Reviewer Comments on "AIMIP Phase 1: Systematic Evaluations of AI Weather and Climate Models"

By Brian Henn, Christopher S. Bretherton, Nikolay Koldunov, Christian Lessig, Maria J. Molina, Troy Arcomano, Oliver Watt-Meyer, Guillaume Couairon, Renu Singh, Robert Brunstein, Yana Hasson, Antonia Jost, Noah Brenowitz, Peter Manshausen, Nathaniel Cresswell-Clay, Dale Durran, Kyle Joseph Chen Hall, Janni Yuval, Dmitrii Kochkov, Stephan Hoyer, and Ignacio Lopez-Gomez

This manuscript presents the first phase of the AI Model Intercomparison Project (AIMIP) and introduces a timely and potentially valuable AMIP-style framework for evaluating AI-based weather and climate models. The common experimental protocol, CMIP-compatible output conventions, and publicly available multi-model dataset provide a useful foundation for systematic evaluation by the climate-science community, and the manuscript presents informative initial results across several aspects of climate behavior. However, important questions remain regarding the scientific scope of Phase 1, the comparability of the participating systems, and the interpretation of the evaluation results and conclusions. Additional clarification and revision are therefore needed before the manuscript can be fully assessed. I recommend major revisions, and my detailed comments are listed below.

Major Comments

- The primary scientific target and common basis for comparison in AIMIP Phase 1 are not sufficiently clear. In contrast to traditional AMIP and other CMIP experiments, which are generally organized around a clearly defined scientific question and associated criteria for success, the manuscript presents AIMIP Phase 1 as serving several purposes: establishing a standardized data and evaluation framework, benchmarking current AI weather and climate models, assessing long-term stability and fidelity to ERA5, testing responses to altered SST forcing, and identifying the influence of AI architectures and modeling frameworks. These are all potentially valuable objectives, but they are scientifically distinct and require different experimental designs and levels of control. The current framework successfully establishes a common prescribed-SST experiment, standardized outputs, and a public multimodel dataset, and it provides a useful exploratory comparison of existing systems. However, it does not yet define clearly what common capability is being evaluated across the broad range of participating model classes or whether the selected metrics have an equivalent interpretation for all systems. Because architecture is strongly confounded with training strategy, temporal resolution, prognostic-variable choice, diagnostic output construction, ensemble generation, preprocessing, post-processing, and intended model purpose, the present design supports comparison of complete modeling systems under a harmonized protocol but does not isolate causal effects of architecture. The authors should state one primary scientific question for Phase 1, identify secondary objectives separately, define criteria for success, explain how each protocol element and metric addresses those objectives, and clarify which diagnostics are meaningful across all model classes. A model-comparability table summarizing the major system differences would be very helpful.
- The training protocol in Section 2.2.1 is not fully consistent with the model-specific configurations described in Section 3. Section 2.2.1 states that participating models must be trained exclusively on ERA5 from 1979–2014 and identifies 2015–2024 as a common out-of-sample test period. However, individual models use different training and validation subsets within 1979–2014, while

DLESyM was trained through June 2016 and therefore includes part of the nominal test period. The authors should clarify whether the protocol requires use of the full 1979–2014 period for training or only prohibits use of data outside that interval. They should summarize the training, validation, tuning, and test periods for every submission in a common table, identify protocol deviations explicitly, and explain their implications for the intercomparison. Strictly compliant and noncompliant models should be distinguished in analyses of test-period performance, and the term "out-of-sample" should not be applied uniformly where that condition is not satisfied.

- The interpretation of model performance and the corresponding conclusions should be stated more cautiously and precisely. Section 5 appropriately acknowledges several limitations, including that the AI systems are evaluated against ERA5, on which they were trained, whereas GFDL-CM4 was not tuned to ERA5. However, these caveats are not reflected consistently in the interpretation of the results or in Section 6. GFDL-CM4 is represented by only one realization and is unavailable for the 2015–2024 test period, so it should be treated as a contextual reference rather than as a representative benchmark for physics-based climate models. Lower error relative to ERA5 demonstrates closer agreement with the training and evaluation target, but it does not by itself establish greater physical fidelity, predictive skill, or climate-projection capability. The trend analysis also requires additional uncertainty assessment because the test period contains only ten annual values and the models omit evolving anthropogenic radiative forcings; confidence intervals, ensemble uncertainty, sensitivity to the analysis period, and statistical significance should be reported. Likewise, the ENSO analysis should be described specifically as the atmospheric response to prescribed ENSO-related SST variability, and the daily-variability metric as an assessment of daily variance magnitude rather than a comprehensive evaluation of weather dynamics or extremes. Accordingly, statements in Section 6 that the AIWCMs more faithfully reproduce ERA5 climate patterns than a conventional CMIP6 model and show a strong ability to simulate the ENSO response remain broader than the evidence directly supports and should be reformulated more narrowly. Selected independent observational products and, where feasible, a broader AMIP multimodel range would provide more appropriate context.
- Protocol differences and physical robustness should be incorporated more explicitly into the evaluation and interpretation. Section 5 appropriately recognizes that the +2 K and +4 K experiments are not implemented identically across models because SST values over land and sea ice are filled, interpolated, merged, and perturbed differently, and it notes that these choices may contribute to the spread in Figure 9. However, the discussion remains qualitative and does not establish how much of the spread arises from implementation differences versus genuine out-of-distribution behavior. The authors should document the effective perturbation received by each model, discuss or quantify the sensitivity of the results to these choices, and distinguish numerical instability from stable but physically unrealistic responses. More broadly, the five core evaluations emphasize fieldwise statistical agreement and provide limited assessment of physical consistency. The manuscript should discuss whether additional checks of conservation, budgets, balance relationships, and long-term drift are needed, particularly because some evaluated variables are generated through diagnostic or downscaling components rather than belonging to the prognostic state. Finally, the effects of native resolution, learned downscaling, and different regridding methods should be evaluated where possible or identified explicitly as sources of uncertainty in the cross-model comparison.
- The evaluation does not yet cover the full range of capabilities stated in the AIMIP Phase 1 goals. Sections 1 and 2 identify weather phenomena and extremes as explicit targets of AIMIP, and

Section 2.3.4 states that the requested daily output is intended to enable assessment of weather variability and extremes. However, the daily evaluation presented in Benchmark E4 is limited primarily to the temporal standard deviation of daily anomalies during 1979 and errors in dry-day fraction. These diagnostics provide useful information about bulk variability amplitude and precipitation occurrence, but they do not directly evaluate extreme-event statistics, weather regimes, event persistence, or the temporal and spatial organization of weather systems. The manuscript should therefore either narrow the stated scope of the evaluation or incorporate a limited set of process- and event-oriented diagnostics, such as upper-tail quantiles of daily temperature and precipitation, exceedance frequency and duration, temporal autocorrelation, or selected event-based analyses. The authors should also distinguish clearly between analyses that are enabled by the released AIMIP dataset and capabilities that are actually demonstrated in this paper.

Minor Comments

- Figure 1: Please clarify the meaning of all empty panels directly in the figure caption. The empty precipitation panels for ArchesWeather and ArchesWeatherGen appear to reflect the absence of submitted surface-precipitation output. The two empty GFDL-CM4 panels appear to correspond to the 2015–2024 test period, for which GFDL-CM4 output is unavailable because the simulation ends in 2014. This should be stated explicitly so that readers can distinguish missing variables from unavailable time periods.
- Section 3.3, final paragraph: In the phrase “with a larger half-lifer corresponding to more variability,” should “half-lifer” be changed to “half-life”?
- Section 4.3, second-to-last paragraph: The sentence “The GFDL-CM4 temperature trend pattern error magnitude are comparable to those of the AIWCMs that have smaller trend pattern errors …” is grammatically awkward. Please revise for clarity. For example: “The GFDL-CM4 temperature trend-pattern error is comparable in magnitude to those of the AIWCMs with smaller trend-pattern errors (NeuralGCM, cBottle1.3, DLESyM, and ArchesWeatherGen).”
- Section 4.6, second paragraph: In the sentence “However, ACE2.1-ERA5, cBottle1.3, and MD-1.5 v0.9 all both underestimate the ocean warming magnitude and implausibly …,” the phrase “all both” is redundant and reads awkwardly. Please revise, for example: “However, ACE2.1-ERA5, cBottle1.3, and MD-1.5 v0.9 all underestimate the magnitude of ocean warming and implausibly …”
- The manuscript should state explicitly and consistently for each figure and metric whether the presented results represent individual ensemble members, ensemble means, ensemble medians, or another summary. At present, ensemble treatment varies across the analysis without a clear rationale: Figure 1 presents a single ensemble member, Figure 3 shows ensemble means, and Figure 8 refers to the median realization for cBottle1.3, while the aggregation used in several other diagnostics, including Figures 4–7 and 9, is not clearly identified in the captions or accompanying text. For Figure 1, the authors should explain why a single realization is shown instead of the ensemble mean, describe how that realization was selected, and confirm that it was not chosen post hoc based on performance. For Figure 3, displaying ensemble spread around the ensemble-mean time series would help readers assess whether differences among models and relative to ERA5 exceed within-model variability. Comparable uncertainty information or summary

measures of ensemble spread should also be provided for other ensemble-based diagnostics where feasible.

- The manuscript should distinguish more systematically among native prognostic variables, vertically interpolated fields, learned diagnostics, and learned downscaled products, because these may not be directly equivalent for evaluation. For example, ACE2.1-ERA5 uses a secondary decoder for several near-surface and pressure-level fields, NeuralGCM-HRD applies learned high-resolution downscaling to a coarser prognostic state, cBottle1.3 interpolates selected pressure levels, and MD-1.5 v0.9 decodes physical variables from a learned latent state. These pathways should be summarized in a model-comparability table, together with discussion of whether they affect reported errors, rankings, variability, or trends. The manuscript should also clarify when a metric primarily evaluates an output-generation component rather than the underlying prognostic evolution. Finally, the treatment of below-ground pressure levels should be explained more clearly, including the implications of applying a common mask based on GFDL-CM4.
- Please confirm that precipitation accumulation, sampling, and temporal-averaging conventions are consistent across submissions. Because the models operate at different temporal resolutions and some diagnose precipitation through separate output components, the manuscript should document whether all submitted daily and monthly precipitation fields represent comparable time-integrated fluxes, how accumulation windows are handled, and whether any conversions or resampling could affect the precipitation bias and variability diagnostics.
- Models that do not fully comply with the training or experiment protocol should be identified directly in relevant figures, captions, or legends. In particular, test-period figures should mark submissions for which 2015–2024 is not a fully independent holdout period, so that readers do not interpret all models as being evaluated under identical out-of-sample conditions.
- To support reproducibility, the data and code availability section should provide persistent dataset identifiers, exact model or checkpoint versions, evaluation-code version information, and a concise workflow for reproducing the principal figures. Where repositories are used, release tags or commit hashes should be reported.