.

Comment 1

1. Abstract, lines 19-26, the results should focus not on how these three methods ranked based on daily precipitation but on what this application revealed about the method. How sensitive is the ranking to the selection of evaluation metrics? To GCM selection? To weighting of uncertainty vs. performance? To different climatological regions (as mentioned briefly at Lines 666 and 722)? Three QM methods is too small a pool to draw useful conclusions, and the vague language leaves too much unanswered ("QDM outperformed other methods in specific regions…" and "DQM…shows the highest uncertainty in certain regions").

Answer

Thank you for your insightful response. We have revised the abstract as follows to fully incorporate your suggestions. Thank you again.

This study proposed a Comprehensive Index (CI) that jointly considers bias correction performance metrics and uncertainty to guide the selection of quantile mapping methods. This approach reveals not only a performance-based ranking of bias correction methods but also how optimal method choices shift as the uncertainty weight varies. This study evaluated daily precipitation performance from 11 CMIP6 GCMs corrected by Quantile Delta Mapping (QDM), Empirical Quantile Mapping (EQM), and Detrended Quantile Mapping (DQM) using ten evaluation metrics and applied TOPSIS (Technique for Order Preference by Similarity to an Ideal Solution) to compute performance-based rankings. Furthermore, Bayesian Model Averaging (BMA) was used to quantify both individual model and ensemble prediction uncertainties. Moreover, entropy based weighting of the ten evaluation metrics reveals that error based measures such as RMSE and MAE carry the highest information content (weights 0.13-0.28 and 0.15-0.22, respectively). By aggregating TOPSIS performance scores with BMA uncertainty measures, this study developed CI. Results show that EQM achieved the best performance across most metrics 0.30 (RMSE), 0.18 (MAE), 0.98 (R^2), 0.87 (KGE), 0.93 (NSE), and 0.99 (EVS) and exhibited the lowest uncertainty (variance = 0.0027) across all continents. QDM outperformed other methods in certain regions, reaching its lowest model uncertainty (variance = 0.0025) in South America. EQM was selected most frequently under all weighting scenarios, while DQM was least chosen. In South America, DQM was preferred more often than QDM when performance was emphasized, whereas the opposite occurred when uncertainty was emphasized. These findings suggest that incorporating uncertainty leads to spatially heterogeneous and parameter dependent changes in optimal bias correction method choice that would be overlooked by metric only selection.

Comment 2

2. Lines 142-144, the three QM methods are outlined. Were these selected because they represent very different approaches to bias correction? Many other bias correction methods exist, from simple delta change to multivariate. Would a broader selection of methods test your ranking methods more robustly?

Answer

We agree that incorporating a broader range of bias-correction methods (e.g., simple delta change, multivariate approaches) would further reinforce the robustness of our CI evaluation. However, because our analysis applies all bias-correction schemes across 11 CMIP6 GCMs, computational costs increase dramatically. In particular, Cannon's multivariate bias correction (MBC), which requires additional climate variables and increases run time by approximately fivefold per model, is especially demanding. Given these constraints, we selected three representative methods (QDM, EQM, DQM), all of which have been widely adopted in leading studies.

https://doi.org/10.1016/j.jhydrol.2020.125685

https://doi.org/10.1016/j.ejrh.2025.102223

https://doi.org/10.5194/gmd-17-191-2024

https://rmets.onlinelibrary.wiley.com/doi/10.1002/joc.1602

https://doi.org/10.2166/wcc.2020.261

Cannon et al. (2015), which introduced QDM, has been cited over 1,400 times to date, and EQM-related publications have been referenced over 100,000 times. By focusing on these three quantile mapping approaches, we span the key axes of bias-correction strategies from rare extreme precipitation to full-distribution fitting to detrended adjustment enabling a comprehensive and balanced assessment. Moreover, comparing these representative methods is likely to attract substantial interest among both academic and industry researchers, thereby broadening the applicability and impact of our study. We appreciate the reviewer's suggestion and will explicitly note these limitations and their rationale in the revised manuscript's conclusion. Thank you once again for your understanding. Please translate the above text into English.:

Furthermore, more bias correction methods should be used to extend the robustness of CI.

Comment 3

3. Line 148-149, the "frequency-adaptation technique" is applied to address potential biases. I did not see this explained further in the methods section. Did this add something beyond what is described by the definitions of the QM methods?

Answer

Thank you for your comment. In response to your comments, we have added the following to Section 2.3 to clarify exactly what the frequency-adaptation technique is and what additional bias-correction benefits it provides.

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

Comment 4

4. Table 2 lists 10 evaluation metrics (though the table caption and line 193 state "seven"). It would help to add columns for the range of values each can take, and what value would indicate a 'perfect' fit. An obvious question is why so many somewhat redundant metrics are used (for example RMSE, MAE, MdAE). Could the method be employed with 2 or 3 metrics and perform as well? That would be a useful detail to explore.

Answer

Thank you for your comments. Error metrics such as RMSE, MAE, and MdAE are widely used to guide 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. For example, as shown in Table 3, metrics with identical error values receive markedly different entropy weights according to their information content. By applying entropy-based weighting to all evaluation metrics, we ensure the reliability and fairness of the CI results.

In response to your feedback, we have revised Table 2 as follows:

Table 2. Information of the ten-evaluation metrics used in this study

Metrics	Equations	Factors	References	Range
RMSE	$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i^{sim} - X_i^{ref}\right)^2}$	X_i^{ref} reference data X_i^{sim} Bias corrected GCM		$[0, +\infty)$ Best value: 0
MAE	$= \sum_{i=1}^{n}\left\|X_i^{sim} - X_i^{ref}\right\|$			

R^2	$= 1 - \frac{\sum_{i=1}^{n}(X_i^{sim} - X_i^{ref})^2}{(X_i^{ref} - \bar{X}_i^{ref})^2}$		Galton, 1886	$(-\infty, 1]$ Best value: 1
Pbias	$= \frac{\sum_{i=1}^{n}(X_i^{ref} - X_i^{sim})}{\sum_{i=1}^{n} X_i^{ref}} \times 100$			$(-\infty, +\infty)$ Best value: 0
NSE	$= 1 - \frac{\sum_{i=1}^{n}(X_i^{sim} - X_i^{ref})^2}{\sum_{i=1}^{n}(X_i^{ref} - \bar{X}_i^{ref})^2}$		Nash and Sutcliffe, 1970	$(-\infty, 1]$ Best value: 1
MdAE	$= median(\lvert X_i^{sim} - X_i^{ref} \rvert)$			$[0, +\infty)$ Best value: 0
MSLE	$= \frac{1}{n}\sum_{i=1}^{n}(\log(1 + X_i^{sim}) - \log(1 + X_i^{ref}))^2$			$[0, +\infty)$ Best value: 0
EVS	$= 1 - \frac{Var(X^{sim} - X^{ref})}{Var(X^{ref})}$			$(-\infty, 1]$ Best value: 1
KGE	$= 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}$	r Pearson product-moment correlation α Variability error β: Bias term	Gupta et al. 2009	$(-\infty, 1]$ Best value: 1
JS-D	$= \frac{1}{2} D_{KL}\left(P \parallel \frac{P+Q}{2}\right) + \frac{1}{2} D_{KL}\left(Q \parallel \frac{P+Q}{2}\right)$	$P(x)$: Probability density distribution of reference data $Q(x)$: Probability density distribution of GCM D_{KL}: KL-D	Lin, 1991	$[0, ln2]$ Best value: 0

Comment 5

5. For clarity, variables should not be used for different quantities. For example, α is used in equation 7 as the scale factor in the GEV distribution, αw is in equation 10 (and doesn't appear to be defined), and α is used in equation 16 as a performance weight.

Answer

Thank you for your comment. We appreciate your insight and have revised Equations 7 and 16 accordingly. Thank you for your valuable comments.

The modified Equations were 7 and 16 as follows:

$$g(x) = \frac{1}{s}\left[1 - k\frac{x-\epsilon}{s}\right]^{\frac{1}{k}-1} exp\left\{-\left[1 - k\frac{x-\epsilon}{s}\right]^{\frac{1}{k}}\right\} \qquad (7)$$

where, k, s, and ε represents a shape, scale, and location of the GEV distribution, respectively.

$$CI = \omega \times C_i - \beta \times UI \qquad (16)$$

where, UI represents the uncertainty indicator. V_w and σ_e represent the normalized weight variance and the normalized ensemble standard deviation, respectively, calculated using BMA. C_i represents the closeness coefficient calculated from TOPSIS. ω represents the weight given to the performance score, β represents the weight given to the uncertainty indicator. Furthermore, by adjusting the weights ω and β, the study evaluated the QM methods under different scenarios. Equal weight (ω = 0.5, β =0.5) balances performance and uncertainty equally, and the emphasized performance weight (ω = 0.7, β =0.3) prioritize performance over uncertainty. The emphasized uncertainty weight (ω= 0.3, β=0.7) prioritize uncertainty over performance. The results from the CI provide a holistic evaluation of the QM methods, considering both their effectiveness in bias correction and the reliability of their predictions.

Comment 6

6. Figure 1, Since all points in all panels are in the first quadrant, just include that in the figure. That will help readers see the individual points better. That is a pretty standard way to present Taylor diagrams. Also, I would assume this is for the validation period (1997-2014) – it should be noted in the caption.

Answer

.

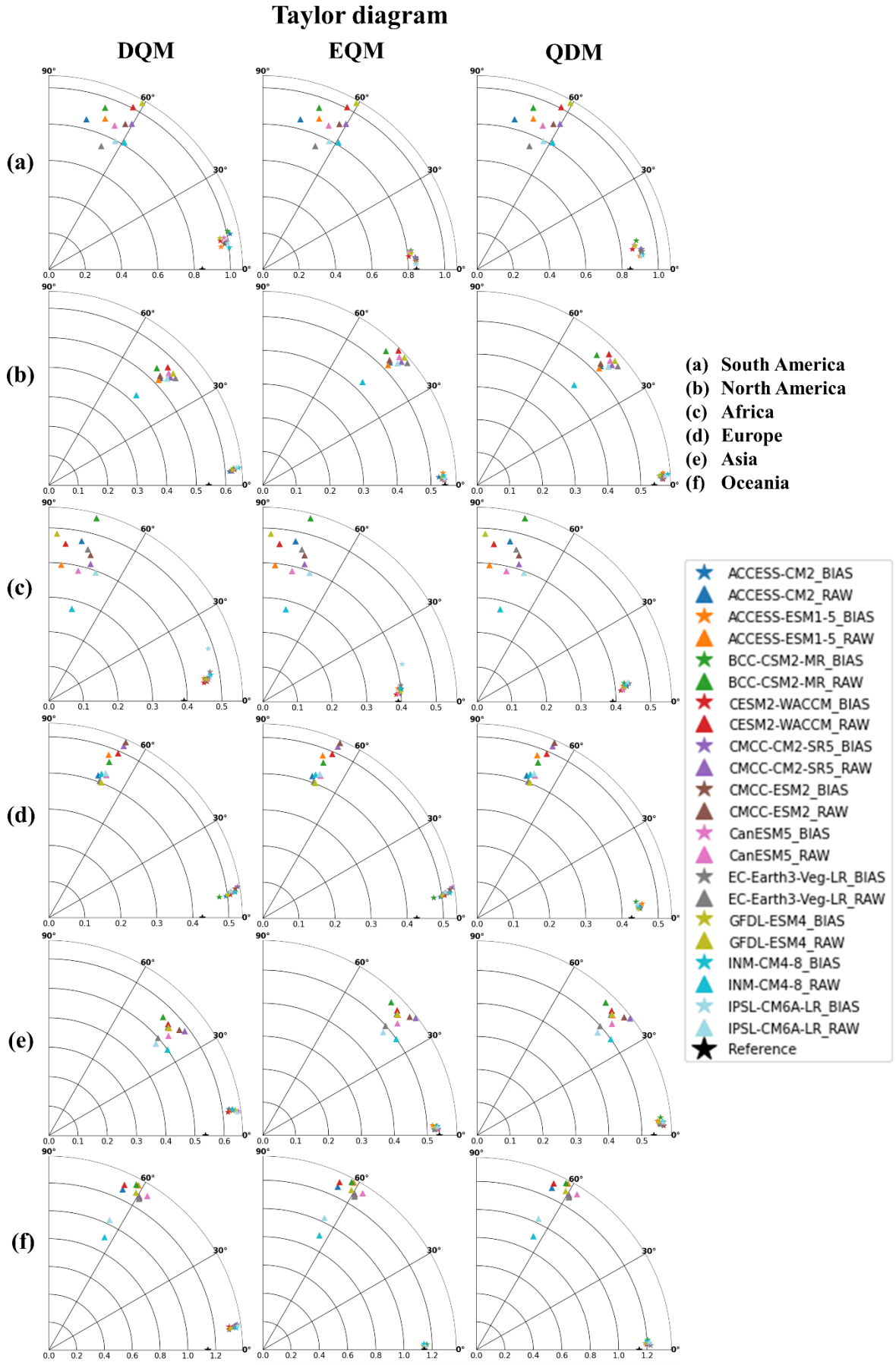
Taylor diagram
DQM
EQM
QDM
(a)
(b)
(c)
(d)
(e)
(f)
(a) South America
(b) North America
(c) Africa
(d) Europe
(e) Asia
(f) Oceania
ACCESS-CM2_BIAS
ACCESS-CM2_RAW
ACCESS-ESM1-5_BIAS
ACCESS-ESM1-5_RAW
BCC-CSM2-MR_BIAS
BCC-CSM2-MR_RAW
CESM2-WACCM_BIAS
CESM2-WACCM_RAW
CMCC-CM2-SR5_BIAS
CMCC-CM2-SR5_RAW
CMCC-ESM2_BIAS
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 corrected daily precipitation on six continents using Taylor diagrams (x-axis: standard deviation; y-axis: the correlation coefficient)

Comment 7

7. Line 319 (the beginning of section 3.1.2). From here to line 421 offer very little to the aim of the paper. The interpretation of the figures is all qualitative (two examples: for South America, Fig 2 "EQM demonstrated lower JSD values, as well as higher EVS and KGE values, compared to other methods." and "QDM and DQM also performed well but exhibited slightly larger errors in some regions than EQM."). First, statistical tests are needed to determine if any differences are statistically significant. Second, a discussion of whether the differences are physically meaningful is needed, such as EVS varying from 0.95 to 0.98 across the region and across methods. Again, if the effort is to rank QM methods, this is far from adequate. If the aim of the study is to demonstrate the application of the method and its sensitivity to methodological choices, then most of this section is not needed.

Answer

Thank you for your comment. The aim of this study is to highlight the limitations of selecting bias-correction methods or GCMs based solely on evaluation metrics, as has been done in previous research. While uncertainty in future climate projections is well recognized, model uncertainty arising from initial and boundary conditions, as well as uncertainty introduced by multiple bias-correction methods, also exists. Therefore, we have proposed a framework that incorporates these uncertainties into the selection process, and we demonstrate that considering uncertainty leads to differences in the optimal bias-correction method chosen at each grid cell. In this context, Section 3.1.2 illustrates the spatial variability in method choice—for example, by the differing area ranges of each category shown in the legend of Figure 2. Furthermore, we assessed statistical significance using the Friedman and Wilcoxon tests. The results of the statistical tests are as follows:

3.1.2 Spatial distribution of bias correction performance

This study used the Friedman method to statistically test the ten-evaluation metrics, as shown in Table 3. Overall, all evaluation metrics showed highly significant differences across the six continents, with p-values below 0.001. Even metrics that exhibited relatively larger raw form p-values such as MdAE in South America and JSD in North America remained well below the 0.05 threshold. Figure 2 further depicts Wilcoxon pairwise comparisons of QM methods for each continent. In South America, the comparisons between EQM and DQM as well as between QDM and EQM yielded large $-\log_{10}(p)$ values, indicating that divergence and fit metrics drive methodological differences in that region. In North America, MSLE and MdAE were significant for QDM versus DQM, while JSD was significant for EQM versus DQM. In Africa and Asia, correlation metrics (NSE and R^2) showed statistically significant

differences in QDM versus DQM comparisons. Finally, JSD and MSLE emerged as the primary metrics distinguishing EQM from DQM.

Table 3. Statistical significance comparison of ten evaluation metrics based on Friedman tests over six continents

Metrics	South America	North America	Africa	Europe	Asia	Oceania
RMSE	0.000	0.000	0.000	0.000	0.000	0.000
MAE	0.000	0.000	0.000	0.000	0.000	0.000
R^2	0.000	0.000	0.000	0.000	0.000	0.000
NSE	0.000	0.000	0.000	0.000	0.000	0.000
KGE	0.000	0.000	0.000	0.000	0.000	0.000
Pbias	0.000	0.000	0.000	0.000	0.000	0.000
MdAE	3.39E-22	0.000	0.000	1.80E-106	0.000	4.28E-172
MSLE	2.21E-09	0.000	1.51E-260	5.14E-109	0.000	7.78E-89
EVS	0.000	0.000	0.000	0.000	0.000	0.000
JSD	5.57E-101	3.19E-50	5.27E-10	1.20E-116	6.29E-79	2.09E-52

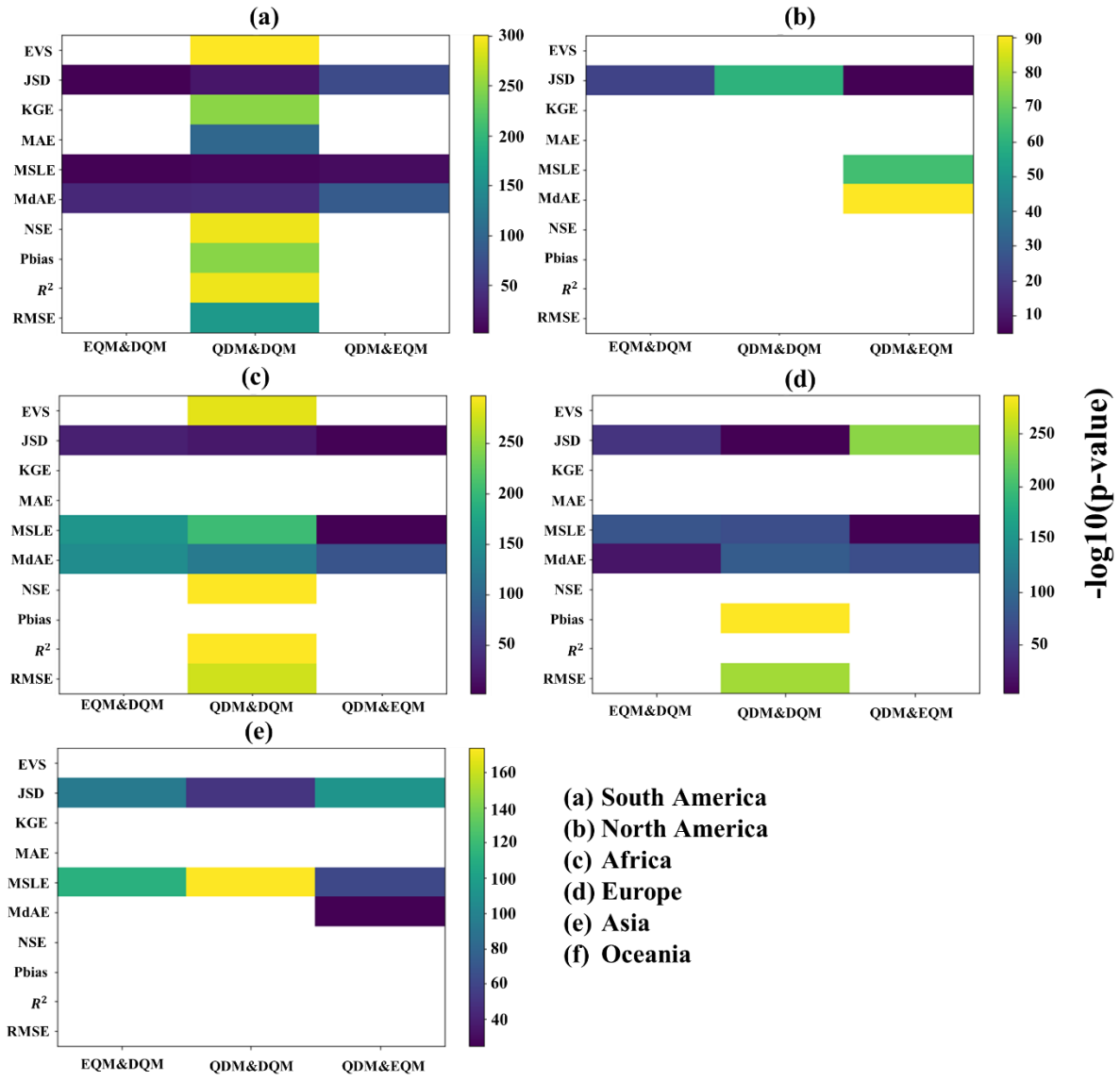

Figure 2. Statistical significance comparison of ten evaluation metrics based on Wilcoxon tests over six continents

We also substantially shortened the text in Section 3.1.2 and provided more quantitative results [Section 3.1.2].

Furthermore, we noted in the methodology that we used the Friedman and Wilcoxon tests as follows:

This study used the Friedman test to perform statistical comparisons among the three bias-correction methods (DQM, EQM, QDM), and when the Friedman test indicated overall significant differences, pairwise Wilcoxon signed-rank tests were conducted between each method pair to determine which specific comparisons differed. The detailed concepts of the two methods can be found in Friedman (1937) and Wilcoxon (1945).

Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance.

Journal of the American Statistical Association, 32(200), 675-701. 1937. https://doi.org/10.1080/01621459.1937.10503522

Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin, 1(6), 80-83. 1945. https://doi.org/10.2307/3001968

Comment 8

8. Following on comment 7, Figures 2-7 show the results for different metrics for different continents (again, presumably for the validation period, but that should be clearly stated in figure captions). Figures 2-7 all suffer from the same issues, but I will mostly focus on Figure 2. The color bars are non-linear – while each color segment is the same length, the interval they represent varies widely. For example, the row for Pbias the yellow represents a range of less than 2% while the purple represents over 20%, so a 3% negative bias is indistinguishable from a 20% negative bias. The scales themselves are confusing: for JSD the red colors are the worst skill, while for EVS red is the best skill. Every row is different in this regard. Some indices are unitless and others have units, and that should be represented. Some rows show a wide range of values (Pbias in Fig 2) while others show virtually identical values (NSE in Fig 3, where the colors vary only from 0.98 to 1.0), so where some differences are shown they may essentially be all the same value.

Answer

Thank you for your comment. The captions for Figures 2-7 now clearly indicate the validation period, and the color-bar scales have been redefined by dividing values into five quantiles to address the nonlinearity issue. Optimal metric values are consistently shown in red and the worst in green, and units have been added to any metrics that require them.

The figure below shows Figure 3 among the revised examples.

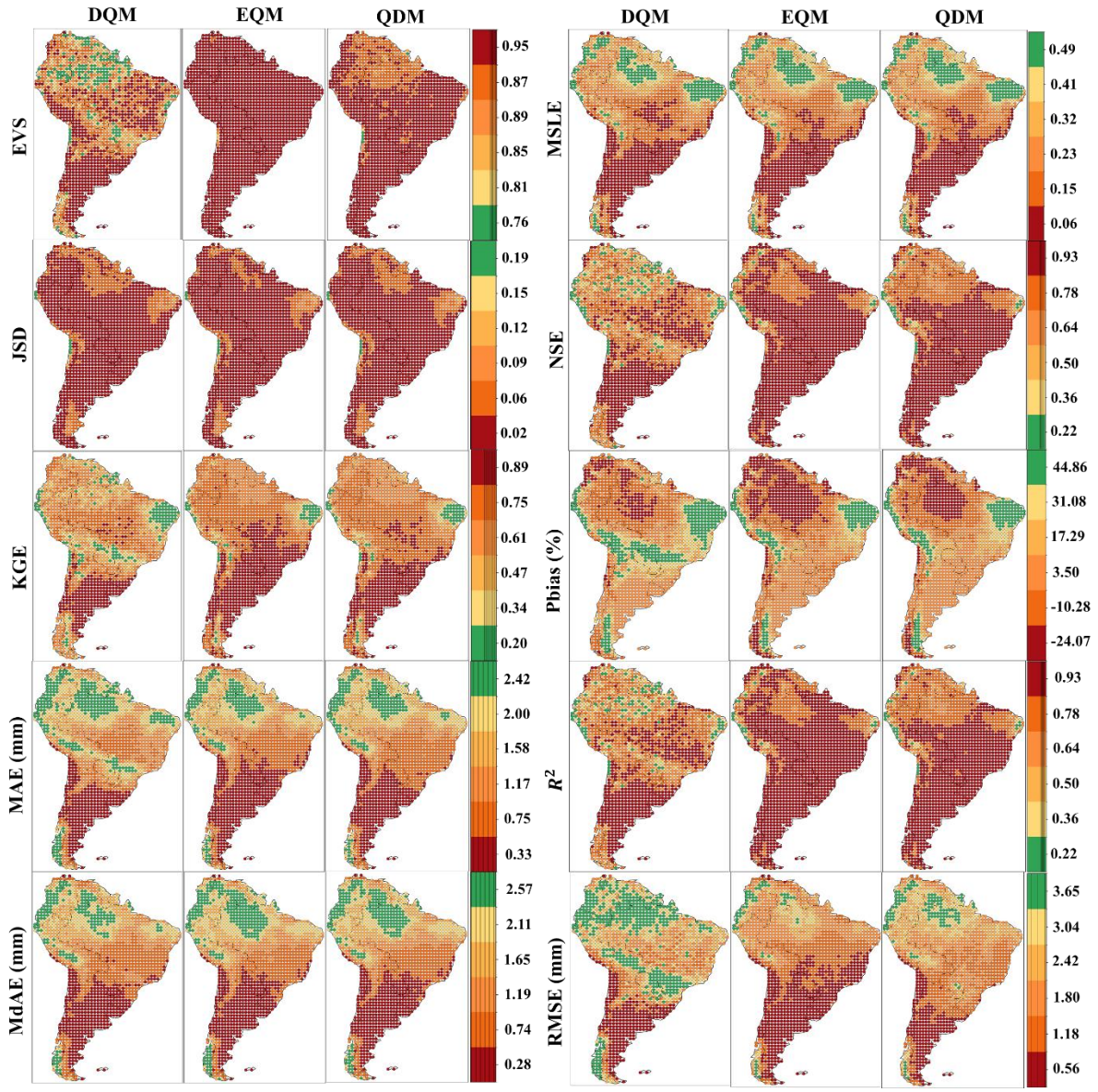

Figure 3. Performance comparison of DQM, EQM, and QDM for the validation period (1997-2014) using evaluation metrics for daily precipitation in South America.

Comment 9

9. The subset of indices shown in Figures 2-7 varies. While the supplemental material may complete the set, it should be consistent.

Answer

Thank you for your comment. To incorporate your comment, we have merged all metric-related figures from the supplemental material into the main text (Figures 3-9).

The figure below shows Figure 4 among the revised examples.

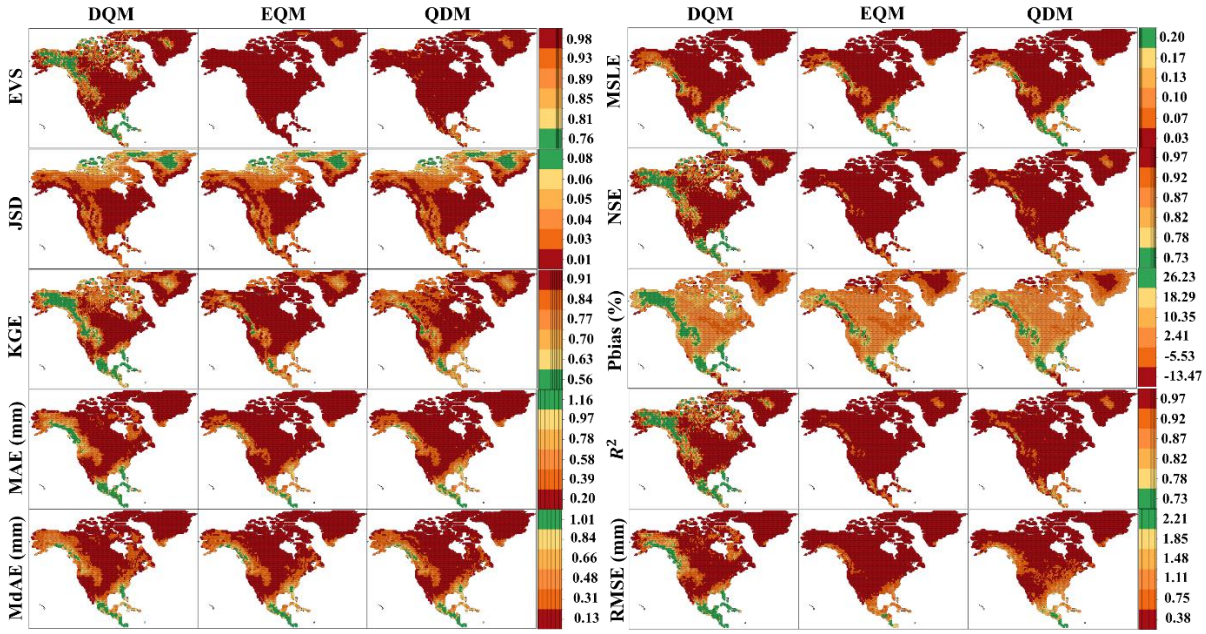

Figure 4. Performance comparison of DQM, EQM, and QDM for the validation period (1997-2014) using evaluation metrics for daily precipitation in North America.

Comment 10

Figure 8 summarizes on a continental scale the performance using each metric. Without any statistical test, the apparent differences cannot be claimed to represent anything. Also, with wide variability (in some metrics) across each continent, a single continent-wide average may not be very meaningful.

Answer

Thank you for your comment. First, we performed statistical tests based on the reviewer's comments using the Friedman and Wilcoxon paired comparison tests. These results are presented in Section 3.1.2. Furthermore, since boxplots display not only the median but also the maximum, minimum, first and third quartiles, and any outliers, they are not based solely on mean values. In fact, each boxplot contains all grid-cell results; using only the mean would produce a single value, making it impossible to generate a boxplot. To enrich the presentation of these results, we have revised the main text as follows.

Figure 9 presents the distribution of ten evaluation metrics for bias-corrected daily precipitation across six continents using boxplots. Each box shows the interquartile range (IQR) and median of the metric values computed over 11 CMIP6 GCMs. Overall, EQM′s boxes generally have lower medians and narrower IQRs for error metrics (RMSE, MSLE, MAE) on most continents, indicating both smaller typical errors and less scatter compared to QDM and DQM. QDM's boxplots lie slightly above those

of EQM but still exhibit relatively tight IQRs, suggesting consistently strong performance. In contrast, DQM often has higher median errors, wider IQRs, and more extreme outliers, reflecting larger and more variable biases relative to the other methods.

Comment 11

Section 3.1.3 looks at extreme precipitation. This is not well integrated into the rest of the paper, and only includes a cursory look at continent-aggregated values. It does not fit into any of the rest of Section 3. Line 445 claims differences are "relatively significant" and that distributions "vary significantly". Since there is no mention of statistical tests, these terms are inappropriate.

Answer

Thank you for your insightful comment. We have revised Section 3.1.3 to incorporate both a formal statistical test and a clear explanation of why extreme value behavior was analyzed separately. The updated paragraph now reads [Section 3.1.1]

This study also compared how well each bias correction method reproduces extreme precipitation by fitting a Generalized Extreme Value (GEV) distribution to the corrected daily values and then quantifying the distributional differences. Figure 10 shows the JSD of GEV fitted daily precipitation for DQM, EQM, and QDM on each continent. Across most continents, the median JSD for all three methods is extremely low (on the order of 10^{-4} to 10^{-5}), and even the interquartile ranges fall within narrow bands indicating that statistically the GEV curves for DQM, EQM, and QDM are almost indistinguishable for historical data.

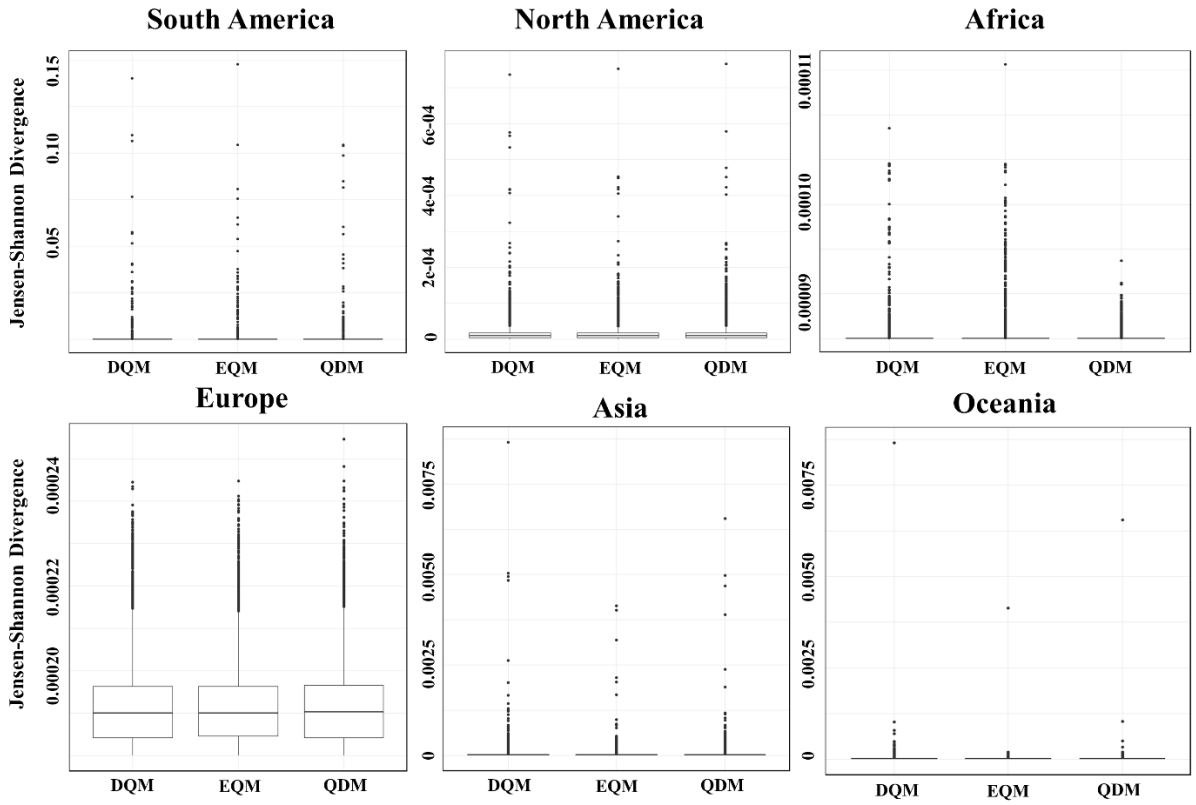

Figure 10. Comparison of distribution differences for GEV distribution using JSD across six continents.

Table 4 shows the results of a Friedman test and subsequent Wilcoxon signed rank pairwise comparisons for the ten highest daily precipitation values exceeding the 95th percentile on each continent. The Friedman test yielded a p-value of 4.5399×10^{-5}, indicating a highly significant difference and that at least one of the three quantile-mapping methods differs systematically. All Wilcoxon pairwise comparisons between methods produced 0.00195 on every continent, demonstrating that no two bias-correction approaches generate equivalent extreme-precipitation estimates.

Table 4. Friedman test and Wilcoxon paired comparison Test (p-values) by continent for precipitation exceeding the 95th percentile based on the GEV distribution

Continent	Friedman	Wilcoxon		
		DQM & EQM	DQM & QDM	EQM & QDM
South America	4.5399×10^{-5}	0.00195	0.00195	0.00195
North America				
Africa				
Europe				

Asia Oceania				

Because the reproducibility of extreme values in the corrected GCM is essential for impact assessments, Figure 11 presents the estimated probability density function (PDF) of precipitation values above the 95th percentile for the same GEV fit. Overall, DQM shows the highest probability density for extreme precipitation across all continents and has the widest tail, indicating that DQM boosts extreme events most aggressively. In contrast, EQM shows the lowest and narrowest density conservatively correcting extremes (often 5-8 % below DQM's values). QDM falls between EQM and DQM in most regions but remains closer to EQM.

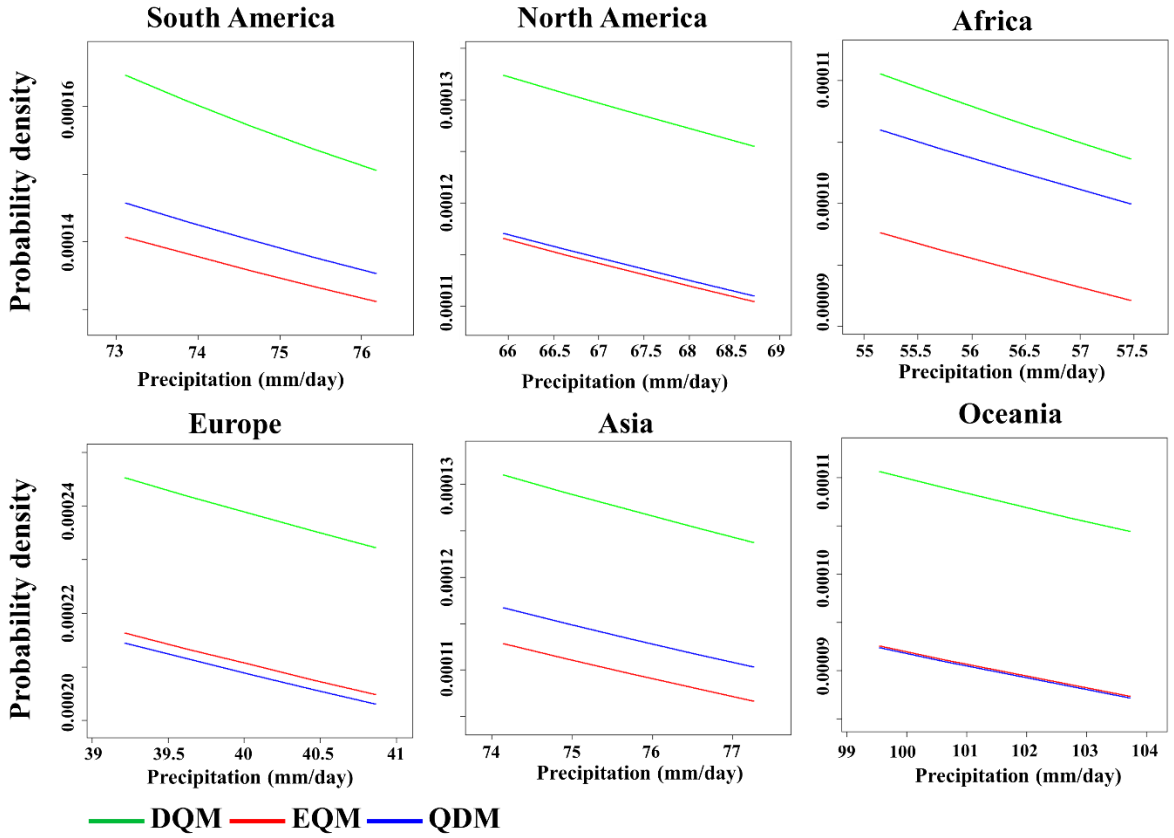

Figure 11. Comparison of probability densities for extreme precipitation values above the 95th percentile using GEV

By adding the Friedman test ($p < 0.001$) to Table 4's description and explicitly explaining why extreme value PDFs (Figure 10) matter, we have addressed the concern about statistical significance and the physical relevance of extreme precipitation differences. Thank you again for helping us strengthen this section.

Comment 12

12. Figure 9, Why do the scales between the top and bottom rows vary so widely (max of 0.15 vs 0.00024)? Does this figure only represent the 95th percentile like Fig 10?

Answer

Thank you for highlighting the apparent discrepancy in scales between Figures 9 and 10. We agree that it is important to clarify why the two figures display such different numerical ranges and how each figure serves a distinct analytical purpose. Figure 9 presents boxplots of the Jensen–Shannon Divergence (JSD) calculated over the entire GEV return-level distribution for each continent and each bias-correction method (QDM, EQM, DQM). By summarizing JSD values at every quantile, Figure 9 captures the full range of distributional differences—from the lowest return levels up to the most extreme values. Because JSD across all quantiles can span several orders of magnitude (especially when comparing methods under different climate regimes), the y-axis scales in Figure 9 vary to accommodate the true spread of values on each panel. This ensures that each continent's boxplot accurately shows the full extent of distributional differences rather than compressing them into a single, uniform scale. Figure 10, on the other hand, focuses exclusively on the upper tail of the GEV distribution, specifically the precipitation values above the 95th percentile. In other words, rather than evaluating JSD at every quantile, Figure 10 isolates the probability-density functions for extreme precipitation events (the top 5% of modeled intensities) and directly compares those densities across the three correction methods. Because probability densities in the extreme tail are inherently much smaller (on the order of 10^{-4}), the plotted values in Figure 10 appear in a much narrower range (with a maximum of around 0.00024). This deliberately zoomed-in view allows readers to see the subtle yet important differences in how each method handles the most extreme precipitation events, which would be difficult to discern if overlaid on the full-distribution scale used in Figure 9.

Comment 13

13. Figure 10, what are the values and units for the x-axis? Again, while these lines appear to be different the densities are extremely close (the y-axis scales cover a very small range) and no statistical test results are presented, so drawing conclusions is limited.

Answer

Thank you for your comment. In response, we have added the x-axis units to Figure 10.

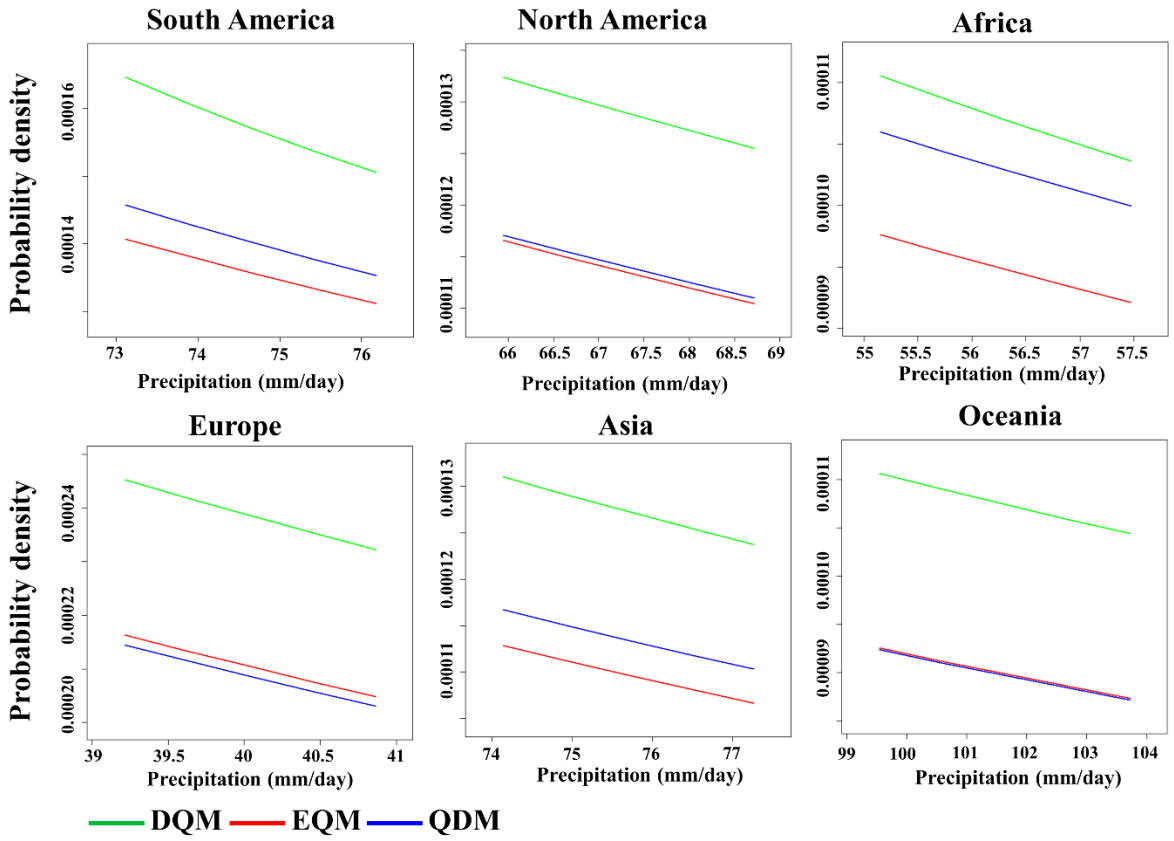

Figure 10. Comparison of probability densities for extreme precipitation values above the 95th percentile using GEV.

We also conducted statistical tests on the GEV 95th-percentile precipitation values using the Friedman and pairwise Wilcoxon tests, as shown in Table S6. Based on these results, we have revised the paragraph as follows:

Table 4. Friedman test and Wilcoxon paired comparison Test (p-values) by continent for precipitation exceeding the 95th percentile based on the GEV distribution

Continent	Friedman	Wilcoxon		
		DQM & EQM	DQM & QDM	EQM & QDM
South America	4.5399×10^{-5}	0.00195	0.00195	0.00195
North America				
Africa				
Europe				
Asia				
Oceania				

Furthermore, the main text has been revised as follows:

This study also compared how well each bias correction method reproduces extreme precipitation by fitting a Generalized Extreme Value (GEV) distribution to the corrected daily values and then quantifying the distributional differences. Figure 9 shows the JSD of GEV fitted daily precipitation for DQM, EQM, and QDM on each continent. Across most continents, the median JSD for all three methods is extremely low (on the order of 10^{-4} to 10^{-5}), and even the interquartile ranges fall within narrow bands indicating that statistically the GEV curves for DQM, EQM, and QDM are almost indistinguishable for historical data.

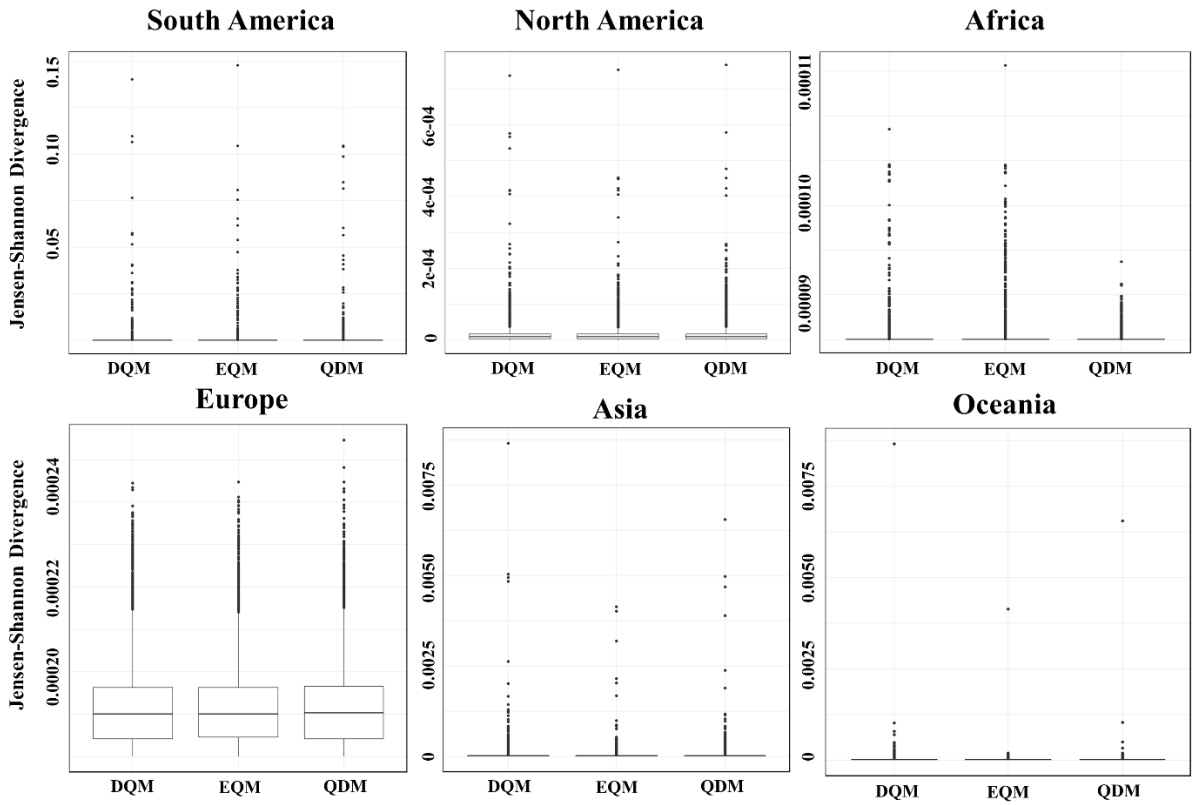

Figure 9. Comparison of distribution differences for GEV distribution using JSD across six continents.

Table 4 shows the results of a Friedman test and subsequent Wilcoxon signed rank pairwise comparisons for the ten highest daily precipitation values exceeding the 95th percentile on each continent. The Friedman test yielded a p-value of 4.5399×10^{-5}, indicating a highly significant difference and that at least one of the three quantile-mapping methods differs systematically. All Wilcoxon pairwise comparisons between methods produced 0.00195 on every continent, demonstrating that no two bias-correction approaches generate equivalent extreme-precipitation estimates.

Table 4. Friedman test and Wilcoxon paired comparison Test (p-values) by continent for precipitation exceeding the 95th percentile based on the GEV distribution

Continent	Friedman	Wilcoxon

		DQM & EQM	DQM & QDM	EQM & QDM
South America	4.5399×10^{-5}	0.00195	0.00195	0.00195
North America				
Africa				
Europe				
Asia				
Oceania				

Because the reproducibility of extreme values in the corrected GCM is essential for impact assessments, Figure 10 presents the estimated probability density function (PDF) of precipitation values above the 95th percentile for the same GEV fit. Overall, DQM shows the highest probability density for extreme precipitation across all continents and has the widest tail, indicating that DQM boosts extreme events most aggressively. In contrast, EQM shows the lowest and narrowest density conservatively correcting extremes (often 5-8 % below DQM's values). QDM falls between EQM and DQM in most regions but remains closer to EQM.

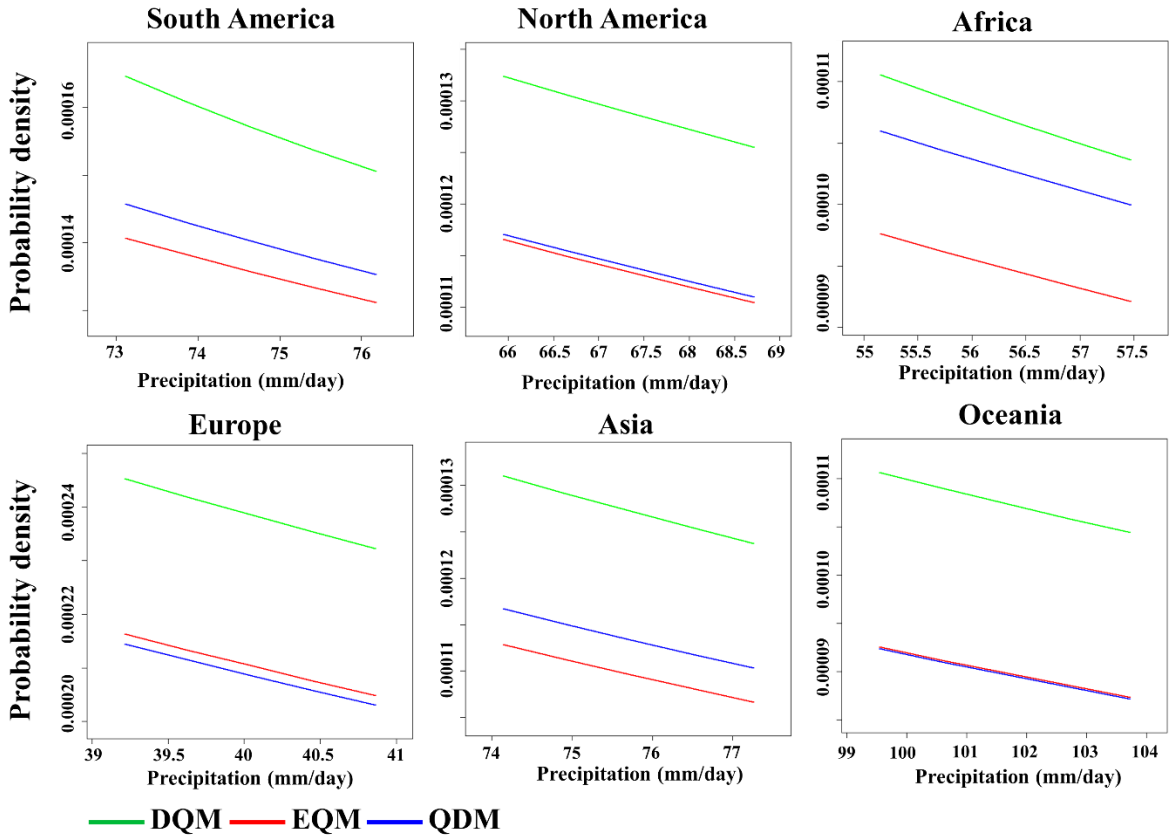

Figure 10. Comparison of probability densities for extreme precipitation values above the 95th percentile using GEV

Comment 14

Line 464, the weights are “calculated by applying entropy theory”, but that does not appear to have been discussed in the methods section of the paper. This should be mentioned there with appropriate citations.

Answer

Thank you for your comment. We have added the following sentence to Section 2.7:

Moreover, this study employed entropy theory to compute objective weights for the evaluation metrics as an alternative to TOPSIS (Shannon and Weaver 1949).

Additionally, we have included the following references.

Shannon, C. E., and Weaver, W.: The mathematical theory of communication. University of Illinois Press. 1949

Comment 15

Line 464, While this seems like the more important part of the effort, higher and lower weights are only discussed qualitatively, and while “significant importance” is mentioned (Line 471) and differences are claimed to be “significant” (Line 499) no assessment of significance is shown.

Answer

Thank you for your comment. As noted in Section 3.1.2, we applied these nonparametric tests to the ten evaluation metrics (RMSE, MAE, R^2, NSE, KGE, Pbias, MdAE, MSLE, EVS, JSD) used in our entropy-weight calculation. The full test statistics and p-values are now provided in Table 4, demonstrating that, for each continent, metrics such as JSD in South America and RMSE/MAE/MSLE in North America differ from lower-weighted metrics (e.g., EVS, NSE) at $p < 0.01$. Because entropy weights directly reflect these underlying metric differences, this analysis confirms that descriptors like significance and differences are grounded in robust statistical evidence rather than qualitative judgment alone.

Below is the revised paragraph including our new sentence:

By conducting Friedman and Wilcoxon tests on the evaluation metrics, this study confirms that the observed differences in entropy-derived weights are statistically significant. In this study, the weights were calculated by applying entropy theory to the evaluation metrics used in the TOPSIS analysis, and the results are presented in Table 5.

We trust these additions fully address your concern by linking our qualitative statements to the quantitative results of the Friedman and Wilcoxon tests. Thank you again for guiding us toward a more rigorous and transparent presentation.

Comment 16

16. Figs. 12, 14, and 16 have many of the same issues as Figure 2-7, noted in comment 8 above.

Answer

Thank you for your comment. We have revised all Figures 12, 14, and 16 in response to your comments, and we have also updated the figures in the supplementary materials accordingly (Figure S2 and S3).

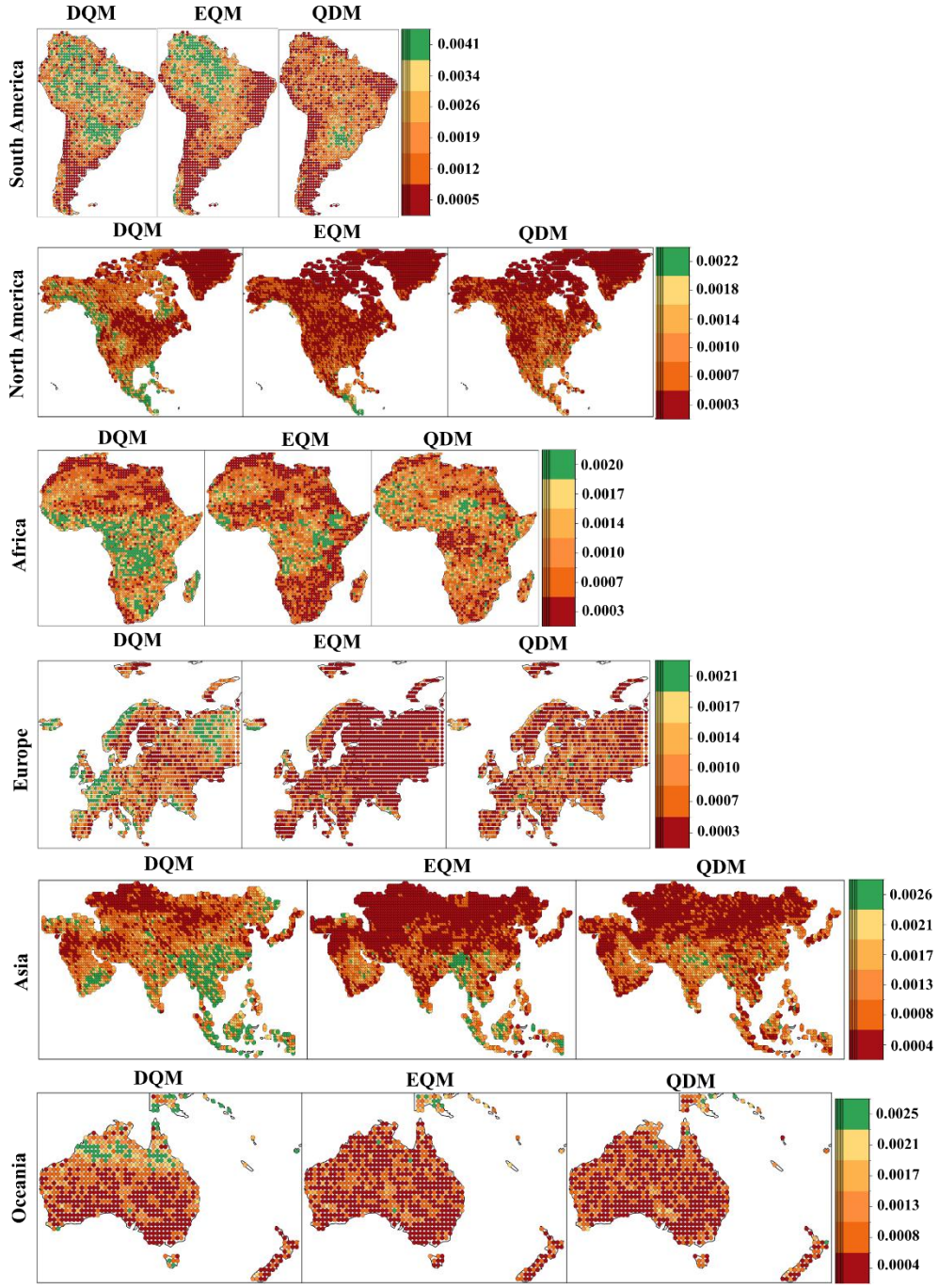

Figure 13. Spatial distribution of weight variance across continents for bias corrected CMIP6 GCMs using BMA

Comment 17

17. Lines 557-563 discuss Fig. 15. It is claimed that "the EQM ensemble showed the lowest standard deviation across all continents." The box and whisker plots clearly shows nearly identical results for all methods, and I would be shocked if any of the differences showed any statistical significance. The

conclusions here are just not supported.

Answer

Thank you for your comment. We have revised the sentence as follows in response to your comments.

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. Overall, 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, indicating the greatest prediction uncertainty.

Comment 18

18. Fig 17 shows the final comprehensive index across all continents. That some apparent differences emerge is due to weighting the uncertainty (as in Fig 15) less. Statistical tests are needed here too, and then it could be explored why the closeness index changes the results so strongly.

Answer

Thank you for your comment. To address your feedback, we applied the Friedman and pairwise Wilcoxon tests, and we have added Table 6 and the following text to the manuscript.

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. 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.

Table 6. 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

Comment 19

1. Line 42, rather than saying the bias corrections differ in “physical approaches”, “statistical approaches” would be more accurate.

Answer

Thank you for your comment. We have revised the text as follows in response to your comment.

Despite these advancements, the suggested bias correction methods differ in their statistical approaches, resulting in discrepancies in the climate variables adjusted for historical periods.

Comment 20

2. Line 47, Correct spelling of Maraun.

Answer

Thank you for your comment. We have revised the text as follows in response to your comment.

Furthermore, the distribution of precipitation across continents and specific locations causes variations in the correction outcomes depending on the method used, which makes it challenging to reflect extreme climate events in future projections and adds another layer of confusion to climate change research

(Song et al., 2022b; Maraun, 2013; Ehret et al., 2012; Enayati et al., 2021).

Comment 21

3. Line 61, Elaborate on “higher performance.” Is this the same as “better skill”?

Answer

Thank you for your comment. We have revised the text as follows in response to your comment.

In recent years, climate studies using GCMs have adopted several improved QM methods that offer better skill meaning reduced bias and more accurate distributional matching than previous approaches to correct historical precipitation and project it into the future.

Comment 22

4. Line 136, the model resolution “was provided by the institution for research availability.” This is confusing. Is there a citation to add?

Answer

Thank you for your comment. This sentence was removed during an earlier review process.

Comment 23

5. Line 212, “JSD” is used, where prior to this JS-D is used. The term should be consistent throughout.

Answer

Thank you for your comment. We have corrected JS-D to JSD.

Comment 24

6. Line 361, The sentence begins with “In this study, the daily precipitation in Africa was corrected using three QM methods.” This sort of recap appears at the beginning of many sub-sections, and is not needed. Search for “This study,” and rephrase.

Answer

Thank you for your comment. We have revised the text as follows in response to your comment.

Daily precipitation in Africa was corrected using three QM methods, and performance is shown in Figures 4 and S3.

Comment 25

7. Line 438-439, ”adjusted by the biased bias correction methods” must be an error.

Answer

Thank you for your comment. This sentence has been removed in response to your comment.