Reviewer #01

This study develops a multi-criteria GCM selection framework across domains for CORDEX-CORE2 dynamical downscaling. The work systematically evaluates CMIP6 models, establishes an evaluation indicator system adapted to regional climate characteristics, and integrates future projections under the SSP3-7.0 scenario. This study provides transparent and reproducible standardized methodological to select GCM candidate pool that balances simulation credibility, ensemble diversity, and practical feasibility. Undoubtedly, this is valuable research with solid methodological design and clear application potential. However, given the complexity of the framework, explanations for some key processes remain insufficient, and the presentation of several results can be further improved. I recommend the following revisions before publication:

We thank the reviewer for the careful reading and the positive assessment of the framework. We agree that some key steps were under-documented and that several figures can be improved, and we have revised the manuscript accordingly. We respond to each comment in turn below.

> Throughout this response, text quoted from the revised manuscript is set in bold italic; everything else is our own explanation. Section numbers are given in place of line numbers, since the line numbering of the revised manuscript is not yet fixed. Figure numbers are those of the version the reviewer read; the revision adds one figure to the Methods section, after which the later figures renumber.

The framework is also available as an interactive toolkit at https://cordex-merit.org, which reproduces every analysis in the manuscript from the same data. Where a comment below asks a question, the toolkit lets a reader answer directly, we say so.

1. **Line 144 states that the framework is built on the five criteria proposed by Sobolowski et al. (2025). Please explain why these five dimensions were selected as screening criteria, and why other potentially important dimensions (e.g., extreme climate simulation capability, model resolution, initial-condition ensemble members) were not included.**

We thank the reviewer for this question. The framework is designed primarily for a specific purpose: selecting Earth system Models (ESMs) to serve as driving models for regional dynamical downscaling. The criteria therefore concentrate on the prognostic fields passed to the regional climate model (RCM) as lateral boundary and surface forcing, and on the dynamic and thermodynamic processes that must be right in the driver before an RCM can add value, including large-scale circulation, moisture transport, thermal structure, and the modes of variability. Deficiencies in these propagate directly into the downscaled simulation and cannot be corrected by the RCM. That focus also explains the three omissions, each of which we would revisit as the framework is extended.

Extremes: The premise of downscaling is that coarse-resolution ESMs do not resolve the processes generating extremes. Screening candidate drivers on a capability the downscaling step is intended to supply, and on which most ESM ensemble members are likely deficient, would

discriminate weakly and for the wrong reasons. What determines whether an RCM can simulate extremes credibly is the fidelity of the large-scale conditions it inherits, and that is what the present suite tests.

Resolution: Resolution is not itself a measure of performance: higher-resolution configurations do not uniformly outperform their lower-resolution counterparts and can amplify existing biases. Few CMIP6 models offer high-resolution configurations, so the criterion cannot yet be applied systematically without arbitrarily restricting the candidate pool. We do regard it as a legitimate tiebreaker at the final stage, where two otherwise equivalent configurations differ only in resolution, and it becomes a more useful criterion as the ESMs archive grows.

Initial-condition ensemble members: These sample a different quantity. Members of one such ensemble share the same physics, biases, feedbacks and sensitivity, differing only in the phasing of internal variability, so their spread estimates unforced variability rather than structural uncertainty. Sub-selection needs the opposite: structurally distinct models, so that regional projections reflect genuine disagreement about how the large-scale state will evolve. Additional initial-condition members contribute nothing on that axis and would be flagged as redundant by our own independence criterion. Practical constraints point the same way: only a few modelling centers produce large initial-condition ensembles, and those that exist cover a limited set of scenarios, so the criterion could not be applied consistently across the models evaluated here. Downscaling cost reinforces this, since each additional member consumes the same computing time as a structurally different model, and spending it on internal variability would narrow the range of responses the experiment is designed to span. We note one consequence: with a single realization per model, our sensitivity and precipitation-spread estimates contain an internal-variability component that we do not separate from the forced response. Nothing in the framework prevents a group from downscaling multiple members where that question is central to its application.

In response to this comment, we have also revised the closing paragraph of the Introduction accordingly:

"This framework does not prescribe a universal set of "best***" models; rather it offers a transparent, reproducible process for evaluating and selecting GCMs based on scientific and practical considerations. Because the selected GCMs provide boundary conditions, the criteria target what a driving model must represent well for downscaling to add value, and deliberately exclude the evaluation of extremes, which coarse-resolution GCMs are known to represent poorly and which the RCM is expected to improve. The framework is intended both to guide the sub-selection of CMIP6 GCMs for CORDEX CORE-2 and to help modeling centers and climate service organizations worldwide identify credible, usable driving models for regional downscaling."***

2. **What is the rationale for selecting the evaluation metrics listed in Table 2? Why are only partial metric categories listed in Line 192?**

We thank the reviewer for both parts of this comment.

On Line 192: that sentence lists the source variables from which the metrics are computed, not the metric categories, and the wording did not make this clear. It was also incomplete, since geopotential height is used (ZG500 DJF and JJA in Table 2) but was omitted. We have corrected and reworded the list and verified that every variable underlying an entry in Table 2 now appears there.

On the rationale, the metrics follow from the purpose of the framework, which is to establish whether a GCM provides credible large-scale forcing rather than to characterize its performance in general. The common set describes the first-order state of the regional climate: seasonal means, amplitude, interannual variability, and the timing of the annual cycle in precipitation, temperature, winds, humidity and sea surface temperature. These fields constitute the boundary and surface forcing, and the corresponding metrics are applied uniformly across domains, which is what makes the cross-domain ranking possible. The region-specific set consists of established diagnostics of the processes that govern climate in each domain, including monsoon onset and demise, jet position and strength, vertical wind shear, the semi-permanent pressure centers, the ITCZ position and the dominant modes of variability. Each has a documented basis in the regional climate literature and was selected because the process it represents controls the regional circulation and moisture supply an RCM inherits from its driver. A model may reproduce domain-mean seasonal fields adequately while misplacing the African Easterly Jet or the Mascarene High, and the consequences for a downscaled simulation differ substantially between those cases.

The Appendix provides each index definition, the process it represents, the supporting reference, and the statistical measures used to evaluate it. We have added a sentence stating that the region-specific metrics are established diagnostics drawn from the regional climate literature, selected because the process each represents governs the circulation and moisture supply that a regional model inherits from its driver.

In response to this comment, the revised text in the section ***2.1 Evaluation Metrics*** is shown in bold:

"The second set consists of region-specific metrics that capture unique dynamical and climatological processes important for individual CORDEX domains. Examples of these include the timing of monsoon onset and withdrawal, the position and strength of jet streams, vertical wind shear in monsoon regions, the mean latitude of the Intertropical Convergence Zone (ITCZ), and the representation of circulation features such as the South Pacific High, the Aleutian Low, and the Azores High. Each is an established diagnostic drawn from the regional climate literature, included because the process it represents governs the circulation and moisture supply that a regional model inherits from its driver. Because these processes are only relevant for certain regions, a single skill score is computed for each metric and then distributed equally across all AR6 regions in the corresponding CORDEX domain to ensure fair weighting."

Across all nine CORDEX domains, a total of 98 evaluation metrics is used, combining both the common and region-specific categories (see Table 2). The definition of each metric, the process

or feature it represents, its source in the literature, and the statistical measures used to evaluate it are documented in the Appendix.**”**

3. **The African domain includes the largest number across all domains, which can be attributed to its *“diverse climate regime”* in L240. However, an alternative interpretation is that limited research or data availability in this region has not yet supported the condensation of more efficient indicators. Could the regional disparity in metric count indirectly undermine the fairness of cross-domain rankings?**

We thank the reviewer for raising this. The comment has two parts: why AFR carries the largest metric count, and whether that count affects the fairness of the global ranking.

On the first, the count follows from the domain's physical geography rather than from data availability or limited process understanding. AFR is among the largest CORDEX domains by land area and is the only one containing three distinct monsoon systems, the West African, East African and Southeast African. Each is diagnosed with its own set of metrics (onset, demise, meridional temperature gradient, and zonal and meridional wind shear), so the domain accumulates three complete sets where others carry one or none. Together with the associated circulation features evaluated here, the African Easterly Jet, the East African low-level jet, the Mascarene High, the Mozambique Channel Trough, the South Atlantic High and the Atlantic ITCZ, this accounts for most of the difference between the 57 metrics in AFR and the counts elsewhere. The larger number therefore reflects the number of distinct processes requiring evaluation, not an inflation of diagnostics for a single process or an inability to condense them into fewer indicators.

On the second point, the default aggregation makes the global ranking insensitive to metric count. Under the domain-rank method (referred to as Method 3 in the original submission, now named descriptively following the reviewer's earlier comment), metrics are averaged within each domain, models are ranked within that domain, and it is the resulting ranks, not the underlying scores, that are averaged globally. Because ranks are scale-free and count-free, a domain evaluated with 57 metrics and one evaluated with 37 contribute equally to the global result. Metric count affects the precision with which a domain's internal ranking is estimated, not that domain's weight in the global aggregation. This is one reason the domain-rank method is preferred over the global-mean-score method (formerly Method 1), which averages raw scores globally and does implicitly overweight metric-rich domains, as noted in the revised Section 2.5 and quantified in the revised Fig. 6.

Two further features limit the influence of metric count. The region-specific metrics are distributed equally across the AR6 regions within a domain rather than added to each, so a domain gains no influence from its subdivision. And the correlation-based weighting prevents closely related metrics from accumulating influence. Had the three African monsoon diagnostics produced similar model-to-model performance patterns, they would have been identified as redundant and down-weighted. Their retention at high weight (Fig. 5) indicates they carry genuinely distinct information rather than triplicating a single signal.

4. **All current evaluation metrics are based on monthly mean states, with no extreme climate indicators included. For end users of climate impact research, the ability to simulate extreme events is a core requirement. It is recommended to add usage guidance for this user group, or may be explicitly acknowledge this setting as a limitation.**

As set out in our response to comment 1, the absence of extreme-event metrics is a deliberate consequence of what the framework is for. We are selecting driving models, and coarse-resolution GCMs are known not to fully resolve the processes that generate extremes, which is precisely why downscaling is undertaken. Screening candidate drivers on a capability the whole ensemble lacks, and that the downscaling step is intended to supply, would discriminate weakly and on the wrong grounds. What determines whether an RCM can simulate extremes credibly is the fidelity of the large-scale conditions it inherits, and that is what the present suite evaluates.

We have already added this reasoning to the closing paragraph of the Introduction in response to comment 1, where the revised text states that the criteria deliberately exclude the evaluation of extremes, which coarse-resolution GCMs represent poorly and which the RCM is expected to improve. The same revision narrowed the closing claim from the original ***"credible and usable for regional downscaling and climate impact applications"*** to credible, usable driving models for regional downscaling, so the manuscript no longer implies that a high ranking here certifies skill at reproducing observed extremes.

On usage guidance, the framework produces a candidate pool rather than a prescribed list, and a user whose application is extremes-driven can apply an additional, application-specific screening step within that pool without forfeiting the cross-domain consistency that motivates it. Such screening is most informative when applied to the downscaled output rather than to the driving GCM, since it is the regional simulation that is the object of the impact assessment.

Both reviewer's suggestions, the explicit limitation and the usage guidance, are addressed in the new Section 5.3, ***"What the framework does and does not establish"*** The third caveat there states the limitation directly: the criteria ***"deliberately exclude the evaluation of extremes, which coarse-resolution GCMs represent poorly and which the RCM is expected to improve"***, so that ***"a high ranking therefore indicates fidelity in large-scale forcing and climatological state, not demonstrated skill at reproducing observed extremes at the GCM scale."*** The same paragraph gives the guidance the reviewer asks for: ***"because the framework yields a candidate pool rather than a fixed list, users whose application is extremes-driven retain the option of an additional, application-specific screening step within that pool, most informatively applied to the downscaled output rather than to its driver."*** Section 5.4 further notes that ***"process-oriented diagnostics such as land–atmosphere coupling, storm-track statistics and teleconnection fidelity"*** are identified for future iterations, though ***"most require sub-monthly output and so lie outside the present configuration"***

5. **Why was the exponential penalty function chosen for the normalization of bias and RMSE? Will it artificially compress performance differences among high-scoring models? Additionally, for models with comparable total scores, does this scoring criterion favor models with balanced overall performance over those with outstanding performance in specific aspects?**

We thank the reviewer for these questions. On the first point we would note that the effect of the transformation runs opposite to the concern raised.

For the bias score $S = 100 \cdot \exp(-|\text{bias}|/N)$ with $N = 2$ for most metrics, the magnitude of the derivative is largest at small error and decays as error grows. Discrimination is therefore greatest among well-performing models, and compression occurs at the poor-performing end. An increase in normalized bias from 0.1 to 0.2 costs 4.6 points, whereas the same absolute increase from 2.0 to 2.1 costs only 1.8 points, a factor of about 2.6 in resolving power in favor of the models that are candidates for selection. The same holds for normalized RMSE. The exponential therefore preserves fine distinctions where they matter for sub-selection while preventing a single catastrophic metric from dominating a model's mean score. We have added this property to the manuscript where the penalty function is introduced: ***"Because the exponential decays fastest at small error, it also discriminates most strongly among well-performing models, with compression occurring at the poor-performing end of the range where differences matter less for sub-selection"***

We would also note that the choice of transformation cannot affect the ordering of models on any individual metric, since the exponential is monotonic in error. Its influence acts only through aggregation, determining how much weight a large error on one metric carries relative to moderate errors on others. This is now stated in the manuscript alongside the scoring equations: ***"The transformations in Equations 1–3 are monotonic in error and therefore cannot alter the ordering of models on an individual metric; their effect is confined to how much weight a large error on one metric carries relative to moderate errors on others once the scores are averaged"***This is also why the transformation is applied to bias and RMSE but not to pattern correlation: PC is already bounded and dimensionless, and an exponential penalty there would compress precisely the high-correlation range where well-performing models cluster, which is the effect the reviewer is concerned about.

On the second question the reviewer is correct, and we do not regard it as a defect. Averaging across metrics favors models with balanced performance over those that excel on a few diagnostics and fail on others, and this is intended. The objective is a single subset that performs acceptably across all nine CORDEX domains, so that inter-regional differences in downscaled projections reflect genuine climatic contrasts rather than inconsistencies in the driving models. The global ranking is therefore designed to reward consistency across domains rather than excellence in a few. We accept, however, that the aggregate score conceals a model's dispersion across metrics and domains, and we have added diagnostics that make it visible: inter-domain rank spread reported alongside the global rank (response to comment 14).

6. **Most general evaluation metrics adopt standard JJA, DJF for regional consistency. But for regions where wet/dry seasons are misaligned (like some tropical regions), the representativeness may be limited**

We agree that calendar-fixed seasonal means are an incomplete descriptor where the wet season does not align with the boreal calendar, which is why the suite was not built on them alone.

Of the 98 metrics in Table 2, only four are fixed season means of the quantities where a displaced wet season matters: PR DJF, PR JJA, TAS DJF and TAS JJA. Precipitation and temperature are each also evaluated in MAM and SON, so no season is left unsampled. Most other fixed-season metrics are winds and geopotential height, for which the boreal seasons remain physically meaningful in the tropics because these features follow the seasonal migration of solar forcing rather than local rainfall regimes.

The timing properties the reviewer identifies as at risk are carried by metrics that make no calendar assumption. Precipitation is described by six of these, namely amplitude, interannual standard deviation, peak month, the seasonality index, relative entropy and the Hovmöller diagnostic, against its four seasonal means; temperature by three. Monsoon onset and demise are computed for every monsoon system evaluated here, so a model that reproduces DJF and JJA means acceptably while shifting the monsoon by a month is still penalized. The redundancy weighting reinforces this, since strongly inter-correlated seasonal means are down-weighted while the calendar-independent diagnostics retain weights near unity (Fig. 5).

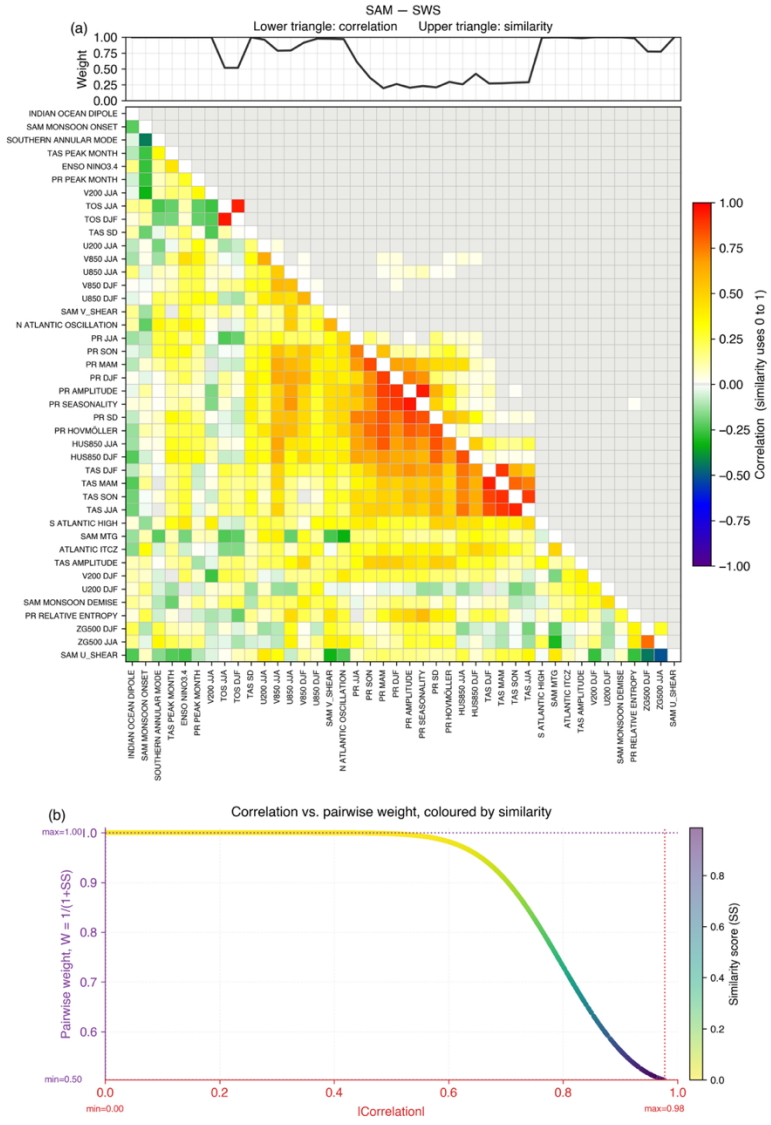
One limitation is worth stating. The difficulty arises not from wet seasons that are merely displaced relative to the boreal calendar, but from bimodal regimes, where a single seasonal mean averages across two rainfall peaks and the dry interval between them. We address this explicitly for Central America through the bimodal rainfall index (Appendix 5.20), but the East African and equatorial Southeast Asian double-peak regimes are represented by their dominant maximum only. Extending a bimodality diagnostic of the Central American type to those regions is a natural refinement of the suite, and we will prioritize it in future updates of the framework. This is now identified as a refinement for future iterations in the new Section 5.4: ***"Regions with two distinct rainfall maxima are currently represented by their dominant peak, except over Central America where a bimodality index is applied. A single seasonal mean averages across both peaks and the dry interval between them, so the diagnostic value of that mean is limited where the regime is bimodal rather than merely displaced relative to the boreal calendar. Extending that diagnostic to the East African and equatorial Southeast Asian regimes would improve the representation of seasonality there"*** We intend to prioritize this in a future update of the framework.

7. **In Figure 2b, the z-axis values and color mapping appear to represent the same thing, could be redundant. Why not simplify the plot to a 2D scatter plot with color only, which would be more readable than a 3D.**

We agree. Because the pairwise weight is a deterministic function of the correlation, the three quantities lie on a single curve, and the vertical axis carried no information beyond the color scale. We have replaced panel (b) with a two-dimensional plot of |correlation| against pairwise

weight, colored by similarity score, with the global minima and maxima of each measure annotated as before. The revised panel shows the same relationship more directly: the weight remains at unity until |r| reaches roughly 0.5, falls steeply between 0.6 and 0.9 as the similarity score rises, and approaches its lower bound of 0.5 only where two metrics are almost perfectly correlated.

Revised caption, Figure 2. *Metric redundancy weighting. (a) Similarity as a function of the correlation between two metrics, for the radius of similarity used here. (b) Correlation against the pairwise weight 1/(1+SS) for every metric pair in every AR6 region and CORDEX domain, coloured by similarity. Because similarity is a deterministic function of correlation, the pairs fall on a single curve; the panel shows the weighting function rather than a spread of outcomes.*



8. **Why were these four aggregation methods tested? Why were other rank aggregation methods (e.g., Borda count, Copeland method) not adopted? And what are the statistical advantages and disadvantages of Method 2 and the default Method 3?**

We thank the reviewer for this question, which prompted us to be clearer about both the naming and the scope of the comparison.

On the naming: The default method is a Borda count: ranks are computed within each CORDEX domain and averaged across the nine. Calling it ***"Method 3"*** obscured that, and we have renamed the four schemes descriptively: global mean score, AR6-region ranks, domain ranks (Borda) and metric-type blocks. The reviewer's suggested method is therefore the scheme we adopt, not an untested alternative.

On why these four: The comparison was not a survey of rank-aggregation rules. It varies the two choices specific to this framework: the level at which scores are aggregated before ranks are taken (globally, at AR6 region, or at CORDEX domain), and whether metrics common to all domains are pooled with domain-specific ones or treated as separate blocks.

On Copeland: We have now tested it, together with the median of the domain ranks. Neither changes the conclusions. HadGEM3-GC31-MM ranks first under all six schemes; relative to Borda, Copeland moves 34 of 45 models by at most 2.5 places and the median moves 37 by at most 4.5. The three CORDEX-CORE2 selected models retain ranks 2–2.5, 2.5–4 and 11–13.5 across all six, and the mean range over the 45 models is 3.1 places.

Section 5.3 now records the magnitude of the effect: ***"Sensitivity analyses demonstrate that rankings are insensitive to the choice of aggregation method, with the three models selected for CORDEX-CORE2 each varying by no more than 2.5 places."***

On the two rank-based schemes: AR6-region ranks uses 92 contributing rankings rather than nine, so it is less sensitive to any single one. But the regions are not independent. Adjacent regions within a domain share circulation feature and therefore share errors, and a domain subdivided into more regions gains more influence: AFR contributes 16 and EUR 7, a weighting by subdivision rather than by anything physical. Borda weights the nine domains equally, which matches the framework's purpose, and because ranks are scale-free a domain with 57 metrics and one with 37 contribute equally. Its costs are that nine is a small number of contributing rankings, and that averaging ranks discards the margins. Copeland worsens the second, reducing each comparison to a win or a loss. The median is more robust to an outlying domain but discards more information and needs a larger fraction of the metric suite to converge.

This is one of the questions the toolkit at https://cordex-merit.org answers interactively: the model ranking page lets a reader switch between all six aggregation schemes and see which models move, so the sensitivity can be examined rather than taken on trust.

Manuscript changes: Section 2.5 has been revised to test six aggregation approaches rather than four, to name them descriptively, and to set out the properties of each. The section now explains

the basis of the comparison: ***"The first four vary the two choices specific to this framework: the level at which scores are aggregated before ranks are taken, and whether metrics common to all domains are pooled with domain-specific ones or treated separately. The remaining two vary the rule by which the resulting ranks are combined."***

The default scheme is now named for what it is: ***"Averaging ranks in this way is a Borda count, and the scheme is named accordingly."*** The section also states why metric count does not confer influence under it: ***"Because ranks are scale-free, a domain evaluated with 57 metrics and one evaluated with 37 contribute equally."***

The limitations of ranking at AR6-region level are stated: ***"It also raises the number of contributing rankings from nine to 92, although those rankings are not independent, since adjacent regions within a domain share circulation features and therefore share errors, and it weights domains by their subdivision rather than by anything physical."***

The two additional rules and the justification for the default are given in a new closing paragraph: ***"Two further rules were tested on the same nine domain rankings. A Copeland count scores each model by the number of domains in which it is ranked above each competitor and is therefore insensitive to the size of each margin. Taking the median rather than the mean of the domain ranks is more robust to one anomalous domain but discards more information. Domain-ranks (Borda) is retained as the default because it weights the nine domains equally, is insensitive to the number of metrics evaluated in each, uses the ordinal information without reducing it to pairwise outcomes, and is the most stable of the six when the metric suite is subset."***

Figure 6 and its caption have been revised to show all six schemes rather than four.

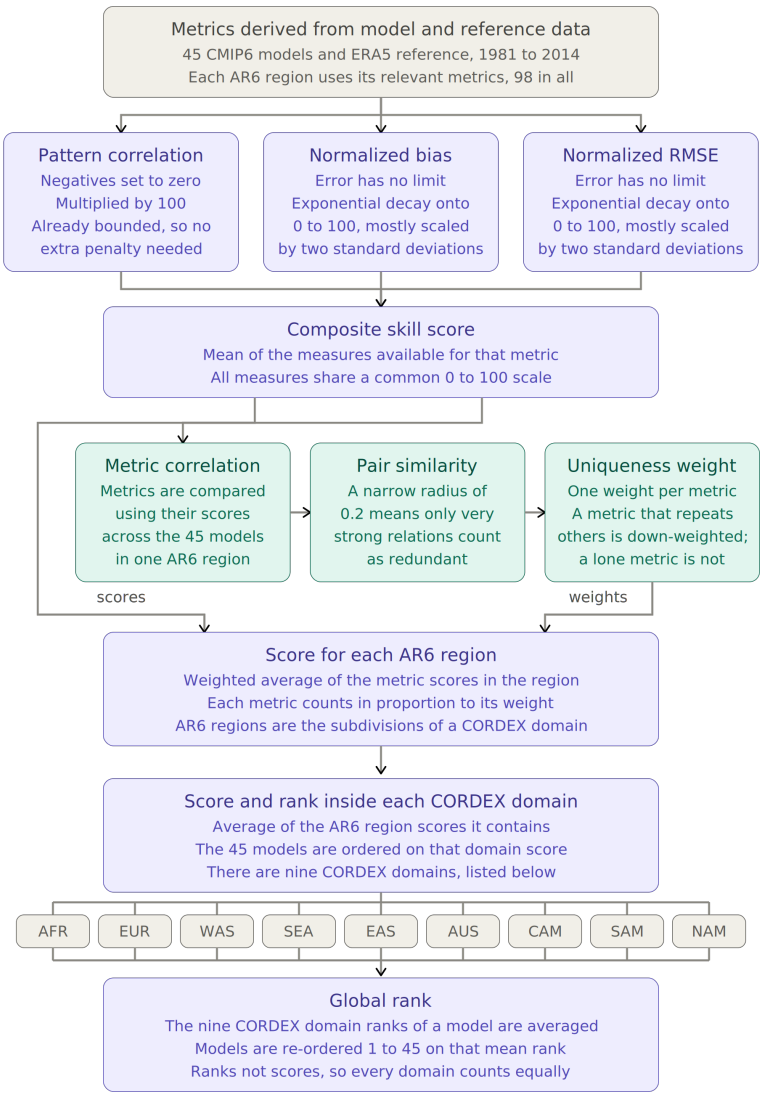
Revised caption, Figure 6. *Global model rankings under six aggregation schemes: global mean score, AR6-region ranks, domain ranks by Borda count, domain ranks by Copeland count, the median of the domain ranks, and metric-type blocks. Rank 1 is at the centre. The fifteen models whose rank varies most across the schemes are drawn in colour and listed with their range in parentheses; the remaining thirty are grey. Domain ranks (Borda) is the default used throughout.*

9. **It is recommended to add a technical flowchart in the Methods section, clearly illustrating the full calculation workflow from single-metric scoring, indicator weight calculation, to aggregation.**

We agree and have added a schematic of the full calculation chain to the Methods section. It follows the workflow in four stages. First, the three statistical measures are computed for each metric within each AR6 region, with the reason each is scaled linearly or exponentially. Second, these are combined into a composite skill score on a common 0 to 100 scale. Third, the uniqueness weight of each metric is derived from the pairwise correlations among metrics. Fourth, the weighted AR6 region scores are aggregated to a rank within each of the nine CORDEX domains, and those ranks to a global rank. The figure also makes two points explicit that were easy to miss in the text. The weighting branch takes the composite skill scores as its input and returns a weight per metric, so the region score draws on both. And the global rank is obtained by averaging domain ranks rather than domain scores, which is why every domain contributes equally regardless of how many metrics it contains.

Caption, new Figure X. *The calculation workflow. Metrics are computed for each AR6 region from the model and reference data, scored by the statistical measures available for that metric and combined into a composite skill score. Pairwise correlations among metrics give each metric a uniqueness weight; scores and weights then combine to an AR6 region score, a CORDEX domain score on which models are ranked, and a global rank averaged over the nine domain ranks.*

This figure is new, so it has no number in the version the reviewer read. We call it Figure X here; it becomes Figure 2 in the revised manuscript, after which the later figures are renumbered.



10. ***"Figure 3b"*** referenced in Line 426 does not appear in the text. Please check and correct it.

The reviewer is correct. Figure 3 has no panel (b); the radial dashboard is Figure 4. We have corrected the reference.

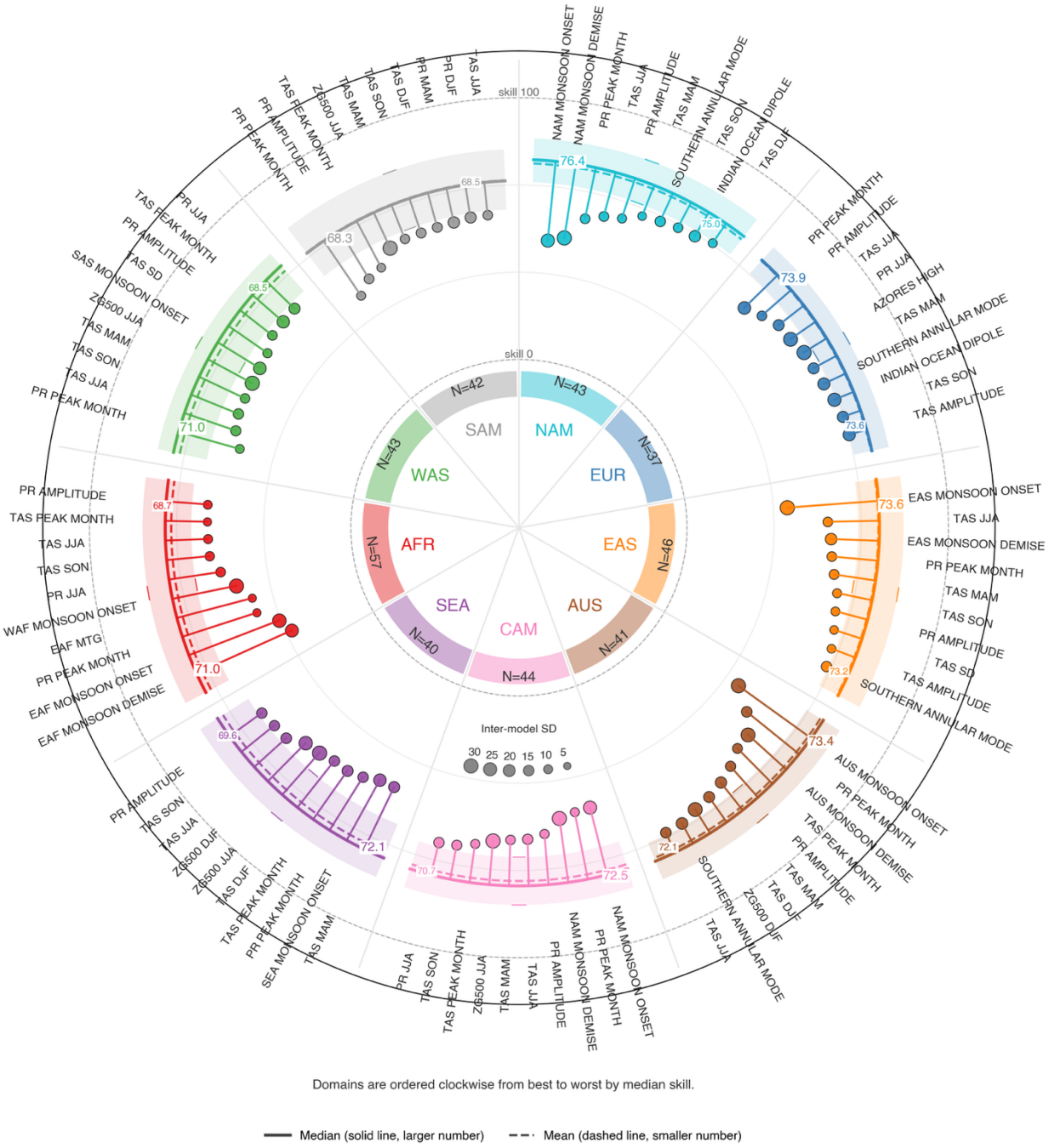
11. **The design of Figure 4 is well conceived, but the comprehensive scores of regions other than NAM are relatively close, making it hard to compare. It is recommended to sort regions clockwise by their medium or mean scores. In addition to the median line, a dashed line for the mean value may provide richer statistical information.**

We agree on both points and have revised Figure 4 accordingly. The domains are now ordered clockwise from highest to lowest median skill: NAM, EUR, EAS, AUS, CAM, SEA, AFR, WAS, SAM. Each domain's annulus carries both statistics, the median as a solid line and the mean as a dashed line, with the corresponding values labelled at the ends of the band. A note on the figure states the ordering rule, and a legend distinguishes the two lines.

The domain positions in the revised figure differ from the original for two independent reasons, both arising from checks we carried out during revision to confirm that the calculations were robust across all indices. First, we identified a bug in the analysis script producing this figure: it was reading inconsistent input data for the NAM domain, so its values were not on the same footing as the other eight. With consistent input, NAM has the highest median skill of the nine (76.4) and an intermediate spread (IQR 16.7), in place of the previously reported median of 61 and IQR of 11, so it is no longer placed at the bottom of the figure but at the top. Its apparent low skill was an artefact of the inconsistency rather than a property of model performance over North America, and the text describing NAM as standing out with the lowest and most uniform skill has been revised accordingly. Second, reviewing the index calculations revealed an error in the monsoon onset and withdrawal diagnostics. Recalculating them lowered the scores of those metrics, which reorders the remaining eight domains, since the monsoon domains carry more of these metrics than the extratropical ones.

The revision also makes a feature of the distributions visible that the original did not. In eight of the nine domains the mean lies below the median, most strongly in SEA and WAS (−2.5 points each) and AFR (−2.3), indicating left-skewed metric-level skill distributions: a minority of poorly simulated metrics, principally the monsoon timing diagnostics, pulls the mean down while the bulk of metrics score higher. SAM is the exception, with the mean marginally above the median (68.5 against 68.3) and the widest spread of any domain, indicating a distribution closer to symmetric about a lower center rather than a high-scoring bulk with a low tail.

Revised caption, Figure 4. *Radial dashboard of metric-level skill by CORDEX domain. The coloured annulus spans the interquartile range of mean model skill across the metrics of each domain, with the median as a solid line and the mean as a dashed line, both labelled. Domains are ordered clockwise by decreasing median skill. The ten lowest-scoring metrics of each domain are named around the rim.*



Domains are ordered clockwise from best to worst by median skill.

Median (solid line, larger number) Mean (dashed line, smaller number)

12. **Metric weights across regions show clear common patterns in Figure 5. For example, PR amplitude and seasonality have low weights in all domains. Then it might be helpful to group metrics with similar physical meanings together, to help readers compare across regions.**

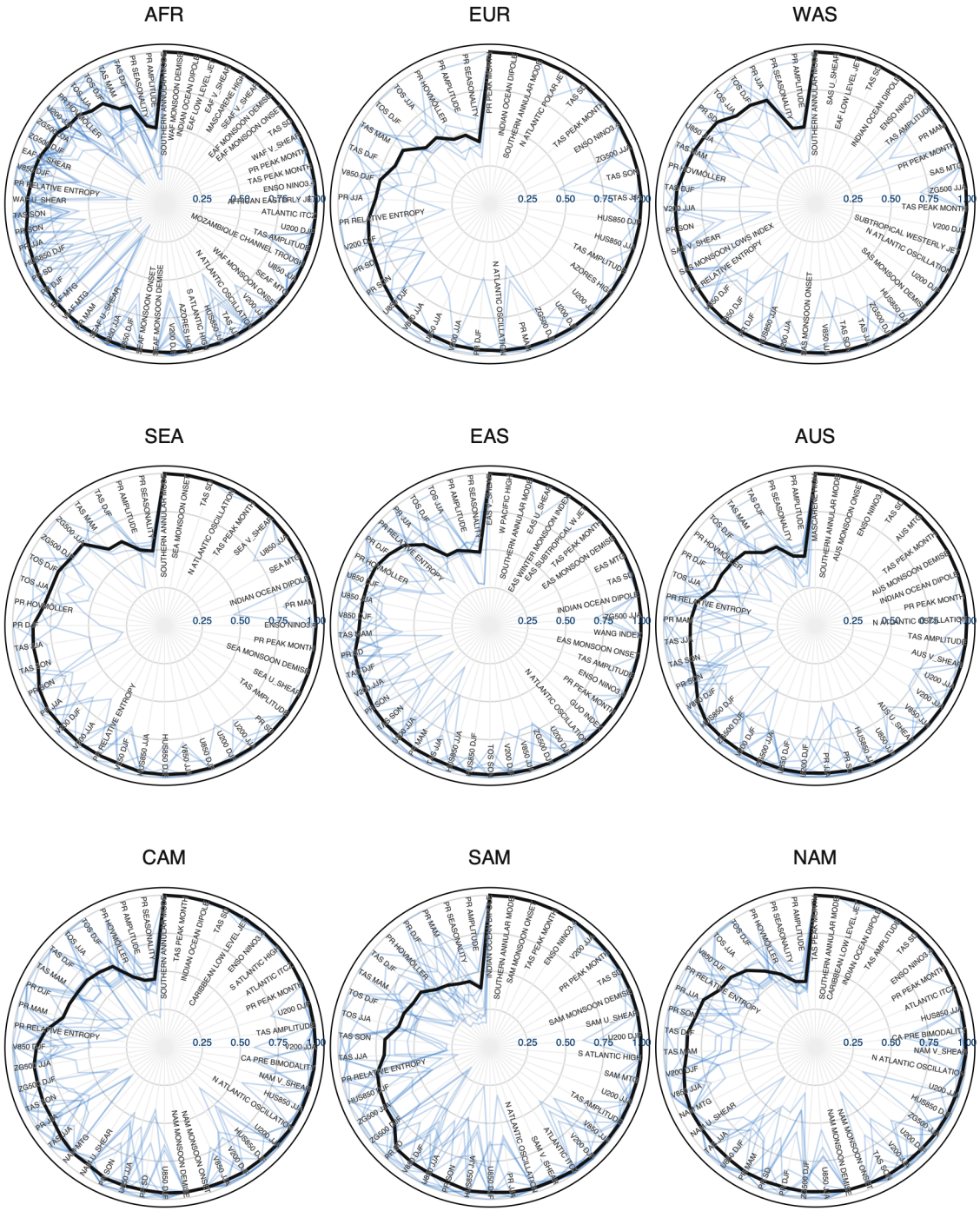
We agree and have restructured the presentation of the metric weights. Figure 5 remains the nine-panel figure showing all metrics in each domain, but the metrics within each panel are now ordered by weight rather than by their position in the metric list, so the high- and low-weight groups fall together and can be compared between panels. To make the cross-domain comparison the reviewer asks for more direct, we have added a supplementary figure (Fig. S1) built on the 35 metrics common to all nine domains, with the domains as rows and the metrics as columns in a single shared order. Its upper panel shows each metric's mean weight across the AR6 regions of a domain, and its lower panel the range of that weight across those regions.

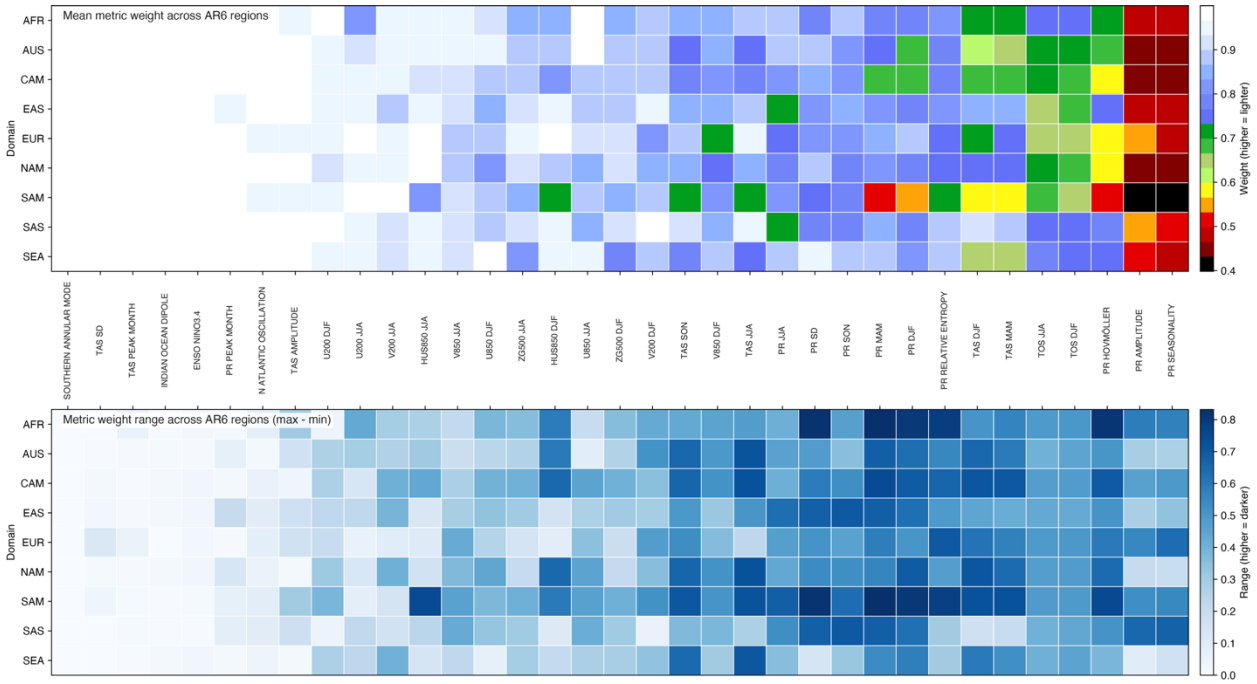
The shared ordering groups the metrics by physical character, which is the comparison the reviewer asked for. The modes of variability (Southern Annular Mode, Indian Ocean Dipole, ENSO Niño-3.4, North Atlantic Oscillation), together with temperature variability and the timing of the annual cycle in temperature and precipitation, hold weights close to unity in every domain (0.96 to 1.00), confirming that they carry non-redundant information. Weights then decline through the circulation and thermodynamic fields, the winds, geopotential height and specific humidity, and are lowest for the precipitation seasonal-cycle descriptors, with PR seasonality (0.43–0.53), PR amplitude (0.40–0.55) and the Hovmöller diagnostic (0.53–0.77). The pattern is consistent across domains, with SAM showing the strongest down-weighting overall and WAS, SEA and EAS retaining the highest effective weights.

The lower panel of Fig. S1 adds information neither the original nor the revised Figure 5 conveys: how much a metric's weight varies between the AR6 regions within a domain. Regional variability is negligible for the high-weight modes of variability, below 0.1 across regions, and grows for the precipitation and seasonal-mean metrics, which is expected since their inter-correlation depends on local climate. WAS, EUR, AFR and SAM show the widest ranges, reaching 0.55 to 0.67 for PR seasonality and PR amplitude, while SEA and NAM are the most uniform.

The toolkit at https://cordex-merit.org also carries these weights: the metrics interdependencies page shows the correlation and similarity matrices for any AR6 region, and the historical performance page can colour its cells by uniqueness weight instead of skill, which makes the grouping by physical character visible in any domain the reader chooses.

Revised caption, Figure 5 and Fig. S1. *Figure 5: metric weights for the nine CORDEX domains, one panel per domain, with metrics ordered by weight. Thin lines are individual AR6 regions and the thick black line the median across regions. Fig. S1: the 35 metrics common to all nine domains, with mean weight across the AR6 regions of a domain in the upper panel and the range of that weight in the lower panel, metrics in a single shared order.*





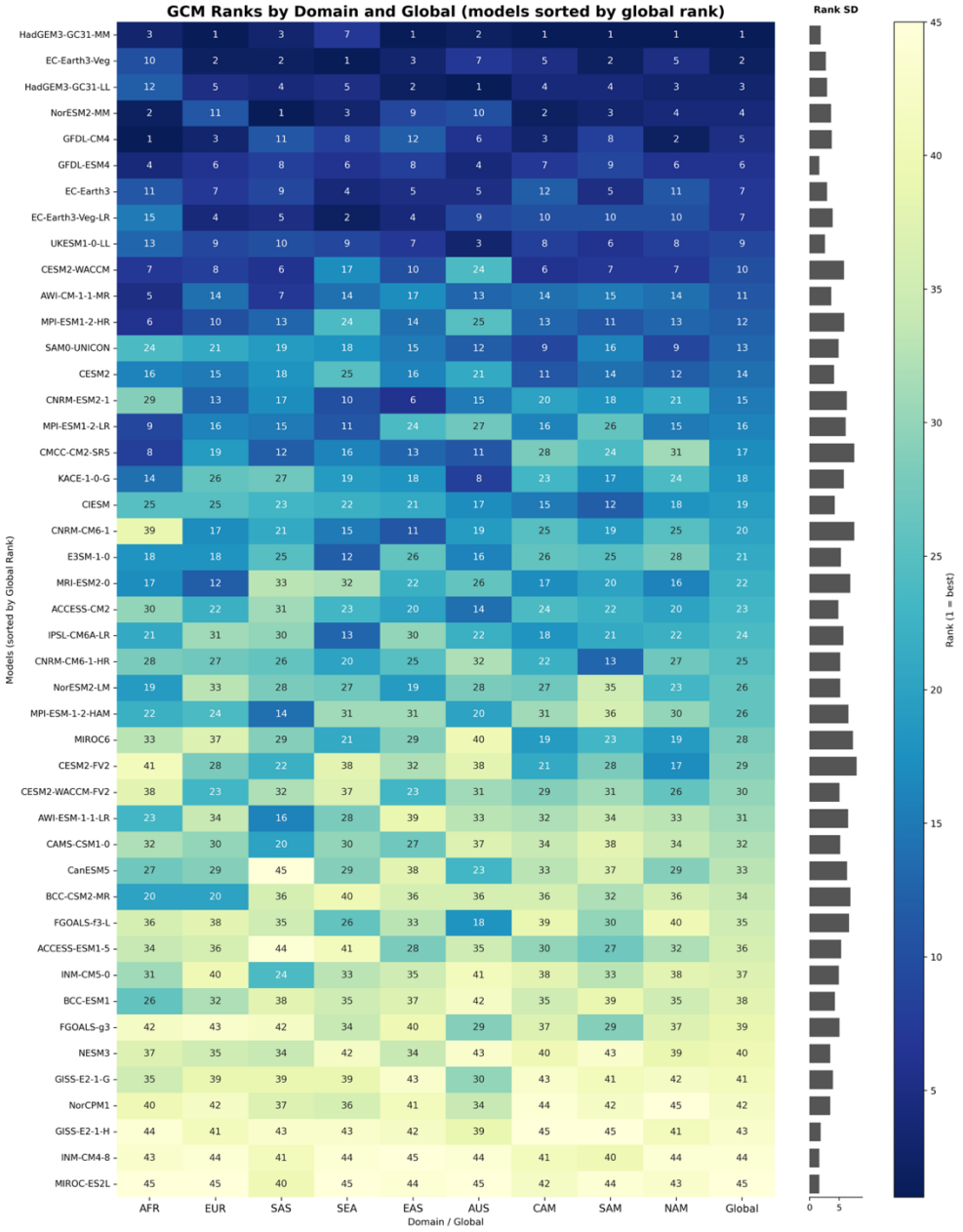
13. Instead of general good models, it could be helpful to additionally identify *"region-specialized"* models that perform outstandingly in a single domain but have lower global rankings. This would be highly valuable for users focusing only on a single region.

We agree this is valuable for single region users and note that Figure 7 already provides it. The heatmap gives each model's rank in every CORDEX domain alongside its global rank, with the rank printed in every cell, so a user interested in one domain can read the best-performing models for that domain directly from the corresponding column, independently of the global ordering. The color scale makes such cases immediately visible: a region-specialized model appears as a dark cell within an otherwise lighter row, marking a domain rank much better than the model's global rank, and the printed value gives the rank exactly. Examples include CNRM-CM6-1, ranked 20th globally but 11th in EAS and 15th in SEA, and KACE-1-0-G, ranked 18th globally but 8th in AUS. The effect is larger further down the table: FGOALS-f3-L ranks 35th globally but 18th in AUS, AWI-ESM-1-1-LR 31st globally but 16th in WAS, and CESM2-FV2 29th globally but 17th in NAM.

The newly added rank standard deviation column (response to comment 14) makes these cases easier to identify, since a large spread flags a model whose performance is domain dependent. Each of the models named above has a rank spread of 5.8 to 8.0 places, against 1.6 to 3.9 for the models in the global top ten. We have added a sentence to the text pointing out that the figure can be read this way.

A reader interested in a single domain can also use the toolkit at https://cordex-merit.org, where the model ranking page can be restricted to one CORDEX domain and the historical performance page reports each model's rank within the domain being viewed, so region-specialized models can be identified without reference to the global ordering.

Revised caption, Figure 7. *Model ranks by CORDEX domain and globally. Each cell gives a model's rank in that domain; the final column is the global rank, on which the rows are sorted. Bars to the right give the standard deviation of each model's nine domain ranks, so a long bar marks a model whose performance depends on the domain.*



14. **What is the quantitative basis for the claim in Line 528 that some models show high performance inconsistency across regions? It is recommended to add a column of SD or range in the last column of Figure 7 to visually represent the *"inconsistency"*. In addition, CNRM-CM6-1 shows large performance variations both across aggregation methods and regions, does it carry clear physical implications?**

We agree and have added a rank standard deviation column to Figure 7, shown as horizontal bars alongside the global rank. This makes the degree of cross-domain consistency directly readable for every model and replaces the qualitative description at L528 with a quantitative one.

The revised figure also shows that the behavior the reviewer highlights is not specific to CNRM-CM6-1. Rank spread is smallest at both ends of the ranking, around 1.6 to 3.0 ranks for the highest- and lowest-ranked models, and largest for models in the middle of the distribution, averaging about 6 ranks and reaching 8.0. This is partly structural: a model ranked near 1 or 45 has little room to vary, so rank standard deviation is compressed at the extremes, and the mid-range spread is not directly comparable with it. CNRM-CM6-1 has a spread of 7.6 ranks, which places it among the four widest but within the range typical of mid-ranked models rather than standing apart from them: CESM2-FV2 (8.0), CMCC-CM2-SR5 (7.6) and MIROC6 (7.4) are comparable. We have rewritten the passage to reflect this, identifying the models with the largest spread from the figure rather than by example.

On the physical implications, CNRM-CM6-1's low global rank is driven by its performance in AFR, where it ranks 39th against a global rank of 20. The domain-level diagnostics indicate which metric families are responsible, but attributing this to a specific process would require a process-level analysis beyond the scope of the present framework, and we prefer not to speculate. We have therefore stated where the deficiency is concentrated rather than proposing a mechanism for it.

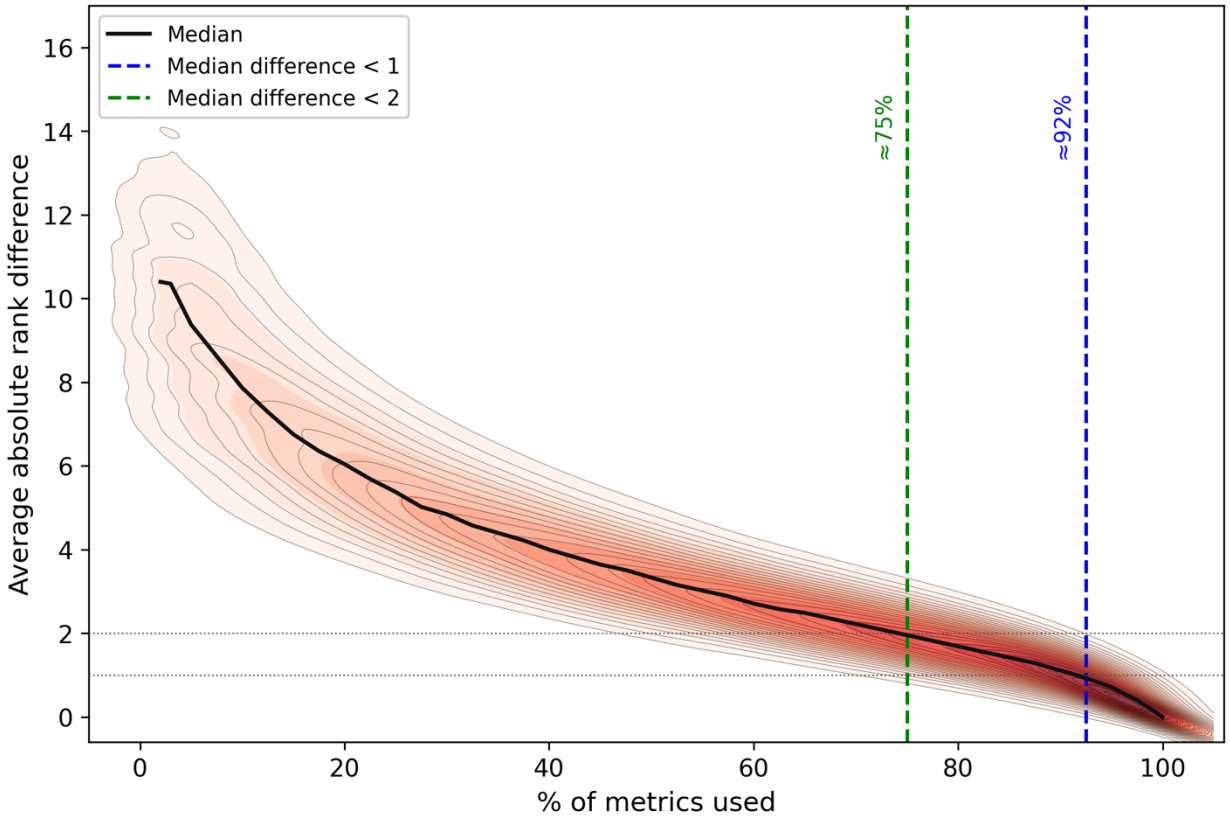
15. **The figure reference in Line 544 appears to be incorrect. please verify and correct.**

The reviewer is correct. Line 544 cites Figure 7 for the Monte Carlo convergence analysis, but that analysis is shown in Figure 8. We have corrected the reference.

16. **The threshold stated in Line 570 is *"75%"*, while the corresponding figure shows *"70%"*. Please make this consistent.**

We thank the reviewer for catching this. The convergence analysis has been rerun as part of the revision, and the text, the figure markers and the figure caption now all report the same values: the median absolute rank difference falls below two ranks at ~75 % of the metric suite and below one rank at ~92 %.

Revised caption, Figure 8. *Monte Carlo convergence of the ranking. Probability density of the average absolute rank difference from the full-suite ranking, against the fraction of the metric suite included, pooled over the nine CORDEX domains. The black line is the median. Dashed verticals mark where the median first falls below two ranks and below one rank.*



17. The study concludes that more than 75% of metrics are required to obtain stable rankings. Is it possible to identify a few metrics that can yield approximately consistent stable rankings?

Answering this requires clarifying a design feature of the convergence experiment that we did not emphasize sufficiently.

Metrics in the Monte Carlo are not sampled uniformly. The selection probability at each step is weighted by the metric's influence, the product of its uniqueness weight and the inter-model standard deviation of its scores, so metrics that carry non-redundant information and discriminate strongly among models are drawn preferentially. Early subsets in each realization are therefore already enriched in the most informative diagnostics available, and the convergence behavior we report is not that of an arbitrary metric list but of one assembled in a deliberately favorable order. Read this way the result is stronger than we stated: even with selection biased toward the most informative candidates, the median global rank deviation does not fall below ±2 until 75% of the suite is included, and not below ±1 until 92.5%. A naive or convenience-based selection would be expected to perform worse.

This bears on the reviewer's question. A small subset that reproduces the full-suite ranking closely can certainly be constructed, by optimizing directly against the reference ranking. But such a subset is identified using the answer it is meant to approximate. It is a post hoc fit to this particular 45-model ensemble evaluated against ERA5, not a criterion that could have been applied in advance, and there is no basis for expecting it to transfer to a different ensemble, a different reference dataset, or the CMIP7 generation for which this framework is intended to remain usable.

We would add that rank agreement is not by itself sufficient grounds for reducing the suite. The metrics test whether a driving model represents the processes that govern regional climate, including monsoon onset and demise, jet position and strength, the semi-permanent pressure centers and the modes of variability. A reduced set that reproduced the ordering while omitting these diagnostics would leave the ranking unsupported by any assessment of the processes it is meant to reflect: two models could be correctly ordered while the evaluation says nothing about whether either of them simulates the West African monsoon credibly.

For these reasons we have not proposed a reduced metric set and have instead made the influence

18. What is the rationale for setting the threshold at 3%?

We thank the reviewer for this question. The 3% threshold is not a statistical significance level but a deliberately conservative choice, made so that the clusters identified correspond to genuine model relatedness rather than to incidental similarity.

The pairwise cosine similarities across the 45 models span a narrow range, roughly 0.68 to 0.99, and their distribution is unimodal with a strong central peak near 0.85 (Fig. 10a). Almost all pairs are therefore moderately similar, which is expected: the models share a common observational target and much of their component physics, so their skill vectors point in broadly the same direction. Any threshold placed near the center of this distribution would group most of the ensemble together and would reflect that shared baseline rather than specific relatedness. Taking the upper 3% of pairs, a similarity above 0.927, isolates the tail of the distribution, where the pairs are similar by substantially more than the ensemble-wide baseline.

The resulting clusters support this reading. The most similar pairs are those already known to share components or lineage: EC-Earth3 and EC-Earth3-Veg (0.988), HadGEM3-GC31-LL and UKESM1-0-LL (0.984), CNRM-CM6-1 and CNRM-ESM2-1 (0.978), and CESM2 and CESM2-WACCM (0.975). The threshold recovers the model families independently of any prior knowledge of their genealogy, which is the outcome the criterion is intended to produce.

A less conservative threshold would merge structurally distinct models into shared clusters and so remove candidates from the selection pool without physical justification; a stricter one would resolve only the closest variant pairs and leave the broader families unseparated. We therefore treat 3% as a choice that has to be justified by what it recovers rather than derived, and we note in the revised text that the clusters are stable to moderate changes in the threshold because the tail of the distribution is sparse.

The toolkit at https://cordex-merit.org shows the effect of this choice: the model independence page lets a reader move the threshold and watch the clusters merge or separate, so the sensitivity of the grouping can be judged directly.

19. **Line 649 describes the warming trajectory of some models as *"steady"*, what quantitative analysis supports this conclusion? If it is based on Figure 11b, it is recommended to cite this figure when it is first introduced, rather than deferring the reference to Line 666. Similarly, is the *"trend"* mentioned in Line 715 supported by quantitative calculation? Maybe it's helpful to add relevant statistics directly to Figure 11.**

We thank the reviewer for both points and have addressed them.

On the quantitative basis for ***"steady"*** (L649): The description rested on a comparison of late-century and full-period warming trends, but the manuscript gave neither value. Figure 11a now reports, for each model, the ratio of its late-century trend (2071–2100) to its full-period trend (2015–2100), in a column beside the existing trend column, and the text states the basis of the comparison: ***"In Fig. 11a the models are ordered by their full-period trend, so a model's position gives its warming rate; a second column reports the ratio of its late-century trend (2071–2100) to that full-period trend, which distinguishes warming that is steady from warming that strengthens toward the end of the century."***

The revised text now defines both terms relative to the ensemble and gives the ratio for every model it names: ***"The late-to-full trend ratio averages 1.14 across the 31 models over EUR and spans 0.65 to 1.66. Several models warm steadily on this measure: CanESM5 (1.02), NorESM2-MM (1.02), CESM2 (1.02) and, warming more slowly still in its closing decades, AWI-CM-1-1-MR (0.70). In contrast, a subset shows clear late-century strengthening: FGOALS-g3 (1.66), MPI-ESM1-2-HR (1.57), EC-Earth3 (1.56), UKESM1-0-LL (1.51) and IPSL-CM6A-LR (1.50). Trajectory and warming rate are largely independent: EC-Earth3 has the tenth-strongest full-period trend of the 31 models but the third-highest ratio, while BCC-CSM2-MR ranks eleventh by trend with a ratio of 0.85."***

Two of the original examples do not survive this check and have been changed. MIROC6, cited as warming steadily, has a ratio of 1.23, above the ensemble mean, and has been removed from that list. EC-Earth3-Veg, cited as strengthening late in the century, has a ratio of 1.12, essentially at the mean; EC-Earth3 at 1.56 is the variant that accelerates, and the list has been corrected accordingly. We have also separated UKESM1-0-LL and CanESM5, which the original text grouped together as showing persistently high warming rates: CanESM5 warms almost uniformly (ratio 1.02) while UKESM1-0-LL strengthens markedly late in the century (1.51).

The same exercise revealed a conflation in the following paragraph, which we have corrected: ***"Their trajectories differ, however, even though their warming rates group together. MRI-ESM2-0 (0.65) and CAMS-CSM1-0 (0.94) combine low warming with no late-century amplification, whereas FGOALS-g3 (1.66) and NorESM2-LM (1.50) have among the strongest amplification in the ensemble and reach low end-of-century warming only because their full-period trends are the weakest of the 31. Level of warming and trajectory of warming***

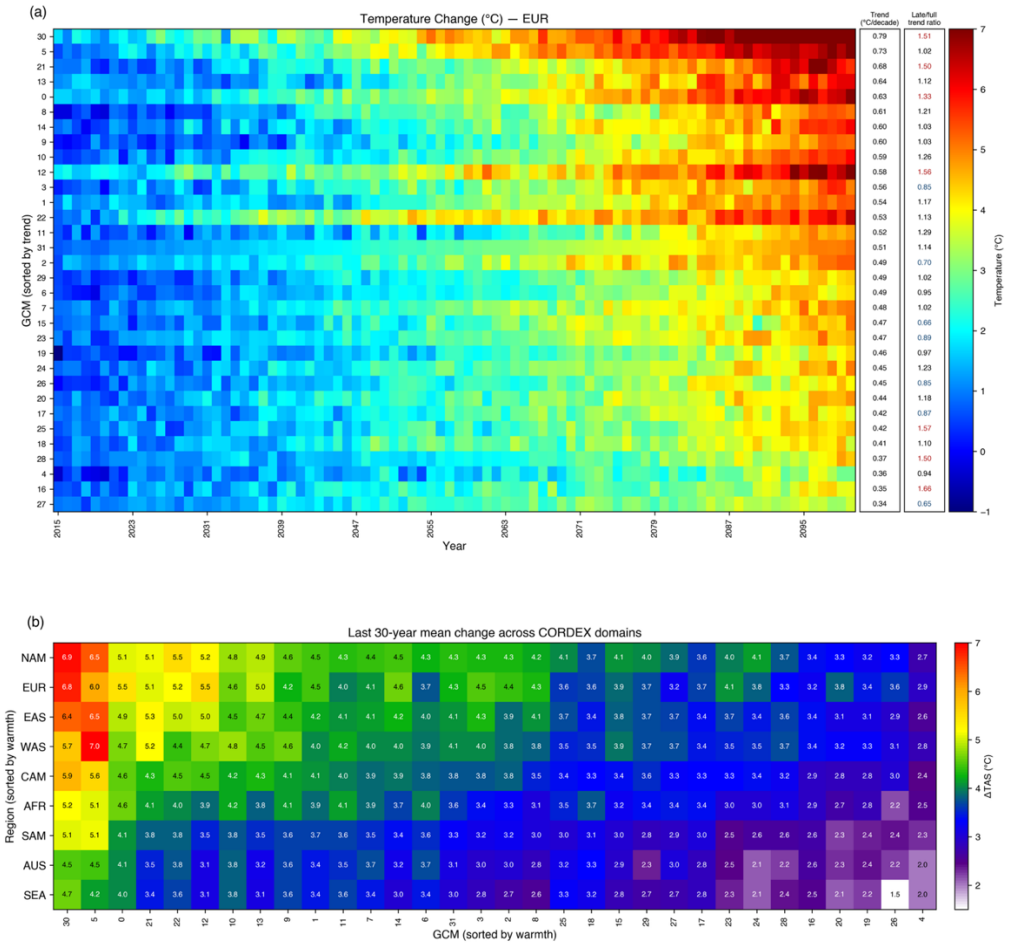
are therefore separate properties, and the ratio column in Fig. 11a allows them to be read independently."

On the figure reference: Figure 11a and 11b are now cited where each is first introduced rather than only at the later mention.

On the ***"trend"*** at L715: This is an ordinary least-squares slope over the period stated. The revised text says so and gives the value rather than referring to a trend in general terms.

On adding statistics to Figure 11: We have added the ratio as a column on Figure 11a rather than in a supplementary table, so the statistic supporting the description sits with the figure the reviewer identified. Because the models in Fig. 11a are ordered by their full-period trend, a model's position already conveys its warming rate; what the ordering cannot show is whether that warming is steady or back-loaded, and the new column supplies exactly that.

> **Revised caption, Figure 11.** *(a) Annual temperature change over EUR for each model, 2015–2100 relative to 1995–2014, with the full-period trend and the late-century-to-full-period trend ratio in columns at the right. Models are ordered by full-period trend. (b) Mean change over the closing 30 years across the nine CORDEX domains. (b) Mean change over the closing 30 years across the nine CORDEX domains.*



20. The WDI is mathematically equivalent to the sum, but this equivalence is not explicitly stated in the paper. This may lead readers to mistakenly regard WDI as an entirely new independent diagnostic. In addition, why choose WDI instead of using the sum or even the mean as in the previous temperature analysis?

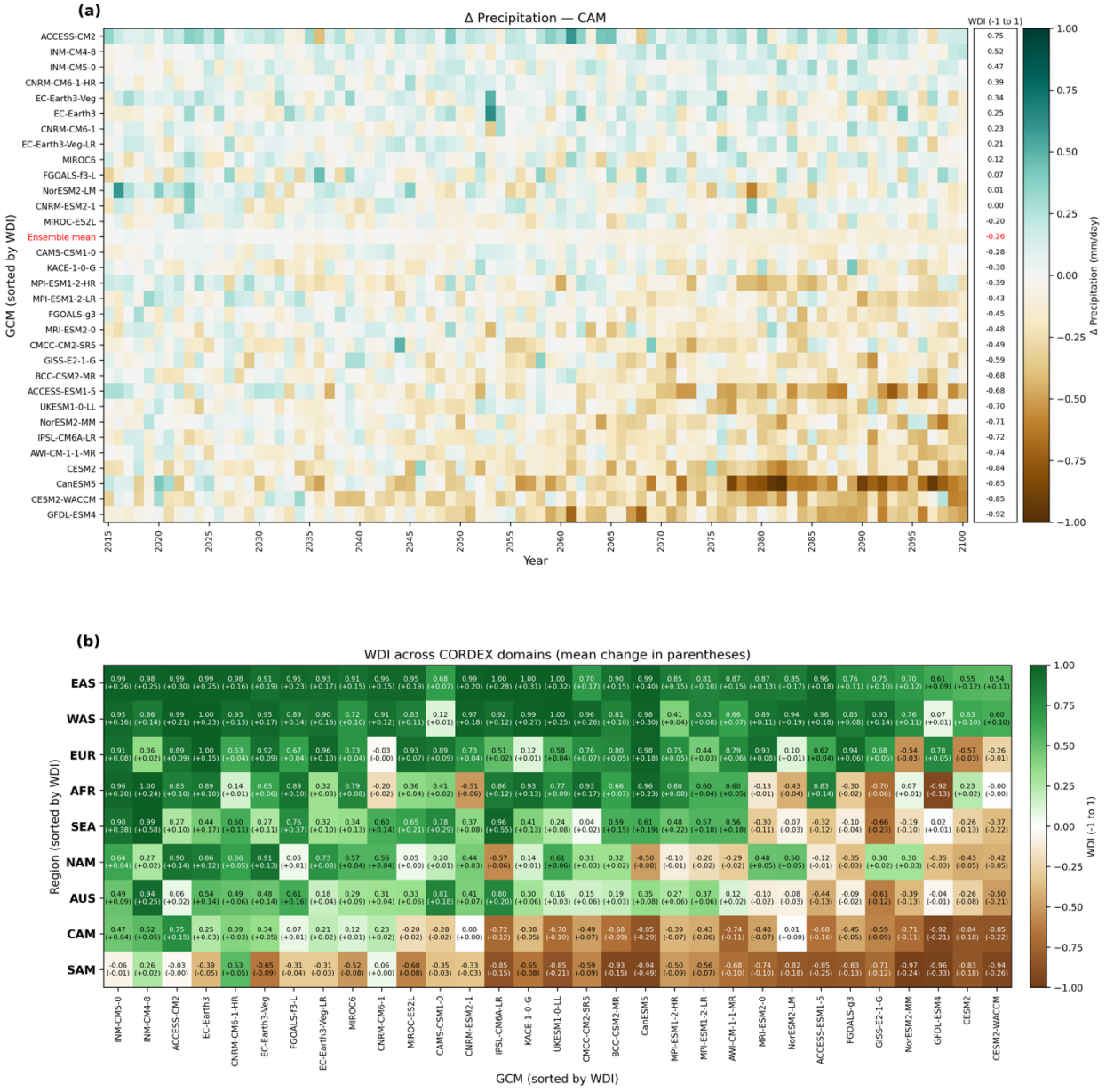
We thank the reviewer for this and agree on both counts.

The reviewer is correct that the index as published was the anomaly sum: Equation 8 gave the numerator only, and the figure it described was computed that way. We have corrected this. All analyses in the revision, including Figure 12, the cross-domain WDI summary and the selection Sankey, use the bounded form, in which the numerator is divided by the total anomaly magnitude $\Sigma^{+} + \Sigma|^{-}|$. Equation 8 has been amended accordingly.

We have also stated the relationship to the sum explicitly, so that the WDI is not mistaken for an independent diagnostic: ***"The numerator is the sum of the annual anomalies, so the WDI always shares the sign of that sum and orders models similarly to it (Spearman $\rho = 0.93$ across the models and domains evaluated here); the index is therefore a rescaling of that sum rather than an independent diagnostic."***

On why a bounded index rather than the sum or the mean, the WDI serves the same purpose for precipitation that late-century mean warming serves for temperature: a single number to support the selection decision. Temperature admits a simple level because the warming signal is monotonic and large relative to internal variability. Precipitation does not: the sign of the response varies between years and between domains, and its magnitude is small relative to its variability, so what a user selecting models needs is how consistently a model leans wet or dry rather than by how much. The denominator also makes the index comparable across domains, since the total anomaly magnitude varies by a factor of twelve across the model–domain combinations evaluated here, from 4 to 51 mm yr^{-1}.The ordering differences between the two measures are consequently modest, $\rho = 0.93$ rather than 1.00, and no model changes the sign of its response between them. We have therefore made the relationship explicit rather than replacing the index, and we note in the text that the sum would give a similar ranking of models within a single domain but not across domains.

Revised caption, Figure 12. *(a) Annual precipitation changes over CAM for each model, with the wet–dry dominance index (WDI) in the column at the right; models are ordered by that index. (b) The WDI index across all nine CORDEX domains, with the mean change in parentheses.*



21. The physical mechanism findings from Section 3.7 should be translated into more actionable guidelines. For example, for thermodynamically dominated regions such as EAS and WAS, covering low, medium, and high-warming tiers in the selected ensemble may naturally capture the main uncertainty in precipitation. But for circulation-dominated regions, the diversity of precipitation responses needs to be preserved separately.

We thank the reviewer for this and agree that the finding should be stated as guidance rather than left as a diagnostic result.

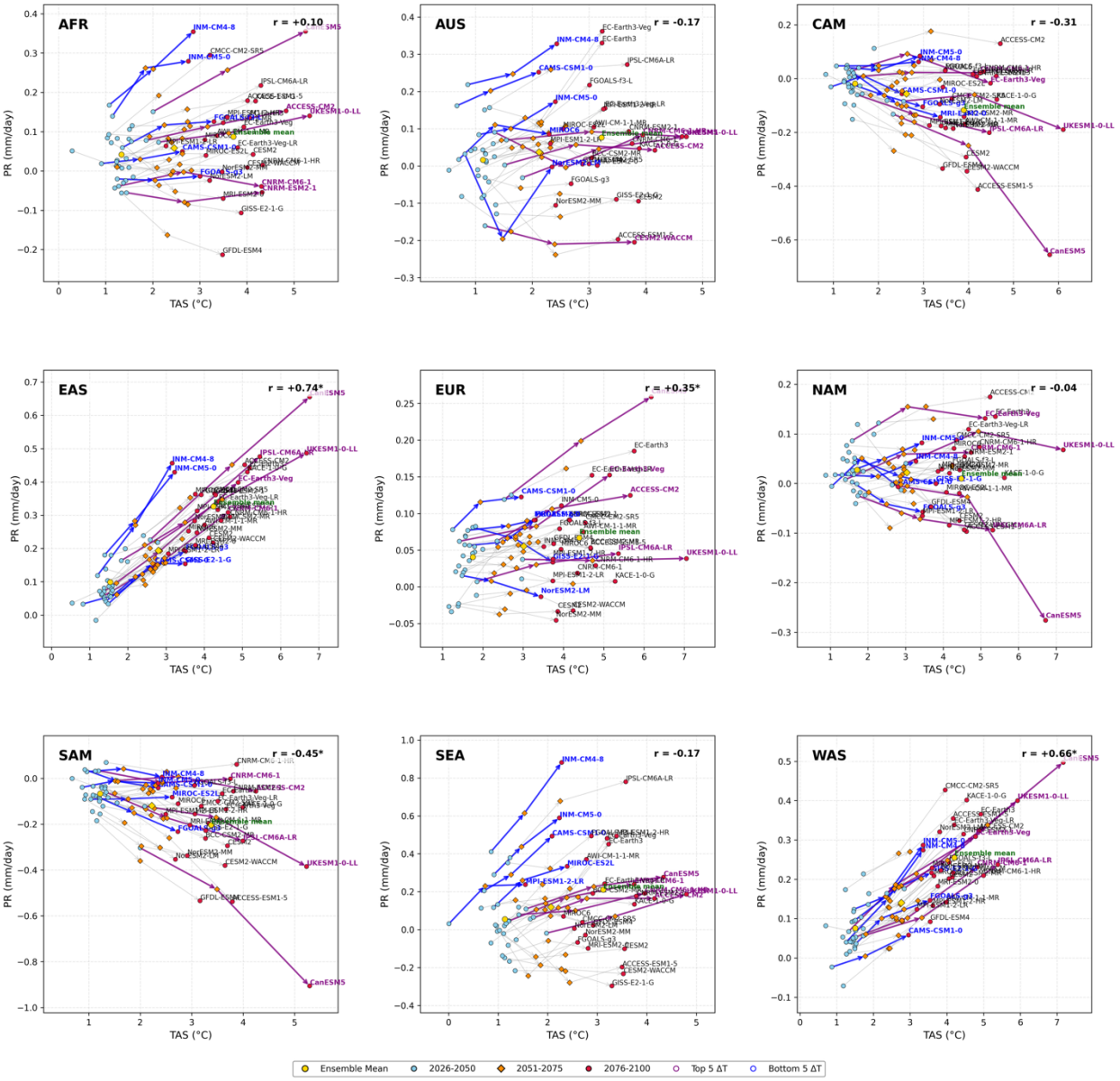
We would first note that the framework does not rely on ΔTAS alone. Precipitation response is already a separate selection criterion, applied independently of the warming tier, precisely because the two cannot be assumed to co-vary. The CORDEX-CORE2 subset illustrates this: it was chosen to contain a dominantly wet model, a mixed one and a dominantly dry one, in addition to spanning the warming range. Section 3.7 explains when that separate treatment is indispensable rather than introducing the need for it.

On the physical result, the coupling is strong in only two of the nine domains, which strengthens the reviewer's point. Correlating the precipitation response against the temperature response across the 31 models at the late-century period (2076–2100) gives +0.74 in EAS and +0.66 in WAS, but |r| at or below 0.35 in the remaining seven: EUR +0.35, AFR +0.10, NAM −0.04, AUS −0.17, SEA −0.17, CAM −0.31 and SAM −0.45, with only EAS, WAS, EUR and SAM significant at the 5 % level. We report the correlation between models at a single period rather than pooling the three 25-year periods of Fig. 13, because the three points belonging to one model trace that model's own warming through the century and are not independent observations. Sampling the warming tiers therefore does not span the precipitation uncertainty in most CORDEX domains. The negative correlations in CAM and SAM matter more still: there the wetter models are the cooler ones, so a subset chosen on warming alone would be systematically biased in precipitation rather than merely uninformative about it.

We have translated this into explicit guidance in the new Section 5.1, which follows the eight procedural steps with the judgement the last of them requires: ***"Where the precipitation response is thermodynamically constrained, as in EAS and WAS, sampling the warming tiers spans most of the precipitation uncertainty as well. Where the response is circulation-dominated or compensating, as in CAM and SAM, or only weakly related to warming, as in AFR, AUS, SEA and NAM, precipitation must be sampled as an independent axis, and the subset should deliberately include both wet- and dry-tending models within the same warming tier."*** Section 5.1 refers the reader to Section 3.7 for the values rather than repeating them, so that the two sections cannot drift apart.

> **Revised caption, Figure 13.** *Joint relationship between projected precipitation and temperature change, one panel per CORDEX domain. Each model contributes three points, for 2026–2050, 2051–2075 and 2076–2100, joined by arrows tracing its trajectory through the century. The correlation annotated in each panel is computed between the 31 models at the late-century period, with an asterisk where it is significant at the 5 % level; values for every*

domain are given in Section 3.7. Axis ranges are set per domain, since the spread of precipitation response differs by up to an order of magnitude between them.



22. The two leftmost columns of Figure 14 are redundant. It can be simplified as models in the upper half can be merged into a single gray branch, while models in the lower half flow into different warming groups. The color of the wet/dry category labels on the far right is not distinguishable, and streamlines already present their categories by color, not sure if labels are still necessary.

We thank the reviewer for these suggestions and have revised Figure 14 accordingly.

On the leftmost columns: The models without SSP3-7.0 output no longer flow into a separate node. In the original figure each of them was connected by a thin ribbon to a ***"No SSP3-7.0***

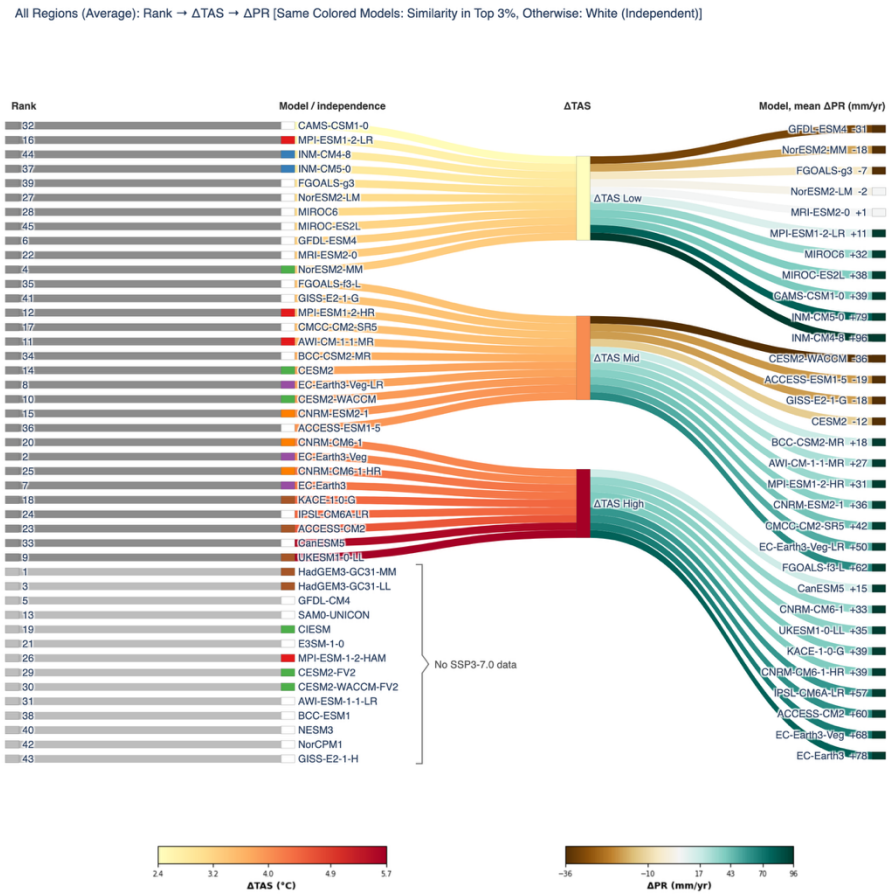
data" node, which added a column of links carrying no information beyond the fact of exclusion. In the revision those models simply terminate at the model column, are grouped together in a block, and are identified by a single brace and label to the right of their names. This removes the redundant ribbons and node while keeping the information that the models were evaluated and excluded only at the data-availability step.

On the ribbon widths: The original figure drew the excluded models with narrower ribbons than the rest, which implied a quantitative difference where none was intended. All ribbons now have uniform width, so width carries no meaning and the reader is not invited to infer one.

On the ordering of the rightmost nodes: The leaves are now grouped by their ΔTAS bin before being ordered within it, so ribbons from one warming group arrive together rather than crossing between groups. This makes the flow from warming tier to precipitation category readable without following individual ribbons.

On the wet/dry labels: We have kept them, and we would argue they remain necessary. The ribbon color encodes each model's ΔPR as a continuous quantity, whereas the label states which of the three terciles the model falls in. These are not the same information: a model near a tercile boundary has a ribbon color close to its neighbor across the boundary but a different categorical assignment, and it is the assignment that determines how the subset samples the precipitation response. We have addressed the legibility problem the reviewer identifies by revising the label text rather than removing it, and we have also corrected an inconsistency in the labelling: a model with a positive mean ΔPR could previously be labelled ***"ΔPR Dry"*** on the basis of its WDI tercile alone, and such cases are now assigned to the middle category so that the label cannot contradict the sign of the mean change.

Revised caption, Figure 14. *Selection flow. Models are ordered by global rank, grouped into independence clusters, then stratified by late-century warming tier and by precipitation response. Models without SSP3-7.0 output terminate at the model column and are marked by a brace. Ribbon colour encodes each model's precipitation change.*



23. **There still should be a concluding section or even paragraph at the end of the paper, simply summarizing the GCM selection into clear step-by-step guidance. Even though a single fixed recommendation list is not required, an example of a standard subset would be greatly helpful.**

We thank the reviewer for this suggestion, which we have adopted. A new Section 5 has been added to the revision, structured around exactly the two things the reviewer asks for.

Step-by-step guidance: Section 5.1 reduces the framework to a numbered sequence of eight steps, from computing the metrics for each AR6 region through to the final tiebreaker: ***"Metrics relevant to each AR6 region are computed for the models and the reference dataset. Each metric is scored using the available statistical measures and combined into a composite skill score. Pairwise correlations among metrics within a region give each metric a uniqueness***

weight. Weighted scores are averaged to an AR6 region score, then to a CORDEX domain score on which models are ranked, and the domain ranks are averaged to a global rank. Cosine distances between models identify independence clusters. Surviving clusters are then stratified by projected warming tier and precipitation response, and the data-availability filter is applied. Where two configurations remain equivalent in performance, independence and projected response, the higher resolution one is preferred."

The same subsection identifies where judgement is required and how it should be exercised, since the steps alone do not determine a subset: where the precipitation response is thermodynamically constrained, sampling the warming tiers also spans most of the precipitation uncertainty, whereas where the response is circulation-dominated or compensating, precipitation must be sampled as an independent axis and the subset should deliberately include both wet- and dry-tending models within the same warming tier.

A worked example: Section 5.2 applies the framework to the CORDEX-CORE2 selection and documents how the three models satisfy each criterion: their global and SSP3-7.0-restricted ranks, their membership of three different independence clusters, their positions in the warming distribution, and the spread of their precipitation responses. It also states what three models cannot do — they do not span the full sensitivity range, and the coolest and warmest parts of that range are unsampled — and names the two models that would be the obvious extensions were the required six-hourly fields available.

We have deliberately not presented this as a recommended list. Section 5.3 sets out what the framework does and does not establish, and Section 5.4 identifies the extensions we regard as most useful, so that the CORDEX-CORE2 subset is read as a worked application of a reproducible procedure rather than as a prescription.

The toolkit at https://cordex-merit.org complements this by making the procedure executable: the framework page presents the calculation as a flow chart in which each step opens the analysis that performs it, and the selection synthesis page traces the CORDEX-CORE2 subset back through the criteria that produced it.