Replies to reviewer #2

"Benchmarking convection-permitting climate simulations for hydrometeorological characterizations: A comparative study of WRF-SAAG and observation-based products in Chile"

Sofía Segovia-Escafi, Pablo A. Mendoza, Miguel Lagos-Zúñiga, Lucía Scaff, and Andreas Prein

We thank the reviewer for his/her time, revision, and suggestions to our paper. We provide responses to each individual point below. For clarity, comments are given in black italics, and our responses are given in plain blue text.

Anonymous Referee #2

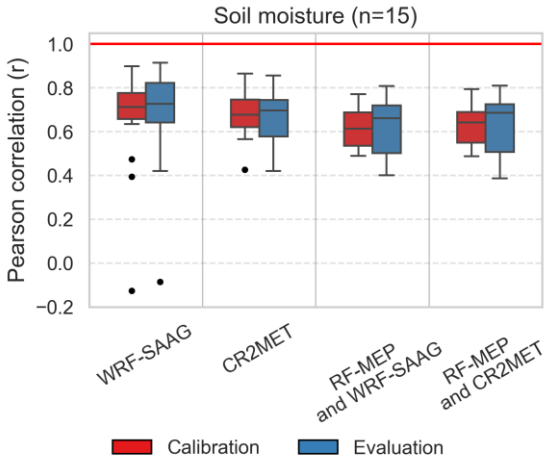
I appreciate that the authors have sincerely addressed my previous comments, and the manuscript has improved as a result.

We thank the reviewer for the positive feedback.

Below are my additional comments based on the current version of the manuscript:

1. Line 360, "volumetric soil moisture was estimated by normalizing simulated soil moisture by the field capacity parameter in mm". This does not seem correct. Converting volumetric soil moisture (m^3/m^3) to a soil moisture depth in mm requires multiplying the volumetric water content (VWC) by the thickness of the soil layer in mm, not normalizing by field capacity. The authors will need to clarify the depth of the soil moisture output from the TUW model and then standardize it to the 2-5 cm depth consistent with the satellite soil moisture (line 237) for comparison.

We thank the reviewer for catching this. In the original manuscript, we normalized simulated soil moisture by the field capacity parameter (FC, mm) to obtain a dimensionless variable. We agree with the reviewer that such ratio is not volumetric soil moisture, but rather the degree of saturation (which ranges 0-1). Therefore, we now evaluate soil moisture simulations by contrasting the simulated degree of saturation (i.e., simulated soil moisture divided by the parameter FC) with the variable "Surface soil moisture (saturation) %" (retrieved from the reference product Copernicus Climate Change Service, Climate Data Store), defined as the content of liquid water in a surface soil layer of 2–5 cm depth expressed as the fraction of total saturation. Since the TUW model is not physically based, the depth of the surface soil layer is not considered, we decided to evaluate the simulated degree of saturation using the Pearson correlation coefficient (i.e., we only assessed the timing of soil moisture simulations). The behavior of this correlation for 15 basins can be seen in the following Figure (Fig 12 on the manuscript)



2. The manuscript includes a lot of details, but dataset performance is simultaneously a function of season, precipitation characteristics (intensity bin, daily amount, and variability), and spatial domain. The current writing makes it easy to lose the main message. The authors are suggested to distill the key findings into high-level summary statements where possible. A couple of places that the details make it difficult to identify the central findings are Figure 4 or Section 5.2.1, Lines 503-521, and Lines 427-434.

We have modified the description of Figure 4 (Section 5.2.1 Ability to simulate precipitation events) as follows to improve clarity:

"Figure 4 shows categorical metrics assessing the ability of WRF-SAAG, CR2MET, and RF-MEP to replicate precipitation events of different magnitudes across the study macrozones. The results show that all datasets achieve high PC values (≥ 0.64) across macrozones and precipitation thresholds, although this metric is strongly influenced by the large number of correct negatives (i.e., all the products are able to correctly simulate non-precipitation events). For the remaining metrics (i.e., POD, CSI and FAR), clearer differences among products are obtained. CR2MET provides the highest POD for precipitation events >1 mm/d, with median values generally above 0.85 across the study macrozones. However, the detection ability of CR2MET and WRF-SAAG decreases as the precipitation threshold increases. For events >5, >10, and >20 mm/d, RF-MEP generally achieves the highest median POD values. This is not the case for events above 1 mm/d, which may be attributed to precipitation distribution, since the fraction of events (i.e., combinations of station locations and days) between 1–5 mm/d dominates among all events with P > 1 mm/d (not shown).

Figure 4 also shows that RF-MEP overall reaches the highest median CSI values for events >5, >10, and >20 mm/d. CR2MET yields the lowest median FAR for the 10 mm/d and 20 mm/d thresholds across the five macrozones, while RF-MEP achieved the lowest FAR at lower thresholds (>1 mm/d and >5 mm/d) in the Far North, Near North and Central Chile."

Regarding Lines 427-434, we have simplified the description of Figure 9 (Section 5.3 Hydrological model performance) as follows:

"Figure 9 compares hydrologic model calibration and evaluation metrics obtained with different forcing combinations. The results show consistently good calibration performance compared to the evaluation period across all macrozones and forcing combinations. Additionally, the results reveal comparable performance metrics for WRF-SAAG and CR2MET within Central Chile, the South and the Austral macrozones. The highest median OF values are obtained in South (0.87–0.89 in calibration, 0.67–0.73 in evaluation) and Central Chile (0.83–0.89 in calibration, 0.58–0.63 in evaluation). In these macrozones, median β and γ values also remain close to the optimum, ranging 0.98–1.02 and 0.95–0.99 (1.01–1.14 and 0.93–1.14), respectively, during the calibration (evaluation) period. Conversely, the poorest performance is obtained in the Far North and Near North, with median OF dropping to very low or even negative values during the evaluation period. In the Austral macrozone, we obtain good calibration performance (OF≈0.71–0.78) and a marked decline during the evaluation period when RF-MEP-based combinations are used (OF≈0.49–0.52, β up to 1.36–1.45)"

Regarding Lines 427-434, please see our response to comment 4.

3. Line 423, "throughout all seasons" seems inconsistent with Figure 5, which shows results for summer only. Additionally, instead of stating "reach >=0.65", it is suggested to report the specific KGE_T values for each of the three datasets, because the text later states that the second best KGE_T for CR2MET is 0.72, which is already higher than the 0.65 threshold cited here.

We have restructured the paragraph as follows (the corresponding figure is attached below):

"Figure 5 shows the spatial distribution of KGE_T (and its components) for summer (DJF) daily precipitation estimates from WRF-SAAG, CR2MET, and RF-MEP. The highest KGE_T values during summer are observed in the Southern macrozone, where median KGE_T values reach $\geq$ 0.65, 0.79, and 0.76 for WRF-SAAG, CR2MET, and RF-MEP, respectively. Median KGE_T values also remain at or above 0.65 in this macrozone throughout all seasons (see Figure 6 for JJA and Figures S.1 and S.2 in the Supporting Information for MAM and SON, respectively). The second best KGE_T for CR2MET is obtained in the Austral region (median KGE_T = 0.72), whereas the second best KGE_T for WRF-SAAG (RF-MEP) is achieved in Central Chile, with a median of 0.59 (0.57). However, there are also negative KGE_T values within this macrozone, near its northern boundary (~ 32.18°S)."

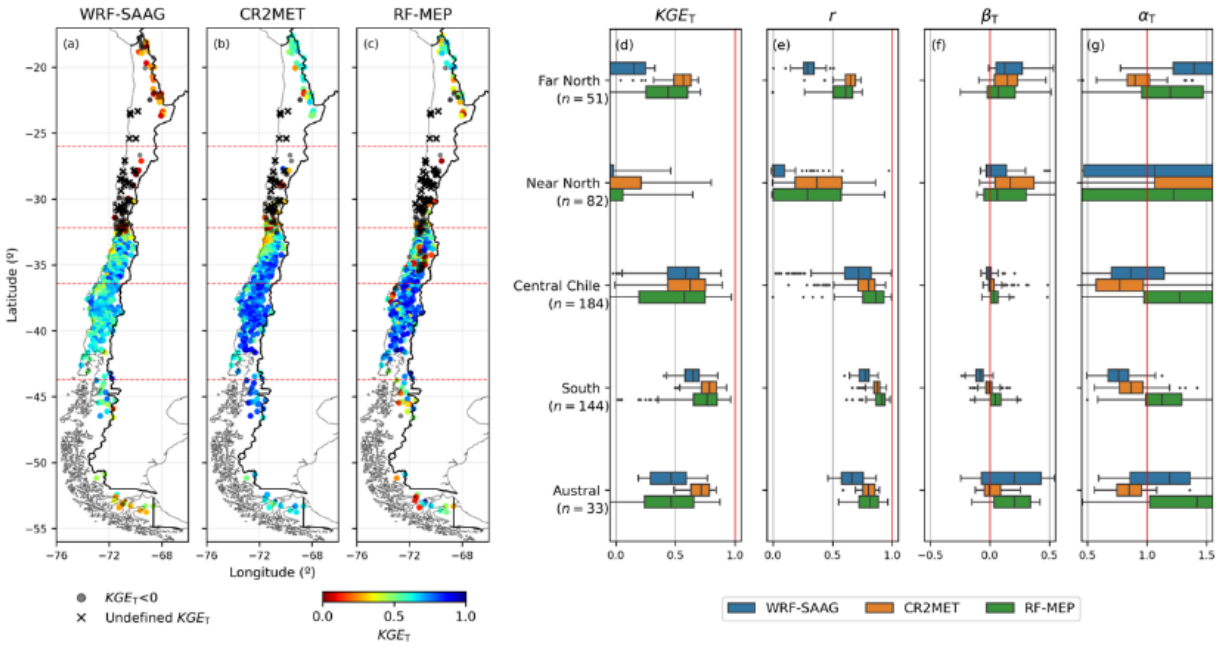


Figure 5. (a)–(c) Spatial distribution of Tang's Kling-Gupta Efficiencies (KGE_T) for summer (DJF) daily precipitation retrieved from WRG-SAAG, CR2MET, and RF-MEP (period April/2001–March/2018) across continental Chile, using ground-based observations as the reference. Red dashed lines indicate the limits of the macrozones analyzed. (d) to (g) boxplots with KGE_T, the Pearson correlation coefficients (r), bias term (β_T), and variability ratios of Tang's Kling-Gupta Efficiency (α_T) obtained with the products examined for each macrozone, with the red line indicating the optimal value. Each boxplot comprises results for all the stations within a specific macrozone. The boxes correspond to the interquartile range (IQR, i.e., 25th and 75th percentiles); the vertical line in each box is the median, and the whiskers extend to the ±1.5· IQR of the ensemble of stations.

4. Line 427-434, this paragraph is difficult to follow. It seems the authors try to make the point that datasets struggle to represent very low precipitation, but the paragraph is a mix of station counts and median KGE_T values. A clearer structure is needed.

We agree with this observation and have revised the paragraph to improve its clarity and structure. We have removed the text about precipitation events detected by the gridded products but not by the meteorological stations (which is addressed in the subsection 5.2.1 "Ability to Simulate Precipitation Events"). Additionally, we have incorporated the median KGE_T values directly into the macrozone description to better highlight dataset performance in arid regions:

"The results in Figure 5 show that the three datasets have limitations in effectively capturing daily precipitation in arid regions, characterized by very low (or zero) precipitation amounts. In the Far North (medians KGE_T of 0.15, 0.44 and 0.57 for WRF-SAAG, RF-MEP and CR2MET, respectively) and Near North (medians KGE_T of -0.18, -0.32 and –0.22 for WRF-SAAG, RF-MEP and CR2MET, respectively), 23 stations did not record any precipitation events during all studied summers, which yields undefined KGE_T at these sites. Additionally, the largest differences among datasets in valid KGE_T are observed in the Far North and Austral macrozones, with CR2MET standing out in both regions, with medians of 0.57 and 0.72, respectively."

5. Line 470 should be Fig. 7e? Line 475 should be Fig. 7f?

Thank you for catching this error; we have corrected the figure references.

6. Figure 10 caption, please capitalize "april" and "march".

Thank you for the correction. The month names in Figure 10 caption are now capitalized accordingly.

7. Figure 10. Several boxes appear to be truncated by the y-axis limits (−100 to 100, or −50 to 50 days). The authors should expand the axis ranges to display the full extent of all boxes so that no information is hidden.

We did try expanding the axis ranges to display the full extent of all boxes, as the reviewer suggests (see figure below). However, it is very difficult to visualize differences among the forcing datasets, especially in the Central, Southern and Austral macrozones. Therefore, we decided to fix the first column ranges from –100 to 100 because it represents percentage biases (%), and the second column ranges from –50 to 50 because it represents differences in days. These limits were intentionally selected to facilitate the visualization of differences in the median values of the boxplots and to enable a clearer comparison among macrozones, while maintaining fixed and consistent axis limits within each column.

Although we appreciate the reviewer's sentiment, we have decided to keep the original Figure 10. Additionally, we have added the following text to the Figure caption: "Note that the y-axis are truncated to facilitate the visual comparison among input forcing combinations."

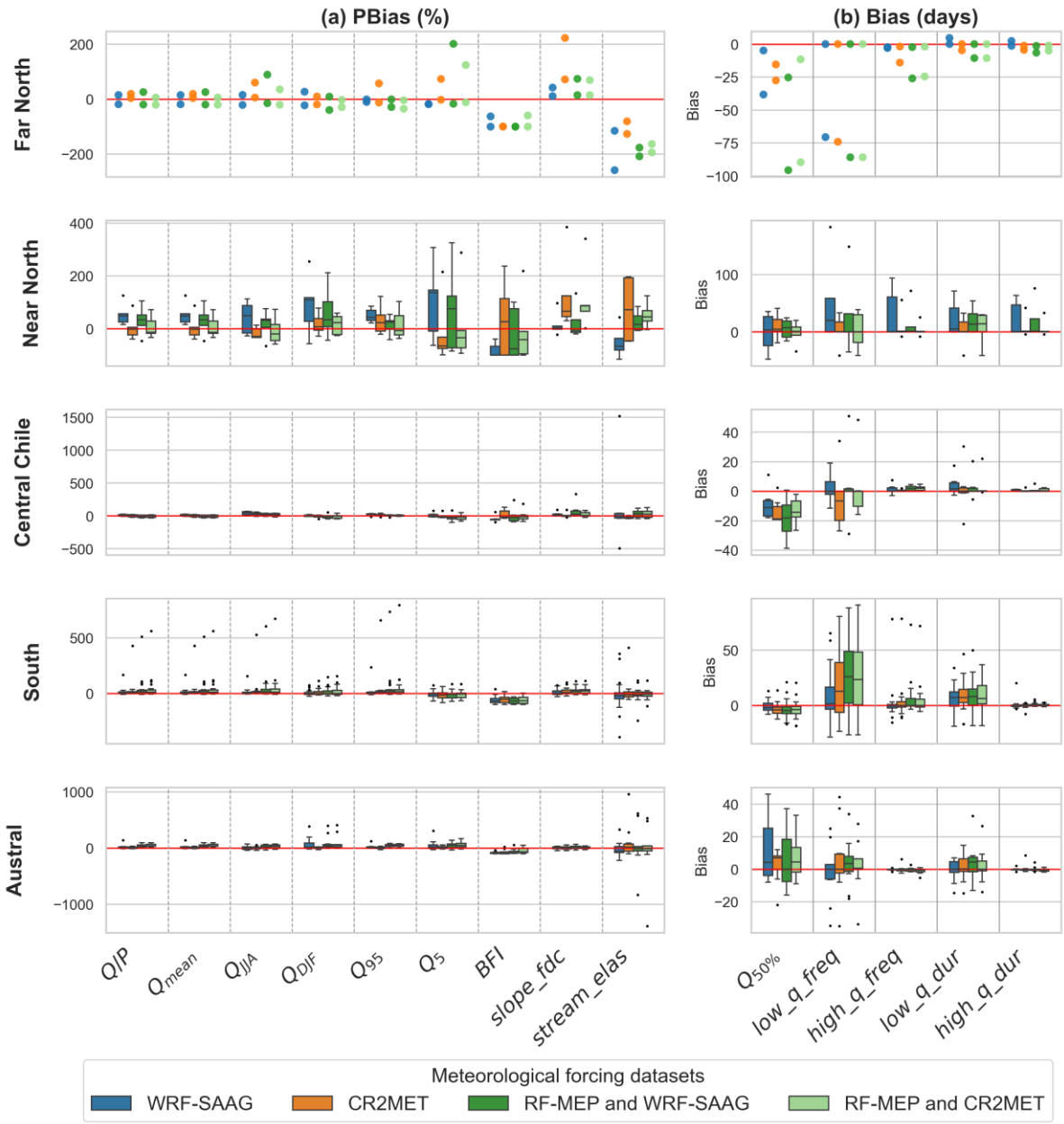


Figure 10 displaying the full extent of all boxes.

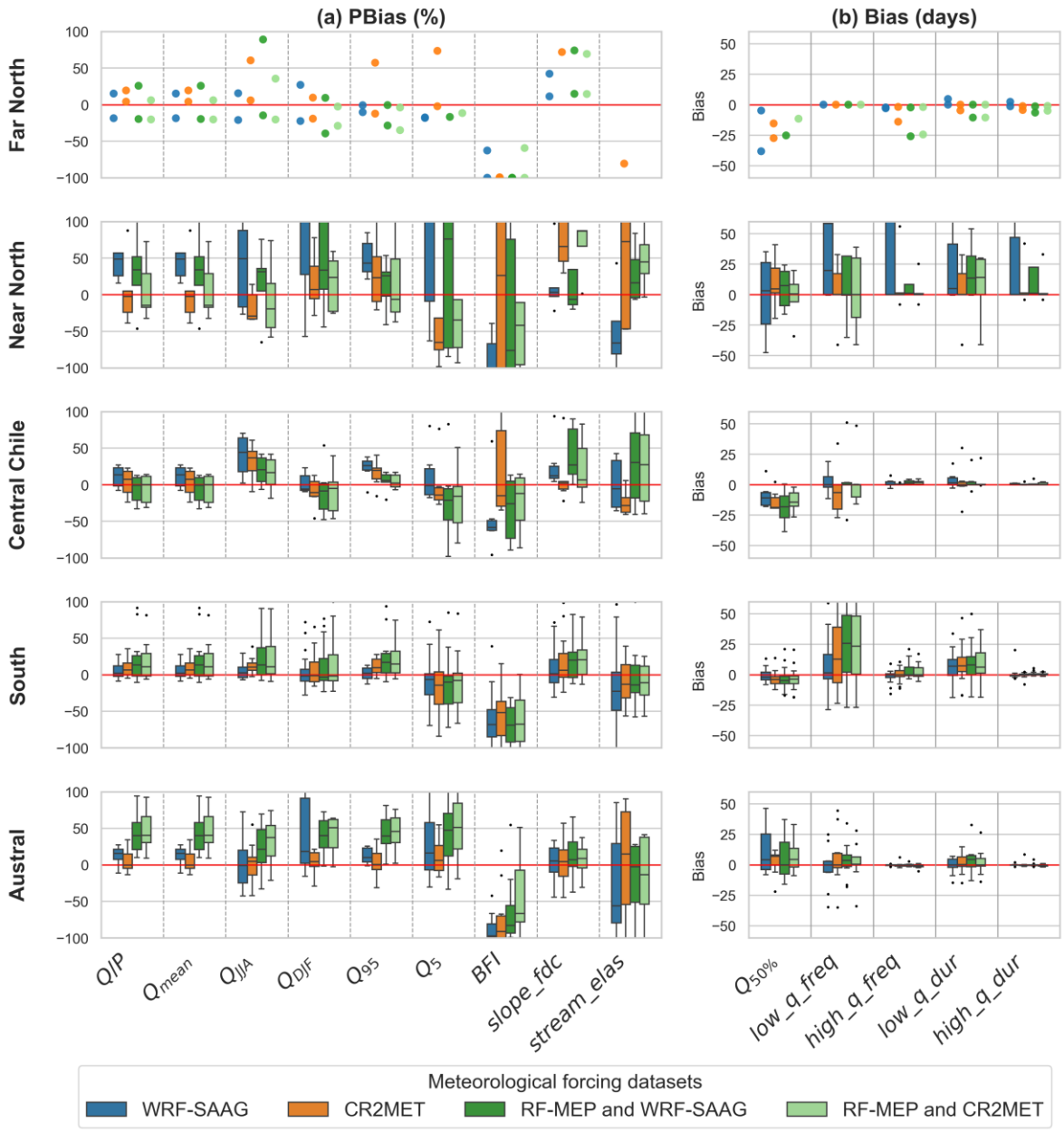


Figure 10 with the truncated y-axis, which is the version we decided to adopt in the manuscript.

8. Figure 13. Why is there no evaluation period shown, in contrast to Figures 11 and 12?

In the previous version of the manuscript, we did not include the evaluation period (defined as April 2011–March 2018) because the SWE reference dataset ends in 2015. In the revised version of the manuscript, we have included the SWE evaluation during both calibration (April/2005–March/2011) and the first four years of the evaluation period (April/2011–March/2015) for 11 case-study basins.

9. Lines 679-680, the statement that "WRF-SAAG delivered performance comparable to observation-based products at the daily time scale" does not seem to be supported by the results. Figures 4–7 consistently show that WRF-SAAG has lower KGE_T and correlation, and larger absolute bias and variability ratio deviations, relative to CR2MET. The WRF-SAAG might have comparable runoff simulation (Figure 9), but it doesn't seem to have comparable performance in daily climatology of precipitation and temperature.

We have modified the referred paragraph as follows:

"The main outcome of this work is that, despite substantial spatial differences in climatological averages and despite not being designed to replicate individual precipitation events in space and time, the performance of WRF-SAAG approaches that of CR2MET (the most used observation-based product in our domain) at the daily time scale in Central and Southern Chile - where the spatial density of stations is higher -, especially during winter.

The hydrological modeling results showed similar streamflow performance metrics and SWE model simulations for the three products, with differences in simulated hydrological signatures, ET and soil moisture that varied with macrozone. For streamflow simulations, the results suggest that in the Southern macrozone all forcing combinations provide reasonable performance. In contrast, in the Austral macrozone, configurations using RF-MEP precipitation as input tend to show lower performance compared to those forced with WRF-SAAG or CR2MET. In the Near North macrozone, the performance of WRF-SAAG-driven simulations suggests that testing alternative hydrological model structures may be necessary to better represent hydrological processes since WRF-SAAG precipitation simulations provided reasonable daily KGE_T results."

10. Lines 685-687, The logic of this sentence is unclear. The opening clause "the results do not allow us to conclude that WRF-SAAG-driven simulations are unsuitable" does not read straightforward and it reads like an endorsement, but the closing clause identifies a limitation (the need to test alternative hydrological model structures). The authors should establish a clearer logical flow.

Regarding lines 685-687, we have modified the referred sentence as follows:

"In the Near North macrozone, the performance of WRF-SAAG-driven simulations suggests that testing alternative hydrological model structures may be necessary to better represent hydrological processes since WRF-SAAG precipitation simulations provided reasonable daily KGE_T results."