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#2 (Dr. Micha Werner)'s Comments:

This manuscript develops and in-depth validation of nineteen sub-seasonal (ensemble) forecasting models across the contiguous United Stated (CONUS). These models are derived from the archives of two initiatives (S2S and SubC) that collated sub-seasonal hindcasts from providers of these modelling systems. Evaluation of the products is developed using standard performance metrics (Anomaly Correlation Coefficient, PBIAS and CRPS) as well as skill using CRPSS. The PRISM dataset, based on interpolated observational data is used here as a reference. The paper addresses the challenge of the disparate frequency as well as forecast start times of these sub-seasonal models, which makes inter-model comparison challenging.

Overall, the results of the study are interesting to the audience of this journal, particularly in the light of the increasing use of extended range (sub-seasonal) forecasts in hydrological streamflow prediction.

There are, however, several comments and suggestions to the authors that could help strengthen the scientific merit of this contribution.

Response: We sincerely thank you for taking the time to provide this thorough and constructive review of our manuscript. We are pleased that you find the study relevant and of interest to the hydrological forecasting community.

We have carefully considered all of your comments and will revise the manuscript accordingly. The major revisions will include clarifying the preprocessing procedures and spatial resampling of the archived forecast datasets, introducing climatology-referenced CRPSS, clarifying the interpretation of the original CFSv2-referenced skill scores, correcting the mathematical formulation of CRPS, expanding the discussion of the application of subseasonal precipitation forecasts in hydrological forecasting, improving the organization and presentation of the seasonal results and figures, and addressing the additional technical comments and editorial suggestions throughout the manuscript. Detailed point-by-point responses are provided below.

The datasets are sourced from two initiatives, the S2S and SubC projects. These two projects established a common dataset to allow comparison of the different models. In establishing this dataset, these projects also applied a number of pre-processing steps. For the models included in the S2S dataset, this included resampling the original model resolution to a common 1.5 degrees resolution. A similar preprocessing may also have been applied to the models participating in the SubC project, where the common spatial resolution of 1 degree was chosen. It may also be that other pre-processing steps were applied to the participating models in each experiment. The authors do not comment on this, but it may be useful to do so. Some of these models may have a much higher resolution, and while it may be that this resampling does not impact forecast skill, it is not entirely clear what influence this has. The authors should to my mind at least acknowledge this resampling in their source datasets. In the preprocessing described in the methods section, the choice is then made to resample all data (forecasts as well as the much finer resolution PRISM data) to the common 0.25 degrees resolution, using nearest neighbour resampling. What was the reason for choosing 0.25 degrees, rather than something closer to the resolution of the S2S and SubC datasets? In some of the analysis it is noted that performance metrics vary from cell to cell. The question would then arise to what extent this is an artefact of this up and down sampling, or

if it truly reflects the spatial variability of model performance. I think that this at least warrants some discussion.

Response: We thank you for this constructive comment. Indeed, the project protocols of S2S and SubC projects do specify standard spatial grids for the submitted forecast fields (i.e., 1.5° for S2S project and 1° for SubC). However, our preliminary review of the available S2S and SubC documentation has not yet identified a clearly documented list of specific preprocessing steps applied uniformly to all participating model products before their release through the project archives. We will conduct a more comprehensive review of the project documentation in search of such information. In particular, we will continue our attempt to clarify the distinction between native model resolution and project-standard archive resolution and to summarize any additional preprocessing steps that can be verified from the official documentation. Regardless of the outcome of on-going documentation review, we will comment on the implications of this spatial standardization for the representation of fine-scale precipitation features following your suggestion in the Discussion during the revision.

Regarding your comment on our selection of 0.25° analysis grid, one of the reasons was to preserve the relatively high spatial detail of the observational reference during 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).

Nevertheless, we agree that resampling hindcast datasets to a uniform 0.25° resolution may contribute to the observed localized cell-to-cell variations in the evaluation metrics. Considering this comment together with Referee #1's concern regarding our choice of the 0.25° analysis grid, we will conduct an additional sensitivity analysis using a coarser common grid and compare the resulting spatial patterns of PBIAS with those obtained at 0.25°. The results of this analysis will be presented in the Supplementary Material and summarized in the main text. We will also expand the Discussion to explicitly acknowledge this limitation and clarify that localized differences should be interpreted with appropriate caution. The sensitivity analysis will further allow us to assess whether the broader regional-scale patterns discussed in the manuscript are robust to the selected analysis resolution.

For the comparison of the forecast models using CPRS, there is no comparison with climatology. As CRPS has dimension (which is noted). It is difficult to interpret the meaning of the values of CRPS provided in for example Figure 6. What would the value have been if the climatology for PRISM had simply been used? I would assume it would be similar to the values found, as these show to become quite stable for most of the models after the 2-3 week lead time. For some values are higher than others (e.g. IAP). What is the implication of this? Is this due to the differences in the comparison (e.g. not the same years used)? To understand the lead time at which there is predictability, the lead time at which skill (CRPSS) goes to zero is considered, using climatology as a reference. CRPSS below zero indicates the forecast is worse than climatology. It would be helpful to understand that for these forecasts. This should at least be discussed.

Response: We thank you for this valuable comment. We agree that the dimensional CRPS values alone are difficult to interpret without a climatological reference, particularly when comparing forecast skill across models and lead times. To address this concern, we will replace the original

CRPS analysis in the manuscript with the Continuous Ranked Probability Skill Score (CRPSS), using a climatological ensemble constructed from the PRISM observations as the reference forecast. The higher values (e.g., IAP) you observe reflect the limitation of using CRPS without reference. Different years, different ensemble sizes are all possible factors contributing to different CRPS values. We therefore believe the revised CRPSS analysis will make the inter-model comparison more interpretable in a way the absolute values are not. The original CRPS results will be moved to the Supplementary Material for reference. This revision will provide a direct measure of forecast skill relative to climatology. In addition, the revised analysis will explicitly identify the lead time at which each forecasting system no longer outperforms climatology, thereby providing a more meaningful assessment of forecast predictability.

Related to the above, CRPSS is calculated here using CFSv2 as the reference. Looking at Fig6 this model seems to have a similar performance (spatially as well as in lead time). For the seasonal comparison it appears that CFSv2 is not one of the best performing models at shorter lead times (particularly for ACC). Using this as a reference, as is then done in the results presented in section 4.3 means that the question being asked is not how skilful a model in question is (using e.g. climatology as a reference) but rather if it is better or worse than CFSv2. This makes Figures 16 and 17 quite difficult to interpret. A high value of ACCSS or CRPSS does not necessarily imply a good performance of the model itself. I can understand the use of CFSv2 as a reference to unify the temporal resolution in comparisons, but the interpretation of what the skill scores say about model performance can only be done if the results of CRPS for CFSv2 shown in Fig6 are considered at the same time. It would have been useful to my mind to calculate at leas the CRPSS of CFSv2 using climatology as a reference, which could then help interpret the skill scores of the other models. I understand that this adds quite some scope to the work presented, but also think this adds to the scientific content of the article. Perhaps this was already done? In any case this should be carefully discussed.

Response: We thank you for this thoughtful comment. As detailed in our response to the previous comment and in line with your suggestion here, we will replace the original CRPS analysis with climatology-referenced CRPSS, using a climatological ensemble constructed from the PRISM observations as the reference forecast. We believe the proposed climatology-referenced CRPSS will complement the original CFSv2-referenced CRPSS by providing a measure of forecast skill relative to climatology. Together, the two skill scores will distinguish whether a forecasting system is skillful relative to climatology and whether it performs better or worse than the CFSv2 benchmark. Accordingly, we will also revise the manuscript to more clearly explain the different interpretations of these two skill scores and clarify that the CFSv2-referenced CRPSS is intended as a relative inter-model comparison rather than an absolute measure of forecast skill.

The results section is quite long, particularly that dealing with the scores for each season. While the maps are well developed, the readability of the paper may increase by making this section a little shorter. As some of the seasons do have similar patterns, it could be an idea to leave only 2 seasons in the main paper and the remainder in the supplementary material. The two remaining are where the patterns are very different (e.g. autumn). The text can of course describe all seasons but focus primarily on the differences between the seasons, using the main one presented as the referenced. That may also raise the interest of this section. This is already done in some sections (e.g. Line 348).

Response: We thank you for this helpful suggestion. We agree that the seasonal results can be streamlined to improve the readability of the manuscript. In the revised manuscript, we will shorten the corresponding text by reducing repetitive descriptions of similar spatial patterns across seasons and placing greater emphasis on the key differences among seasons. We will also move the results for one or two seasons to the Supplementary Material where the spatial patterns are sufficiently similar to those presented in the main manuscript, while ensuring that the main findings remain clearly presented.

It is in any case important to include high resolution versions of these images in the digital supplementary material, allowing more detailed inspection.

Response: We thank you for raising this point. The original figures were prepared and submitted at high resolution; however, their quality appears to have been reduced during the generation of the review PDF. For the revised manuscript, we will further explore the best way to preserve the original image quality throughout the submission process and will also include high-resolution versions of the relevant figures in the digital Supplementary Material to facilitate more detailed inspection.

It would be good to expand a little how these sub-seasonal datasets could be used to support hydrological sub-seasonal forecasting. In using such forecasts as a forcing to a hydrological model, it would be commonplace to use a bias-correction method, of which there are several examples in literature. This may be less the case for sub-seasonal model than for seasonal models. Mention of this is made in lines 522-24 as a recommendation, but it is somewhat brief. Would such a bias correction, developed for each model mean that the differences found between the models diminish? Many bias correction methods deal well with systematic biases, but other properties such as ensemble overdispersion or underdispersion are more challenging. There are also studies that show that bias correction can indeed be detrimental to skill. It would add value to the paper to expand this discussion. Several papers discuss bias correction of precipitation forecasts for hydrological models (e.g. https://doi.org/10.1016/j.jhydrol.2023.129322, several others). Recommendation to extend this work to study the hydrological predictability is given in lines 581 – 587, but the discussion can be extended to my mind.

Response: We thank you for this insightful comment. Indeed, bias correction is widely considered an essential step in the downstream hydrological application of dynamical precipitation forecasts. The present study, however, focuses on evaluating and comparing the raw precipitation forecasts prior to any post-processing. We therefore agree that a more comprehensive discussion of the implications of these results for hydrological subseasonal forecasting would strengthen the manuscript.

In the revised Discussion, we will expand this section by discussing the role of bias correction in operational hydrological forecasting, including its ability to reduce systematic precipitation biases while recognizing that it may not fully address other forecast characteristics (e.g., reliability, sharpness, and spatial structure). We will also discuss that (1) differences among forecasting systems may be reduced following model-specific bias correction; (2) bias correction does not necessarily improve forecast skill in all situations and may even degrade skill under certain conditions, as reported in previous studies; and (3) meaningful differences in forecast performance may still remain after bias correction because many post-processing methods primarily target systematic biases rather than all sources of forecast uncertainty. Finally, we will expand the discussion on the application of subseasonal precipitation forecasts in hydrological forecasting

workflows and incorporate relevant literature on precipitation forecast bias correction, including the study suggested by the reviewer. We believe this expanded discussion will better place our findings in the context of practical subseasonal hydrological forecasting.

Specific comments:

L74-L83: The authors state that they aim to something, that in fact they actually did! In particular the aim to assess 19 models was achieved, so perhaps reformulate.

Response: We thank you for pointing this out. The corresponding sentences will be revised accordingly by replacing the objective-oriented wording (e.g., "we aim to") with wording that describes the work carried out in this study.

L160: I can understand the sampling approach, with CFSv2 being the reference as it has the higher (daily) resolution. Please check the number of dashes per week – to mind there should be seven and not six!

Response: We apologize for this careless mistake and thank you for pointing it out. Each week should be represented by seven days rather than six, and this will be corrected in the revised Figure 2.

L168: complex terrain due

Response: We thank you for pointing out this wording issue. The sentence will be revised to read "complex terrain due to the presence of the Sierra Nevada, Rocky, and Appalachian Mountains."

L184: Please check the formula for CRPS. CRPS is the difference (in area) between the cumulative distribution function (CDF) of the forecast and the CDF of the observation, represented by the Heaveside function if the observation is deterministic. This means it integrates the probability values for the predictand value from -Inf to +Inf. In my understanding of the formula as presented here the actual values are integrated (summed) rather than their (cumulative) probabilities. Perhaps this is due to the notation used, which may then be not be consistent with how it is used in PBIAS and ACC. Please check, and also that a correct implementation of the calculation is used.

Response: We thank you for this careful observation. You are correct that the notation used in Equation (3) is incorrect. The equation will therefore be corrected by adopting the standard ensemble representation of CRPS following Gneiting and Raftery (2007), with the ensemble forecasts denoted explicitly rather than using the symbol (F). We will also verify that the numerical implementation used in this study is correct and is consistent with the standard CRPS formulation. The revised equation and accompanying description will be updated accordingly.

L255: check spelling of models – here ECMWF is missing the F.

Response: We thank you for pointing out this error. It will be corrected in the revised manuscript.

L612: In line with the reflection on using CFSv2 as a reference, I believe that this third conclusive point should be revised, as it does not mention the benchmark used when considering a model as skilful.

Response: We thank you for this comment. We agree that the original concluding statement could be misleading. The concluding statement will therefore be revised to explicitly indicate that the comparison is based on CFSv2-referenced skill scores.

(Please note that this review was provided by the handling editor due to inability to find a second reviewer for this article after inviting over 20 potential referees).

Response: We sincerely thank you for taking the time to provide this thorough and constructive review, particularly given the difficulty in securing additional referees. We greatly appreciate your careful evaluation and insightful comments, which have helped us improve the clarity and quality of the revised manuscript.

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.

Gneiting, Tilmann, and Adrian E. Raftery. "Strictly proper scoring rules, prediction, and estimation." *Journal of the American statistical Association* 102.477 (2007): 359-378.

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.