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

Response to Reviewer 1

Replies to general comments

This paper describes a standardized benchmarking framework for selecting CMIP6 GCMs for CORDEX downscaling over Southeast Asia. The topic is important because Southeast Asia faces a high risk of flooding due to climate change, yet fewer models or frameworks are available for characterizing regional changes in precipitation compared to other regions such as Europe and the US. The authors did a great job highlighting the differences between their approach and those in the literature, which mainly rank GCMs according to specific evaluation matrices. The logic of this paper is very clear, and it is very well written. I only have a few minor points for the authors to consider.

Thank you for your positive feedback on the manuscript. We appreciate your recognition of the importance of improving regional climate modelling for Southeast Asia, given the area's vulnerability to flooding due to climate change. The comments have made substantial improvements to this manuscript. Please see the point-by-point responses below regarding your concerns.

Replies to specific comments

Technical corrections

1. L44: GCMs'

Thank you. Revised (L44)

2. L51, L55: should be 'WCRP'?

Revised (L51,55)

3. Section 1: an overview of the paper structure should be added to the end of this section, so the readers know what they expect in each section.

Thanks for suggesting this improvement. We have added the structure-related paragraph below (L116-120)

"The structure of the paper is as follows: Section 2 introduces the data and the benchmarking framework employed in this study. The results are presented in three subsections: Section 3.1 focuses on model assessment using the benchmarking framework; Section 3.2 examines the spread of future climate change among models;

and Section 3.3 assesses model dependence through cluster analysis. Finally, we conclude with a discussion of our results in Section 4 and a summary of the main conclusions in Section 5."

4. L120-121: do you mean 'We *do not* consider models which have a horizontal grid spacing greater than…'. Or by 'greater' do you mean finer resolution than 2 degrees?

Thanks for pointing to the need for enhanced explanation here. Yes, we are referring to models with a finer resolution than 2 degrees. We have modified the text accordingly (L125-126).

"We consider only models which have a horizontal grid spacing finer than 2°× 2° which are likely to be more suitable for dynamical downscaling."

5. L123-124: incomplete sentence.

Revised the text accordingly (L129).

"At the time of this analysis, the first member of some models (e.g., CNRM-family models, UKESM1-0-LL and HadGEM3-GC31-MM) was not available so another member was utilized."

6. L174: you may want to remove theta from the first half of the sentence and explain it as wind direction(?)

Good point. We have revised the manuscript accordingly (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".

7. L198: did you define DMI somewhere above?

Thank you for raising this point. We have now moved the paragraph that defines DMI (L209-216).

"To track ENSO variability, the Niño3.4 index (5°S-5°N and 160°E-120°W) (Trenberth and Hoar, 1997; Shukla et al., 2011) derived for the 1951-2014 period as area-mean monthly SST anomalies with respect to a 1961-1990 climatology is used. For IOD, we use the Dipole Mode Index [DMI; (Saji et al., 1999; Meyers et al., 2007)] DMI measures differences in monthly SST anomalies between the west equatorial Indian Ocean (50-70° E, 10°S-10°N) and those in the east (90-110°S, 10°S -0°N).

We use a 5-monthly average Niño3.4 and IOD index to remove seasonal cycles. The resulting month time series are detrended using a fourth-order polynomial fit to remove the possible influence of a long-term trend and to better preserve high amplitude (<10 years) variability (Braganza et al., 2003)."

8. L218: what do you mean by 'significant sign'?

We agree that this could be confusing. We are simply referring to the significant correlations that would be obtained from the application of the metrics for assessing agreement in teleconnections. Following the comment, we have amended the text accordingly (L242-243).

"For high-level qualification, we employ spatial correlation and simplified metrics to assess whether there are significant correlations in teleconnections, as recommended by Liu et al. (2024)."

9. L320-322: not sure if I follow the definition or description of the benchmarking threshold. Do you find the six wettest and driest modelled months and require the four wettest and driest months from observations to be within those six modelled months? Then how is the threshold determined?

We appreciate the spirit of this comment. We have defined the benchmarking threshold such that the four driest and wettest observed months must fall within the six driest and wettest months simulated by the models. A model meets the benchmark if the four driest observed months rank between 1 and 6 in the model's simulation. We have addressed your concern and revised the text for a clearer clarification(L357-361).

"According to the benchmarking threshold definitions, all models meet the benchmark for simulating the four wettest observed months. However, six models do not pass the benchmark for simulating the four driest observed months, as highlighted in orange in Fig. 4. Specifically, one of the four driest months according to the APHRODITE dataset (December through March) is ranked as the sixth wettest month (ranked 7th in Fig. 4) by these models."

10. L340-342: did you show the observational trend somewhere or can you cite references for this claim?

The observed trends (e.g., in APHRODITE) are presented in the top panel of Figures 5-6. We have included this information in the text for clearer illustration (L367-368).

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

Response to Reviewer 2

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)."

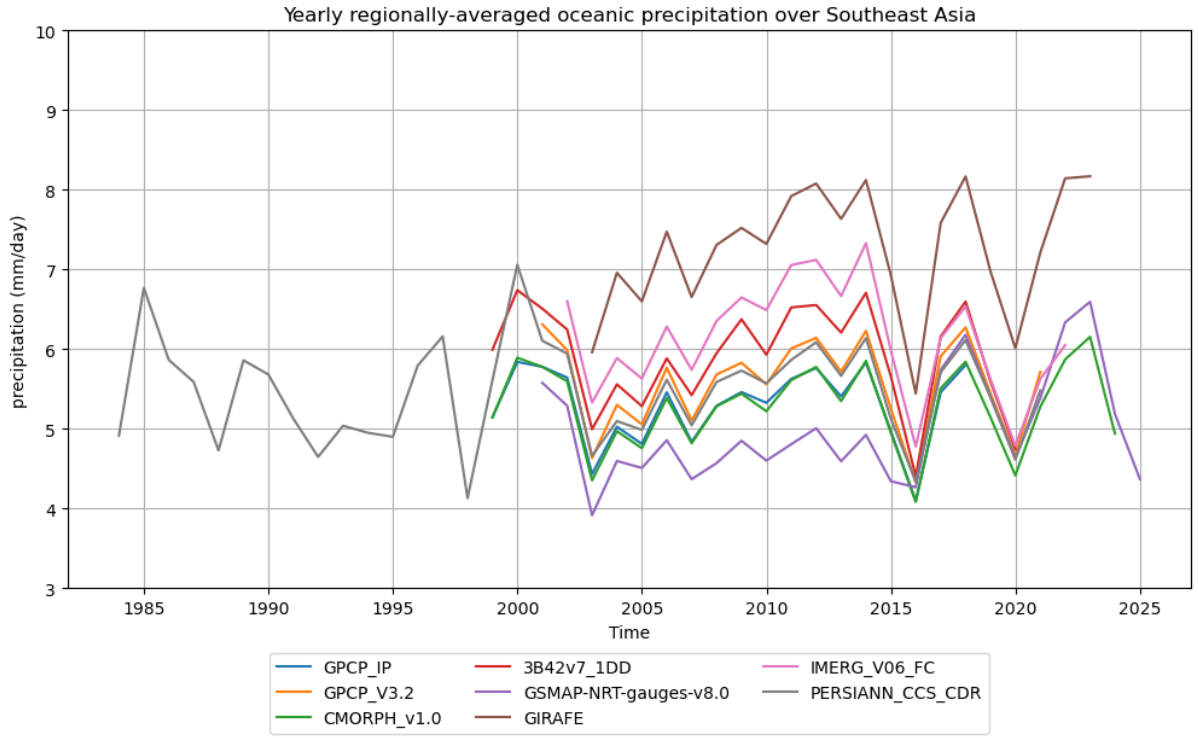

Figure s1. Time series of yearly regionally averaged oceanic precipitation (in mm/day) over Southeast Asia domain from multiple satellite precipitation products extracted from the Frequent Rainfall Observations on GridS (FROGS, Roca et al., 2019).

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.

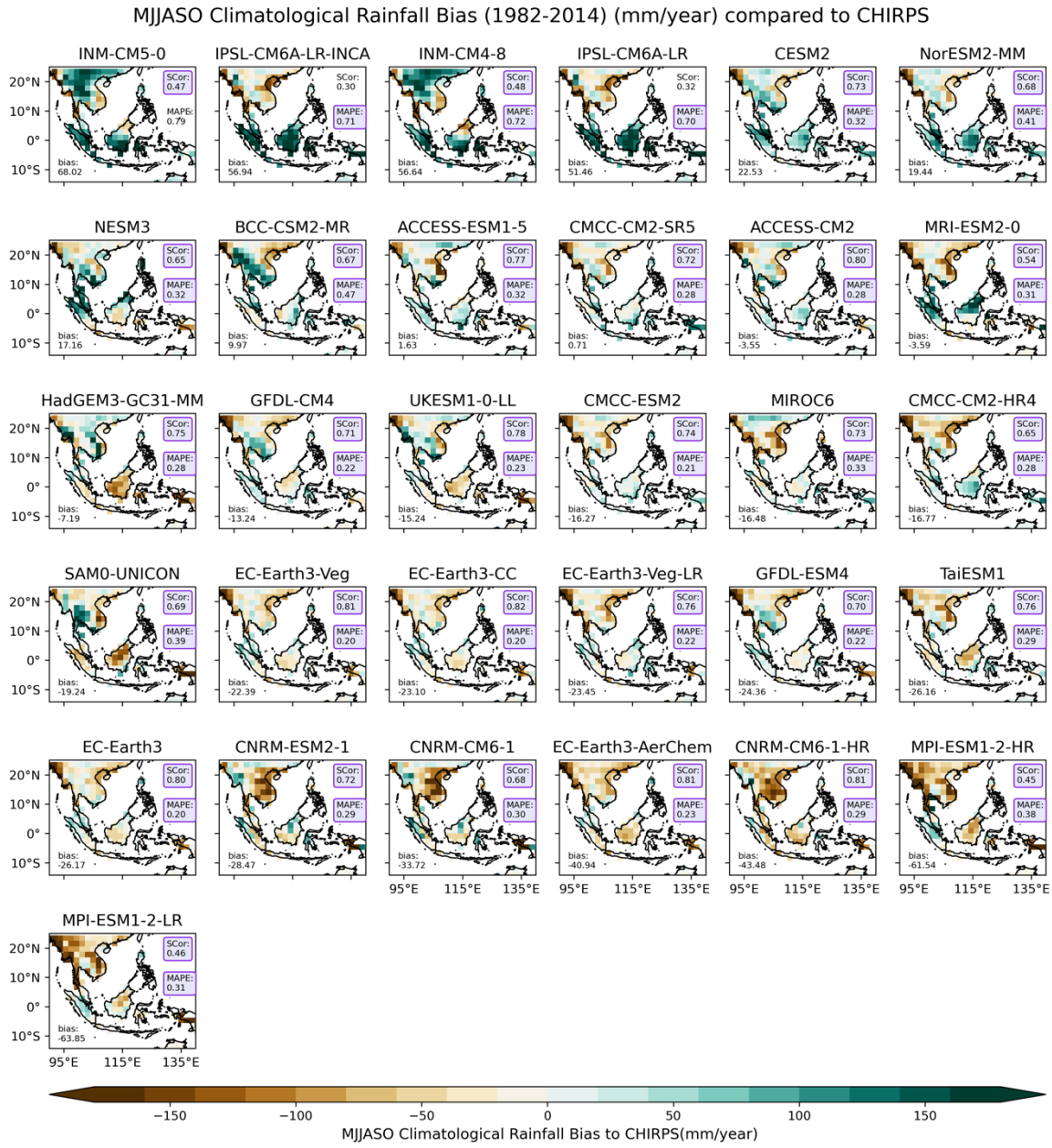

Figure 1. The seasonal climatological (1982-2014) bias (in mm/year) for each model against the CHIRPS_v2 observational product during the wet season (May-October; MJJASO), ranked wettest to driest based on regionally-averaged bias. The mean absolute percentage error (MAPE) and spatial correlation (Scor) 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).

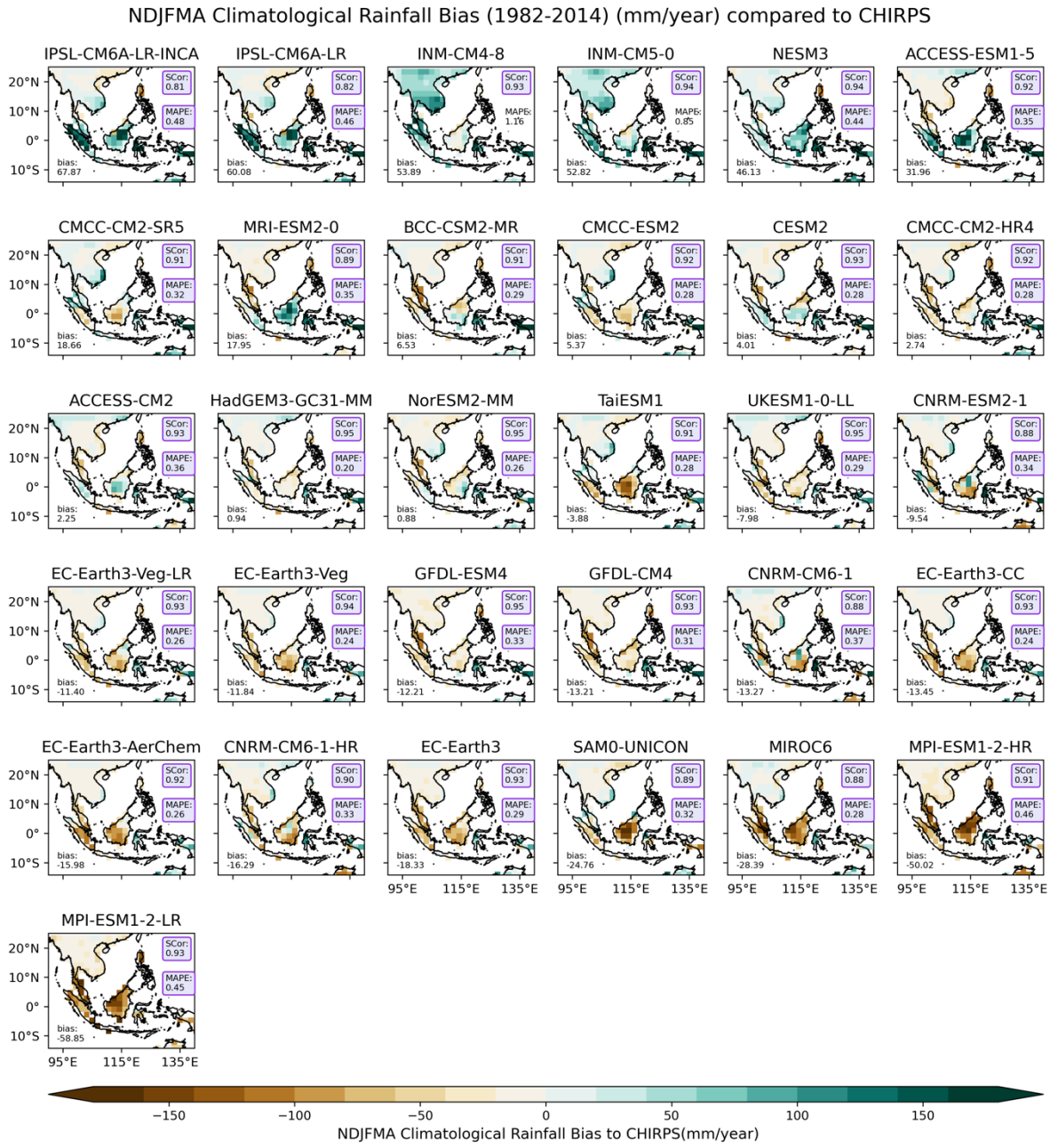

Figure 2. Same as Figure 1 above but for the NDJFMA season.

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.

Response to Reviewer 3

Replies to general comments

Review: “"Selecting CMIP6 GCMs for CORDEX Dynamical Downscaling over Southeast Asia Using a Standardised Benchmarking Framework ".

The manuscript proposes a hierarchy of statistical indices and precipitation features (drives, teleconnection pattern, and climate change signal) with the aim of selecting the most suitable CMIP6 global climate models (GCMs) to be used for dynamical downscaling in Southeast Asia (SEA). The proposed methodology evaluate simulated precipitation following the steps: a) first, GCMs ability to simulate precipitation in SEA is statistically checked considering their mean absolute relative error, spatial correlation coefficient, annual cycle and time trends; b) second, the 850 hPa winds are used to discuss the ability of GCMs in reproduce observed monsoon characteristics, while the teleconnections are evaluated by considering the time correlation with two SST indices for ENSO and IOD; 3) third, GCMs are checked considering their independence and climate change signal in future SSP-3.70 scenario. Overall, the methods used are appropriate to reach the aims of selecting GCMs, with an abstract/conclusion reflecting the main results that recommends two independent groups of GCMs to dynamical downscaling in SEA. However, I have some minor comments before the acceptance of the manuscript.

Thank you for your constructive feedback on our manuscript. We appreciate your positive assessment of our proposed methodology for selecting CMIP6 global climate models (GCMs) for dynamical downscaling in Southeast Asia (SEA). Your comments have helped to substantially improve our manuscript. Please see below for our detailed responses to each of your comments:

Replies to specific comments

Minor points

1. I would like to know from the authors if all models in Table 1 have enough data available (atmospheric variables in three dimensions at each 6 hours) for dynamical downscaling. I did not have time to do this check.

Thanks for pointing out the potential value of including data availability in Table 1. We have reviewed the tables and included the requested information by highlighting the models that offer atmospheric variables in three dimensions at 6-hour intervals. These models are now marked with an asterisk in their names (Table 1 and L130-132).

2. Section 3.3 - I do not understand the criteria of analyzing the "... GCMs that simulated at least monthly tas (near-surface air temperature) and pr (precipitation) for the SSP-3.70 scenario

only …" since simulations having only these two variables are not appropriated to dynamical downscaling. We know that atmospheric variables in three dimensions at each 6 hours are required for dynamical downscaling. In my opinion this should be the first criteria to select GCMSs, **being essential to exclude models that do not pass this criterion of the manuscript analysis**. Please, clarify.

Thank you for your thoughtful comments. We understand your concern regarding the criteria for selecting GCMs for dynamical downscaling. In our manuscript, we have used a parallel approach to assess GCMs based on different criteria, including (1)Model Performance (2)Future Climate Change and (3)Model Dependency. We recognize that for effective dynamical downscaling, models must indeed provide atmospheric variables in three dimensions at 6-hour intervals. However, to maximize the number of CMIP6 GCMs we can evaluate, we initially included models that offer at least basic precipitation and temperature data at the SSP 370 as required from the CORDEX CMIP6 experiment guide. This approach allows us to assess a broad range of models' future climate responses.

We agree that the availability of 6-hourly atmospheric variables is crucial and will clarify this in our manuscript. We will revise our criteria and discussions to better emphasize this requirement for dynamical downscaling and to highlight models that meet this criterion in our analysis (L531-536 and L588-589). Thank you again for bringing this to our attention.

"In this section, we examine the climate change signals from CMIP6 GCMs that provide at least mean temperature and precipitation data for the SSP3-7.0 scenario across two distinct seasons (see Fig. 12). Note that some models, such as CNRM-CM6-1-HR and EC-Earth3-Veg-LR (listed in Table 1), do not offer the sub-daily data (e.g., atmospheric variables in three dimensions at 6-hour intervals) required for dynamical downscaling at the time of writing. Nevertheless, we include these models in our analysis to gain insights into the future climate change responses of CMIP6 GCMs."

"Models from these two groups also offer atmospheric variables in three dimensions at 6-hour intervals required for dynamical downscaling (Table 1)."

3. L530-531 - Please, remove the affirmation "historical simulations were constrained by … observed SST)" since it is not correct because all GCMs listed in Table 1 are coupled GCMs having an oceanic component. Therefore, even in the present climate the SST is a model product without any "constraint" with observation.

Good point. We have now removed this sentence. Thanks.

4. L184-202 - The description of how ENSO/IOD indices are correlated with seasonal precipitation is hard to follow. I suggest to the authors to include a diagram to make this point clear.

Thank you for the excellent suggestions for enhancing the paper. We have acted on this and added a diagram in Supplementary Material for better clarification.

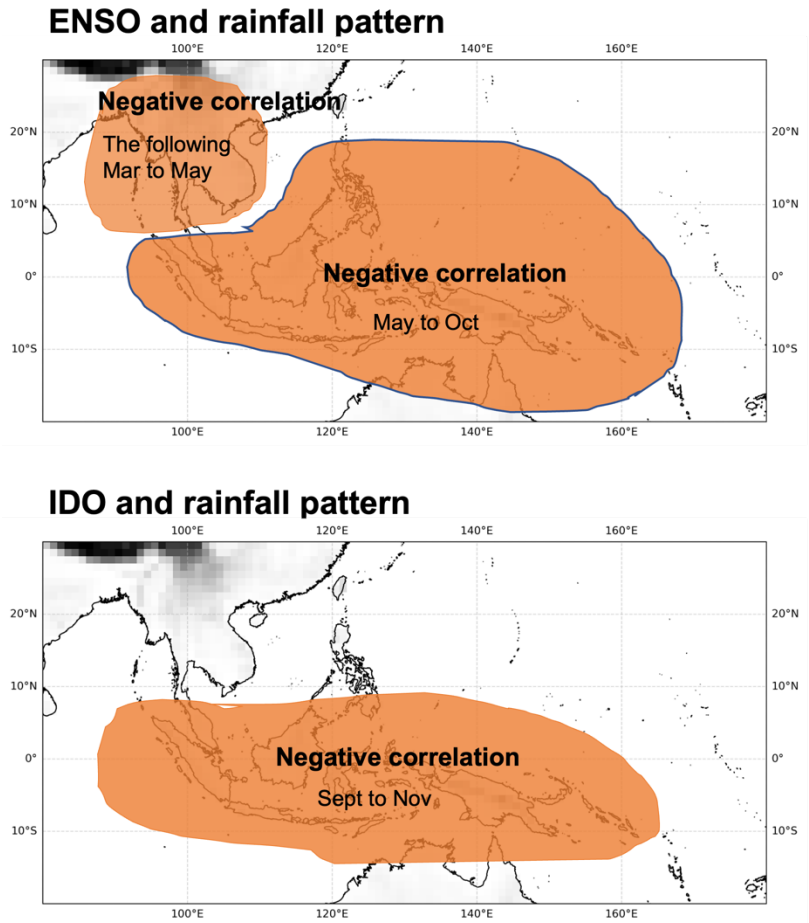

Figure s2*. The schematic shows the impact of ENSO and IOD on the rainfall pattern over Southeast Asia. The correlation coefficients are calculated between DJF Nino3.4 or SON DMI indices and each regionally averaged precipitation anomaly during the corresponding marked period.*

5. L120 - should be "grid spacing smaller than 2 x 2 “ since there is no model with grid spacing "greater than 2x 2" in Table 1.

Revised (L125-126).

“We consider only models which have a horizontal grid spacing finer than 2°× 2° to avoid the impact of the coarser GCMs on dynamical downscaling.”

6. L123 - The phrase "At the time …" is incomplete. What happens with the first member of some models?

Revised (L128-129).

“At the time of this analysis, the first member of some models (e.g., CNRM-family models, UKESM1-0-LL, and HadGEM3-GC31-MM) was not available so another member was utilized.”

7. L144-147 - Please, use these lines to define BMF and MSM. They only are defined in the legend of Figure 1.

Thanks, we added the section “2.2.1 Minimum standard metrics” to define the BMF and MSMs (L156-179).

“**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.”

8. L174 - Please, write that theta is the wind direction … and only "ui refer to simulated wind speed …"

Thanks. Revised (L199-200).

9. L183 - should be " … ENSO/IOD indices"

Revised (L207-208).

10. L198 - "DMI” is not yet defined. Move its definition in L205 to L198.

We have now moved the paragraph defining DMI and Nino3.4 index to L209-216.

11. L207 - Please, move "We use a … " as a new paragraph.

Revised.

12. L260 - Write out “MAPE" to be coherent with "Spatial correlation (Scor)".

Revised (L284).

13. L264 - Please, use the correct symbol for "greater or equal".

Revised (L285)

14. L281 - Please, to make clear what is "regionally-averaged climatologies". Is it referring to the average over all grid points over the continent inside the domain?

Thank you for your clarification. Added (L306-307).

15. L337 - should be "... the signal of statistically significant … trends using the wet (Fig. 5) and dry (Fig. 6) seasons accumulated precipitation”.

Revised (L364-365).

16. L366 - remove "of variable sign"

Revised (L392).

17. L369 - should be "temperature annual cycle"

Revised (L395-396).

18. L403 - Please, to improve the description of the winds. I am seeing "easterly-northeasterly winds in the North Hemisphere" crossing the Philippines.

Revised (L420-421).

19. L405 - should be "westerly winds predominate between …"

Revised (L422).

20. L410 - should be " … all MSM-selected models for precipitation …"

Revised (L427).

21. L417 - should be " … the extended summer season …"

Revised (L444).

22. L414-415 - change to "To benchmark CMIP6 GCMs, three metrics (HR, MR and FAR, see section 2.2.3) are calculated for each GCM considering the thresholds ≥50% for HR and ≤65% for MR and FAR, given the limited number of simulations used at this stage".

Changed (L446-448).

23. L433 - to refer to Fig. 10 " …CMIP6 GCMs (Fig. 10)"”

Changed (L469).

24. L463 - should be "... metrics stages (Fig. 11)."

Revised (L497).

25. L509 - change " … both signal and magnitude …"

Revised (L537).

26. L533 - remove the last "MJJASO”.

Revised (L560).

Figures

27. Figures 4,5 - It is hard to see the wind direction. Please, improve these Figures, maybe using a less intense shading for wind speed.

Thank you for your suggestion. We have updated the colorbar for wind speed and changed the colour of the wind vectors to white to enhance visibility and clarity (L429 and L437).

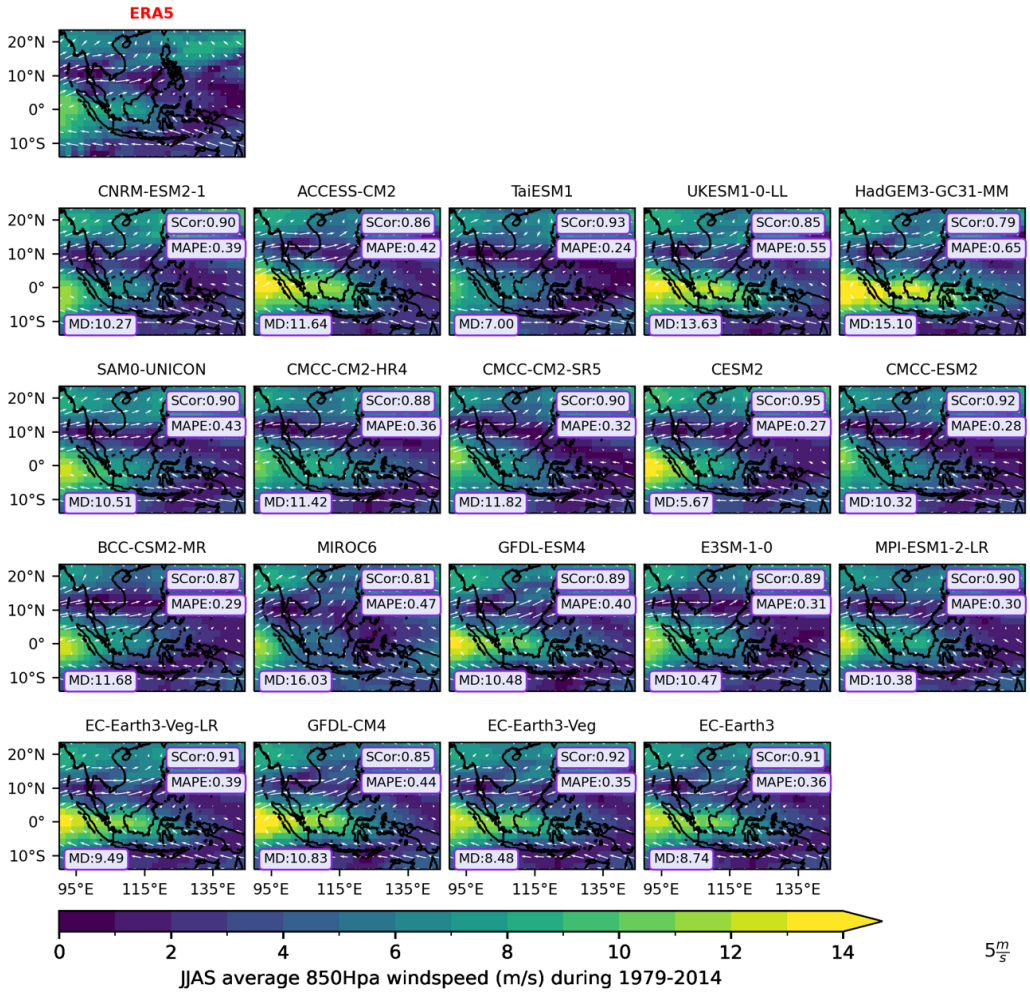

Figure 5. The spatial distribution of the climatology (1979-2014) of low-level wind circulation during the summer (JJAS).

28. Figures 2,3 - the scale in the bottom is in %, but the mean values in the boxes synthesizing MAPE in SEA are in hundredths. I would like to ask the authors to use only one unity for the same variable, for example, changing the values inside the box to %. A similar problem occurs in Figures 7 and 8, L295-296.

Thank you for your thoughtful suggestion. We understand your concern regarding the consistency of metric units. In Figure 2, the scale at the bottom is presented in mm/year, which reflects the difference between simulated and observed accumulated precipitation. The values for MAPE and Scor are expressed in hundredths, as per the design by Isphording et al. (2024).

In response to your comment, we have revised the units for the Hit Rate (HR), Miss Rate (MR), and False Alarm Rate (FAR) to also be expressed in hundredths (L246-248). We have updated the corresponding interpretation in the text to ensure consistency throughout the manuscript.

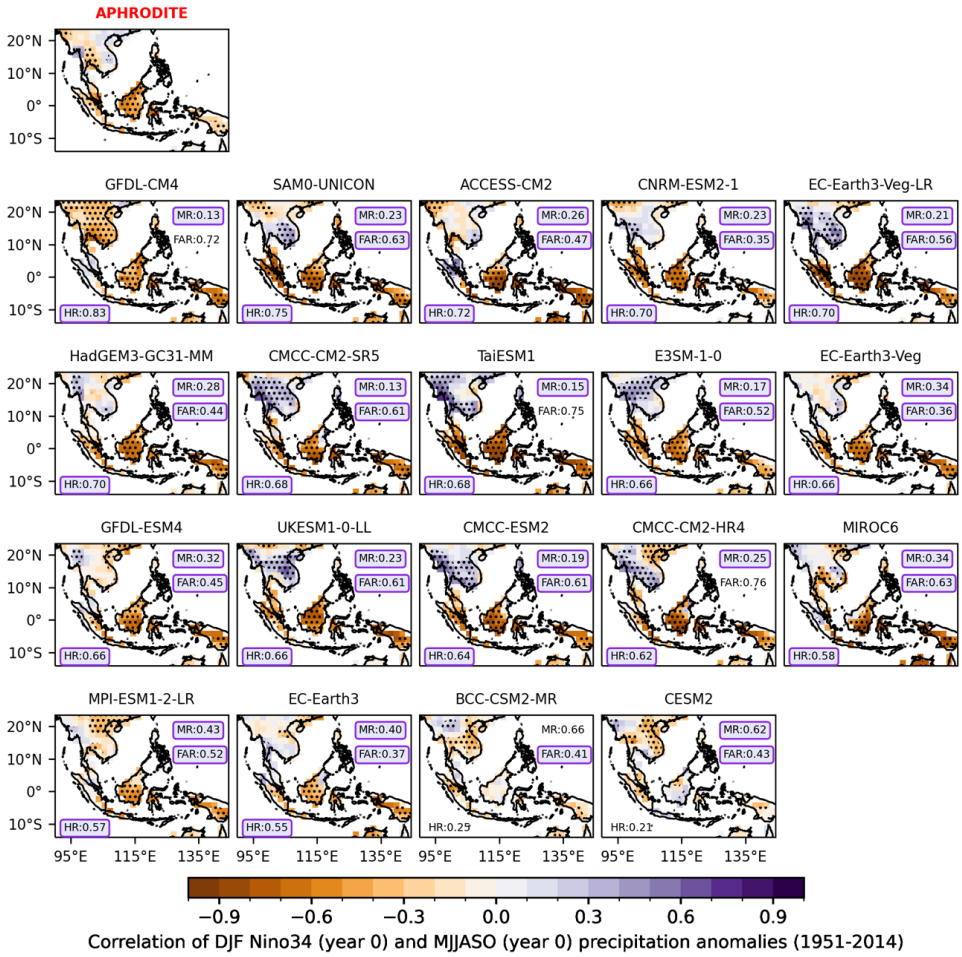

Figure 9. Lead correlation coefficients of the boreal summer (May-October, MJJASO year 0) rainfall with the mature phase of ENSO (December-January-February, DJF year 0 of Niño3.4 indices) in observation and models.

29. Figure s2 - should be "The annual climatological (1960-2014) bias of temperature …" Revised.