In terms of the weighting method, the authors did not answer my previous question correctly. I want to know why the weight values of some metrics (e.g., Mean, SDII) are 0.05, and some are 0.15, 0.1, or 0.2. Why not are 0.01, 0.02, 0.03, ...(Just an example, not meaning that this weighting method is correct)? This requires robust theoretical support.

Dear referee,

Thank you again for your insightful comments to our paper, and also your question about the RF score weighting methodology. We apologize for any confusion caused by our previous response.

To clarify, the weights assigned to each metric in our validation methodology are designed to sum to one, ensuring a balanced and normalized approach, as indicated in eq. 3 in our manuscript:

$$RF_j = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot Z'_{i,j} \quad \text{where} \quad \sum_{i=1}^{N} w_i = 1$$

Where RF is the "ranking framework" score for calibration *j*, *N* is the number of validation indices considered in the validation procedure, $Z'_{i,j}$ is the (absolute) normalized bias for index *i* and calibration *j* and w_i is the weight assigned to each validation index. In our manuscript we have considered a set of N=10 indices widely used to describe different aspects of precipitation, and also some arbitrary weights in the case of the weighted RF scheme, for illustration purposes.

Normalization of weights is a straightforward method where we assign weights based on criteria such as the importance or impact of the validation measure within the validation procedure. For example, we might assign a higher value to performance in extreme indices. Each weight is then divided by the total sum of all weights to ensure they sum to one. This method is often used in multi-criteria decision analysis, providing a clear and intuitive approach for weighted validation schemes.

The specific values (e.g., 0.05, 0.15, 0.1, 0.2) were chosen to illustrate the potential for varying weights based on different criteria. These values are illustrative and do not serve any specific purpose beyond demonstrating the possibilities of the weighted approach. We acknowledge that an equally weighted scheme is often a suitable alternative, particularly when aiming to avoid arbitrary choices.

For instance, the weight of 0.05 for the SDII bias reflects its relative importance in our analysis. Biases in this index are less penalized than biases in other indices with higher weights, such as 0.2 for the P98 wet amount. The value of 0.05 indicates that the performance of the calibration method in this specific index accounts for 5% of the total RF score, which is appropriate for a calibration procedure that does not prioritize this indicator highly. Similarly, the weights of 0.15, 0.1, and 0.2 for other metrics were chosen to reflect their respective contributions to the overall validation process.

We have introduced an explanatory sentence in the new revised version of our manuscript (Sec. 2.4.2) to better explain this point. We hope this explanation provides the robust theoretical support you requested. Please let us know if further clarification is needed.