Review #1:
This manuscript describes a global bias-corrected seasonal forecast dataset based on ECMWF SEAS5 and ERA5. While the dataset concept has merit, the manuscript suffers from critical methodological gaps, circular evaluation design, and insufficient rigor in several areas. I have several substantive concerns and recommendations that should be addressed.

1) A fundamental concern is that ERA5 is both the correction target and the verification benchmark. The BCSD maps SEAS5 quantiles onto ERA5 by design. Evaluating skill against the same ERA5 guarantees apparent improvement. This is methodologically trivial, not evidence of forecast skill. The authors must verify against independent observations (GPCC, CRU, station networks), not simply against ERA5. And the brief discussion in Appendix A1 does not resolve this fundamental problem.

We thank the reviewer for raising this important point regarding the use of ERA5 as both reference for bias correction and verification. We acknowledge that ERA5 is not an independent observational dataset but a reanalysis product, and that verification against the same reference system primarily assesses consistency with the chosen target climatology rather than absolute forecast skill in an observational sense.

However, ERA5 provides a physically consistent, spatially complete and widely used benchmark for large-scale evaluation of meteorological fields, which is particularly relevant for global gridded applications and impact modelling (see, e.g., Lorenz et al. 2021). While independent observational products such as GPCC or CRU exist, these are also subject to interpolation and methodological uncertainties and are not fully independent ground truth datasets. In addition, they are not fully consistent with the joint representation of variables required for impact modelling applications. Such models can be initialized (spun up) with ERA5, which is available in near real time, and subsequently driven seamlessly using SEAS5-BCSD forecasts.

We have clarified this limitation in the revised manuscript and explicitly state that ERA5-based skill scores should be interpreted in terms of consistency with the reference system. Nevertheless, ERA5 remains a standard benchmark in seasonal forecast evaluation, and we additionally discuss limitations and potential sensitivities to the choice of reference dataset in the revised discussion section.

We further note that the primary objective of this study is the evaluation of the bias-corrected SEAS5-BCSD dataset against a consistent reference climatology, rather than a comprehensive assessment of raw SEAS5 forecast skill. Comparisons with the uncorrected SEAS5 forecasts are included only to provide context for the effect of the bias correction, whereas the main analysis focuses on climatology-relative performance, which is most relevant for impact-oriented applications.

We further note that bias correction affects the full forecast distribution and therefore does not guarantee uniform improvement across all verification metrics. While empirical quantile mapping improves the marginal calibration of the forecast distribution and reduces systematic biases, threshold-based probabilistic scores such as the Brier Skill Score or ROC-based measures depend nonlinearly on exceedance probabilities and event definition. Consequently, improvements relative to raw SEAS5 are not straight-forward across all

categorical skill measures, as changes in the corrected distribution may affect tail probabilities differently from mean-error-based metrics.

2) Line 404 references "six variables" processed operationally, but only two variables (tp and t2m) are documented in this manuscript. What are the other four? The authors must provide systematic analysis for all six variables in the paper. Documenting only two variables is very far from sufficient for a dataset paper.

We have corrected the wording in the manuscript to avoid the incorrect impression that all six variables are fully analyzed within this study. The present manuscript focuses on precipitation (tp) and 2-m temperature (t2m), which represent the primary variables of interest for hydrological and impact applications and for which a comprehensive evaluation is provided. The remaining variables are part of the operational processing chain but are not analyzed in detail here. We have clarified this throughout the revised manuscript to ensure consistency between dataset description and evaluation scope.

3) The authors must provide systematic comparison with existing bias-corrected seasonal products (MSWX, C3S products, etc.). A comparison table (resolution, variables, ensemble size, methods, skill) is necessary to substantiate claims of novelty and complementarity. The authors should also conduct a direct skill comparison at least for selected regions and lead times between SEAS5-BCSD and one or two competing products using the same verification framework. The authors must demonstrate the dataset's added value over what is already publicly available.

To better place the dataset in the context of existing efforts, we expanded the introduction. Regarding the request for direct skill comparisons, we note that the primary objective of this ESSD manuscript is the description and documentation of the dataset and its characteristics, rather than a comparative assessment of competing products. Such intercomparison studies represent a valuable but separate scientific question and would require careful harmonization of variables, initialization strategies, ensemble sizes, hindcast periods, and verification frameworks. In addition, only a limited number of global bias-corrected seasonal forecast datasets have been described in the literature, and several products differ substantially in scope and intended applications. We therefore believe that a comprehensive product intercomparison is beyond the scope of the present data descriptor. Nevertheless, we have strengthened the discussion of related datasets and clarified the complementary role of the SEAS5-BCSD dataset.

4) The authors note the ensemble size inhomogeneity (25 members for 1981-2016 vs. 51 for 2017-2024) and state that "including the operational period in the statistical analysis would therefore introduce an inhomogeneity" (Lines 68-69). Yet the dataset spans both periods. How is the bias correction applied to the 2017-2024 operational forecasts? Are CDFs from 1981-2016 used to correct 2017-2024 data without updating?

Bias correction is performed using cumulative distribution functions derived exclusively from the 1981–2016 hindcast period, which are subsequently applied unchanged to the 2017–2024 operational forecasts. This fixed calibration strategy ensures temporal consistency and avoids introducing inhomogeneities associated with differing ensemble sizes and changes in the forecast system configuration. We have clarified this procedure in the revised manuscript. Alternative approaches, such as including all operational ensemble members in the calibration or subsampling to match ensemble sizes, were not adopted, as they would either introduce temporal weighting inconsistencies or require additional assumptions regarding ensemble representativeness. The chosen approach follows standard practice in seasonal forecast post-processing, where calibration is performed on a fixed reforecast period and applied to subsequent operational forecasts.

5) The SD component is described in only two sentences (Lines 101-103). The authors must provide substantially more technical detail. What exactly are these "relative differences"? How is the "coarse-grid ERA5 climatology" constructed (long-term mean, monthly climatology, or daily climatology)? What interpolation method is used (bilinear, conservative)? How are the relative differences applied back to obtain the downscaled field? For precipitation, are multiplicative factors used?

The description of the spatial downscaling method has been expanded in the revised manuscript, including clarification of the interpolation approach (bilinear interpolation of raw fields) and the post-processing quality control step to ensure physically consistent precipitation values.

6) The authors state that precipitation extremes are handled via "linear extrapolation/scaling" and temperature via an "additive delta approach" (Lines 124-125). These are described in one sentence each with no equations or further detail. The authors must add more technical details.

The description of the treatment of precipitation and temperature extremes has been expanded, and the corresponding mathematical formulations have been added to the revised manuscript.

7) None of the skill scores (CRPSS, BSS) include confidence intervals or significance tests. The authors must provide bootstrapped confidence intervals on the global/regional mean skill scores, or apply a field significance test for the spatial maps. Without this, it is impossible to determine whether the reported positive skill at longer lead times is statistically significant.

We have addressed this comment by incorporating bootstrap-based uncertainty estimates throughout the analysis. Confidence intervals based on 1000 bootstrap resamples have been added for global and regional mean skill scores, and spatial maps now include significance information based on bootstrap confidence bounds (see Figures CRPSS_bootstrap and ROC). These additions provide a more robust assessment of the statistical significance of the reported positive skill, particularly at longer lead times.

8) The 1981-2016 period contains a strong warming trend and regional precipitation shifts. Empirical quantile mapping assumes stationary bias. The authors acknowledge this only tangentially (Line 496) and argue it away by assertion. The authors must show that the bias structure is stable over time, or acknowledge and quantify the resulting uncertainty.

We agree that the stationarity assumption underlying empirical quantile mapping represents an important limitation, particularly in the presence of long-term climate trends. In response to this comment, we expanded the discussion and explicitly examined the temporal evolution of SEAS5 biases. While ERA5 exhibits a pronounced warming trend over the hindcast period, the large-scale temperature bias of SEAS5 remains comparatively stable, and no clear evidence of a substantial temporal drift was identified. For precipitation, stronger interannual variability was found, but no pronounced trend in the globally aggregated mean bias was apparent. We emphasize, however, that residual non-stationarity cannot be ruled out and represents an inherent source of uncertainty. We further note that seasonal forecasts are initialized from the contemporary climate state and therefore already reflect much of the underlying warming signal, implying that the stationarity assumption is generally less restrictive than in applications involving long-term climate projections.

9) The bias correction method employed here called pixel-wise Empirical Quantile Mapping (EQM) differs fundamentally from the calibration approaches widely adopted in the initialized prediction community (e.g., Doblas-Reyes et al., 2013, Nat. Commun., https://doi.org/10.1038/ncomms2704). Can the authors discuss the advantages and disadvantages of both approaches, and clarify why EQM is more suitable than mean-variance calibration for the intended applications of this dataset?

We thank the reviewer for highlighting the distinction between distribution-based bias correction and the calibration approaches commonly used in the initialized prediction community. We expanded the discussion to clarify the advantages and limitations of both approaches. While methods such as mean-variance calibration are primarily designed to improve forecast skill and ensemble reliability while preserving the characteristics of the original prediction system, the objective of SEAS5-BCSD is the generation of locally consistent meteorological fields that are particularly well suited for drought analysis and other impact-oriented applications. Because precipitation and drought indicators are strongly influenced by distribution tails, dry-day frequencies, and percentile-dependent biases, correcting the full local distribution was considered more important than preserving the raw ensemble characteristics. We therefore regard pixel-wise EQM as an appropriate compromise for the intended use of the dataset, while acknowledging the limitations associated with observation-based bias correction and the assumption of temporal stationarity.

10) The current evaluation is limited to statistical skill metrics. Can the authors provide application case study (e.g., historical drought or extreme precipitation event) to demonstrate the dataset's practical utility?

We agree that application-oriented case studies provide valuable insight into the practical use of the dataset. However, the primary objective of this ESSD manuscript is the description and

comprehensive evaluation of the dataset using established verification metrics across variables, regions, and lead times. Event-based analyses are inherently selective and address a different scientific question than the systematic assessment presented here. We therefore consider detailed case studies to be beyond the scope of the present data descriptor. To clarify this, we have added a statement in the discussion highlighting that application-specific analyses of individual drought events constitute a natural direction for future work.

11) – 13) small grammatical errors:

We thank the reviewer for pointing out these issues. The grammatical errors have been corrected, and redundant phrasing has been revised for clarity.