Comment 1

The authors went to great lengths to improve the manuscript and revised it significantly in response to my comments and suggestions. I have no further comments and recommend for possible publication in the present form.

Answer

Thanks to the interest you have shown in our paper, our manuscript has been further improved. We are grateful for this.

Comment 1

1) The revised abstract is improved.

Answer

Thanks to your detailed comments, the quality of our manuscript has improved considerably. Thank you

Comment 2

2) The authors provide a long response to my prior comment, but only add a short sentence at the end of the manuscript to address this, which is inadequate. It would be better to include some of the logic in the authors' response into the manuscript. One way would be to add at the beginning of Sec 2.3 something like "Because our analysis applies all bias correction schemes across 11 CMIP6 GCMs, computational demands only permitted us to select three bias correction methods (QDM, EQM, DQM). Additional bias correction methods should be used to extend the robustness of this analysis."

Answer

Thank you for your comments. In response, this study has updated Section 2.3 (Quantile Mapping) by adding and revising the following sentences. Thank you.

The global application imposed substantial computational demands. Consequently, the scope was limited to these three techniques, and incorporating additional bias-correction methods in future study would further strengthen robustness. For calibration and evaluation, the dataset was divided into a training period (1980–1996) and a validation period (1997–2014).

Comment 3

3) typo in the inserted sentence: if -> it

Answer

Thank you for your comments. In response, the sentences have been revised as follows:

This technique removes the systematic wet bias caused by the model's overestimation of dry days relative to observations. Based on this procedure, it effectively corrects the underestimation of excessive dry days during the summer and ensures stable performance even under rigorous cross validation.

Comment 4

4) Line 218 still says "seven" metrics when ten are used.

Also, the authors' response to my comment is not reflected in the revised manuscript. Some version of the response should be added to the first paragraph of section 2.4:

″GCM selection and bias correction method decisions. We employed ten evaluation metrics in this study because these indicators are commonly cited when selecting optimal models and techniques, thereby enhancing the credibility of our results. However, recognizing that some metrics may provide redundant information and introduce bias in multi criteria decision making, we incorporated entropy theory to strengthen objectivity by assigning weights based on each metric's distribution."

Answer

Thank you for your comments. In response, the manuscript has been revised as follows: all instances of "seven evaluation metrics" have been corrected to "ten" throughout.

The equations of ten evaluation metrics are presented in Table 2.

Furthermore, in response to your comment, the manuscript now includes the following addition:

This study evaluated the performance of three quantile-mapping methods against reference data during the validation period (1997-2014) using ten metrics commonly employed in climate research, and used these metrics to identify the optimal GCMs and bias-correction techniques. Recognizing that redundancy among metrics can bias multi-criteria decision making, this study applied an entropy-based weighting scheme that assigns weights according to each metric's distribution to enhance objectivity.

Comment 5

5) Thank you for correcting those equations.

Answer

Thanks to your detailed comments, the quality of our manuscript has improved considerably. Thank you.

Comment 6

6) My prior comment was apparently missed that for Figure 1 "I would assume this is for the validation period (1997-2014) – it should be noted in the caption."

Answer

Thank you for your comments. In response, the caption of Figure 1 has been revised as follows.

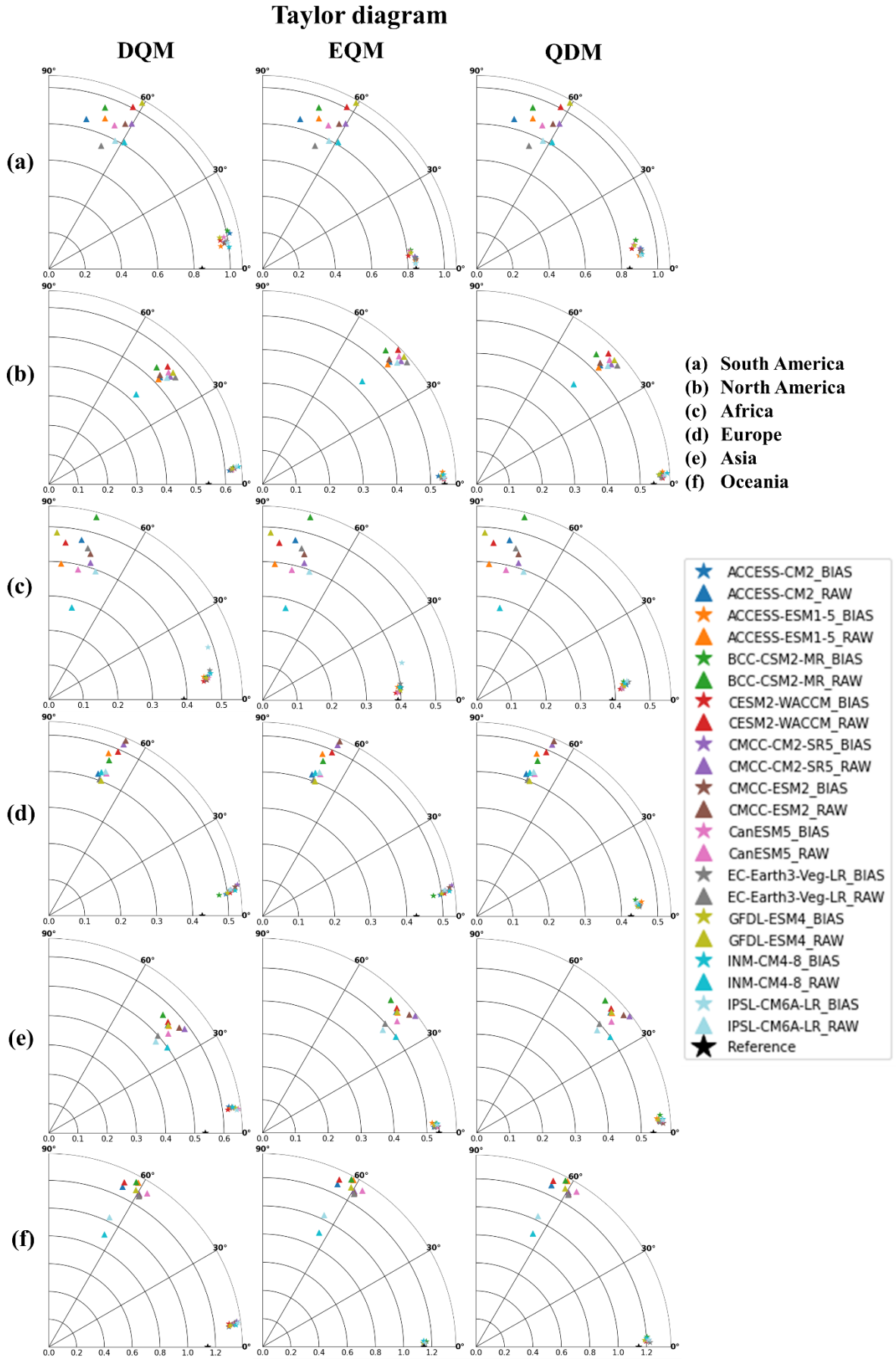
Taylor diagram
DQM
EQM
QDM
(a)
(b)
(c)
(d)
(e)
(f)
(a) South America
(b) North America
(c) Africa
(d) Europe
(e) Asia
(f) Oceania
ACCESS-CM2_BIAS
ACCESS-CM2_RAW
ACCESS-ESM1-5_BIAS
ACCESS-ESM1-5_RAW
BCC-CSM2-MR_BIAS
BCC-CSM2-MR_RAW
CESM2-WACCM_BIAS
CESM2-WACCM_RAW
CMCC-CM2-SR5_BIAS
CMCC-CM2-SR5_RAW
CMCC-ESM2_BIAS
CMCC-ESM2_RAW
CanESM5_BIAS
CanESM5_RAW
EC-Earth3-Veg-LR_BIAS
EC-Earth3-Veg-LR_RAW
GFDL-ESM4_BIAS
GFDL-ESM4_RAW
INM-CM4-8_BIAS
INM-CM4-8_RAW
IPSL-CM6A-LR_BIAS
IPSL-CM6A-LR_RAW
Reference

Figure 1. Comparison of raw and bias-corrected daily precipitation across six continents during the validation period (1997–2014) using Taylor diagrams (x-axis: standard deviation; y-axis: correlation coefficient)

Comment 7

7) The revised Sec. 3.1.2 now includes some statistical measures, which helps. The interpretation could use a little more clarity.

First, clarify what was tested and what each p-value means. Perhaps "A p-value < 0.05 indicates that for the 11 downscaled GCMS, the three downscaling methods rank differently for the metric."

Table 3 can be removed, since all values (to 3 significant figures) are zero. Your statement in the second sentence that all p-values are less than 0.001 for all continents and metrics is adequate.

Likewise, Figure 2 also essentially shows that most or all p-values are < 0.05. Fig 2 can be removed, with revisions to the text in Sec 3.1.2 mentioning the few cases where p > 0.05. Adding a discussion of "large -log(p)" values along with small p-values is confusing.

Answer

Thank you for your comments. Section 3.1.2 has been revised to clarify the hypotheses, tests, and the meaning of p-values. Table 3 and Figure 2 were removed from the main text and placed in the Supplementary Information as Table S1 and Figure S1 to aid interested readers. The manuscript now states that the Friedman test evaluates whether the three methods have equal ranks across the 11 downscaled GCMs for each metric within each continent and that a p-value $<$ 0.05 indicates different ranks. When the Friedman test is significant, pairwise differences are assessed with the Wilcoxon signed-rank test, for which a p-value $<$ 0.05 indicates a significant difference between a given method pair. In this section, all Friedman p-values were $<$ 0.001; no method pairs failed to reach p $<$ 0.05. -log10(p) magnitudes are not discussed further to avoid confusion.

This study used the Friedman test to evaluate whether the three quantile-mapping methods (QDM, DQM, EQM) rank differently across the 11 downscaled GCMs for each of the ten-evaluation metrics within each continent. A p-value $<$ 0.05 indicates that the methods rank differently for the metric, and in this section all Friedman p-values were $<$ 0.001 (Supplementary Table S1). When the Friedman test was significant, pairwise differences were examined with the Wilcoxon signed-rank test. The results, summarized in Supplementary Figure S1, show that most method pairs are significant across continents.

Comment 8

8) the revised figures (Fig 3-8 in the new version) are better as re-plotted. I notice some odd differences in the values between the revised and prior versions that the authors should double-check. For example, the range of values for NSE for North America (now Fig 4) is 0.73-0.97, where in the earlier version it was 0.978-0.998. Many other panels show similar discrepancies. Please ensure the correct data are being shown.

Answer

Thank you for pointing out the noticeable differences between the previous version and the revised Figures, current Figures 2-7. After rechecking the data and the plotting pipeline, this study confirms that the underlying values are identical to those in the earlier submission. The observed differences arise from the upper and lower limits of the color bars and the binning settings. The earlier figures used panel-specific upper breaks concentrated at high values for NSE in North America, for example 0.978-0.998, whereas the revised figures adopt fixed per-metric ranges applied to all panels to improve comparability, for example 0.73-0.97 for NSE. As a result, under the standardized legend many grid cells fall into the highest class, so the color distribution can look different even though the numbers are the same. The evaluation metrics for the other continents also use the same data, and differences in emphasis reflect visual effects introduced by the legend ranges, particularly the choice of the lower bound for each metric.

Comment 9

9) The new panel arrangement is better.

Answer

Thanks to your detailed comments, the quality of our manuscript has improved considerably. Thank you.

Comment 10

10) My prior comment that "with wide variability (in some metrics) across each continent, a single continent wide average may not be very meaningful" appears to have been misunderstood. The first sentence of the paragraph (lines 470-471) discussing Fig 9 could be revised to something like "Figure 9 presents the distribution of ten evaluation metrics for bias-corrected daily precipitation averaged across six continents." Another sentence could be added that performance of any metric for continent-wide average daily precipitation smooths the high spatial variability in metric values (Figs. 3-8).

Answer

Thank you for your comments. In response, the sentence has been revised as follows:

Figure 8 presents the distribution of the ten-evaluation metrics for bias-corrected daily precipitation

averaged over each continent, summarized as boxplots.

Comment 11

11) While continent-aggregated extreme precipitation is not a very meaningful variable, its use to demonstrate the method here is fine. That the statistical test results for all continental areas are identical seems curious – add an explanation why the same value is obtained for all continents. Also, the revised text is adequate, and Table 4 can be deleted.

Answer

Thank you for your comments. Continent-aggregated extreme precipitation is used in this study only to illustrate the method, and its limited interpretability is acknowledged. The reason the statistical test results are identical across continents is as follows. In the Friedman test, the three methods had the same ranking in each continent, so the rank sums were identical. In the Wilcoxon signed-rank test, the sign pattern and the rank order of the absolute paired differences across the 11 GCMs were essentially identical, yielding the same p-value. This indicates that the direction of the differences was consistent across all GCMs. To make this explicit, the following sentence has been added to the manuscript.

Furthermore, the fact that both tests yielded identical results across continents indicates that the sign and rank structure of the three methods was the same in every continent, which in turn shows that the direction of the differences was consistent for each GCM.

Furthermore, Table 4 has been moved to Supplementary Information and renumbered as Table S2. Thank you.

Comment 12-16

12-16) Revised manuscript is fine for this comment.

Answer

Thanks to your detailed comments, the quality of our manuscript has improved considerably. Thank you.

Comment 17

17) The revised text discussing Fig 16 still fails to refer to a statistical test to support the assertions that "EQM exhibited the lowest median standard deviation across all continents. QDM showed slightly higher median standard deviations than EQM across most continents, though the difference was minimal. In contrast, DQM showed the highest median standard deviation…" A statistical test is needed to support any of these claims.

Answer

Thank you for the suggestion. In Figure 16 the quantities are grid-cell standard deviations computed from the same BMA ensemble, and they exhibit strong spatial dependence. Applying pairwise tests to all grid cells would inflate the effective sample size and can overstate significance. Our intent here is to summarize predictive uncertainty across continents rather than to make inferential claims about method superiority. We therefore revised the text to provide a qualitative description only. Additionally, this study conducted statistical tests on the CI results that account for both uncertainty and performance (Table S3). The revised text is as follows:

Figure 15 shows the standard deviation of daily precipitation for the ensemble forecasted by BMA using three methods, DQM, EQM, and QDM, in a boxplot for each continent. Visually, EQM tends to show the lowest medians across continents, QDM appears slightly higher, and DQM tends to show the highest medians. The interquartile ranges overlap broadly within most continents and the differences in medians are small in magnitude.

Comment 18

18) The revised text discussing Fig 16 still fails to refer to a statistical test to support the assertions that "EQM exhibited the lowest median standard deviation across all continents. QDM showed slightly higher median standard deviations than EQM across most continents, though the difference was minimal. In contrast, DQM showed the highest median standard deviation…" A statistical test is needed to support any of these claims.

Answer

Thank you for the suggestion. Table 6 has been removed from the main text. To maintain transparency, a minimal summary is provided in the Supplementary Information as Table S3, which lists the single non-zero case and confirms that all other values round to zero to three significant figures.

The added Table S3 in the Supplementary Information is as follows:

Table S3. P-values of Friedman and Pairwise Wilcoxon tests for DQM, EQM, and QDM across continents under different α/β weightings

α: 0.5, β: 0.5				
Continents	Friedman	Wilcoxon		
		DQM & EQM	DQM & QDM	EQM & QDM
South America	1.01E-160	1.99E-135	4.53E-16	2.07E-191
North America	0.00E+00	0.00E+00	0.00E+00	0.00E+00
Africa	0.00E+00	0.00E+00	1.75E-234	3.20E-211

Europe	0.00E+00	0.00E+00	1.07E-106	0.00E+00
Asia	0.00E+00	0.00E+00	0.00E+00	0.00E+00
Oceania	0.00E+00	1.76E-266+	3.93E-01	6.93E-287
α: 0.7, β: 0.3				
South America	2.33E-150	2.69E-120	2.17E-18	4.40E-189
North America	0.00E+00	0.00E+00	0.00E+00	0.00E+00
Africa	0.00E+00	0.00E+00	6.40E-226	3.76E-200
Europe	0.00E+00	0.00E+00	6.42E-129	0.00E+00
Asia	0.00E+00	0.00E+00	0.00E+00	0.00E+00
Oceania	0.00E+00	1.78E-295	1.63E-02	6.49E-285
α: 0.3, β: 0.7				
South America	3.73E-175	2.50E-160	1.01E-11	8.12E-189
North America	0.00E+00	0.00E+00	1.28E-277	0.00E+00
Africa	0.00E+00	0.00E+00	2.47E-249	2.90E-232
Europe	0.00E+00	0.00E+00	2.08E-59	0.00E+00
Asia	0.00E+00	0.00E+00	0.00E+00	0.00E+00
Oceania	6.63E-250	4.75E-193	3.39E-02	6.27E-285

In addition, the main text has been revised as follows:

Under the three weighting scenarios defined in the main text, the Friedman test produced p-values effectively rounded to zero for every continent, indicating highly significant differences among DQM, EQM, and QDM (Table S3 in the Supplementary Material). Subsequent pairwise Wilcoxon tests showed that most method comparisons remained significant across all regions. The only notable exception occurred in Oceania under equal weighting, where the p-value of 3.93×10^{-1} failed to reach significance at the 0.05 level. These findings demonstrate that, aside from that single case in Oceania, the choice of scenarios exerts a statistically significant impact on composite scores across all continents.