Paper on methods for assessment of model: response to reviewers

Nguyen et al.

August 16, 2024

Response to Reviewer 2

We are grateful to the reviewers for their insights and believe their comments have substantially improved our manuscript. We address all their comments below, point-by-point, in blue. We trust that our changes to the manuscript will satisfy the reviewers and the Editor.

Replies to general comments

In this study, the authors proposed an approach to select suitable GCMs for dynamical downscaling. This approach includes a standardized benchmarking framework that consists of two steps. One is based on minimum performance requirements in terms of the reproducibility of simulated precipitation. The other is associated with the representation of simulated key precipitation drivers and teleconnections. The second step seems to be unique and reasonable. However, there are some concerns as mentioned comments written below. The most important one may be the method for determining threshold values of metrics to judge whether a model well reproduces precipitation itself, key precipitation drivers, and teleconnections.

Thank you for your thorough review and constructive feedback on our study. We appreciate your positive remarks regarding the uniqueness and rationale of our two-step benchmarking framework for selecting suitable GCMs for dynamical downscaling.

We fully understand your concern about the method for determining the threshold values for metrics. The benchmarking framework (BMF) was designed to identify "fit-for-purpose" models, with thresholds based on strong scientific reasoning, the specific research question, the region or sector of interest, and the general purpose of benchmarking model performance (Isphording et al. 2024). In this research, we aim to identify models that perform well in simulating precipitation over land, key precipitation drivers, and teleconnections. We utilized various metrics, considering different seasons, and the thresholds for each metric were determined based on our understanding of observational uncertainties. In addition, we also provide each model with the "benefit of the doubt," allowing us to include as many models as possible in the initial selection before further refinement.

We discussed our strategies for determining these thresholds in Section 4 (Discussion, L603-618) of the manuscript, to provide readers with a clearer understanding of our methodology. We appreciate your insights and will ensure that this section clearly communicates the rationale behind our approach.

Isphording, R. N., Alexander, L. V., Bador, M., Green, D., Evans, J. P., and Wales, S.: A Standardized Benchmarking Framework to Assess Downscaled Precipitation Simulations, Journal of Climate, 37, 1089-1110, https://doi.org/10.1175/JCLI-D-23-0317.1, 2024.

Replies to specific comments.

Major comments

1. L147: Perhaps the authors forgot to put section 2.2.1 just after this line. Putting here an explanation of fundamental metrics, such as MAPE and Scor, would be preferable.

Thank you so much for your thorough review. You are correct and we apologise for the omission. We just added section 2.2.1 Minimum standard metrics (MSMs) in the manuscript, which explains the fundamental metrics of MAPE, Scor, Scyle, and Trend, back at lines (L156-167).

"2.2.1 Minimum standard metrics

The BMF introduces a set of minimum-standard metrics (MSMs): 1. mean absolute percentage error (MAPE), 2. spatial correlation (Scor), 3. seasonal cycle (Scyc) and 4. significant changes (SigT) (Isphording et al., 2024) to assess the skill of climate models in simulating very fundamental characteristics of precipitation (e.g., magnitude of biases, spatial distributions, annual cycles and temporal variability). Before exploring complex processes, a model should meet performance expectations for these MSMs. Therefore, we initially calculate the MSMs for precipitation. In addition, we acknowledge that models should produce adequate present-day simulations of other fundamental climate variables like near-surface temperature. Hence, we also apply the MSMs for near-surface temperature in the supplementary information. Given the strong seasonality of precipitation in the region (Juneng et al., 2016), the analyses related to precipitation are conducted at a seasonal scale (e.g., the dry season November-April – NDJFMA and the wet season May-October – MJJASO). Meanwhile, temperature analyses are conducted at the annual scale."

2. L158: Maybe a good model performance based on key physical process in the historical climate does not always guarantee a good performance in terms of future climate. This is the same situation as the case of MSMs, as the authors mentioned.

Thank you for highlighting this important point. We acknowledge that a model's good performance in simulating historical climate conditions does not necessarily guarantee similar accuracy in future climate projections, a well-recognized issue in climate modelling. However, there is no evidence in the literature suggesting that models with weaker skills in simulating historical climatology perform better in future projections. On the contrary, we believe that models demonstrating good performance in both statistical and process-based metrics are more likely to provide credible future projections. This confidence is based on their proven ability to accurately simulate the historical physical mechanisms responsible for generating rainfall in the region.

We have thought carefully as to how we might accommodate your comment by adding this discussion into section 3.3 which related to future climate change signals (L579-585) to highlight our point of view.

"We acknowledge that a model's good performance in simulating historical climate conditions does not necessarily guarantee similar accuracy in future climate projections, a well-recognized issue in climate modelling (Herger et al., 2019). However, there are no arguments in the literature suggesting that models with weaker skills in simulating historical climatology perform better in future projections. On the contrary, we believe that models demonstrating good performance in both statistical and process-based metrics are more likely to provide credible future projections given their proven ability to accurately simulate the physical mechanisms responsible for generating rainfall in the region."

3. L244: Maybe relative change would not always be a good indicator. Wouldn't it be OK if the authors could also check the difference between the two (future minus historical), in particular, in a dry season?

We appreciate this point. We used relative changes since it can help facilitate a fair intercomparison of changes among models that have different precipitation climatology so that we can identify a subset of model coverage with different ranges of future change spread (low-middle to high changes). Additionally, using relative change makes it easier to compare the precipitation response per degree of global mean surface temperature warming, providing a more standardized way to assess future climate responses.

L251: Using satellite data, such as TRMM and CMORPH, enables the authors to validate simulated precipitation over ocean as well.

Thank you for your suggestions on conducting the assessment of precipitation over the ocean. We do not consider ocean precipitation over Southeast Asia for two reasons. First, there is a lack of in situ reference datasets over oceanic regions. Meanwhile, the satellite-derived products have a much shorter (e.g., most cover from 1998 forward) temporal coverage and are inhomogeneous due to different instruments used through time and potential algorithm change. Second, oceanic precipitation in satellite products exhibits significant variability with discrepancies reaching up to 4 mm/day. We have highlighted this issue by providing the additional discussion in Section 2.2.1 and Figure s1 in supplementary (L175-179).

"Note that in this research, we focus only on precipitation over land given the lack of in situ reference over the ocean. Some satellite-derived products provide oceanic precipitation data but most of their temporal coverage is not sufficiently long to use as a reference. In addition, the observational uncertainties among satellite clusters in estimating oceanic precipitations over SEA are quite substantial, with discrepancies reaching up to 4 mm/day (Figure s1)."

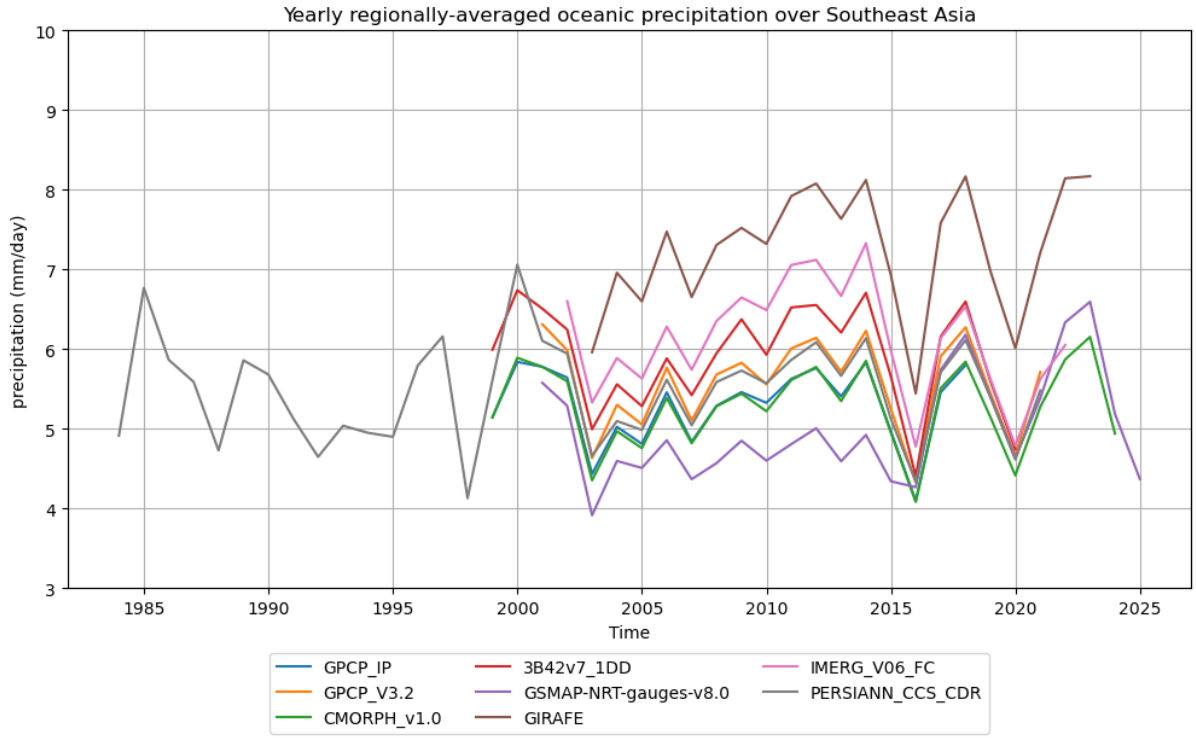

Figure s1. Time series of yearly regionally averaged oceanic precipitation (in mm/day) over Southeast Asia domain from multiple satellite precipitation products extracted from the Frequent Rainfall Observations on GridS (FROGS, Roca et al., 2019).

4. L289: How about using RMSE as a metric to validate simulated precipitation. What do the authors think about it?

To evaluate model performance in simulating precipitation intensity, Isphording et al. (2024) proposed using Mean Absolute Percentage Error (MAPE) instead of Root Mean Squared Error (RMSE). MAPE offers a metric that is more robust to large biases in small regions of the study domain. Additionally, MAPE reflects the relative error of model simulations compared to observations so that this metric ensures that contributions from locations with different climatological values are treated equally. Please reference Isphording et al. (2024) for more details.

Isphording, R. N., Alexander, L. V., Bador, M., Green, D., Evans, J. P., and Wales, S.: A Standardized Benchmarking Framework to Assess Downscaled Precipitation Simulations, Journal of Climate, 37, 1089-1110, https://doi.org/10.1175/JCLI-D-23-0317.1, 2024.

5. L313: Do the authors think that further validation is needed by using another observational product, such as CHIRPS?

Thanks for pointing out the potential value of including CHIRPS for further validation. Since observational uncertainties in estimating precipitation over SEA are large, our objective is to incorporate the observational uncertainties into the model assessment. Therefore, further validation is needed and conducted by using other observational products, including REGEN-

ALL, CHIRPS, and GPCC_FDD. The results are presented in Table s2 and figures below. In general, INM- and IPSL-family models still fail the MAPE or Scor criterion since they exhibit much higher precipitation intensity than other observational products.

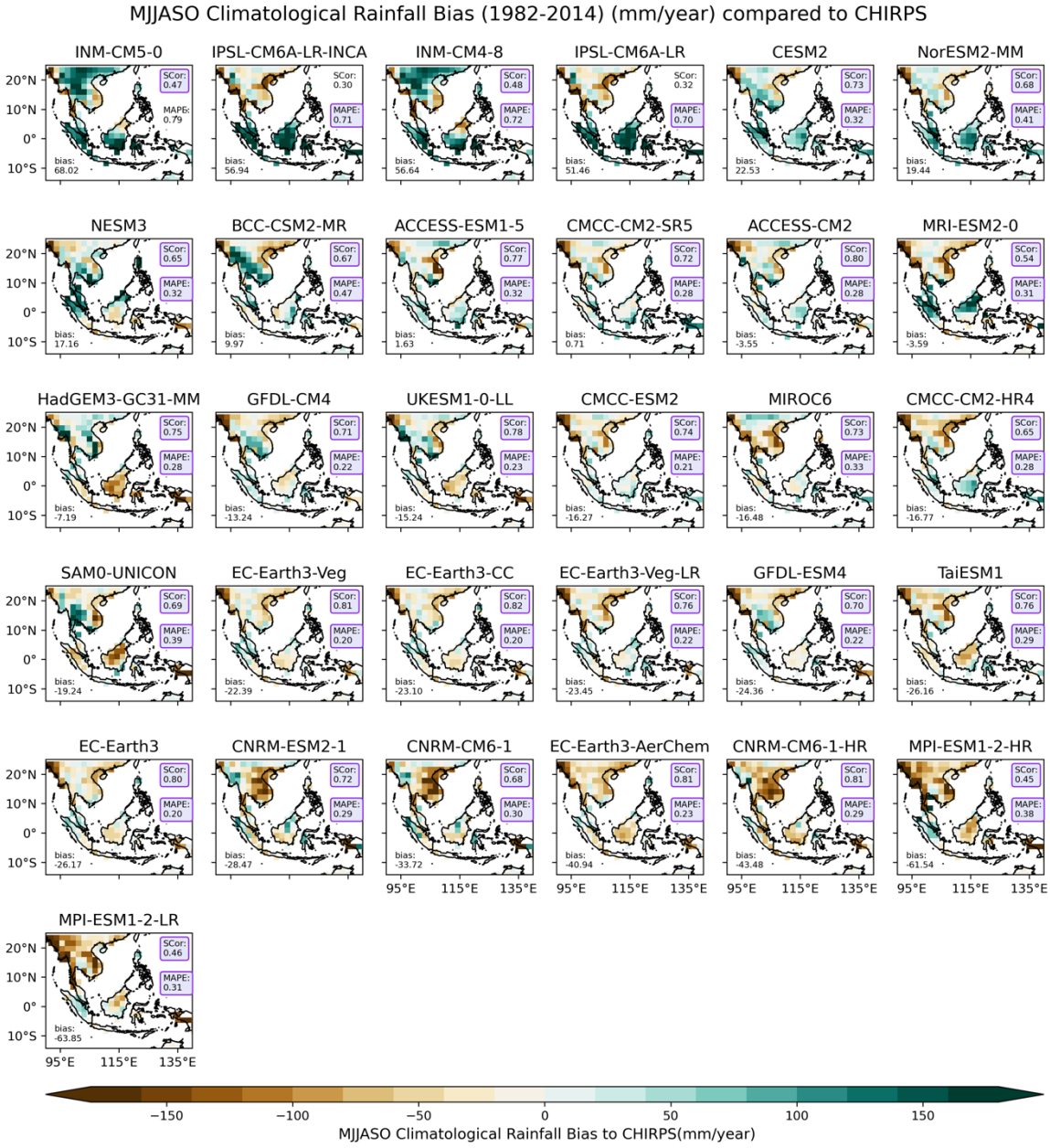

Figure 1. The seasonal climatological (1982-2014) bias (in mm/year) for each model against the CHIRPS_v2 observational product during the wet season (May-October; MJJASO), ranked wettest to driest based on regionally-averaged bias. The mean absolute percentage error (MAPE) and spatial correlation (Scor) calculated against CHIRPS are shown in the upper right corner. Values highlighted in purple-coloured boxes indicate values that meet our defined benchmarking thresholds. All analyses are considered at the resolution of the coarsest CMIP6 GCM (i.e., NESM3, ~ 216km).

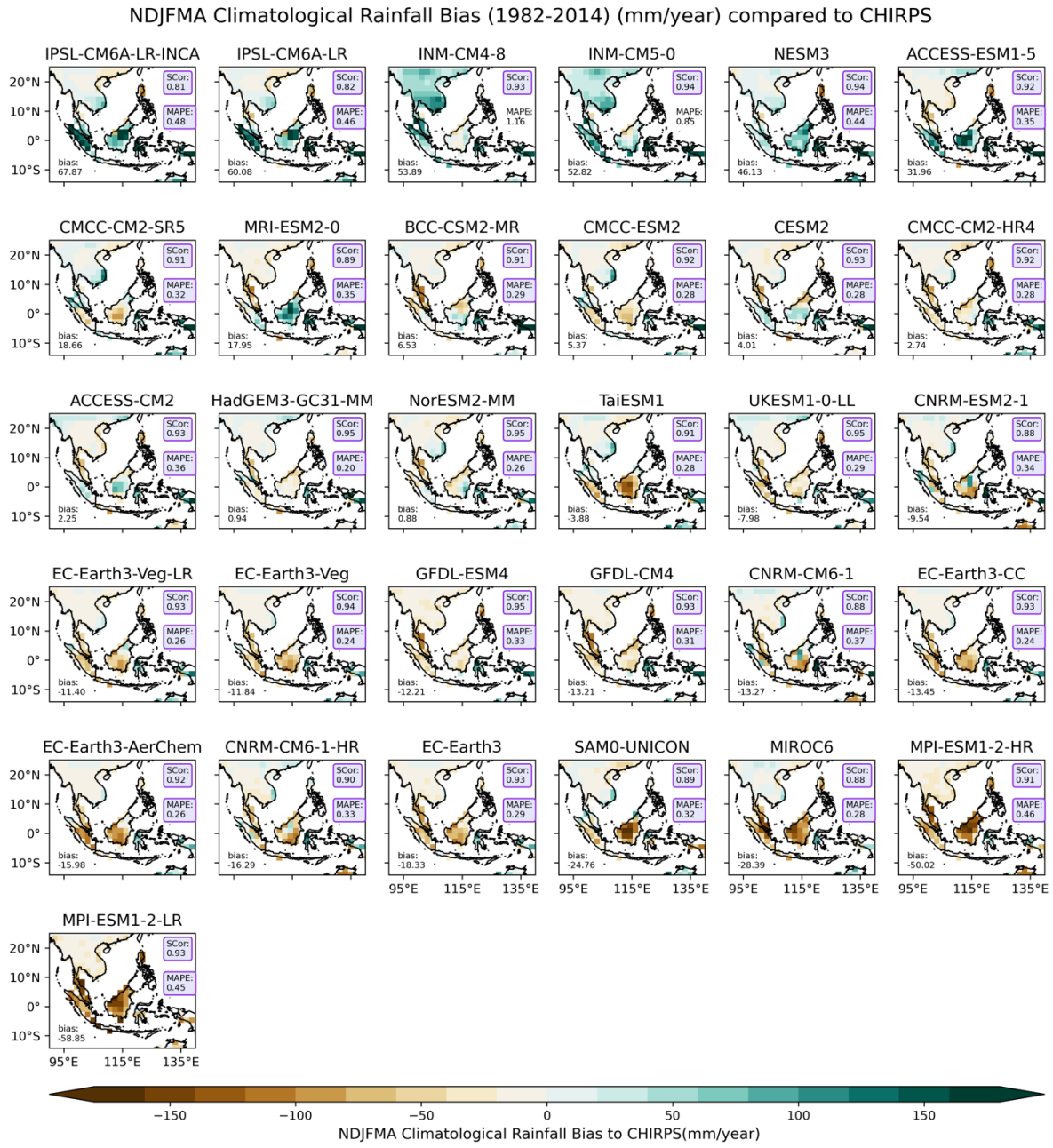

Figure 2. Same as Figure 1 above but for the NDJFMA season.

6. L315: The method for determining threshold values seems to be important, as the authors mentioned here. Wouldn't it better to determine the number of models that would be used for downscaling first, and then, to choose models in order of better performance? In this case, the authors do not need to determine threshold values.

We fully understand the reviewer's concern regarding the method to determine the thresholds. Please refer to our response to your general comment.

Minor comments

1. L131: It would be better to write the resolution of ERA5 here, which would be helpful for readers.

   Added (L139).

2. L170: There seems to be no description of the abbreviation of MAPE.

You are correct and we apologise for the omission. As mentioned above, we added the section "2.2.1 Minimum standard metrics" (L156-178) in which MAPE was described as the Mean Absolute Percentage Errors.

3. L173: Could you explain the advantage of this metric? How about a metric as follows:

Sqrt((Ui-Ui,ref)**2+(Vi-Vi,ref)**2)

Thank you for your suggestion relatated to the magnitude of the difference of the wind vectors. In our research, we also use the metric evaluates the difference in wind direction between observational data and model predictions at grid point i. Additionally, our metrics also accounts for the effects of high wind speeds, which places greater emphasis on the errors in wind direction than your suggestion.

4. L174: Typo? Should we delete "theta i theta ref"?

Revised (L199-200).

"where u_i refers to the simulated wind speed at the grid i, θ_i *and* $\theta_{i,ref}$ are the wind direction at grid i in the simulated and reference data respectively"

5. L274: The threshold values seem to be somewhat subjective. What made the authors deduce these values.

Thank you for your concerns regarding the method to identify the thresholds. Please refer to our response to your general comments on this matter.

6. L307: "Consequently" would not the right word here because the performance of biases does not always result in that of correlation.

Revised (L332).

7. L340: There seems to be a decreasing trend.

Thanks. The sentence is now corrected as follows (L367-368)

"There is a significant decreasing trend in observed total precipitation during the wet season (Figure 5 – top panel) while the dry season sees a significant increasing trend (Figure 6- the top panel)."

8. L407: Figures in bias seem to be preferrable for clear understanding of this discussion: overestimation of the wind intensity relative to ERA5.

We appreciated this comment. We agreed these figures clearly illustrate the overestimation of wind intensity related to ERA5. This can help to explain why we observed wet biases in CMIP6 models.

9. L509: The linear relationship is not necessarily needed because it is between the changes of temperature and precipitation, not between temperature and precipitation themselves.

We appreciated and agreed with your comment. The scatter plots presented in Figures 12 and 13 illustrate the relative changes in regional total precipitation (expressed as a percentage) with changes in global near-surface temperature. While a linear relationship among models is anticipated, such a trend is not observed over Southeast Asia (SEA).

10. L533, L543: The number of clusters seem to be somewhat subjective. It would be preferrable to describe what is behind these specific numbers.

Thank you. We understand your concern regarding the number of clusters. Indeed, the number of clusters is not predefined before clustering is performed as suggested by Tellaroli et al. (2016). Please refer to this reference for details.

Tellaroli, P., Bazzi, M., Donato, M., Brazzale, A. R., & Drăghici, S. (2016). Cross-clustering: A partial clustering algorithm with automatic estimation of the number of clusters. PLoS One, 11(3), e0152333. https://doi.org/10.1371/journal.pone.0152333.