Reviewer #02 — Prof. Subimal Ghosh

This manuscript presents a timely, comprehensive, and transparent framework for selecting CMIP6 GCMs for CORDEX-CORE2 dynamical downscaling. While the individual evaluation criteria are not novel, their systematic integration into a reproducible multi-criteria framework is a significant contribution. The rigorous assessment of model performance, independence, metric weighting, and ranking sensitivity makes this work particularly valuable ahead of IPCC AR7. The manuscript is technically sound, well-organised, and clearly written. I particularly commend the authors for their transparent evaluation strategy, which provides an objective and robust foundation for future coordinated regional climate modeling efforts. I suggest minor revision following some discussions on the following points:

We thank Prof. Ghosh for the careful reading and for the positive assessment of the framework, and for reviewing openly under his own name. We are grateful for the observation that the contribution lies in the systematic integration of the criteria rather than in their novelty, which is how we see it too. The four points raised are all ones we agree with, and each has been addressed by new text in the revision. We respond to them 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 that the toolkit lets a reader answer directly, we say so.

1. **Process-based evaluation: Consider briefly discussing if future work on this framework could incorporate process-based evaluation (e.g., monsoon dynamics, land–atmosphere coupling, ENSO, storm tracks) in addition to reproducing observed climatology.**

We agree and note that the framework already contains a process-oriented component. Monsoon onset and demise, jet position and strength, vertical wind shear, ITCZ latitude, and the major modes of variability (ENSO Niño-3.4, NAO, SAM, IOD) are process diagnostics rather than climatological means, and they receive the highest uniqueness weights, close to unity in every domain (Fig. 5), so they exert strong influence on the rankings. We agree this could be extended: land–atmosphere coupling, storm-track statistics, teleconnection fidelity and moisture-transport diagnostics are natural additions, most of which require sub-monthly data and so sit outside the present configuration.

We have added this to the new Section 5.4, which reads:

"Two refinements to the metric suite are identified for future iterations. 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. And process-oriented diagnostics such as land–atmosphere coupling, storm-track statistics and teleconnection fidelity would strengthen the process-based content of the evaluation, though most require sub-monthly output and so lie outside the present configuration."

2. ***"Right evaluation for the right reason"***: A brief acknowledgement (if true) that good historical performance may sometimes arise from compensating errors or unrealistic process representations.

We accept this. Skill scores computed against ERA5 cannot distinguish a model that is right for the right reasons from one whose errors cancel, and a high composite score is compatible with offsetting biases in individual processes. Our multi-metric design reduces but does not eliminate the risk: a model must perform well across many partly independent diagnostics, which makes systematic compensation less likely than under a single-metric ranking, but process-level attribution would be required to rule it out.

We have added this to the new Section 5.3. While doing so we also noted a related caveat Prof. Ghosh did not raise but which bears on the same argument, namely that ERA5 is our sole reference dataset, which limits how far compensating errors can be detected and matters particularly for precipitation over the tropics. The added passage reads:

"First, skill scores computed against a single reference dataset cannot distinguish a model that is right for the right reasons from one whose errors offset. A high composite score is compatible with compensating biases in individual processes. The multi-metric design reduces this risk, since a model must perform well across many partly independent diagnostics, but process-level attribution would be required to exclude it. The use of ERA5 as the sole reference compounds this, particularly for precipitation over the tropics."

3. **Historical performance versus future credibility: It would be useful to note that strong historical performance does not necessarily guarantee more credible future projections, as future responses also depend on the representation of climate feedbacks and other emergent processes.**

We agree and are happy to state it plainly. Present-day skill is a necessary but not sufficient condition for credible projections, since the future response depends on feedbacks and processes that the historical record constrains only weakly.

This is in fact the reason our framework does not select on performance alone. Independence, temperature sensitivity and precipitation spread are included precisely so that the retained subset spans the response space rather than collapsing onto the best historical performers, and the

CORDEX-CORE2 selection now described in Section 5.2 shows what this means in practice. The three models chosen are one representative from each of the three highest-performing independence clusters, and they were chosen because together they span the response space: EC-Earth3-Veg in the upper part of the sampled warming range with wetting in eight of nine domains, MPI-ESM1-2-HR in the middle with a mixed response, and NorESM2-MM at the lower end drying in five domains. The framework's design is therefore a direct response to the limitation Prof. Ghosh identifies rather than something we add as a caveat.

We have made this logic explicit in the new Section 5.3, which reads:

"Second, present-day skill is necessary but not sufficient for credible projections, because the future response depends on feedbacks that the historical record constrains only weakly. This is precisely why the framework does not select on performance alone: independence, temperature sensitivity and precipitation spread are included so that the retained subset spans the response space rather than collapsing onto the best historical performers. The CORDEX-CORE2 selection illustrates the point, since the three models are not simply the three best performers but one representative from each of the three highest-performing clusters, chosen to span the range of projected response."

4. **Emergent constraints: The authors may briefly comment on whether future work on this framework could benefit from emergent constraints or other physically based approaches linking present-day model performance to future climate projections.**

We agree this is a promising complement. The Introduction already motivates the framework partly by reference to emergent constraints (Hall et al., 2019; Brient, 2020; Shiogama et al., 2022), but we do not return to the idea. We now close the loop: emergent constraints could be integrated as an additional criterion that weights or filters models by their position relative to an observationally constrained relationship, most readily for regional precipitation (Shiogama et al., 2022) and climate sensitivity. We note the practical caveats, namely that constraints are typically variable- and region-specific, can conflict, and would compromise the single cross-domain subset that our design requires, and therefore propose them as a refinement within an already-screened candidate pool rather than as a replacement criterion.

We have added this to the new Section 5.4, which reads:

"Emergent constraints offer a natural complement. The framework already rests on the premise that observable behavior carries information about future response, and constraints could be integrated as an additional criterion, weighting or filtering models by their position relative to an observationally constrained relationship, most readily for regional precipitation and climate sensitivity. The practical obstacles are that constraints are variable- and region-specific and can conflict, which sits awkwardly with the requirement for a single cross-domain subset. We therefore see them as a refinement applied within an already-screened candidate pool rather than as a replacement criterion."