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

**Community comments (Dr. Nima Zafarmomen):**

---

This study provides a comprehensive assessment of 19 subseasonal precipitation hindcast datasets from the Subseasonal Consortium (SubC) and the Subseasonal-to-Seasonal (S2S) Prediction Project. The evaluation covers the Contiguous United States (CONUS) and utilizes the PRISM dataset as a high-resolution reference. The paper provides high value to both forecast end-users and model developers. By identifying specific regions, seasons, and lead times where certain models excel, it helps practitioners move beyond "one-size-fits-all" model selection. The systematic use of both deterministic (ACC, PBIAS) and probabilistic (CRPS) metrics ensures a robust evaluation of model performance.

Response: We appreciate the time and effort you spent reviewing our manuscript and thank you for your positive assessment of our work. We have carefully considered your comments and suggestions, and our point-by-point responses are provided below.

1. The authors correctly note that CRPS is sensitive to precipitation magnitude and should be interpreted alongside ACC. However, the discussion could be slightly strengthened by more explicitly discussing how "dry masks" might impact the perceived skill in arid regions like the Southwest during summer, as briefly mentioned in the discussion of future work.

Response: We thank you for this comment and suggestion. We agree that the interpretation of absolute CRPS values over climatologically dry regions requires caution, as CRPS is influenced by precipitation magnitude. As discussed in our responses to Referee #1 and Dr. Werner, the revised manuscript will replace the main-text CRPS analysis with climatology-referenced CRPSS, which evaluates each forecast against a location- and season-specific climatological benchmark. The climatology-referenced CRPSS will reduce the influence of regional differences in precipitation magnitude and provide a more comparable measure of probabilistic skill across seasons and regions. For completeness, the original CRPS results will be moved to the Supplementary Material.

2. The study re-grids all datasets to a uniform 0.25-degree resolution. While this is a standard approach, it would be beneficial to add a brief comment on whether models with inherently finer native resolutions (like the 1-degree SubC models vs. the 2.5-degree BOM model) showed a consistent advantage in complex terrains due to better representative physics.

Response: We thank you for this comment. Jointly considering this comment with similar comments made by Referee #1 and Dr. Werner, the revised manuscript will include an expanded discussion of the distinction between native model resolution, project-standard archive resolution, and the common 0.25° analysis grid used in this study, together with the potential influence of spatial remapping on forecast evaluation.

We will examine whether forecast products with finer archived spatial resolutions, such as the 1° SubC products, show consistently better performance than coarser products, such as the ~2.5° BOM product, particularly in regions with complex terrain. The corresponding results and discussion will be added to the revised manuscript. In addition, we will conduct a sensitivity analysis using a coarser common analysis grid to assess the influence of the selected analysis resolution on the evaluation results. We believe that these additions will help clarify the interpretation of model performance across forecasting systems with different spatial resolutions.

3. In the results, the authors identify "ACC hotspots" where accuracy remains higher for longer lead times (e.g., the West Coast and Florida Peninsula). It would be helpful to explicitly state in the conclusion if these hotspots are primarily driven by large-scale teleconnections (like ENSO or MJO) which are mentioned in the discussion, to provide a more process-based takeaway for the reader.

Response: We thank you for this helpful suggestion. We agree that relating the identified ACC hotspots to potential large-scale climate drivers provides useful context for interpreting the results. As discussed in our response to Referee #1, the revised manuscript will include an expanded discussion noting that these persistent ACC hotspots may be associated with slowly varying large-scale climate drivers, such as ENSO and MJO. We will also clarify that establishing the physical mechanisms responsible for these patterns would require dedicated process-based analyses. The revised manuscript will further emphasize the importance of investigating the physical mechanisms underlying these persistent skill hotspots and highlight it as a promising direction for future research.

4. The authors transparently acknowledge the limitations of extending pairwise logic to multi-model comparisons. To add even more value, a sentence could be added to the Discussion recommending a specific "best practice" for future researchers who might have the computational resources to run a fully synchronized multi-model experiment.

Response: Thank you for this comment. We agree that a brief recommendation for future benchmark studies would strengthen the Discussion. The revised manuscript will therefore include a statement recommending that, where computational resources permit, future multi-model intercomparison studies adopt fully synchronized experimental designs with consistent initialization dates, forecast configurations, and evaluation protocols to facilitate more direct and interpretable comparisons among forecasting systems.

5. I do strongly recommend the authors consider incorporating a comparison with recent findings in diverse hydroclimatic regions. Specifically, the study 'Analysis of historical global warming impacts on climatological trends for the partially gauged Hirmand river basin based on multiple data products and bias correction methods' demonstrates how similar multi-product evaluations and bias correction techniques perform in data-scarce, partially gauged environments.

Response: We thank you for this helpful suggestion. Indeed, comparing the findings of the present study with similar studies conducted in diverse hydroclimatic regions may provide additional insights. We will carefully consider incorporating the suggested study into the revised manuscript.