

## Reviewer 1

The authors introduce the Python-based software package called SDMBCv2, which is designed for bias-correcting global climate models (GCMs) prior to input into regional climate models (RCMs) for dynamical downscaling. Using the ERA5 reanalysis dataset as their "observation" dataset, they were able to show general improvements to MAE and K-S scores after bias correcting the ACCESS-ESM1.5 GCM. The SDMBCv2 software can thus be useful for researchers conducting dynamical downscaling. In general, the article was well written and presented, and I recommend it for publication after Minor Revisions, as well as addressing some questions about the overall utility/performance of the software.

### Scientific Questions:

1.) How long does it take to run SDMBCv2? I do wonder about its performance as a Python package. Though I understand the accessibility of a Python package for a wider audience, most researchers working with GCMs/RCMs would already have familiarity with Fortran. From a scalability standpoint, especially across ensembles of GCMs and for different CMIP6 pathways, wouldn't a Fortran package be more useful? Though I understand that xESMF and CDO form the core of SDMBCv2, I would appreciate if the authors could comment on performance improvements (or penalties) vs. a pure Fortran approach (e.g. SDMBCv1).

At the individual grid-cell level, the bias correction requires approximately 3 to 4 seconds per cell, with no meaningful difference in execution time between SDMBCv2 and SDMBCv1. This is because SDMBCv2 is a Python-based workflow that integrates xarray and Dask for data management, while keeping the essential numerical correction kernel in Fortran. The key improvement is not in replacing Fortran for computation, but in restructuring how data is processed.

The performance differences arise primarily in large-scale execution across spatial domains. For example, for a single vertical level at global resolution (i.e.,  $192 \times 145$  grid cells), SDMBCv2 completes processing in approximately 2–3 hours, whereas SDMBCv1 typically requires 6–7 hours under comparable computational conditions. This improvement is attributable to Dask-enabled parallelization in Python, which distributes grid-cell computations efficiently across multiple cores and avoids the bottlenecks associated with bulk memory allocation and serial execution.

Therefore, SDMBCv2 improves computational efficiency in two ways:

1. Memory efficiency: chunked, out-of-core processing dramatically reduces RAM requirements compared to bulk in-memory array allocation.

2. Parallel scalability: Dask enables efficient multi-core execution for preprocessing and data transformation steps, which are often the dominant cost in practical workflows (e.g., interpolation, I/O, regridding).

2.) Have the authors used SDMBCv2 for other GCMs other than ACCESS? Especially in regards to the reduction in pass rate for q in hour 18 (Table 1); if this issue would persist with other GCMs. More generally, whether or not the SDMBCv2 could be generalized to successfully bias correct other GCMs.

Although the application example presented in the manuscript uses ACCESS-ESM1.5, SDMBCv2 is designed to be model-agnostic. The framework operates on user-provided atmospheric fields and does not rely on model-specific assumptions. In principle, the same workflow can be applied to other CMIP6 GCMs, provided that the required variables are available and appropriately pre-processed.

We agree that applying more cases across various GCMs would better enhance the assessment of generalizability. Future evaluations will examine climatological statistics before and after bias correction for different models and investigate whether behavior similar to that observed for specific humidity occurs systematically across GCMs.

3.) Have the authors tried running an RCM using the bias-corrected ACCESS? From a historical climate downscaling perspective, what advantage would this have over say downscaling from a reanalysis dataset directly?

End-to-end RCM applications using bias-corrected GCM boundary conditions have been conducted in our previous studies, in which improvements were demonstrated across multiple aspects of regional simulations, including mean climate, variability, and extreme behavior. These studies provide evidence that correcting large-scale boundary biases can improve regional simulations.

This has been stated in Section 1 as below.

“These methods correct RCM input boundary conditions and have been shown to improve the accuracy of output variables, particularly for extreme events (Kim et al., 2023a). Furthermore, Kim et al. (2023b) demonstrated that multivariate bias-corrected boundary conditions better represent compound events where multiple extremes occur simultaneously at the same location.”

- Kim, Y., Evans, J. P., and Sharma, A.: Multivariate bias correction of regional climate model boundary conditions, *Climate Dynamics*, 10.1007/s00382-023-06718-6, 2023a.

- Kim, Y., Evans, J. P., and Sharma, A.: Correcting biases in regional climate model boundary variables for improved simulation of high-impact compound events, *Iscience*, 26, ARTN 107696  
10.1016/j.isci.2023.107696, 2023b.

From a historical downscaling perspective, we agree that reanalysis-driven RCM simulations are often preferable when the objective is to reconstruct past climate as realistically as possible. However, the primary motivation for pre-downscaling bias correction arises in the context of future climate projections. Reanalysis data are unavailable for future periods, making GCMs the sole feasible source of large-scale boundary conditions. In this context, inherent biases in GCMs can influence RCM simulations. SDMBCv2 is designed to reduce these large-scale systematic errors prior to downscaling, thereby improving the physical consistency of the boundary forcing and reducing inherited biases in future regional projections.

4.) Related to 3.), one of the big assumptions with bias correcting a climate dataset is stationarity of bias into the future. Given that the improvement in metrics in 1990–2020 are fairly moderate after bias correction, what assumptions can be made regarding the performance of this approach for future years (either for the GCM directly or after downscaling with an RCM)? Corollary to this, have the authors tried different calibration periods, and/or testing periods, to confirm any temporal aspects to their bias correction approach?

Thank you for this comment. The application of bias correction to future periods assumes that model biases estimated in the historical period are sufficiently stable over time (i.e., approximate bias stationarity). While this assumption cannot be fully verified for the future, the independent validation (1990–2020), Section 4.3, shows that the correction retains skill beyond the calibration period, providing some evidence of temporal transferability.

The moderate improvements in the validation period suggest that biases are only partially stationary and may vary over time. Therefore, the method is expected to reduce systematic errors in future simulations but not eliminate them entirely. This limitation is acknowledged, and the sensitivity to different calibration periods is highlighted as an area for future work (Section 4).

“Although the bias-corrected fields mostly outperform the raw GCM, future work should investigate robustness across different calibration periods.”

5.) Also related to 3.): there was an emphasis on being able to represent extreme events in the paper. Has this been verified, i.e. that this bias-correction approach could improve the representation of 95+-percentile events, particularly after downscaling with an RCM?

The correction methodology included in the package has been evaluated in previous studies using RCM simulations driven by bias-corrected boundary conditions (See the references noted in comment 3).

In those applications, historical downscaling experiments demonstrated statistically significant improvements in the representation of extreme events. Furthermore, the approach has been shown to enhance the representation of compound extreme events by preserving intervariable dependence structures that are critical for physically consistent extremes.

Details can be found in the references mentioned earlier in Question 3.

Minor Comments:

- Recommend that any reference to "Observations" or "observed dataset" should be switched to "Reanalysis"

Thank you for your comment. "Observations" and "observed dataset" have been changed to "Reanalysis".

- e.g. Line 140: "raw" ---> "GCM"

Thanks. Changed.

- Why is SST evaluated on a seasonal timescale (Figure 2) while the variables are evaluated on a daily timescale (Figure 3)?

SST is evaluated on the seasonal timescale in Figure 2 because GCMs' dominant biases are expressed primarily in the mean state and low-frequency variability rather than in daily fluctuations. In our analysis, daily SST biases were comparatively small and spatially uniform, making daily-scale diagnostics less informative for illustrating the practical impact of the correction. In contrast, seasonal statistics reveal more pronounced systematic biases and variability patterns, which are more relevant for large-scale ocean-atmosphere coupling and boundary forcing in RCM applications.

Therefore, seasonal evaluation provides a clearer and more meaningful assessment of SST correction performance.

To clarify this rationale, we have revised the sentence in Section 5.1 from:

“Figure 2 presents bias maps of the seasonal statistics used to evaluate seasonal variability in SST, comparing uncorrected and bias-corrected GCM outputs against ERA5 for the calibration period.”

to:

“Figure 2 presents bias maps of the seasonal statistics used to evaluate seasonal variability in SST, where seasonal-scale diagnostics highlight the dominant systematic biases, comparing uncorrected and bias-corrected GCM outputs against ERA5 for the calibration period.”

- In Figure 5, the bias-corrected plots show sizeable biases. How is the computed MAE 0.0?

Thank you for this comment. The MAE is computed as the spatial average of absolute errors across all grid cells. While some localized regions show visible residual biases in Figure 5, the majority of grid cells have errors very close to zero. As a result, the domain-averaged MAE is very small (on the order of  $\sim 0.001$ ) and appears as 0.00 due to rounding at the reported precision.

To clarify, a brief note on spatial averaging and rounding has been added to the caption.

“Figure 5 As in Figure 3, but for lag-1 auto-correlation. Domain-averaged MAE values are shown; very small errors ( $\sim 10^{-3}$ ) appear as 0.00 due to rounding.”

- Table 1: SD MBC ---> should be SD MBCv2

Thanks. Changed.

- Line 440: "It's possible" ---> "It is possible"

Thanks. Changed.

## Associated Editor

Dear authors,

I would like to note that in the "Code and Data Availability" section of your manuscript, to obtain the SDMBCv2 code, you point out the reader to a GitHub website. GitHub websites are not acceptable to store assets in scientific publication, and GitHub itself instructs users to use long-term repositories for it, instead of citing GitHub sites. Fortunately, in the internal records for the submission of your work, you have provided an acceptable long-term repository, in this case hosted by Zenodo, namely <https://zenodo.org/records/17707370>. In this regard, I have to request you that if the Topical Editor of your manuscript request you additional reviews or decides to accept it for publication, in any reviewed version you must include the link to the Zenodo repository, and not the one to GitHub.

Thank you for your comment. We have revised the "Code and Data Availability" section accordingly. The sentence has been modified from:

"The bias correction framework developed in this study, SDMBCv2, is openly available at [https://github.com/young-ccrc/sdmbc\\_v2](https://github.com/young-ccrc/sdmbc_v2), including all scripts required to reproduce the methodology."

To

"The SDMBCv2 source code corresponding to the version used in this study is archived on Zenodo (<https://doi.org/10.5281/zenodo.17707370>). The development repository is maintained on GitHub ([https://github.com/young-ccrc/sdmbc\\_v2](https://github.com/young-ccrc/sdmbc_v2)), which contains the actively updated source code and related materials."

## Reviewer 2

The manuscript introduces SDMBCv2, an updated Python-based package designed for the bias correction of GCMs. The transition from a Fortran-based legacy script to a parallelized Python architecture is timely and useful. The authors demonstrate the tool's efficacy across multiple timescales and variables, showing significant improvements in capturing the mean state and inter-variable correlations. While the validation framework is robust, using a long (31-year) out-of-sample period, the manuscript requires further clarification of the technical justifications for physical consistency, potential overfitting, and the preservation of climate change signals. I recommend it for publication after Minor Revisions.

### Specific Comments:

**Overfitting and Generalizability.** As shown in Table 1, there is a striking discrepancy in the Kolmogorov-Smirnov (K-S) test pass rates between the calibration period (100% for all variables) and the validation period (dropping to ~60% for wind speed and ~80% for temperature). A 100% pass rate in calibration often indicates that the statistical mapping is "tightly fitted" to the specific noise of the training period. The significant drop in the independent validation period (1990–2020) suggests a degree of overfitting. The authors should discuss whether the complexity of the nesting framework or the quantile mapping frequency contributes to this drop. Please also revise the language in Section 4 (e.g., Line 456), as "consistently high rates" is somewhat misleading given the ~40% failure rate for wind speed in validation.

Thank you for this important comment.

The perfect K–S test performance during the calibration period reflects the construction of the quantile mapping framework, which explicitly aligns the model distribution with the reference dataset. In contrast, reduced pass rates in the independent validation period are expected and indicate the temporal transferability of the correction rather than overfitting. The decrease can be attributed to temporal variability in large-scale circulation, synoptic-scale variability, and potential non-stationarity of model biases between periods. Despite this reduction, the bias-corrected fields generally outperform the raw GCM across the variables, demonstrating that the method retains robust skill in an independent period. It should also be noted that the K–S test is highly sensitive to small distributional differences and may overemphasize minor deviations that have limited practical impact on climatological statistics.

The “consistently high rates” has been replaced with “generally high rates,” which reflects the validation results.

The following paragraph has been added to Section 4.

“Also, the perfect K–S test performance in the calibration period reflects the construction of the quantile mapping framework, while reduced pass rates in the independent validation period highlight the limited temporal transferability under changing conditions, including synoptic variability and potential non-stationarity of model biases. Although the bias-corrected fields mostly outperform the raw GCM, future work should investigate robustness across different calibration periods.”

**Preservation of Climate Change Signals (Signal Smearing).** A critical concern in bias correction is whether the process artificially modifies the GCM’s intrinsic climate change signal (e.g., the warming trend or intensification of the hydrological cycle). Since SDMBC v2 applies correction factors derived from a historical period (1959–1989) to a future or more recent period (1990–2020), there is a risk of "signal smearing." If the GCM predicts a legitimate climatic shift that wasn't present in the historical observations, the bias correction might erroneously treat this shift as a "bias" and remove it. Please clarify if the methodology is trend-preserving.

In the present study, SDMBCv2 is designed to correct climatological statistics based on a historical calibration period and does not explicitly incorporate a trend-preserving constraint.

Our approach is motivated by the fact that many GCMs exhibit substantial systematic biases in the historical period across multiple variables and statistical properties. Such biases reduce confidence in both the baseline climatology and the associated projected changes. Therefore, correcting these structural biases is a necessary step to improve the physical consistency of boundary conditions prior to downscaling, and in this context, retaining the raw model trend without addressing underlying biases does not necessarily ensure a more reliable climate change signal.

We also note that if a GCM predicts a climatic shift that is smaller than the GCM biases, and hence the bias correction, then there would be low confidence in this shift and whether such a shift could be considered “legitimate” is unclear.

We recognize, however, that applying correction functions derived from a historical period may influence projected trends, particularly under non-stationary conditions. The balance between bias reduction and signal preservation depends on the intended application and the reliability of the underlying model. Accordingly, while trend-preserving approaches are not adopted in this study, they could be considered in future work where maintaining specific aspects of the climate change signal is required.

To clarify this point, we have added the following statement to Section 4:

“It should also be worth noting that this study applies bias correction without explicitly preserving long-term trends, prioritizing the reduction of systematic biases in the



baseline climatology. Future work may explore the integration of trend-preserving approaches, where appropriate, depending on the reliability of model-simulated climate change signals and the requirements of specific applications.”

How does the quantile mapping algorithm handle values in the validation or future periods that fall outside the range of the calibration period? Does it use linear extrapolation, constant shifting, or another method? The treatment of 'new' extremes is a well-known source of uncertainty in bias correction that should be explicitly documented.

Quantile mapping in this study is applied to sub-daily fractions (SFs) rather than directly to raw variable values. Six-hourly values are first converted into daily fractions that sum to one, and QM is then applied to these fractions using the calibration-period relationship with the reference dataset. The corrected SFs are subsequently used to disaggregate the bias-corrected daily values back to 6-hourly resolution.

Because SFs are bound between 0 and 1 and constrained to sum to one, the method inherently limits the occurrence of unbounded extrapolation. Values outside the calibration range are therefore not treated using a separate linear extrapolation scheme; instead, they are adjusted through the same QM transformation applied to SFs, followed by a rescaling step to ensure physical consistency of the daily cycle. This approach preserves sub-daily structure while maintaining consistency with the corrected daily totals.

We refer readers to the detailed methodological descriptions in Kim et al. (2023c, 2023d).

Discussion on Low-Variability Regions (e.g., Africa). The authors’ explanation regarding multivariate coupling inducing indirect changes in low-variability regions (Lines 437-445) is insightful. For the benefit of the users, could the authors clarify if SDMBC v2 provides a "safety toggle" or threshold to limit corrections in regions where the observed variability is near zero, to avoid over-amplification of noise?

SDMBCv2 incorporates several built-in safeguards to prevent unrealistic corrections and the over-amplification of noise, particularly in low-variability regions.

Specifically, corrections are not applied to specific humidity when values are below

$10^{-4}$ , and variance correction is only performed when the standard deviation exceeds  $10^{-10}$  to prevent unrealistic variability that might be due to missing data. In addition, variance adjustment is skipped when both the model and observational standard

deviations are very low (below 0.1), as such conditions are particularly sensitive to numerical instability and noise amplification.

These thresholds act as a “safety mechanism” by limiting corrections in regions where variability is minimal. This is especially important in a multivariate correction framework, where adjustments in one variable could otherwise induce artificial changes in another variable with low intrinsic variability. By constraining corrections under such conditions, SDMBcV2 reduces the risk of introducing spurious variability and helps maintain physically realistic behavior.

To clarify, the following sentence in line 449 of Section 4 has been revised from

“These aspects warrant further exploration, especially in areas with low variability that may diminish the effectiveness of multivariate bias correction performance.”

To

“While SDMBcV2 incorporates threshold-based safeguards to limit corrections under low-variability conditions (e.g., constraints on minimum variable magnitude and variance, and conditional skipping of variance adjustment when both model and observed variability are low), residual biases may still occur due to multivariate coupling effects, and these aspects require further exploration.”

#### Minor Comments:

Line 68: "Fortran-based script, which consumed significant memory." Please clarify whether the memory issues were inherent to the Fortran language itself (which is typically highly efficient for numerical arrays) or due to the data-handling logic/structures in the legacy version. As written, the phrasing is somewhat ambiguous. I suggest changing this to "legacy implementation" or "inefficient data handling in the previous version" to avoid implying a limitation of the language itself.

Thank you for this comment. The high memory usage in SDMBcV1 was primarily due to the data-handling strategy in the legacy implementation, where large multi-decadal datasets were processed in bulk rather than in a memory-efficient, chunked manner.

To clarify this, we have revised the sentence from:

“Fortran-based script, which consumed significant memory”

to:

“Fortran-based script with inefficient data handling in the previous implementation, resulting in high memory usage when processing large datasets”

Line 79: "core bias correction process remaining in Fortran." It is recommended that the authors specify the interface method used between Python and the Fortran cores (e.g., via f2py, ctypes, or subprocess calls). This technical detail is crucial for users configuring the software environment on High-Performance Computing (HPC) clusters.

While the interface method is documented in our previous work and GitHub repository, we have added the following sentence to improve clarity in the manuscript:

“... core bias correction process remaining in Fortran due to its superior numerical calculation capabilities. The Fortran subroutines are compiled into a Python-accessible module via F2PY, included with NumPy, facilitating efficient interaction between Fortran and Python.”

Line 180: "eliminating the need for additional processing." This statement may be slightly too absolute, as users still need to manage configuration files (e.g., config.yaml). I suggest softening the phrasing to "simplifying the integration into existing RCM workflows" to more accurately reflect the software's advantage.

Thanks. We have revised the sentence from:

“eliminating the need for additional processing”

To

“simplifying the integration into existing RCM workflows”

Line 197-198: Briefly justify the choice of conservative remapping for specific humidity vs. bilinear for other variables (e.g., to ensure moisture mass conservation).

Thanks. We have revised the sentence from:

“conservative remapping was used for specific humidity, while bilinear interpolation was applied to the remaining variables”

To

“conservative remapping was used for specific humidity to ensure moisture mass conservation, while bilinear interpolation was applied to the remaining variables”

Line 204-205: "1959–1989 (calibration) and 1990–2020 (validation)" The choice of 31 years is slightly unconventional compared to the standard 30-year WMO climate normal. While not a major issue, a brief mention of why 31 years were chosen (e.g., to

include a specific leap year or alignment with ERA5 availability) would show attention to detail.

Thank you for this comment. In this study, a 31-year period was selected to account for a one-year spin-up in the RCM simulations, ensuring that the effective analysis period corresponds to a full 30-year climatology. This choice is aligned with the intended application of SDMBCv2 for dynamical downscaling, where a spin-up period is typically required to stabilize model states.

We have revised the sentence from:

“1959–1989 (calibration) and 1990–2020 (validation)”

To

“1959–1989 (calibration) and 1990–2020 (validation), with 31-year periods selected to allow one year for spin-up in downscaling applications, resulting in a 30-year climatology.”

Line 312-313: "levels where specific humidity approaches zero... were excluded." This is a sound technical decision. However, the authors should specify the exact threshold used for "approaching zero" to allow for exact numerical replication by other researchers.

In this study, bias correction is not applied when the specific humidity is below  $10^{-4}$ , as values in this range approach numerical limits and can introduce instability. This threshold typically corresponds to upper atmospheric levels (around the 24th–25th levels in the tested GCMs, including ACCESS-ESM1-5), although the exact level may vary slightly depending on the model.

To clarify this, we have revised the sentence from:

“levels where specific humidity approaches zero—”

To

“levels where specific humidity approaches zero (i.e., here below  $10^{-4}$ ), ”