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.

Review #2:

This paper presents a valuable and well-documented global bias-corrected and spatially downscaled SEAS5 seasonal forecast dataset. The dataset is likely to be useful for hydrological, agricultural, drought-monitoring, and climate-risk applications, particularly because it preserves the full ensemble structure over the hindcast and operational periods. I found the paper to be well-written, technically sound, and I appreciated the comprehensive analyses of the dataset.

I have only minor comments. Because the dataset is motivated in part by applications to droughts, heat waves, heavy precipitation, and other climate extremes, I recommend slightly expanding the discussion of performance for extremes. The paper already includes threshold-based Brier Skill Score analyses, including 10th and 90th percentile categories, but a more explicit synthesis of how reliable the product is for rare/extreme events would strengthen the manuscript.

Third, the statement that Empirical Quantile Mapping is well suited for precipitation extremes may require clarification. Standard EQM can have limitations for rare or out-of-sample extremes, particularly beyond the empirical calibration range. Additionally, the authors state they apply a linear extrapolation/scaling approach for precipitation extremes outside the historical CDF range, but a few more sentences explaining exactly how this works would be clarify understanding for readers.

Apart from these minor points and a few typographical issues (see attachment), I think this is a strong and useful contribution suitable for publication.

We thank the reviewer for this constructive comment and for highlighting the need for further clarification regarding the representation of extremes in the dataset and the behavior of the empirical quantile mapping (EQM) approach.

With respect to extreme events, we note that the evaluation already includes threshold-based probabilistic skill measures (including Brier Skill Score and ROC analysis for the 10th and 90th percentiles), which provide information on the reliability of forecasts for moderately rare conditions. We have now added a more explicit synthesis in the revised manuscript clarifying that forecast skill is generally higher for moderate anomalies, while performance decreases for increasingly rare extremes, particularly in arid and highly variable regions. This reflects the inherent reduction in sample size and increased uncertainty in the tails of the distribution. Nevertheless, the ensemble-based formulation retains useful probabilistic information for extreme event likelihoods, even where deterministic accuracy is limited.

Regarding EQM, we have clarified that empirical quantile mapping improves the marginal calibration of the forecast distribution but does not guarantee improved performance for all tail-dependent or threshold-based metrics. In particular, standard EQM is constrained by the empirical calibration range and may exhibit reduced robustness for rare or out-of-sample extremes. To address this, the methodology includes a linear extrapolation (scaling) approach for precipitation and an additive delta correction for temperature outside the empirical distribution range, which preserves relative anomalies at the boundaries of the observed

distribution. We have expanded the methodological description in the revised manuscript to make this treatment of extremes more explicit.

Overall, both EQM and the evaluation framework should be interpreted in the context of a probabilistic, impact-oriented dataset rather than a purely deterministic forecast verification framework.

Review #3:

The paper describes a dataset of bias adjusted seasonal forecasts of temperature and precipitation at monthly timescale with lead times up to seven months. The paper is describing the methodology well, besides some minor remarks, and provide relevant analyses to present the forecast skill and remaining deficiencies. I find some sections poorly described and at times confusing and in need of reformulation. I add minor remarks below:

L107: How is the +/- 15 d time window applied at the end points?

We thank the reviewer for this interesting question. The treatment of the ±15-day moving window near the beginning and end of the annual cycle required special consideration and was discussed during manuscript preparation. To clarify this aspect, we have expanded the revised manuscript and added an explanation of how the window is handled at the start and end points. For completeness, the original discussion is reproduced below.

Choice of window size

The use of a time window when calculating the shape of the cumulative distribution functions (CDFs) is particularly important for the effective correction of extreme events. As outlined in the Methods section, the window size plays a critical role in the success of the bias correction. A window that is too narrow may not sufficiently represent the local climatology, especially if extreme events are rare and thus either over- or underrepresented. Conversely, overly broad windows may blend data from different seasons, introducing inconsistencies, such as incorporating wet-season characteristics into dry-season corrections. However, for region-specific applications or certain use cases, a broader window may be advantageous. For example, Sangelantoni et al (2018) employed a 91-day sliding window to better reflect the seasonal characteristics and extremes of Central Italy, a region with relatively smooth transitions between seasons due to its maritime climate. Such a wide window would be less appropriate in a global context, where many regions experience sharp seasonal contrasts, particularly in continental climates. On a global scale, Gergel et al. (2024) used a 31-day window, emphasizing the need to preserve the shape of the distribution's tails.

The 15-day window used in each direction in this study (resulting in a 31-day window in total) was chosen as a compromise. It aligns with previous work by e.g., Thrasher et al. (2012), Themeßl et al. (2012) and is supported by Bedia et al. (2018), who showed that a 31-day window provides sufficiently smooth transitions between days while still being narrow enough to avoid the confounding influence of model drift in seasonal forecasts.

The use of a moving window approach in bias correction requires special handling of the first and last forecast lead days, since the full window cannot be constructed due to the start and end dates of the available SEAS5 data. Several strategies exist to address this limitation. Taking the example of a January forecast, the correction of January 1st values could be approached using one of the following methods:

1. Calculate the SEAS5-CDF with 16 days (0+1+15) and the ERA5-CDF with 31 days (15+1+15).
2. Calculate both the SEAS5-CDF and ERA5-CDF with 16 days each.

3. Use the SEAS5 values from the preceding December forecast to extend the SEAS5 days to 31.
4. Use the ERA5 values from the preceding December reanalysis to extend the SEAS5 days to 31.
5. Use the January 1st value 16 times to obtain 31 SEAS5 days.

As the forecast proceeds, the available window gradually expands, for example, January 2nd can be corrected using a 17-day window, until the full 31-day window is available by January 16th. The same logic applies in reverse to the last forecasted days (e.g., lead days 200 to 215).

Each method carries specific drawbacks. Method 1 leads to mismatched CDF constructions between SEAS5 and ERA5. Lorenz et al. (2021) reported statistical inconsistencies with this approach due to spillovers from prior months, resulting in biases in ERA5 that are not present in SEAS5. Method 5 suffers from a similar issue: overrepresentation of a single day, which distorts the distribution.
Method 3, which involves supplementing with data from the previous SEAS5 forecast, is sensitive to month-to-month drift effects in the model, potentially introducing systematic errors. Furthermore, this approach relies on outdated forecasts that may not represent the actual forecast conditions at initialization.
Gergel et al. (2024) addressed similar edge-window issues by extending the data range using preceding days, corresponding roughly to Method 4. However, since their work was based on climate projections, it is questionable whether this method is transferable to the context of seasonal forecasting. In addition, using ERA5 values to extend SEAS5 distributions compromises the independence between the forecast and reference datasets; essentially, identical values appear in both CDFs, effectively canceling out their impact. Mathematically, this is equivalent to reducing the number of independent observations, aligning with Method 2. Although Method 2 reduces the sample size for the CDFs, and thus slightly limits the capacity to correct extreme values, it avoids the major drawbacks of the other options. Therefore, the global bias correction method presented in this study adopts Method 2 as a conservative yet consistent solution.

L124: Please explain how the extrapolation is performed. How is the linear or additive methods determined? Is it trained on part of the tail?

The description of the treatment of precipitation and temperature extremes has been expanded in the revised manuscript. In particular, we now provide additional details on the extrapolation procedure and include the corresponding mathematical formulations.

L125: The data set shows some issues for dry regions. I attach two figures from a single forecast, where it is seen that an area with precipitation is constant over all bias adjusted ensemble members (looking at file SEAS5_BCSD_tp_1993.nc). This is a recurring problem with quantile mapping approaches, and requires special attention. I note this, but for the forecasts of monthly mean conditions, it is likely of little consequence, but for forcing hydrological models it might have impacts. When studying the data, you'll notice that the pattern over

Sudan/Tchad, and south Algeria is constant across the members, whereas the patterns elsewhere changes as expected.

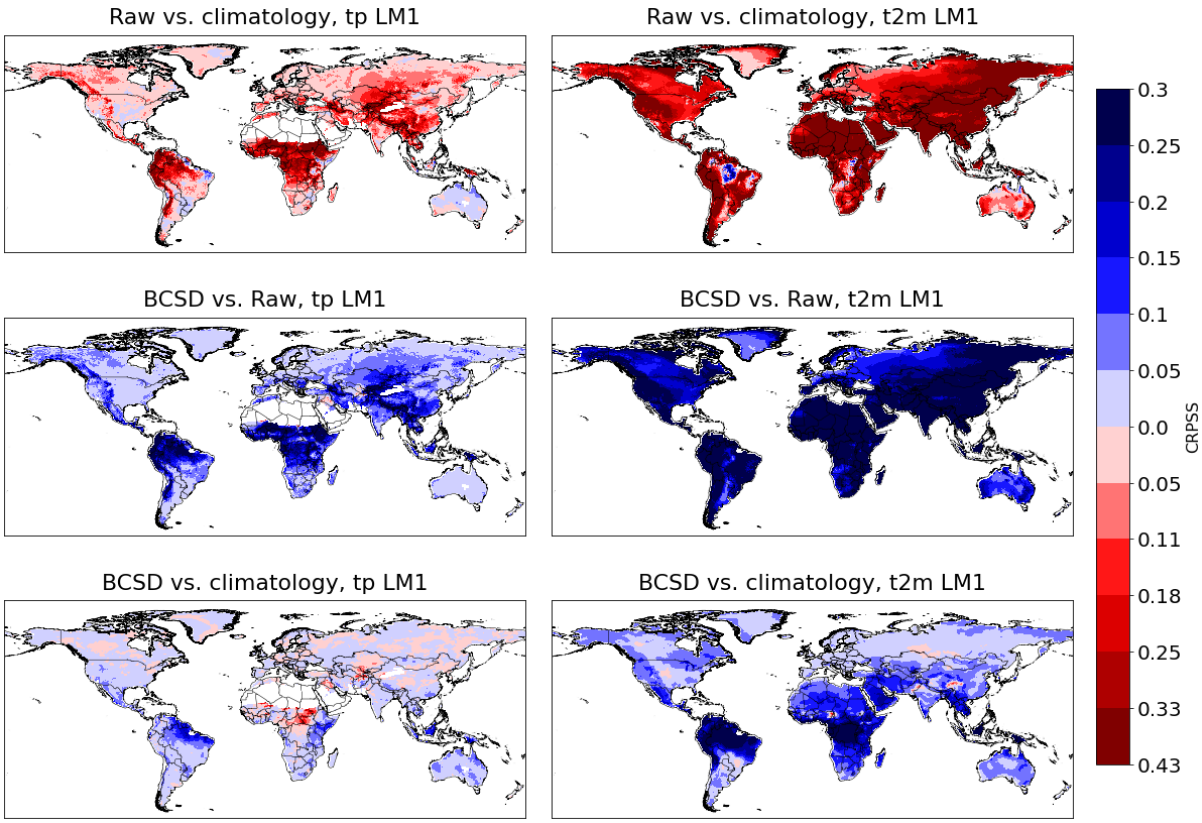
This is a very interesting find and we thank the reviewer for discovering the anomaly. Upon further investigation, we found that the pattern is indeed highly similar across ensemble members, but not identical. The feature originates from a precipitation band extending from southwest to northeast across the Sahara, which is also present in the corresponding ERA5 data for this month, albeit with lower intensity, indicating that SEAS5 slightly overestimated the event.
As January is climatologically very dry in this region, with a mean monthly precipitation of only ~3 mm during 1940–2024, even a single precipitation event can dominate the monthly total. Although SEAS5 ensemble members are generated from perturbed initial conditions, a strong signal already present at the beginning of the forecast can lead to very similar monthly precipitation patterns across ensemble members, with only relatively small differences between them.
We therefore conclude that the apparent similarity does not result from the bias-correction procedure itself, but reflects a physically plausible forecast of a rare precipitation event in an otherwise extremely dry region. Examination of the underlying daily data confirms the existence of differences between ensemble members. If the reviewer is interested, we would be happy to provide access to the daily data via an anonymous THREDDS link.

L192-202: It is not clear if a particular case is shown, such as the bias in a forecast for a specific month, or if this is presenting the bias assessed from a longer time period of multiple years. L194 mentions “loss of skill”, but the figure shows bias and not skill, which should not be confused. Was temperature first “downscaled” by lapse rate correction, or is that included in the “bias”?

The bias is calculated as the absolute/relative deviation of the mean conditions from the reference period 1981 to 2016, using one month would of course not be robust. We clarified this in the caption of Fig. 1 and added it in text.

L205: “BCSD outperforms the uncorrected forecasts”, but only the BCSD result is shown. The uncorrected forecasts or the difference between the two should be shown to assess this statement. Figure 5 would better emphasize the deviations if values from 0.04-0.06 are shown, or simply the values minus 0.05 as deviations from the expected value.

The comparison between raw SEAS5 and SEAS5-BCSD is already illustrated in Fig. 1, which was designed to highlight the substantial biases present in the uncorrected forecasts. We therefore chose to focus Fig. 2 on the performance of the corrected product and to extend the analysis to additional lead months. Nevertheless, we agree that a direct comparison could provide additional insight (see figure below). If the reviewer and editor consider this useful, we would be happy to include the figure as supplementary material.



Regarding the scale of Fig. 5, we intentionally retained the full range of values to avoid overemphasizing relatively small deviations around the expected frequency.

L271: I find this sentence confusing, please reformulate. I do not see anything of linear behaviour in Fig 7, nor do I see any results for “share 33%”. The use of “ensemble share” probably refers to the horizontal axis in the figure, but the figure also shown “Share of events”, which I think adds to the confusion.

The sentence has been reformulated in the revised manuscript to improve clarity and now includes more explicit references to Fig. 7.