

Review of “Evaluating Skill and Stability of ArchesWeather and ArchesWeatherGen under Multi-Decadal Climate Simulations” by Renu Singh, Robert Brunstein, Antonia Jost, Yana Hasson, Thomas Rackow, Claire Monteleoni, Christian Lessig, Guillaume Couairon

Recommendation: **Major revision**

This manuscript extends the ArchesWeather AI weather models, chosen due to their relatively efficient training costs compared to other AI models, to execute as climate models on long historical and warmed-SST simulations. The ML architecture is not significantly changed from weather prediction applications, with the major differences being training on “forcings” (SST and sea ice). The resulting climate model produces a highly-realistic emulation of the ERA5 climate in terms of both means and variability, and impressively produces viable +2K and +4K SST simulations despite these being well out-of-sample.

The model is limited by the fact that, like many other climate emulators, it is only 1-degree resolution, does not include GHGs as a forcing, and lacks precipitation and radiation outputs, which leading climate emulators are now producing with a good level of fidelity. This does mean that the design and training of the emulator is relatively simple, and so is useful for a wide range of climate research that may not need high resolution or additional climate variables. There are a few explanations that were unclear to me, and some additional model realism checks should be added to the paper before it is published.

A challenge for emulated climate models is being able to maintain mass and energy conservation over a long integration. While a total energy budget is beyond the scope of this paper or ability of the emulator, it would be useful to ensure that (a) dry atmospheric mass is roughly conserved during a multi-year simulation; and (b) precipitable water trends are within expected bounds. Could the authors look into both of these?

I didn’t entirely understand the methodology for ArchesWeather-Mx4. For both training and rollout, four forecasts (each deterministic but using a different initial seed) are averaged after each daily step, which is used to compute the residuals (for training) or the inference (during rollout). It is said that doing this averaging at each step reduces the biases of the individual forecasts; how then does this average get used to take the following daily step, without losing the information from the individual forecasts?

I also didn’t understand the difference between ArchesWeather and ArchesWeatherGen, the latter being said to be “inherently stochastic”. What is the difference between the two—does it depend on an ensemble loss function, or is stochasticity included through other means?

Minor comments:

- Section 2: One coupled model worth mentioning is SamudrACE, which couples the ACE atmospheric emulator to the Samudra ocean emulator:

Duncan, J. P. C., Wu, E., Dheeshjith, S., Subel, A., Arcomano, T., Clark, S. K., et al.

(2026). SamudrACE: Fast and accurate coupled climate modeling with 3D ocean and atmosphere emulators. *Geophysical Research Letters*, 53, e2025GL119340.
<https://doi.org/10.1029/2025GL119340>

- Section 3.3: What are the vertical levels used by ArchesWeather? What is done in areas where the pressure surface is below topography? Also, are the output daily SST and sea ice prognostic variables or just interpolated from input monthly means?
- Section 3.4: “We compute RMSE on an aggregated climatology rather than on daily snapshots...” is a roundabout way of saying that you are simulating the climate and not the weather due to the limits of predictability of the latter.
- Figure 3: The large biases over the southern ocean and Antarctica is a bit disappointing due to the relatively good simulation of the tropics and northern extratropics, and for the realistic SAM from ArchesGen. This should be relatively easy for data-driven models to pick up and is also easily tunable in physical climate models (typically through a gravity-wave drag parameterization). Can the authors comment why this bias is emerging?
- Figure 5, and subsequent figures: I liked the addition of the “unforced” model to show the impact of adding forcings to ArchesWeather.
- Section 4.4: Would we ever expect the unforced model to capture ENSO responses in SLP?
- Figure 11: I didn’t particularly see the value of the monsoon analysis especially given the lack of precipitation output. It might be easier to see the result of at least the flow field if errors with respect to ERA5 were shown.
- Section 4.7: Out-of-sample results are impressively stable, but take care that there aren’t odd responses in precipitable water or stratospheric temperatures.