional reference and calculate CRPSS relative to climatology. In the revised manuscript, the climatology-referenced CRPSS will replace the original absolute CRPS analysis in the main text and will be used to assess whether, and for how long, each forecasting system retains probabilistic forecast skill beyond climatology. The original absolute CRPS results will be moved to the Supplementary Material, where they will remain available as a complementary measure of absolute probabilistic forecast error across the evaluated systems. We will also revise the corresponding discussion to clarify that the CRPS- and CRPSS-based analyses are intended solely to evaluate forecast performance, rather than to explain the underlying causes of performance differences among forecasting systems.

e. Analysis of these model predictions under certain initial condition constraints (for example ENSO) could be separated to add additional value to the publication. My guess is that while there may be challenges in improving the overall predictability at these time scales, we may be able to find the contained time and spatial domains that would be more predictable under these special initial conditions.

Response: We thank the reviewer for this valuable suggestion. We agree that evaluating forecast performance under specific large-scale climate conditions or initial-condition constraints, such as different ENSO or MJO phases, could provide additional insight into the conditional dependence of subseasonal forecast skill. For example, previous studies have identified sources of subseasonal predictability for wintertime California precipitation and have examined the roles of ENSO and MJO variability in North Pacific subseasonal predictability (L'Heureux et al., 2021; Mayer et al., 2024). Similarly, atmospheric-river-related precipitation and associated flood-relevant events in coastal western U.S. watersheds may represent another forecast opportunity, as prior work has evaluated subseasonal flood forecast skill associated with atmospheric rivers in this region (Cao et al., 2021).

These studies suggest that the generally modest unconditional skill reported here may mask more useful “forecasts of opportunity” under particular combinations of region, season, lead time, and climate state. Therefore, we agree that a rigorous assessment of such conditional skill would be highly informative. However, such an assessment would require a dedicated analysis that stratifies hindcast cases by ENSO phase, MJO phase, atmospheric-river conditions, or other relevant circulation regimes, while carefully accounting for sample-size limitations across models, regions, seasons, and lead times. This would substantially broaden the scope of the present study, whose primary objective is to provide a comprehensive benchmark evaluation of publicly available subseasonal precipitation hindcast products over the CONUS. We therefore consider this an important direction for future research. To acknowledge this point, we will add additional discussion in the revised manuscript to highlight the value of investigating forecast performance under different large-scale climate modes and initial-condition constraints.

f. The link to the ECMWF site for data doesn't work.

Response: We thank the reviewer for pointing this out. We apologize for the broken link to the ECMWF data portal. The URL has been updated in the revised manuscript to the current, functional link.

Reference

Aerts, Jerom PM, et al. "Large-sample assessment of varying spatial resolution on the streamflow estimates of the wflow_sbm hydrological model." *Hydrology and Earth System Sciences* 26.16 (2022): 4407-4430.

Cao, Qian, et al. "Evaluation of the subseasonal forecast skill of floods associated with atmospheric rivers in coastal western US watersheds." *Journal of Hydrometeorology* 22.6 (2021): 1535-1552.

European Centre for Medium-Range Weather Forecasts (ECMWF): S2S Database: Models and model descriptions, available at: https://confluence.ecmwf.int/spaces/S2S/pages/40796876/Models, last access: 26 July 2026.

Guan, Hong, et al. "GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications." *Monthly Weather Review* 150.3 (2022): 647-665.

L’Heureux, Michelle L., Michael K. Tippett, and Emily J. Becker. "Sources of subseasonal skill and predictability in wintertime California precipitation forecasts." *Weather and Forecasting* 36.5 (2021): 1815-1826.

Mayer, Kirsten J., William E. Chapman, and William A. Manriquez. "Exploring the relative importance of the MJO and ENSO to North Pacific subseasonal predictability." *Geophysical Research Letters* 51.10 (2024): e2024GL108479.

Morcrette, Jean-Jacques, CH JAKOB, and J. TEIXEIRA. "European Centre for Medium range Weather Forecasts." *Numerical Modeling of the Global Atmosphere in the Climate System* 550 (2000): 263.

Pegion, Kathy, et al. "The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment." *Bulletin of the American Meteorological Society* 100.10 (2019): 2043-2060.

Rajulapati, Chandra Rupa, et al. "The perils of regridding: Examples using a global precipitation dataset." *Journal of Applied Meteorology and Climatology* 60.11 (2021): 1561-1573.

Richter, Jadwiga H., et al. "Subseasonal prediction with and without a well-represented stratosphere in CESM1." *Weather and Forecasting* 35.6 (2020): 2589-2602.

Saha, Suranjana, et al. "The NCEP climate forecast system version 2." *Journal of climate* 27.6 (2014): 2185-2208.

Vitart, Frederic, et al. "The subseasonal to seasonal (S2S) prediction project database." *Bulletin of the American Meteorological Society* 98.1 (2017): 163-173.