Dr. Jesse Norris, Handling topic editor

Geoscientific Model Development

July 03, 2026

Re: Decision on Manuscript ID egusphere-2026-1325| Model evaluation paper

Dear Dr. Jesse Norris,

We sincerely appreciate the opportunity to revise our manuscript, *"The CMIP6-downscaled CORDEX-Southeast Asia (SEA) ensemble: evaluation and benchmarking for megacities of SEA"* (egusphere-2026-1325). We thank you and the reviewers for the constructive feedback, which has helped us substantially improve the manuscript.

We have submitted a revised manuscript with tracked changes, together with detailed point-by-point responses. The major revisions include:

- Revised introduction to improve the overall narrative and to clarify the objectives of the study.
- Updated description of the CORDEX-SEA ensemble, including names and clarification of the simulations available at the time of analysis (as of September 2025) and the inclusion of newly available simulations produced as part of the ongoing modelling effort (available as of July 2026 and not included in this study).
- Explanation of the characteristics of the seasonal (precipitation and temperature) cycle over the Maritime Continent
- Discussion of a potential hypothesis on the "pair model biases" in daily maximum temperature and precipitation
- Updated and additional explanation on dendrograms from cluster analyses

We believe these revisions have strengthened the manuscript and addressed the reviewers' concerns. We hope that the revised version is now suitable for further consideration in Geoscientific Model Development.

Sincerely,

Phuong Loan Nguyen

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's comment

1. The introduction is comprehensive and well supported by relevant literature. However, the overall narrative could be clearer. The progression from CORDEX to CMIP5 and CMIP6 is somewhat non-linear. The distinction between background, previous studies, and the specific contribution of this study is not always clear. As a result, the main focus of the paper, standardized evaluation and model selection for future city-scale downscaling, is not sufficiently emphasized. A clearer structure would improve readability and highlight the key contribution.

We thank the reviewer for this insightful comment. We agree that the chronological progression and ultimate focus of the study could be framed more clearly.

In accordance with the reviewer's recommendations, we have reorganised and rewritten the Introduction section to ensure a logical, linear structure that clearly separates global background information, previous regional studies, and our novel contributions. The specific modifications are detailed below:

- Linear progression: We have restructured the narrative timeline to prevent overlapping transitions. The text now introduces the initial global CORDEX framework and then sequentially moves to the historical CMIP5-driven CORDEX-SEA experiments, contextualizing their performance limitations in the literature. After that, we introduce the transition into the next-generation CMIP6/AR6 models and the updated experimental designs.
- Explicit demand for next-generation projections with a cross-model evaluation: To bridge previous studies with our current work, we added text that explicitly highlights systematic wet biases and inter-model variability in legacy CMIP5 runs, as well as the demand for a new generation of CMIP6-downscaled CORDEX simulations.

Please refer to Section 1 (L57-104) in the revised manuscript to view these tracked changes.

"Since its launch in 2009, CORDEX has facilitated the generation of regional climate simulations across 14 continental-scale domains, moving in parallel with the development of the Coupled Model Intercomparison Project (CMIP) phases (e.g., CORDEX-CMIP5). To provide a more consistent modelling framework across regions, the CORDEX Coordinated Output for Regional Evaluation (CORDEX-CORE) further delivered a standardised set of projections for all CORDEX domains using three global climate models (GCMs) from CMIP5 (Giorgi et al., 2021).

Within this global framework, Southeast Asia (SEA), a home to more than 700 million people, is one of the key domains (Fig. 1). The Southeast Asia Regional Climate Downscaling (SEACLID)/CORDEX-SEA project, *launched in November 2013, represents the most comprehensive regional climate modelling initiative for this region, aiming to provide high-resolution climate information to support mitigation and adaptation planning. In its first phase, fourteen dynamical downscaling experiments with a grid spacing of 25 km were conducted over the SEA domain (89.5°E–146.5°E, 14.8°S–27°N, Fig. 1) through a global collaboration among seven modelling centres (Tangang et al., 2020). Subsequently, the WCRP's CORDEX-CORE initiative contributed three additional CMIP5-driven simulations for the SEA domain. Evaluation of CMIP5-driven CORDEX-SEA simulations indicated that while their multi-model ensemble (MME) generally improves representation of climatological precipitation compared to their driving GCMs, significant systematic wet biases remain, particularly within specific regional climate models (Tangang et al., 2020). However, CMIP5-forced regional climate models (RCMs) exhibit larger inter-model variability than their driving counterparts. Furthermore, Nguyen et al. (2022) demonstrated that these RCMs are not necessarily closer to the observed precipitation distribution than their driving GCMs. This highlights the ongoing challenge of regional climate modelling with a demand for a new generation of regional climate projections that provide more reliable and actionable data for climate risk assessments.*

Building on these legacy developments, CMIP Phase 6 introduced a new generation of GCMs integrated with the Shared Socioeconomic Pathways (SSPs) (O'Neill et al., 2016), with their key results presented in the Intergovernmental Panel on Climate Change (IPCC) Sixth Assessment Report (AR6) (IPCC, 2021). These advances enabled more comprehensive scenario-based climate projections. In response, the CORDEX community released an updated experimental design tailored for the dynamical downscaling of CMIP6 models (CORDEX, 2021), and modelling groups worldwide have since contributed simulations to provide well-grounded climate information for regional climate resilience and adaptation planning.

Following the CMIP6 CORDEX guidance, several modelling centres from SEA, Hong Kong, Australia, and the United Kingdom have recently initiated new downscaling experiments using CMIP6 GCMs for an expanded CORDEX-SEA domain (17.3°S–31.9°N, 86.5°E–149.3°E; Fig. 1), which now also fully spans Myanmar. This coordinated effort establishes a valuable foundation for generating updated and improved regional climate projections for impact and adaptation studies (Table 1). It is worthy to note that a fundamental requirement for the reliable use of dynamically downscaled climate projections is that a model can adequately simulate the observed climate, including the extreme events that threaten vulnerable natural and socioeconomic systems."

- Sharpened study focus: We have isolated our specific objectives into an independent, clear paragraph at the end of the section. This ensures that our primary focus (e.g., utilizing a standardized benchmarking methodology to isolate a "fit-for-purpose" model selection tailored for kilometer-scale urban climate downscaling under the CARE for SEA Megacities project) is properly emphasized as the core contribution of this paper. Please refer to Section 1 (L143-162) in the revised manuscript to view these tracked changes.

2. The description of the study objectives is somewhat unclear. In Section 2.3.1, model evaluation is referred to as the "second objective", and in Section 2.3.2, model selection is described as the "third objective". However, the first objective is not explicitly defined. It seems that the construction or documentation of the CMIP6 CORDEX-SEA ensemble (Section 2.1) may correspond to the first objective, but this is not clearly stated. It would help to clearly define all study objectives earlier in the introduction. In addition, the distinction between Sections 2.2 (data) and 2.3.1 (evaluation methodology) could be made clearer to avoid possible overlap.

Thank you so much for these helpful suggestions. In response, we have revised the closing paragraphs of Section 1 (Introduction) to explicitly define all three core objectives in the manuscript (Lines 143–162).

The revised closing paragraph now explicitly states the three objectives as follows:

"To bridge those gaps and to support the robust application of dynamically downscaled datasets in climate impact studies, this study presents the first comprehensive cross-ensemble evaluation of CMIP6 CORDEX-SEA simulations. Specifically, we assess their capacity to reproduce historical climate characteristics, thereby benchmarking the ensemble for subsequent kilometre-scale dynamical downscaling. The objectives of this research are threefold:

(1) to document the experimental design for producing the updated CMIP6 dynamical downscaled climate projection over SEA;

(2) to present the first assessment of CMIP6 CORDEX-SEA simulations, which serves as a standard reference for CMIP6-downscaled CORDEX simulations over Southeast Asia, specially by providing a quantitative evaluation of how well each GCM–RCM combination reproduces different characteristics (e.g., climatological mean, spatial patterns, seasonal cycle and the whole distribution) of maximum, mean, and minimum temperature as well as precipitation for the entire SEA region and its two sub-regions: the Mainland and the Maritime Continent (MC);

(3) to benchmark precipitation and near-surface temperature from CMIP6 CORDEX-SEA ensembles using the BMF developed by Isphording et al. (2024) for subsetting models that can be considered for further kilometre-scale downscaling over the SEA megacities under the CARE for SEA Megacities project and to also consider other factors, including model dependence and spread in future response, given the fact that a balanced selection of models across the different model "families" spanning a reasonable range of future projections is desirable (Brunner et al., 2020; Di Virgilio et al., 2022; Nguyen et al., 2024; Grose et al., 2023; Sobolowski et al., 2025)."

As correctly noted by the reviewer, the first objective corresponds to documenting the updated experimental setup and domain expansion of the new CMIP6 CORDEX-SEA ensemble. We have revised the text in the subsequent sections (Lines 168 and 241) to sharpen the distinction between the baseline data descriptions (Section 2.2) and our evaluation and benchmarking methodologies (Sections 2.3.1 and 2.3.2). These modifications eliminate any potential thematic overlap and ensure a much cleaner structural flow.

3. Around line 321, the text refers to Fig. 1 for tas biases, while the results appear to correspond to Fig. 2.

Revised (L368).

"These are the major contributions to the cold biases in tas found in Fig. 2."

4. The discussion of the seasonal-cycle benchmark over the Maritime Continent is reasonable. However, a brief explanation of why the unimodal benchmark is not applicable in this region would improve clarity. Adding a supporting reference may also help.

Thank you for this constructive suggestion. We agree that adding an explanation of why the unimodal benchmark does not apply to the Maritime Continent (MC) improves clarity. MC spans the equator; the semi-annual overhead passage of the sun naturally drives a bimodal surface temperature regime. This is characterised by two periods of rapid warming in March–

April and October–December, respectively, and a period of rapid cooling in June–August (Li et al., 2018). In addition, changes in oceanic dynamics (including ocean heat content changes), interactions with localised monsoonal wind shifts, and regional cloud-convection feedback further complicate this bimodal cycle (Chang et al., 2005; Aldrian and Susanto, 2003).

We have revised the text in Section 3.3.2 (L623-631) accordingly to incorporate this explanation and have included the foundational references (Aldrian & Susanto, 2003; Chang et al., 2005; Li et al., 2018)

"Note that the MC spans the equator, where the semi-annual overhead passage of the sun naturally drives a bimodal surface temperature regime (Li et al., 2018). This regime is characterised by two periods of rapid warming in March–April and October–December, respectively, and a period of rapid cooling in June–August (Li et al., 2018). In addition, changes in oceanic dynamics (including ocean heat content change) and the interaction with localised monsoonal wind shifts and regional cloud-convection feedbacks complicate this bimodal cycle (Chang et al., 2005; Aldrian and Susanto, 2003). Given that the original benchmark assumes a strict unimodal annual cycle, this benchmark is not strictly applicable there. Thus, all models are considered to pass for the BMF of the temperature seasonal cycle over the MC, although 11 of 21 simulations fail to capture the observed solar-irradiance peak in Oct–December (ranks 10–11–12) over the MC, instead simulating cooler conditions around ranks 5–6."

*Aldrian, E., Dümenil-Gates, L., Jacob, D. et al. Long-term simulation of Indonesian rainfall with the MPI regional model. Climate Dynamics **22**, 795–814 (2004). https://doi.org/10.1007/s00382-004-0418-9*

*Li, Y., Yang, S., Deng, Y. et al. A process-level attribution of the annual cycle of surface temperature over the Maritime Continent. Clim Dyn **51**, 2759–2772 (2018). https://doi.org/10.1007/s00382-017-4043-9*

Chang, Chih-Pei, Zhuo Wang, John McBride, and Ching-Hwang Liu. "Annual cycle of Southeast Asia—Maritime Continent rainfall and the asymmetric monsoon transition." Journal of climate 18, no. 2 (2005): 287-301.

Overall, this is a solid and valuable study with clear relevance to regional climate modelling. The methodology is sound, and the results are useful, although the presentation and narrative structure could be improved to better highlight the key contributions.

We sincerely thank the reviewer for these encouraging and positive remarks and valuable structural recommendations. We have thoroughly revised the presentation and narrative flow throughout the manuscript in light of the reviewer's comments, which we believe has significantly sharpened the clarity of the manuscript and effectively highlighted the key contributions of our study. We hope these comprehensive updates meet your approval.

Response to Reviewer 2's comments

The manuscript describes an ensemble of CMIP6 downscaling simulations and evaluates them to identify a subset that can be further downscaled to city scales for climate impact assessment. As per the GMD Review Criteria, my Scientific Significance score is Excellent (it furnishes a wealth of information on the performance of different regional climate models and their configurations); Scientific Quality is Excellent (the simulations have huge potential for further research leading to significant scientific results); Scientific Reproducibility is Excellent (data and models are sufficiently described and all links in the Code and Data Availability section work); and Presentation Quality is Good (figures are clear, manuscript is well organised, but some descriptions are a little confusing). Overall, the manuscript is perfectly suited for GMD and the authors have done a good job evaluating such a large ensemble of data and distilling actionable information from it.

My comments mainly concern the organisation of the manuscript and ensuring clarity in the communication of results, where I found some descriptions confusing. There are also a few topics I would like to see more discussion on. These issues should be straightforward to address, and I expect the manuscript can be published thereafter. I recommend major revisions rather than minors because of the number of points I've raised.

We thank the reviewer for their rigorous evaluation and constructive feedback. We are pleased that the reviewer found the manuscript's scientific significance, quality, and reproducibility to be "Excellent" and its presentation quality to be "Good".

Below, we address each of the reviewer's specific comments point by point in blue text, and corresponding modifications to the manuscript are indicated.

Major Comments

1. The objectives of the study could be made clearer which aren't stated up front but referred to throughout the manuscript. In the benchmarking section, line 256 states "the third objective of this study, which is to identify a subset of CMIP6- downscaled RCMs that meet predefined performance criteria so as to be considered for further high-resolution dynamical downscaling". On line 235: "the second objective of this study, which is to evaluate the ability of the CMIP6 CORDEX-SEA ensemble in simulating climatology of four model core variables". And later on in line 695: "the objective here is to select suitable RCMs for km-scale downscaling". Please make these objectives explicit in the introduction (e.g. the penultimate paragraph).

Thank you for this constructive recommendation, which echoes Reviewer 1's general comment regarding the explicitness of the study's objectives. As suggested, we have revised the objectives paragraph in the Introduction Section to explicitly outline all three main objectives of this study in sequential order (Lines 143–162).

The revised closing paragraph of the Introduction now explicitly states the three objectives as follows:

"To bridge those gaps and to support the robust application of dynamically downscaled datasets in climate impact studies, this study presents the first comprehensive cross-ensemble evaluation of CMIP6 CORDEX-SEA simulations. Specifically, we assess their capacity to reproduce historical climate characteristics, thereby benchmarking the ensemble for subsequent kilometre-scale dynamical downscaling. The objectives of this research are threefold:

(1) to document the experimental design for producing the updated CMIP6 dynamical downscaled climate projection over SEA;

(2) to present the first assessment of CMIP6 CORDEX-SEA simulations, which serves as a standard reference for CMIP6-downscaled CORDEX simulations over Southeast Asia, specially by providing a quantitative evaluation of how well each GCM–RCM combination reproduces different characteristics (e.g., climatological mean, spatial patterns, seasonal cycle and the whole distribution) of maximum, mean, and minimum temperature as well as precipitation for the entire SEA region and its two sub-regions: the Mainland and the Maritime Continent (MC);

(3) to benchmark precipitation and near-surface temperature from CMIP6 CORDEX-SEA ensembles using the BMF developed by Isphording et al. (2024) for subsetting models that can be considered for further kilometre-scale downscaling over the SEA megacities under the CARE for SEA Megacities project and to also consider other factors, including model dependence and spread in future response, given the fact that a balanced selection of models across the different model "families" spanning a reasonable range of future projections is desirable (Brunner et al., 2020; Di Virgilio et al., 2022; Nguyen et al., 2024; Grose et al., 2023; Sobolowski et al., 2025)."

2. I'm not familiar with dendrograms (Figure 14) and would appreciate a little more description on how to read this figure. What does it tell us? I can see that CCAM models are yellow and RegCM models are green, and they are so similar (which makes sense). However, I'm not seeing what the manuscript says, which is that there are three "clusters" in Figs 14a, c in contrast to two "branches" in Figs 14b. I see two branches in each panel (yellow and green). Moreover, I'm not sure how this figure informs us about which models to select for further dynamical downscaling. Unless its utility to this task can be clarified, I recommend dropping this figure.

We appreciate this comment and agree that using only two colours (yellow and green) for the RCM families made it confusing to isolate and interpret the distinct clusters and branches discussed in the text. To address this, we have revised Figure 14 to make the hierarchical clustering visually clear.

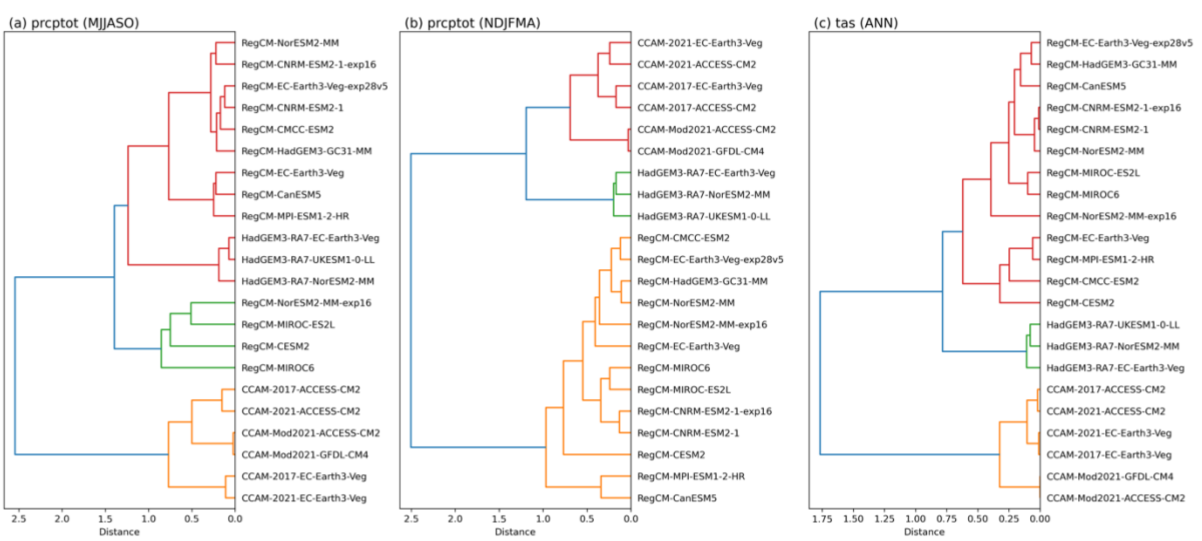

Figure 14. Dendrogram with hierarchical clustering of spatial patterns of historical seasonal total wet-day precipitation (a-b) and annual daily mean temperature (c). Colours indicate the different cluster assignments. The horizontal axis on the dendrogram is a measure of similarity between individual models and clusters, where models aligned on the same branches have similar spatial patterns.

We have also added accompanying text in the revised manuscript to better explain how to read the dendrogram (L702-716) as follows:

"In the next step, the seasonal spatial bias patterns (e.g., Figs. 2, 6, and 7) were subjected to hierarchical clustering using Pearson's correlation matrix as the distance measure between models. This aims to group similar spatial patterns of bias. Clusters shown in the dendrogram (Fig. 14) are designated by different colours. The horizontal axis on the dendrogram is a measure of similarity between individual models and clusters, where models positioned on the same branches share spatial pattern similarities.

Hierarchical clustering of historical NDJFMA total wet-day precipitation and annual mean temperature reveals three dominant spatial clusters (Figs. 14b-c). These clusters, distinguished by three distinct colours, broadly correspond to the RCM families: CCAM, HadGEM3-RA7, and RegCM (Figs. 14b, c). In contrast, the MJJASO precipitation pattern (Fig. 14a) exhibits a strong seasonal dependence, forming three distinct clusters that separate CCAM, specific RegCM simulations (e.g., RegCM4 forced by NorESM2-MM using experiment No.16, see Ngo-Duc et al., 2024; MICROC-ES2L, CESM2, and MIROC6), and HadGEM3-RA7 simulations grouped alongside the majority of RegCM. This seasonal variation in clustering behaviour aligns with the findings of Nguyen et al. (2024) and Gibson et al. (2023)."

We would like to gently emphasise that retaining the dendrogram is crucial for this study because it represents our independence analysis, where assessing for model independence is one fundamental step in selecting the subset of models for further dynamical downscaling. Evaluating model interdependence is a vital step in climate ensemble design because selecting several closely related GCMs can result in an over-representation of particular climate responses and a biased characterisation of future climate change. By identifying and avoiding highly dependent models, we ensure that the subset of models selected for subsequent km-scale dynamical downscaling is genuinely independent.

3. I think some more discussion on the potential causes of the models' inabilities is warranted. A principal finding is that downscaled simulations substantially overestimate precipitation over the maritime continent, especially during the monsoon season. I think it's worth adding some hypothesising to why the models have large wet cold biases. My first thought on the wet bias was perhaps the models are not adequately simulating the temperature gradient between the continent and the ocean. However, another principal finding is that the simulations tended to underestimate near-surface temperature (particularly over the mainland). Perhaps the models do not represent regional SST patterns and magnitudes. I think this intrigue deserves some discussion.

We thank the reviewer for this insightful comment and for highlighting this interesting "wet-cold" bias pairing. We agree that exploring the underlying physical mechanisms behind these systematic model errors adds significant depth to the manuscript and provides valuable guidance for future model development in the region.

Note that previous studies have also reported that cold biases of daily maximum temperature (*tasmax)* strongly correlate with wet precipitation biases in RCMs. One potential underlying explanation is the uncoupled configuration of the RCMs that results in the overestimation of high clouds. RCMs were also overestimating soil moisture. Another reason might be the overactive convective scheme, which triggers massive moisture dumps and leads to overestimation of daytime cloud blocks.

Following the reviewer's suggestions, we have expanded the Discussion section to include a dedicated analysis of the potential causes behind the coupled overestimation of precipitation and underestimation of near-surface temperatures. Specifically, we have incorporated the following hypotheses into the revised text (L762-785):

"Previous studies also reported persistent cold biases in CORDEX-Australasia simulations of tasmax, most notably over northern Australia during the wet season (Schroeter et al., 2024; Chapman et al., 2023; Di Virgilio et al., 2019), which strongly correlate with wet biases in precipitation.

These biases might be associated with an overestimation of high-level cloud cover inherent to the uncoupled, atmosphere-only configuration of the RCM used (Van der Linden et al., 2019). Unlike coupled systems, where intense rainfall and heavy cloud cover induce a negative feedback loop by cooling the underlying ocean mixed layer, the prescribed SST boundaries in these simulations act as an infinite thermodynamic reservoir of heat and moisture. Lacking interactive ocean-cooling feedback, the uncoupled RCMs continuously simulate excessive daytime maritime convection. This generates persistent regional cloud decks that steadily block daytime shortwave solar radiation over land, driving down tasmax. Di Virgilio et al. (2019) further suggested that excessive soil moisture might contribute to this cold bias through enhanced evaporative cooling. Another complementary hypothesis points to convective parameterisations that trigger premature, massive moisture dumps. As suggested by Howard et al. (2024), this process leads to a persistent cold bias in tasmax and a compressed diurnal temperature range due to overestimated daytime cloud block.

While these physical processes in RCMs could help to explain the wet biases, part of the apparent model bias may also reflect the uncertainty in the reference dataset used for evaluation. Notably, observational coverage over

Southeast Asia is relatively sparse, particularly over MC, and both precipitation and temperature are subject to uncertainties (Fig. S1) associated with gauge density and quality (Nguyen et al., 2020), interpolation method, and retrieval technique. For example, ERA5 as a reference displays an opposite pattern, with biases mainly arising from the underestimation of tasmin, rather than tasmax (Figs. S2 and S7). Reanalysis products assimilate a range of observations within a modelling framework and therefore contain their own structural uncertainties. Taken together, those differences highlight that part of the diagnosed bias is sensitive to the choice of reference dataset. Further investigation is therefore warranted to better disentangle model-related deficiencies from observational uncertainties and to fully explore the physical mechanisms behind the cold biases in RCMs and their potential link with wet biases in precipitation over SEA."

Di Virgilio, G., Evans, J. P., Di Luca, A., Olson, R., Argüeso, D., Kala, J., Andrys, J., Hoffmann, P., Katzfey, J. J., & Rockel, B. (2019). Evaluating reanalysis-driven CORDEX regional climate models over Australia: model performance and errors. Climate Dynamics, 53(5), 2985–3005. https://doi.org/10.1007/s00382-019-04672-w

Howard, E., Su, C.-H., Stassen, C., Naha, R., Ye, H., Pepler, A., Bell, S. S., Dowdy, A. J., Tucker, S. O., & Franklin, C. (2024). Performance and process-based evaluation of the BARPA-R Australasian regional climate model version 1. Geoscientific Model Development, 17(2), 731–757. https://doi.org/10.5194/gmd-17-731-2024

Van der Linden, R., Christensen, J. H., & Drews, M. (2019). The impact of ocean-atmosphere coupling on regional climate model simulations over Europe and the Mediterranean. Climate Dynamics, 52(9), 5615–5632. https://doi.org/10.1007/s00382-018-4467-3

Minor Comments

1. Line 28. General Circulation Model.

Thank you for this note. While "GCM" historically stands for "General Circulation Model," we have chosen to use the more modern and comprehensive term "Global Climate Model" throughout the manuscript. This choice reflects the fact that the CMIP6-generation models used to drive our downscaling simulations incorporate fully coupled Earth system components (such as biogeochemical cycles and land-surface chemistry) rather than just atmospheric and oceanic circulation.

2. Line 51. Missing word between in "accurately mesoscale"?

Revised (L51)

"GCMs often struggle to simulate sub-grid weather (e.g., local land use, complex topography, etc.) and therefore cannot accurately simulate mesoscale and local-scale processes and feedbacks (e.g., deep convection, land–atmosphere interactions, etc.) (Douville et al., 2021; Maraun and Widmann, 2018)."

3. Line 54. There is a mix of British English (e.g., modelling, standardise) and American English spellings (e.g., standardize, urbanization, parameterization). The choice is unimportant, of course, but I wanted to point it out if the authors wish to consistently adopt one over the other.

Thank you for your thoughtful comment. We agree it is best to use consistent spellings throughout. We adopted British English style (e.g., modelling, standardise, urbanisation...) throughout the manuscript.

4. Line 70. Resolution and grid spacing are not technically the same thing. Recommend replacing “at 25-km resolution” with “with a grid spacing of 25 km”.

Revised (L66)

5. Line 74–75. This sentence is nonsensical as it compares a multi-model ensemble with itself: The multi-model ensemble provides an improved representation of climatological precipitation but also exhibits larger inter-model variability compared with multi-model ensemble.

Thank you for raising this point. The sentence aims to compare a multi-model ensemble mean (MME) of CMIP5-driven CORDEX-SEA RCM ensemble with its counterpart GCMs. The sentence (L68-71) has been revised to make this point clearer as follows:

“Evaluation of CMIP5-driven CORDEX-SEA simulations indicated that while their multi-model ensemble (MME) generally improves representation of climatological precipitation compared to their driving GCMs, significant systematic wet biases remain, particularly within specific regional climate models (Tangang et al., 2020).”

6. Line 84. “the observed and present-day climate” are the same thing. Better to say “reproducing the historical climate”.

   Revised (L145)

7. Line 90. The previous sentence specifies what the authors will present. The first sense on line 90, however, I think is a general statement. When it says “Cross-ensemble evaluations of model ensembles over SEA have been conducted using various approaches” I think this is a statement about what’s in the literature, not what the authors have done. Potential confusion could be avoided here by inserting “using various approaches in the literature …”.

   We agree that the previous phrasing could cause potential confusion. The sentence (L110-111) is now revised for better clarification as follows:

   “Cross-ensemble evaluations of regional climate models over SEA have been conducted using various approaches in the literature.”

8. Line 90–91. The sentence “The most common…specific applications” seems like it’s missing a word or was maybe meant to be tied to the previous or next sentence?

   Revised (L111-112)

   “The most common method, often applied by research groups targeting specific applications, is ranking models based on their ability to reproduce key climatological variables such as near-surface temperature and precipitation.”

9. Line 95. I think this is the first use of “RegCM4” but has not been defined.

   Revised (L1115-116)

10. Line 129. I think this sentence flows better when deleting the “and” before “spanning”.

Revised (L160)

11. Line 253. Since the area score metric (ASM) is not a traditional metric and won't be familiar to readers, I think more details are warranted (e.g. how it is calculated, what a perfect score would be, and appropriate equations). Since ASM is the area between the model distribution and the observed one, is it always positive i.e. can it differentiate between model underestimation and overestimation?

    Thank you for your suggestion. The ASM calculates the absolute geometric area between the RCM simulation curve and the corresponding observational reference lines. By applying numerical integration via the trapezoidal rule, the metric quantifies cumulative deviations across the evaluated distribution (e.g., across percentiles). A lower ASM value indicates higher model fidelity, representing a tighter agreement between the downscaled output and historical observations. We have added this information in the revised main text (L292-299):

    "The RCM's ability to simulate observed climate variables at daily time scales was also assessed by comparing each quantile in the daily mean observational distribution with those of the RCMs through quantile-quantile (Q-Q) plots. Then the distributions of RCMs and observations are compared using the area score metric (ASM) (Nguyen et al., 2022). In particular, the ASM calculates the absolute geometric area between the RCM simulation curve and the corresponding observational reference lines. By applying numerical integration via the trapezoidal rule, the metric quantifies cumulative deviations across the evaluated distribution (e.g., across percentiles). A lower ASM value indicates higher model fidelity, representing a tighter agreement between the downscaled output and historical observations."

12. Line 275. Since one aim of the article is to identify models for further downscaling, I think this sentence would work better if it defines the models for inclusion, rather than exclusion.

    Thank you for your suggestion. We have revised this sentence (L318-322) as follows:

    "Given the substantial observational uncertainties associated with temperature and precipitation (Nguyen et al., 2020), assessment against a single reference dataset is not considered sufficient justification for model selection; instead, a model is included only if it meets our model performance expectations (e.g., predefined thresholds) for at least half (e.g., three out of five in this study) of the reference datasets."

13. Line 297. Should this be "It's worth noting"?

    Yes. Revised (L343)

14. Line 298. "also exhibits a cold biases".

    Revised (L344)

15. Line 299. "exhibits".

    Revised (L345)

16. Line 308. Is this a two-sided test on the observed and model mean bias errors (MBE)?

Yes, this is a two-sided T-test on the observed and modelled annual mean of daily temperature during the 1982-2014 period. This information is now added in the caption of Fig. 2 (L354).

17. Line 351. "ASM" undefined.

Revised. "ASM" for "Area Score Metric" was defined in the previous section (L294)

"Then the distributions of RCMs and observations are compared using the area score metric (ASM) (Nguyen et al., 2022)."

18. Line 379. "performance over the mainland is somewhat improved compared with that in the Mainland." Wrong word?

The sentence was removed and revised (L423-425)

"These biases are particularly pronounced in the mid-range of the distribution (50th–90th percentiles, corresponding to approximately 32–34°C), indicating that models struggle to reproduce typical daytime temperature conditions, particularly over the mainland (Fig. 5e)."

19. Line 392–394. "Despite this, several simulations (e.g., CCAM-Mod2021-ACCESS-CM2, CCAM-2017-ACCESS393 CM2, CCAM-2017-EC-Earth3-Veg, and HadGEM3-RA7-EC-Earth3-Veg) show relatively good agreement with observations 394 across the full distribution and across multiple reference datasets (Figs. S8-10)." This seems like an important point when identifying simulations for further downscaling. Maybe make sure this is highlighted in the conclusions.

Thank you so much for your suggestions. We have now highlighted this important finding in the conclusion part (L828-830) as follows:

"It is worth noting that among these simulations, a few, namely CCAM-2017-ACCESS393 CM2, CCAM-2017-EC-Earth3-Veg, and HadGEM3-RA7-EC-Earth3-Veg, exhibit distributions consistently close to multiple observed references."

20. Line 404. Should this sentence end "tas and tasmin" rather than "tas and tasmax"?

Correct, now revised (L448)

21. Line 406–408. I think it could be beneficial to mention here which models did actually do a good job at capturing the spatial temperature distribution.

Thanks for your suggestion. An additional sentence was added (L451-453).

"In summary, all simulations actually capture well the spatial distribution of temperature across SEA. However, cold biases of CMIP6-downscaled RCMs in *tas* primarily originate from deficiencies in simulating *tasmax*, which exhibits larger magnitude errors and stronger regional and seasonal variability than those in *tas* and *tasmin*."

22. Line 420. "the bias- and variance-corrected SST forcing". Does HadGEM3-RA7 also use bias- and variance-corrected SST like mentioned for CSIRO-CCAM mentioned on line 174? If so, please note it in the methods section.

Thank you for raising this point. HadGEM3-RA7 didn't use the bias- and variance-corrected SST, as mentioned for CSIRO-CCAM. For better clarification, we have revised these sentences (L469-471) as follows:

"Similarly, biases in HadGEM3-RA7 and CCAM-Mod2021 simulations remain consistent regardless of the driving GCM. For CCAM-Mod2021, this similar behaviour is likely due to the use of bias- and variance-corrected SST forcing, as previously noted by Evans et al. (2020). The underlying cause of the comparable behaviour in HadGEM3-RA7, however, remains unclear."

23. Lines 423–433. When describing model performance please cite the appropriate figure panels.

Thank you for your suggestion. The corresponding figures have been cited throughout the paragraph (L465-481).

"Despite generally reasonable spatial correlations, precipitation bias patterns are highly spatially heterogeneous, characterised by a combination of wet and dry regions. Distinct groups of RCM-type-related biases can be identified. For instance, CCAM-2021 and CCAM-2017 simulations exhibit wet biases over topographically complex regions, particularly the MC, where numerous islands of varying sizes are present. Similarly, biases in HadGEM3-RA7 and CCAM-Mod2021 simulations remain consistent regardless of the driving GCM. For CCAM-Mod2021, this behaviour is likely due to the use of bias- and variance-corrected SST forcing, as previously noted by Evans et al. (2020). The underlying cause of the comparable behaviour in HadGEM3-RA7, however, remains unclear. In contrast, RegCM's simulations show pronounced wet biases over high wind-speed regions such as Thailand, along with dry biases over Borneo. Comparison of RegCM5 and RegCM4 simulations forced by EC-Earth3-Veg (Figs. 6d and 6g, respectively) shows an improvement in the presentation of precipitation with substantial reductions of wet biases over Thailand. Interestingly, although RegCM-MIROC6 (Fig. 6m) and RegCM-CanESM5 (Fig. 6n) exhibit similar domain-mean biases (approximately 83 mm yr^{-1}, Fig. 6), their spatial bias structures differ markedly. RegCM-CanESM5 shows a predominantly wet bias across most of the domain, except over the north-western part of SEA, whereas RegCM-MIROC6 exhibits dry biases over much of the domain, with wet biases confined primarily to north-western Myanmar and southern Thailand. Using the same GCM nudging technique with different convective parameterisations, CCAM-2017 simulations (Fig. 6c, k) show significant improvement compared with CCAM-2021 (Fig. 6p, u). These findings highlight the critical role of RCM setup, particularly the convective parameterisation in shaping precipitation biases over the SEA domain."

24. Line 427–428. Why is the GCM nudging technique mentioned when it's the same for both models and CCAM-2017 improved over CCAM-2021? Wouldn't it be better to cite differences in the model configurations?

Thank you for raising this point. CCAM-2017 and CCAM-2021 used the same technical nudging, but they used different convective schemes (Table 1). This difference has now been noted in the revised main text (L478-480) as follows:

"Using the same GCM nudging technique with different convective parameterisations, CCAM-2017 simulations (Fig. 6c, k) show significant improvement compared with CCAM-2021 (Fig. 6p, u)."

25. Line 430. "The pattern of the NDJFMA reveals…" Missing word?

    Revised (L482)

"The rainfall pattern of the NDJFMA reveals some interesting features (Fig. 7)."

26. Line 430–431. "RCMs show better skill in capturing the spatial distribution of precipitation, with dry biases over the mainland and wet biases over MC." The first time I read this sentence it sounded perfectly straightforward, but on a second reading I'm a little confused. What is being compared to when you say "better"? It doesn't seem to follow that the models capture the spatial distribution of precipitation better than they do that of temperature because Scors are substantially higher in Figs 3 and 4's annotations than in Figs 6 and 7's. Please clarify.

Thank you for raising this point. We agree that this sentence could be confusing. In this context, we are comparing the model performance in simulating precipitation between two considered seasons: NDJFMA and MJJASO. The text (L482-487) was revised for better clarification as follows:

"The rainfall pattern of the NDJFMA reveals some interesting features (Fig. 7). RCMs show better skill (e.g., higher spatial correlation with APHRODITE) in capturing the spatial distribution of precipitation than that during the MJJASO season, with most models exhibiting dry biases over the mainland and wet biases over MC. Interestingly, the higher biases compared with those during the MJJASO season are more prevalent in MC in all simulations. Specifically, the RegCM simulations cannot capture the spatial distribution of precipitation well over MC, with all models showing a low correlation coefficient of less than 0.2 (figure not shown)."

27. Lines 446–448. "The RCMs generally reproduce the peak rainfall over the Mainland; however, most RegCM simulations fail to adequately capture the precipitation peaks over the MC." Is this true? Looking at Fig S12, there is indeed a large wet bias like over the Mainland, but most models seem to dip during June–October. Also, rather than "peaks" isn't there only a single peak over the maritime continent (Nov–April) that only appears as two because of how it's plotted (with Jan and Dec on opposite ends)?

Thank you for your thoughtful comments. We agree that this could be confusing. In this instance, we would like to mention the timing of the rainfall peak. RCMs generally capture the peak time of rainfall over the Mainland well, but not over the Maritime Continent (MC). In particular, the MC exhibits multiple distinct climate regimes, including monsoonal regions (e.g., Java) with a dominant rainfall peak during the austral summer (December–February); semi-monsoonal regions (e.g., parts of Borneo and Sumatra) with equatorial bimodal peaks during the transitional seasons (March–May and September–November); and anti-monsoonal regions (e.g., eastern part of the MC) characterized by an opposite seasonal phase that peaks during the boreal summer (June–August) due to their local topography, specific mountain ranges, and ocean-atmosphere interactions (Tangang et al., 2020; Aldrian and Susanto, 2003). These regimes are further modulated by synoptic processes such as cold surges and the Borneo vortex over Peninsular Malaysia and surrounding areas, leading to substantial spatial variability in rainfall seasonality. As a result, rainfall peaks occur at different times across sub-regions, and no single unimodal seasonal-cycle representation is appropriate for the entire MC. We have added this discussion to the main text accordingly (L499-515):

"The seasonal cycle of monthly mean total rainfall over two sub-regions of SEA is presented in Fig. S12, confirming the substantial overestimation of precipitation intensity in both regions as discussed previously. In terms of timing, observed rainfall peaks during May–October over the Mainland, while rainfall minima is observed during the boreal winter (December–February). RCMs generally capture the peak timing of rainfall over the Mainland. However, most RegCM simulations fail to adequately capture the precipitation peak over the MC. It is important to note that different sub-regions of the Maritime Continent (MC) exhibit multiple distinct climate regimes, including monsoonal regions (e.g., Java) with a dominant rainfall peak during the austral summer (December–February); semi-monsoonal regions (e.g., parts of Borneo and Sumatra) with equatorial bimodal peaks during the transitional seasons (March–May and September–November); and anti-monsoonal regions (e.g., eastern part of the Maritime Continent) characterised by an opposite seasonal phase that peaks during the boreal summer (June–August) due to their local topography, specific mountain ranges, and ocean-atmosphere interactions (Tangang et al., 2020; Aldrian and Susanto, 2003). These regimes are further modulated by synoptic processes such as cold surges and the Borneo vortex over Peninsular Malaysia and surrounding areas, leading to substantial spatial variability in rainfall seasonality. As a result, rainfall peaks occur at different times across sub-regions, and no single unimodal seasonal-cycle representation is appropriate for the entire Maritime Continent. This deficiency highlights the model's inability to correctly simulate the complex circulations and precipitation dynamics over the MC, thereby warranting further sub-regional analyses of atmospheric circulation patterns across localised zones within the MC."

28. Lines 478–480. Initialisms help us say things more efficiently, but there are nine in this sentence. I think it's easier to understand when writing out MSM and BMF.

Thanks for your suggestion. This sentence has been modified to be more efficient (L545-548).

"In this section, we move to the next step of subsetting the 21 RCM simulations from the CMIP6 CORDEX-SEA ensemble for further dynamical downscaling at the kilometre-scale over SEA megacities, under the CARE for SEA megacities project using the minimum standard metrics (MSMs) from the benchmarking framework (BMF) (Isphording et al., 2024) as mentioned in Sect. 2.2."

29. Line 492. Annual average of daily mean near-surface temperature?

Added (L561)

30. Line 497. "testing" or further dynamical downscaling for CARE for SEA cities?

The sentence (L564-566) has been revised for better clarification as follows:

"At this MSM stage, we would like to retain as many models as possible to maintain a sufficiently large ensemble for the next steps (e.g., future response spread and model dependence) during the RCM selection for the CARE for SEA cities project."

31. Line 500. This sentence is a little confusing. What is the "strategy" mentioned here to identify thresholds? They simply seem chosen. I think the manuscript can be clearer about why these thresholds were chosen.

Thank you for your point. In this research, we applied similar strategies from Nguyen et al. (2024) in selecting subset models for further dynamical downscaling. In particular, at the first step of the selection process (e.g., minimum standard metric – MSM), we would like to retain as many models as possible to maintain a sufficiently large ensemble for further steps (e.g., future response spread, model dependence). Therefore, relaxed thresholds were purposefully applied to traditionally challenging metrics, such as wet-season performance. These specific thresholds were informed by a combination of observational uncertainty and expert scientific judgement. Rather than applying statistical constraints, models were given the benefit of the doubt to maximise model retention and maintain a reasonably large sample size, thereby preserving a wide range of plausible climate futures. The text (L564-577) has been modified accordingly:

"At this MSM stage, we would like to retain as many models as possible to maintain a sufficiently large ensemble for the next steps (e.g., future response spread and model dependence) in the RCM selection process. Therefore, relaxed thresholds were purposefully applied to traditionally challenging metrics. To account for seasonal variations in model performance, different fixed thresholds were prescribed for wet and dry seasons. In particular, thresholds of 0.85 (MAPE $\leq$ 0.85) and 0.75 (MAPE $\leq$ 0.75) were applied to the precipitation benchmark for the MJJASO and NDJFMA seasons, respectively, while a threshold of 2.5 (MAE $\leq$ 2.5) was used for the temperature benchmark. These specific thresholds were informed by a combination of observational uncertainty and expert scientific judgement. Rather than applying statistical constraints, models were given the benefit of the doubt to maximise model retention and preserve a wide range of plausible climate futures. These are similar to the selection strategies from Nguyen et al. (2024), who balanced model performance and ensemble representativeness in selecting CMIP6 models for dynamical downscaling purposes over SEA. It is worth noting that more relaxed thresholds were applied here due to the smaller number of RCMs considered compared with the number of GCMs. Similarly, for the spatial correlation metric (Scor), thresholds of 0.3 and 0.5 were applied for precipitation during the MJJASO and NDJFMA seasons, respectively, while 0.8 was applied for annual mean temperature."

32. Line 503. "for [the] spatial correlation metric"?

Revised

33. Line 506. Floating &?

Removed (L579)

34. Line 509–510. In what sense are gauge stations drier? Relative to what? Is this sentence trying to say that gauge stations in regional datasets tend to be from climatologically dry regions? Please clarify.

We thank the reviewer for pointing out this lack of clarity. The sentence was not intended to suggest that gauge stations are located in climatologically dry regions. Rather, we meant that the gauge-based, high-resolution regional gridded dataset (i.e., APHRODITE) exhibits lower precipitation intensity (being systematically "drier", both when regionally averaged and at a single grid-cell level) relative to the global gridded datasets used in this study (e.g., REGEN_ALL, CHIRPSv2, and GPCC_FDD). This systematic underestimation in gauge-based gridded datasets across Southeast Asia is well known and primarily driven by two factors: topographic wind-undercatch (particularly across complex mountainous terrains) and station sparsity and spatial interpolation (the smoothing effects) (Yatagai et al., 2012, and Nguyen et al., 2020). We have revised this section (L582-586) to clearly define this relative comparison and avoid any misinterpretation regarding the physical climatology of the station locations:

"It is worth noting that although regional datasets incorporate a greater number of gauge stations than global products, they tend to be drier in both mean and extreme precipitation (Alexander et al., 2025; Nguyen et al., 2020; Yatagai et al., 2012) compared with the global precipitation dataset due to topographic wind-undercatch (particularly across complex mountainous terrains) and station sparsity and the smoothing effects of spatial interpolation (Yatagai et al., 2012 and Nguyen et al., 2020).'

35. Line 511–512. Delete the period. I'm confused by the word of "exceeding thresholds for at least three reference datasets. – regardless of the reference used". (presumably the period also wasn't meant to be there).

Revised (L586)

36. Line 519–520. This sentence is a bit confusing to me. I can see that RegCM-CESM2 and RegCM-MIROC6 are below the critical thresholds across almost all reference datasets. But Fig. 10 condenses the spatial correlation into a single number. So when the manuscript says the models fail to capture the "spatial contrast", I feel like that references Figures 6 and 7, and I'm not sure what's being highlighted. Please clarify.

Thank you for this point. Yes, RegCM-CESM2 and RegCM-MIROC6 are below the critical thresholds across almost all reference datasets. Therefore, they fail this BMF. In this sentence, we would like to mention the spatial distribution of precipitation. The sentence is now revised accordingly (L595-596):

"Only two models, RegCM-CESM2 and RegCM-MIROC6, fail to meet our expectation in capturing the spatial distribution of precipitation in either or both seasons, regardless of the reference dataset used."

37. Line 536. Orange cells are very hard to see and too similar to the other colors in the colorbar. Recommend boxing out simulations that don't meet the benchmarking expectations, as in Figure 10.

Thank you for your suggestion. Figure 11, along with its caption, has been updated (L604-607)

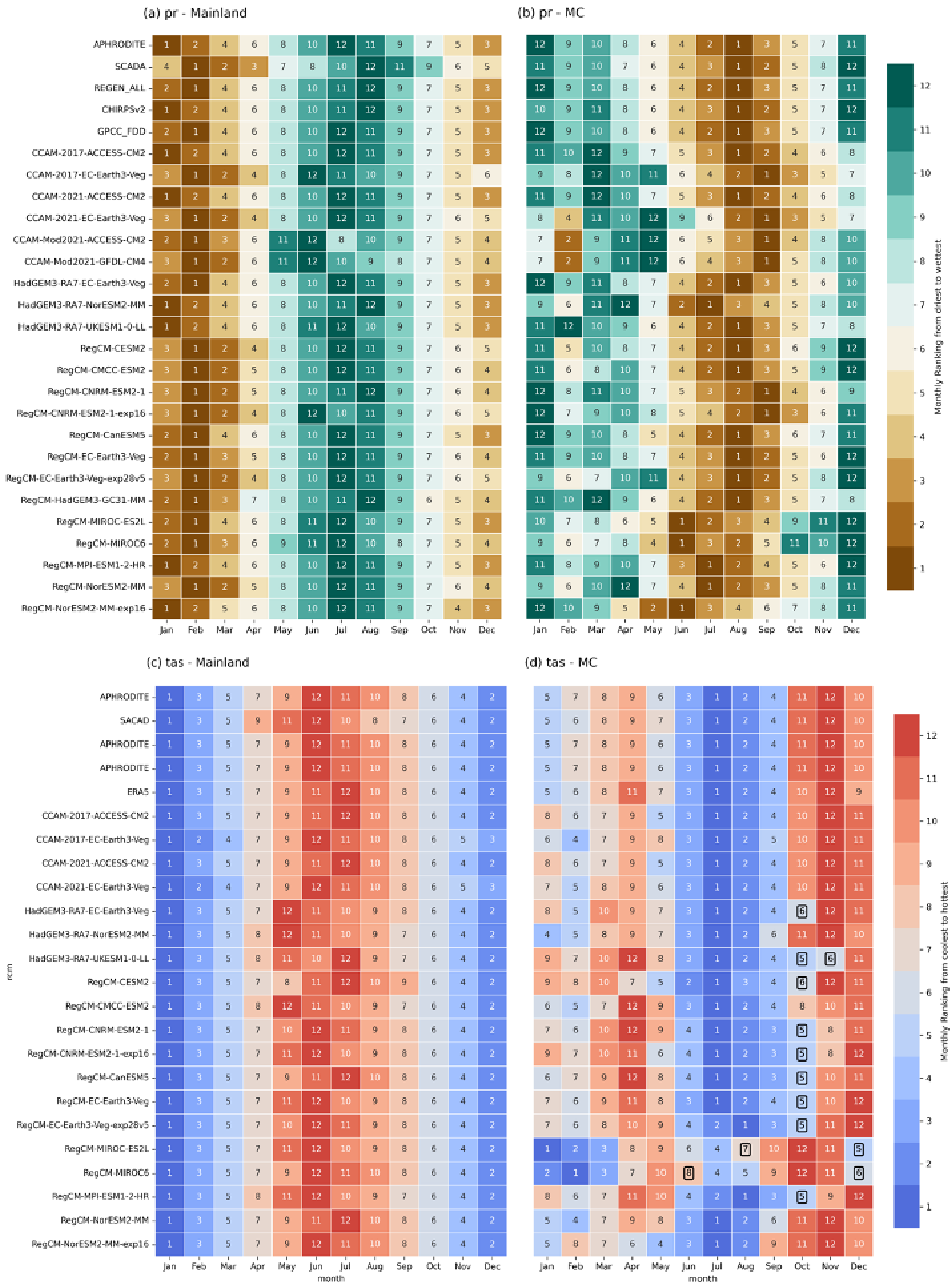

Figure 11. Seasonal cycle benchmark following Isphording et al. (2024) for the unimodal seasonal cycle of (a-b) precipitation (pr) and (c-d) near-surface mean temperature (tas) in the Mainland and MC. Colours and numbers indicate the ranking for latitude-weighted monthly total precipitation/daily temperature, with 1 being the driest/coolest and 12 being the wettest/hottest. The values highlighted in the black box indicate that the simulation does not meet our benchmarking expectations.

38. Line 559. Please add how confidence intervals are calculated to the caption.

Added (L643-646)

39. Line 587. Please add the meaning of blue highlighted numbers to the caption.

Added (L675)

40. Lines 627–629. Do high range, mid-range and low range correspond to emission scenarios? Please clarify.

We thank the reviewer for pointing out the need for clarification. The high-, mid-, and low-range categories do not correspond to different greenhouse gas emission scenarios. Instead, our future response analysis focuses exclusively on a single emission pathway of SSP3-7.0. The high, middle, and low ranges represent the statistical spread of the climate model responses (the projection envelope) within that single scenario, categorised based on the magnitude of their future temperature and precipitation changes as described in the text (Lines X–Y). To prevent any confusion with emission scenarios, we have revised L722-724 to explicitly specify that these ranges refer to the model projection/response spread under SSP370.

"Overall, based on model performance, future spread under the SSP3-7.0 scenario, model dependency, and data availability, we select at least two experiments from three independent model groups to be considered for further km-scale dynamical downscaling over SEA megacities."

41. Line 639. Period instead of a comma?

Revised

42. Line 649. I think this sentence sounds can be simplified to give readers an easier time understanding it. By "spatial contrast" and "temporal mean precipitation distribution" do we mean 'spatial distribution' and 'seasonality'? The first sentence of (2) is very clear and a good template.

Thank you for your point. The sentence (L747-749) has been revised accordingly:

"Despite the large observational uncertainties in precipitation intensity, the CMIP6-downscaled CORDEX-SEA simulations capture the spatial distribution and seasonality of precipitation over SEA reasonably well. However, they tend to substantially overestimate observed precipitation (e.g., at least 14 out of 21 models regardless of the choice of observed reference)."

43. Line 650. "tends to substantially overestimate observed precipitation intensity".

Revised (L749)

44. Line 654. What is a "process-based metric"? Please clarify.

Added. We now include process-based metrics based on large-scale atmospheric circulation and climate drivers, specifically 850 hPa winds and teleconnections with ENSO and the IOD, following the recommendations of Nguyen et al. (2024). These are considered *process-based metrics* because they assess the model's ability to represent the physical mechanisms and climate processes that drive regional climate variability and extremes, rather than simply evaluating the agreement between simulated and observed surface climate statistics. By examining circulation patterns and teleconnections, these metrics provide additional information on whether the underlying processes governing Southeast Asian climate are realistically represented (L752–756).

"This requires further investigation using process-based metrics, such as 850 hPa wind and teleconnection of precipitation with the IOD and ENSO, which assess the ability of the RCM ensemble to present the key physical processes driving precipitation variability and extremes over SEA (Nguyen et al., 2024)."

45. Line 658. There's a space in CRUTS.

Revised (L760)

46. Line 681. Models with 25 km grid spacing will assuredly have difficulty predicting rainfall over mountain terrain. But how does the complex terrain result in an inability to capture the spatial variability? At 25 km, don't the models have sufficient resolution to show where the topography actually is (as we see in the wet biases)? Can you clarify? This makes me wonder again about the ability of the models (GCMs and RCMs) to represent regional SST.

Thank you for raising this point. Please see our response to the main comment no. 3.

47. Line 708. Replace "a range of climate variables" with "temperature and precipitation".

    Revised (L836)

48. Line 712. What case study? A case study is not mentioned anywhere else in the manuscript.

    Revised (L841)

"We also introduce a benchmarking approach in which we utilise the assessment results to select models for further downscaling over SEA megacities, supporting climate risk assessment for climate adaptation and mitigation over these urban areas."