==

RESPONSE TO EDITOR

==

Dear Editor,
We hereby include a detailed response to the referee, hoping to adequately attend his/her questions, some of them already discussed and addressed in the previous iteration. There seem to be two main concerns that we again try to clarify, and that are now emphasized and further elaborated in this third manuscript review. For your convenience, we next describe the two main concerns raised and a summary of our reply and steps followed to clarify these issues.

1. **Regarding the use of an older precipitation product**. We would like to address the referee's concern regarding our use of an older satellite product (TRMM) instead of the currently available IMERG database. To verify the referee's claims about the absence of biases in IMERG, we conducted an additional analysis. While IMERG indeed inherits many attributes from TRMM, including the use of multiple satellites and sensors, and offers improved accuracy and geographical coverage, our analysis indicates that the biases observed in TRMM persist in the newer IMERG products. Given that these biases remain at the locations of our study, we believe that the choice of data source is of secondary importance for illustrating our calibration methodology. The primary focus of our study is on the nature and application of the calibration methodology itself. This approach not only highlights the importance of calibration but also demonstrates the applicability of our methods to both older and newer datasets. Therefore, we consider both the calibration methodology and the results obtained to be perfectly valid and relevant. Additionally, our methods are easily applicable to new datasets, as we have included an open-source notebook to facilitate the reproducibility of the results.

2. **Regarding the limited number of calibration sites**. As we indicated in the manuscript, this limitation is due to the availability of high-quality records over a sufficiently long temporal period, which is essential to ensure the robustness of calibration. Our study focuses on an area where a previously established weather typing exists, a necessary condition for the application of the methodology. This weather typing is crucial for describing the seasonality and other precipitation patterns in the region, providing a relevant context for our analysis, in which calibration is conditioned to specific situations. Despite the limited number of stations, our results are consistent and clearly demonstrate the validity and usefulness of the adaptive calibration approach. The consistency of our findings across the available stations supports the reliability of our methodology. Additionally, the high-quality, long-term data we used enhances the credibility of our results. We believe that the primary objective of our research is methodological. Therefore, it is not essential to include a large number of locations. Instead, we aim to demonstrate the methodology's application and its performance across well-scattered locations within the weather typing domain, each with

varying conditions. Following the referee's concern about lack of robustness of the results, it is important to note that the indicated P95 in manuscript's Table B1, does not depend on the number of stations but on the length of the time series, which is sufficient to obtain robust results, as already explained in the manuscript. This is precisely why we selected a reduced set of locations, applying a stringent criterion to ensure long time series for a robust conditioned calibration.

We hope this clarifies our position and addresses the referee's concerns. With this thorough revision, we sincerely hope that our study meets the high-quality standards of the HESS Journal and will be accepted for publication.

Yours sincerely,

J. Bedia,
On behalf of the authors

==

RESPONSE TO REFEREE

==

This is my second review for this manuscript. Significant improvements have been made by the authors. Thank you very much. Here are some comments that I hope will help the author improve this manuscript.

Thank you very much for your positive feedback. We have made significant efforts to enhance the original manuscript, incorporating the valuable suggestions provided by the referee in his/her first review. We also appreciate the referee's new comments, which we believe will help to better communicate the significance of our study and the relevance of its main findings.

Throughout our response, we will interweave the referee's comments (highlighted in italics) with our replies.

We thank the referee again for their time and positive feedback, and sincerely hope that our responses will meet his/her expectations.

Comments

1. Line 12: please delete "(WT)".

Done.

2. Lines 41-47: the explanation about the selection of the TMPA product is not convincing and misleads readers. The IMERG inherits all attributes of TMPA. In addition, the precipitation estimates of TMPA came from multiple satellites with different orbits and satellite sensors, the same as IMERG. I think the effectiveness and applicability of the proposed approach should be based on the data that is being updated and better, rather than the data that has stopped updating, which does not make much sense.

Thank you for your insightful comments regarding the selection of the TMPA product. We understand your concerns about using data that is no longer updated. However, we included the TMPA data specifically to illustrate the strong biases present in older datasets, which effectively highlight the impact of different calibration methods.

While it is true that IMERG inherits many attributes from TMPA including the use of multiple satellites and sensors, and improves its overall representation of precipitation, our analysis shows that the biases observed in TMPA are still present in the newer IMERG products. Our analysis shows that the biases observed in TMPA are still present in the newer IMERG products at the selected localities. In Fig.1, we depict the biases of both TMPA and IMERG using a quantile-quantile plot against the rain gauge records. In Fig. 2, these biases are more accurately quantified in terms of the different validation indices used in this study (Table 2 of the manuscript). The new IMERG product shows a moderate overall improvement over the old TMPA product in some sites (Alofi, Raoul Island); however, biases are still present, and in some locations are even higher that TMPA (Port Vila) and therefore also require calibration against a 'ground-truth' reference.

Therefore, for the purpose of illustrating our calibration methodology, the choice of data source is, in our view, of secondary importance, as the emphasis of our study is on the nature and application of the calibration methodology itself. We believe this approach not only underscores the importance of calibration but also validates the applicability of our methods to both older and newer datasets.

We appreciate your feedback and hope this explanation clarifies our rationale for including the TMPA data. We have included a mention to this rationale in the new revised version of our manuscript (Lines 43-49), and also included the figures above in the new Appendix C to showcase the remaining biases in the IMERG dataset.

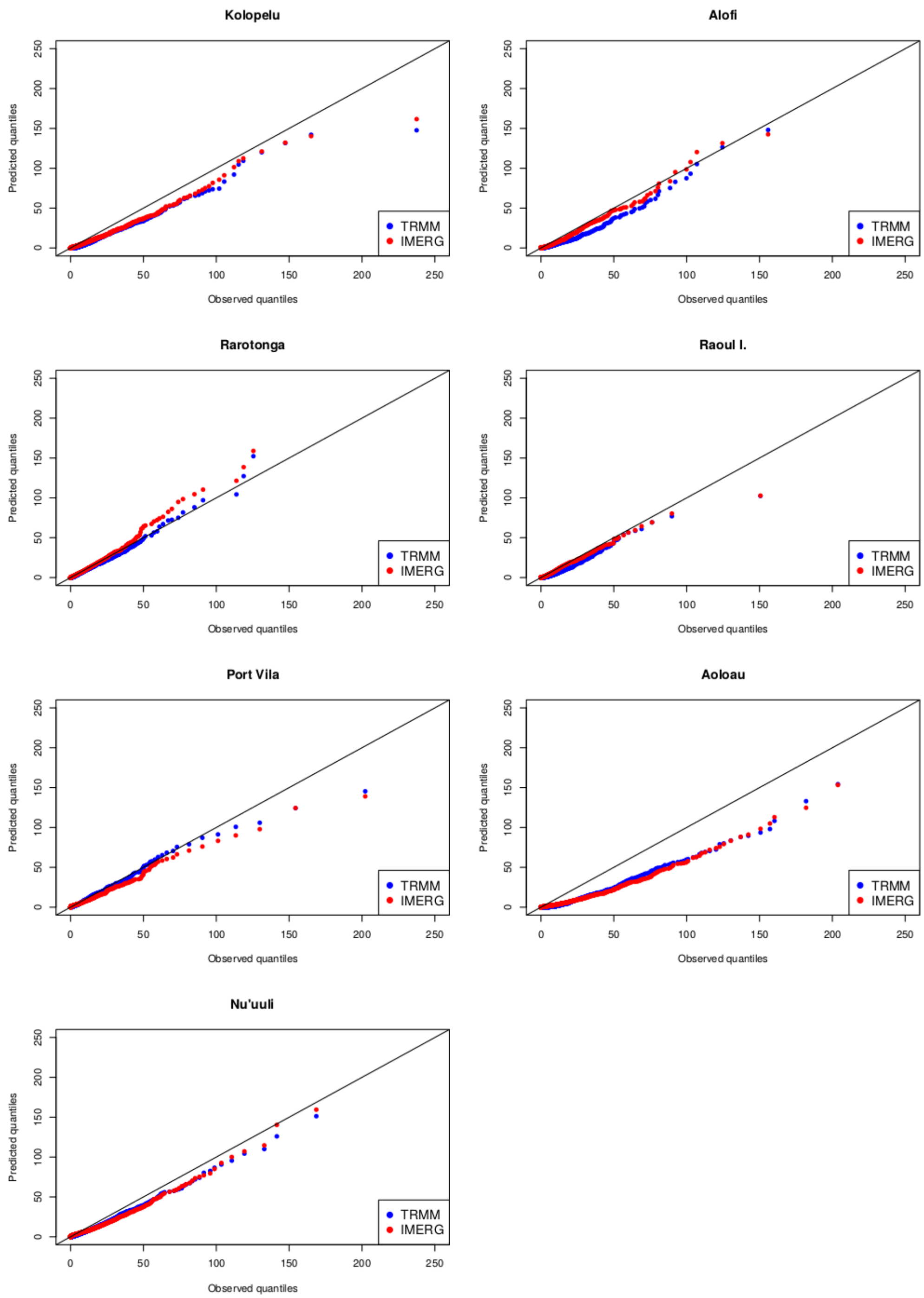

Figure 1. Quantile-quantile plots of TRMM (blue), and IMERG (red), against the rain gauge daily precipitation records of the stations indicated.

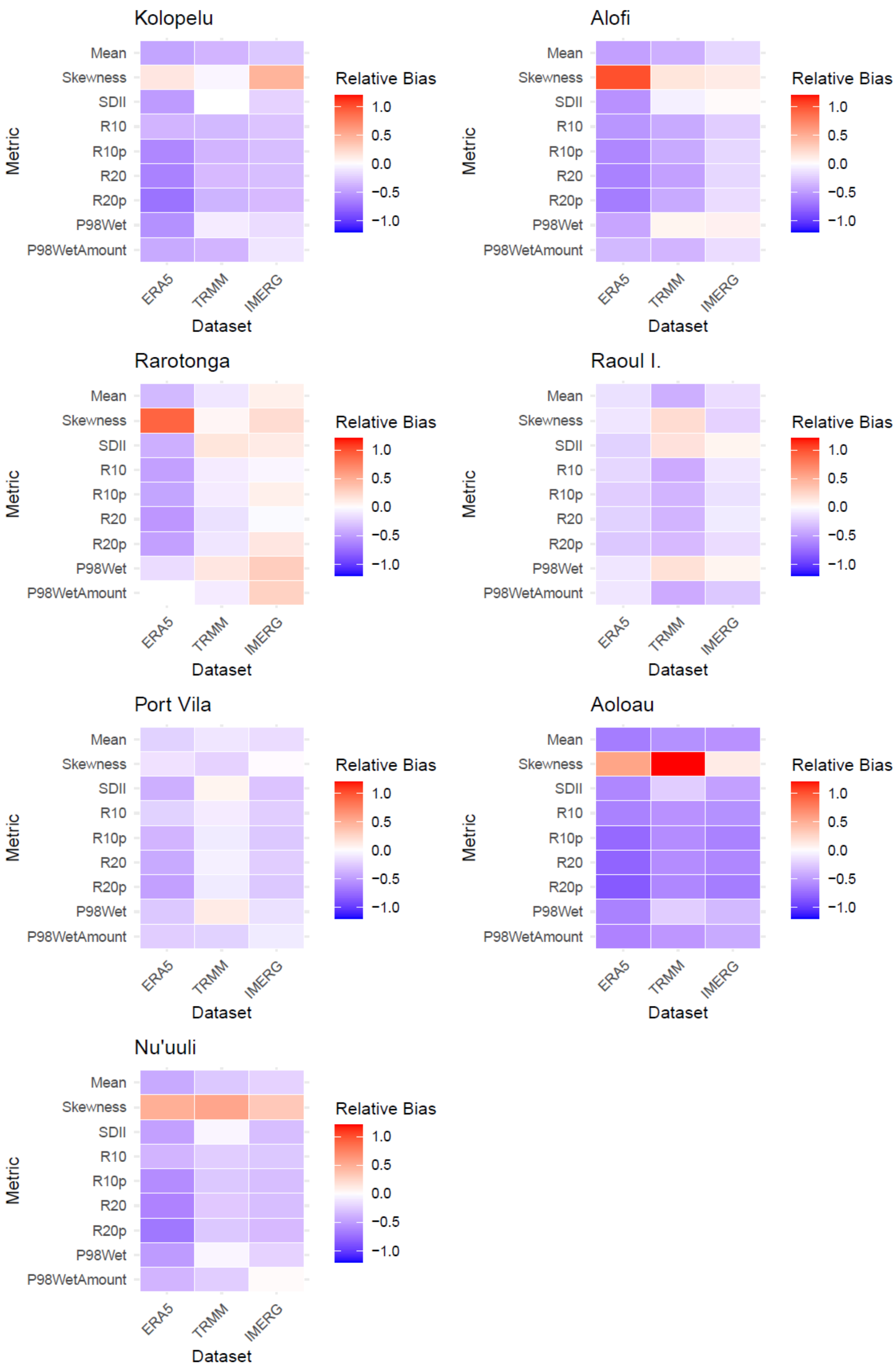

Figure 2. Relative Biases of the climate indices used for validation (Table 2 of the manuscript) of raw TRMM and IMERG datasets, at the closest grid points of the locations of each reference station used in the study (Table 1 of the manuscript).

3. The reliability of the results is doubtful, as the total number of rain gauges is too small, especially for P95 (see Table A1). Appropriate discussion for this should be Added.

Thank you for your valuable feedback regarding the number of rain gauges used in our study. We acknowledge that the total number of gauges is limited. However, as we already indicate in the manuscript, this limitation is due to the availability of high-quality records over a sufficiently long temporal period, which is essential to ensure the robustness of our calibration results.

Furthermore, our study is focused on an area where a previously weather typing exists (Mirones et al. 2023). This weather typing is crucial for describing the seasonality and other precipitation patterns in the region, providing a relevant context for our analysis, in which calibration is conditioned to specific situations.

Despite the limited number of stations, our results are consistent and clearly demonstrate the validity and usefulness of the adaptive calibration approach. The consistency of our findings across the available stations supports the reliability of our methodology. Additionally, the high-quality, long-term data we used enhances the credibility of our results. We believe that the primary objective of our research is methodological. Therefore, it is not essential to include a large number of locations. Instead, we aim to demonstrate the methodology's applications and its performance across a few locations, well scattered across the weather typing domain, with varying conditions. Following the referee's comment, please note that the indicated P95 in the above-mentioned Table, on the other hand, does not depend on the number of stations, but on the length of the time series, one of the reasons for choosing such a reduced set of locations, containing a sufficient amount of data for a conditioned calibration.

We believe that our study provides a valuable contribution to the field, illustrating the effectiveness of our calibration approach even with a limited number of stations. We appreciate your consideration of these points and hope this explanation addresses your concerns. The points above addressed are already indicated in the manuscript (Sec. 2.1)

4. It is not clear that the TMPA used in this study is calibrated against a standard satellite product or uncalibrated or native product.

Thank you for pointing out the need for clarity regarding the TMPA product used in this study. The TMPA product used is the 3B42_7B version (as already outlined in Sec. 2.1), a research-grade dataset designed for climatological and hydrological analyses. This product is calibrated by combining satellite-based precipitation estimates with ground-based gauge data to enhance accuracy. Specifically, it applies bias corrections using monthly gauge analysis from the Global Precipitation Climatology Centre (GPCC), which helps reduce systematic errors present in the raw satellite estimates. Unlike the real-time 3B42RT product, the 3B42_7B version undergoes additional processing to ensure consistency and reliability, making it a suitable choice for retrospective scientific studies.

Despite these calibrations, the product still exhibits significant biases, as demonstrated in our study. We have included extended information in the revised text (Sec. 2.1, Lines 107-115) to include further technical details of this TRMM version.

5. Line 111: please provide the full name of PCA at the first appearance.

Done

6. Lines 115: ‘in order”>>in order to.

Done

7. Line 127: please provide the full name of EOF.

Done

8. I suggest providing the flowchart of the method used in this study.

Thanks for this suggestion aimed at improving clarity. While we acknowledge that our approach is methodologically more complex than a direct calibration, we believe it is not so intricate as to necessitate a flowchart representation. We are committed to making our study as clear and comprehensive as possible, and we believe that the detailed descriptions and the prepared data and code we have included already ensure the reproducibility of each step in our study. However, we are open to further discussion if you believe that a specific aspect of our method could benefit from additional visual representation.

9. Lines 199-202: The description of the weighting method is very crude and vague. Please provide detailed weighting criteria or theoretical support for Table A2. In my opinion, the weighting method used in this study is highly arbitrary.

Firstly, we are unsure what is meant by "crude." Regarding the term "vague," we believe our method is quite detailed and addresses key aspects of precipitation necessary for an adequate assessment of the calibrated series. Our approach leverages well-defined validation indices that consider various statistical aspects of precipitation (e.g., frequency, mean amount, extremes, overall distribution shape) and associated performance measures (e.g., biases—relative and/or absolute—and correlation). These are clearly indicated in Table 2 and supported by the references listed in our manuscript.

Moreover, we believe the validation performed in our study surpasses the typical evaluations conducted in the calibration of satellite products, which often focus on simpler indicators like monthly accumulations. This is also highlighted in the references of our manuscript.

We have already introduced several clarifications in response to your initial comments. It is important to emphasize that the validation of a calibration method is inherently multi-faceted, with results varying based on the aspects considered. Therefore, we have adopted a rigorous approach to ensure a comprehensive validation, following the VALUE framework, which is widely used in the Euro-CORDEX community's coordinated downscaling activities.

Given the multivariate nature of the validation, our weighting method provides users with maximum flexibility to incorporate various metrics into the overall evaluation. This method allows for the assignment of greater weight to certain aspects over others, depending on the context (e.g., emphasizing extremes for hydrological studies or frequencies for agricultural studies). Calibration can alter these distributional aspects of precipitation, necessitating a comprehensive evaluation method.

The introduction of the ranking framework (RF) enables us to rank different calibration methods by summarizing individual results for each validation measure into a single metric. This approach is not new and has been used in previous model validation work, such as the observational uncertainty analysis in Euro-CORDEX (Kotlarsky et al., 2019). We have adopted this robust and widely accepted methodology in our study.

While there is some degree of arbitrariness in selecting validation metrics and assigning weights, we view this as a strength rather than a weakness. This flexibility allows the method to be tailored to different needs. The specific set of validation indices and weights used in our study serves as an example that users can modify as needed, following the reproducibility examples provided.

We hope this clarifies our approach and addresses your concerns.

10. Why introduce ERA5 as a reference to compute the bias? It makes no sense for validation. If necessary, please provide appropriate reasons.

Thank you for raising the issue of using ERA5 as an additional reference to compute biases. We understand your concern about its role in validation. However, please note that the use of ERA5 here is purely comparative, and it is not used elsewhere in the calibration process. In any case, we think that the inclusion of these data adds value to the results presented, helping to better contextualize the magnitudes of the biases found. We deemed it relevant to use ERA5 as a reference to compare its biases with satellite data for several reasons:

1. High-Quality Data: ERA5 is a state-of-the-art reanalysis product that integrates a vast array of observations from satellites, ground stations, and other sources using advanced data assimilation techniques. This makes ERA5 data high-quality, consistent, and reliable, widely used by the community in many impact applications, despite some remaining biases.
2. Detailed Resolution: ERA5 offers high spatial (~31 km) and temporal (hourly) resolution, which is crucial for capturing detailed precipitation patterns. This comprehensive

coverage allows for a more meaningful comparison with satellite data across different regions and time periods.

3. Proven Accuracy: Studies have shown that ERA5 has smaller biases compared to its predecessors and other reanalysis products. For example, ERA5 exhibits lower bias in tropical regions compared to ERA-Interim, MERRA-2 and JRA-55 (Hassler and Lauer 2021), and improves global-mean correlation of precipitation with monthly-mean GPCP, making it a more accurate reference for precipitation data (Hersbach et al. 2020).
4. Consistency with Previous Work: We used ERA5 for a previous weather typing (Mirones et al. 2023), which we later applied to condition our satellite precipitation calibration. Presenting the biases of ERA5 in this context provides continuity and consistency with our earlier work, further validating our approach.

By using ERA5 as a reference, we aim to provide a high-quality benchmark to contextualize the biases of satellite data. This comparison helps to highlight the strengths of the satellite products while also underscoring the importance of calibration against a reliable reference, as satellite product biases remain an issue.

We agree with the referee that providing a rationale for this choice would enhance the overall quality of our manuscript. We have succinctly addressed this in L231.

11. The unweighted RF means equal weight for each metric?

Yes, we have included a sentence to clarify this term when it first appears in the text.

References

Hassler, B., Lauer, A., 2021. Comparison of Reanalysis and Observational Precipitation Datasets Including ERA5 and WFDE5. Atmosphere 12, 1462. https://doi.org/10.3390/atmos12111462

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J., 2020. The ERA5 global reanalysis. Q.J.R. Meteorol. Soc. 146, 1999–2049. https://doi.org/10.1002/qj.3803

Kotlarski, S., Szabó, P., Herrera, S., Räty, O., Keuler, K., Soares, P.M., Cardoso, R.M., Bosshard, T., Pagé, C., Boberg, F., Gutiérrez, J.M., Isotta, F.A., Jaczewski, A., Kreienkamp, F., Liniger, M.A., Lussana, C., Pianko-Kluczyńska, K., 2019. Observational uncertainty and regional climate model evaluation: A pan-European perspective. Int J Climatol 39, 3730–3749. https://doi.org/10.1002/joc.5249

Mirones, Ó., Bedia, J., Fernández-Granja, J.A., Herrera, S., Van Vloten, S.O., Pozo, A., Cagigal, L., Méndez, F.J., 2023. Weather-type-conditioned calibration of Tropical Rainfall Measuring Mission precipitation over the South Pacific Convergence Zone. International Journal of Climatology 43, 1193–1210. https://doi.org/10.1002/joc.7905