Application of the AI2 Climate Emulator to E3SMv2's global atmosphere model, with a focus on precipitation fidelity

James PC Duncan[1], Elynn Wu[2], Jean-Christophe Golaz[3], Peter Martin Caldwell[3], Oliver Watt-Meyer[2], Spencer Koncius Clark[4], Jeremy McGibbon[2], Gideon Dresdner[2], Karthik Kashinath[5], Boris Bonev[5], Michael S Pritchard[6], and Christopher S. Bretherton[2]

[1]University of California, Berkeley
[2]Allen Institute for Artificial Intelligence
[3]Lawrence Livermore National Laboratory (DOE)
[4]Allen Institute for Artificial Intelligence / NOAA-GFDL
[5]NVIDIA
[6]University of California, Irvine and NVIDIA

April 05, 2024

Abstract

Can the current successes of global machine learning-based weather simulators be generalized beyond two-week forecasts to stable and accurate multiyear runs? The recently developed AI2 Climate Emulator (ACE) suggests this is feasible, based upon 10-year simulations trained on a realistic global atmosphere model using a grid spacing of approximately 110˜km and forced by a repeating annual cycle of sea-surface temperature. Here we show that ACE, without modification, can be trained to emulate another major atmospheric model, EAMv2, run at a comparable grid spacing for at least ten years with similarly small climate biases. ACE accurately reproduces EAMv2's frequency distribution of daily-mean precipitation, its time-mean spatial pattern of precipitation, and its space-time structure of tropical precipitation, including the Madden-Julian Oscillation. Moreover, ACE's climate biases with respect to EAMv2 are substantially smaller than EAMv2's own biases compared to the observed historical average surface precipitation rate and top-of-atmosphere radiative fluxes.

Application of the AI2 Climate Emulator to E3SMv2's global atmosphere model, with a focus on precipitation fidelity

James P. C. Duncan[1], Elynn Wu[2], Jean-Christophe Golaz[3], Peter M. Caldwell[3], Oliver Watt-Meyer [2], Spencer K. Clark[2], Jeremy McGibbon[2], Gideon Dresdner[2], Karthik Kashinath[4], Boris Bonev[4], Michael S. Pritchard[4], and Christopher S. Bretherton[2]

[1]University of California, Berkeley
[2]Allen Institute for Artificial Intelligence (AI2), Seattle, USA
[3]Lawrence Livermore National Laboratory, Livermore, California, USA
[4]NVIDIA, Santa Clara, California, USA

Key Points:

- The ACE weather-climate emulator yields an accurate climate when trained on EAMv2, E3SMv2's global atmosphere model.
- Time-mean biases vs. EAMv2 in diverse atmospheric fields are similar to those seen before for ACE applied to the FV3GFS atmospheric model.
- ACE captures the space-time organization of EAMv2 precipitation well, with a much smaller time-mean bias than EAMv2's observational bias.

Corresponding author: James Duncan, jpduncan@berkeley.edu

Abstract

Can the current successes of global machine learning-based weather simulators be generalized beyond two-week forecasts to stable and accurate multiyear runs? The recently developed AI2 Climate Emulator (ACE) suggests this is feasible, based upon 10-year simulations trained on a realistic global atmosphere model using a grid spacing of approximately 110 km and forced by a repeating annual cycle of sea-surface temperature. Here we show that ACE, without modification, can be trained to emulate another major atmospheric model, EAMv2, run at a comparable grid spacing for at least ten years with similarly small climate biases. ACE accurately reproduces EAMv2's frequency distribution of daily-mean precipitation, its time-mean spatial pattern of precipitation, and its space-time structure of tropical precipitation, including the Madden-Julian Oscillation. Moreover, ACE's climate biases with respect to EAMv2 are substantially smaller than EAMv2's own biases compared to the observed historical average surface precipitation rate and top-of-atmosphere radiative fluxes.

Plain Language Summary

Traditional methods to predict the weather use mathematical models of the Earth's atmosphere that are costly to run. However, "data-driven" weather prediction methods, which learn to predict future weather directly from data on past weather, have come to match or even beat traditional methods and do so with much less running cost. In contrast to weather prediction where the goal is to predict the weather in the near future, in *climate modeling* the goal is to study the Earth's long-term weather trends under different possible future scenarios for many years into the future. Until the introduction of the AI2 Climate Emulator (ACE), a recent data-driven method for climate modeling, no data-driven method could match traditional climate models. In this work we test ACE's climate modeling skills and find that it is able to faithfully mimic a traditional model of the climate when looking at patterns of rainfall around the globe and in the tropics. With ACE, we can study the potential future of Earth's climate under many more scenarios and with much lower cost than ever before.

1 Introduction

In recent years, the field of numerical weather prediction has undergone a significant transformation, with researchers and institutions worldwide embracing machine learning (ML) based techniques to make weather forecasts (Pathak et al., 2022; Lam et al., 2023; Bi et al., 2023; Ben-Bouallegue et al., 2023). Notably, the European Centre for Medium-Range Weather Forecasts (ECMWF) unveiled an Artificial Intelligence based Forecasting System (AIFS) as a new companion to their physics-based numerical weather prediction model (IFS). The shift from solely physics-based numerical weather prediction to integrating ML-based systems has sparked considerable excitement within the scientific community. While most studies have focused on short to medium-range weather forecasts (up to 14 days), the AI2 Climate Emulator (ACE) has demonstrated the ability to emulate an existing global atmosphere model, FV3GFS, at climate timescales (Watt-Meyer et al., 2023) by accurately simulating weather variability and deriving climate from the statistics of the simulated weather, as do conventional global climate models. For this reason we call ACE a weather-climate emulator, to distinguish it from much simpler surrogate models that bypass weather simulation. Such models can instead be based on global or large-scale budget equations, e.g. the Model for the Assessment of Greenhouse-Gas Induced Climate Change (MAGICC) (Meinshausen et al., 2011) used in IPCC assessment reports (e.g. Sec. 8.8.2 of IPCC (2013)), in which a few parameters are tuned to give the same climate sensitivity, ocean heat uptake, and other salient global properties as a target global climate model. Alternatively, ML-based surrogate models such as ClimaX (Nguyen et al., 2023) directly predict monthly climate evolution.

ACE approximately conserves mass and moisture, and accurately predicts the climatology of key variables throughout the depth of the atmosphere. ACE can make a decade-long simulation in one hour of wall clock time of one A100 GPU, making it 100 times faster and more energy-efficient than FV3GFS run at a similar grid spacing.

Inspired by the achievements of ACE, in this paper we investigate its generalizability to emulating a different global atmosphere model, the E3SM Atmosphere Model version 2 (EAMv2). EAMv2 is the atmospheric component of the U.S. Department of Energy's Energy Exascale Earth System Model version 2 (E3SMv2) (Golaz et al., 2022). As configured for this study, EAMv2 fluid dynamics uses a grid spacing of approximately 110 km, like the FV3GFS implementation used for ACE. While FV3GFS is based on a finite-volume dynamical core with 64 vertical layers, EAMv2 uses a spectral-element approach with 72 layers while other processes use a finite-volume grid that divides each element into 2×2 cells of equal size, giving a horizontal resolution of 165 km (Hannah et al., 2021). The physical parameterizations of EAMv2 are also substantially different than those of FV3GFS.

We also analyze the emulation of precipitation in more detail than Watt-Meyer et al. (2023), including its time-mean geographic distribution, its frequency distribution of daily variability, and its organization in the tropics. A final goal of this work is to bring awareness of ACE and ML-based climate emulation into the traditional climate modeling literature.

2 Data and Methods

2.1 EAMv2 Dataset

Our training data is derived from 6-hourly outputs of a 73-year simulation of EAMv2, a model described in detail in Section 2.1 of Golaz et al. (2022). The simulation is configured to run with the "F2010" component set[1], forcing the model with perpetual 2010 greenhouse gas concentrations and emissions of aerosols and precursors, along with an annually repeating cycle of sea surface temperature and sea ice derived from the observed 2005-2014 average. The initial 11 years are discarded as spinup because the EAMv2 stratosphere is equilibrating; the following 42 years are used for training; the subsequent 10 years are used for validation; and the final 10 years are reserved for evaluating EAMv2's internal decadal variability. This simulation is performed on the E3SM Chrysalis cluster, achieving 24 simulated years per day using 30 nodes. See Text S2 for a comparison of the computational efficiencies of EAMv2 and ACE.

We make several other design choices following ACE (Watt-Meyer et al., 2023). First, we perform a conservative regridding from the native EAMv2 output to a 1° Gaussian grid to ensure compatibility with the underlying Spherical Fourier Neural Operator (SFNO) architecture (Bonev et al., 2023). Second, we filter the data with a spherical harmonic transform (SHT) round-trip to help eliminate artifacts in the high latitudes. Third, to reduce the emulator's memory footprint, we coarsen the vertical model-level coordinate from the native 72 down to 8 layers. For more details see Table S2.

2.2 ACE Training Overview

As described by Watt-Meyer et al. (2023), ACE is a modified version of NVIDIA's open-source FourCastNet global atmospheric emulator (Pathak et al., 2022) that employs the SFNO architecture for efficient spatial information exchange (Bonev et al., 2023). Much as traditional physics-based numerical models of atmospheric dynamics recursively step forward the atmospheric state X_t at time t, ACE is trained to autoregressively gener-

[1] `https://acme-climate.atlassian.net/wiki/spaces/DOC/pages/961250902/F2010C5-CMIP6-LR`

ate predictions of the atmospheric state at time $t + \delta t$: $\hat{X}_{t+\delta t}$. We use $\delta t = 6$ hours and minimize the average "one-step" loss over a random batch $\mathcal{B}$ of initial condition times t:

$$\frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \frac{\|\hat{X}_{t+\delta t} - X_{t+\delta t}\|_2}{\|X_{t+\delta t}\|_2}$$

Whereas FourCastNet uses identical input and output variables and trains a separate model to predict diagnostic variables (Pathak et al., 2022), ACE uses a set of prognostic variables which are both inputs and outputs, a set of specified forcing input variables such as insolation and sea surface skin temperature which are exogenous to the dynamical system, and a set of diagnostic variables which are incorporated in the training loss but are output-only. This and a variety of other improvements enable ACE, unlike past weather emulators, to produce stable, skillful, more interpretable multiyear emulations of the target model. For more details see Table S3, Watt-Meyer et al. (2023), and Bonev et al. (2023).

3 Results

Watt-Meyer et al. (2023) provide a holistic evaluation of ACE's physical consistency when trained on 100 years of FV3GFS simulation outputs in terms of physical budgets and time- and global-mean biases and pattern errors.

Section 3.1 shows a similar analysis of ACE's global- and time-mean absolute bias and root mean square error (RMSE) metrics on EAMv2. This analysis shows that ACE produces a similarly high-quality emulation of the climatology of EAMv2 as for FV3GFS, demonstrating that ACE's training methodology generalizes across reference models of comparable grid resolution with different dynamical cores and physical parameterizations. In the remainder of Section 3, we present some key metrics of how well ACE emulates EAMv2's precipitation variability over the 10 year validation period, a topic not documented in detail by Watt-Meyer et al. (2023).

3.1 Global- and time-mean biases and RMSE

In Figure 1, we compare ACE's climatological skill to that of an unseen EAMv2 reference dataset, years 64–73 of the EAMv2 simulation run. Both ACE and the reference are evaluated against the validation target years 54–63. The reference values give a 'noise floor' estimate, computed as the difference of time means from a single pair of ten-year segments of the reference simulation. Different pairs of ten-year periods would give different estimates for each output, with a scatter of positive-definite RMSEs and zero-centered biases. For every output variable, we compute global-mean bias and spatial RMSE as in Watt-Meyer et al. (2023) equations (6) and (7), respectively. Figure 1 also includes the previously reported values for ACE trained and evaluated on FV3GFS simulation outputs.

ACE's time-mean RMSEs are comparable to the estimated noise floors for the reference set, falling within a factor of two for many important lower-tropospheric fields and within the same order of magnitude in all but a handful of cases. Global- and time-mean biases are also quite small in real terms and fall within one to two orders of magnitude of the single-pair estimates of the EAMv2 reference dataset biases, with some noted exceptions such as surface pressure (top row in Figure 1). Global-mean surface pressure is the sum of dry air mass (which should be conserved) and a much lesser water mass (which is exchanged with the underlying ocean and land surface). In EAMv2, the 10-year mean of this quantity is tightly constrained, varying little between different decadal samples (i.e. small absolute bias in Figure 1). The current version of ACE does not enforce exact global conservation equations for dry air and water and this causes larger temporal drifts in global mean surface pressure when emulating both EAMv2 and FV3GFS. Nev-

Figure 1. Global- and time-mean absolute bias (left panel) and RMSE (right panel) metrics for all output variables, averaged over the 10 year validation period. From top to bottom, prognostic variables are listed first with diagnostic variables starting with RSW. Metrics computed on ACE EAMv2 outputs ("ACE-EAMv2") are compared against: equivalent metrics for the "ACE-FV3GFS" model of (Watt-Meyer et al., 2023) with respect to the 10-year FV3GFS validation set; the best-case scenario EAMv2 metrics ("Reference"), as in Figure 3. Metrics are plotted with log scaling and units are given on the right margin for clarity.

ertheless, ACE produces a realistic time-mean map of surface pressure (not shown). With a 10 year global-time-mean of -11 Pa the magnitude of ACE's surface pressure bias is only around 0.01% of the typical surface pressure on Earth.

Overall, we find that with 42 years of training data, ACE is able to learn a representation of EAMv2 in terms of these metrics that is of similarly high quality to the results obtained for FV3GFS using 100 years of training data. In what follows, we analyze the frequency distribution of daily precipitation and time-mean spatial bias patterns of precipitation together with highly correlated top-of-atmosphere radiative fluxes. Then we examine the spectrum and temporal evolution of tropical precipitation variability between 15°S and 15°N.

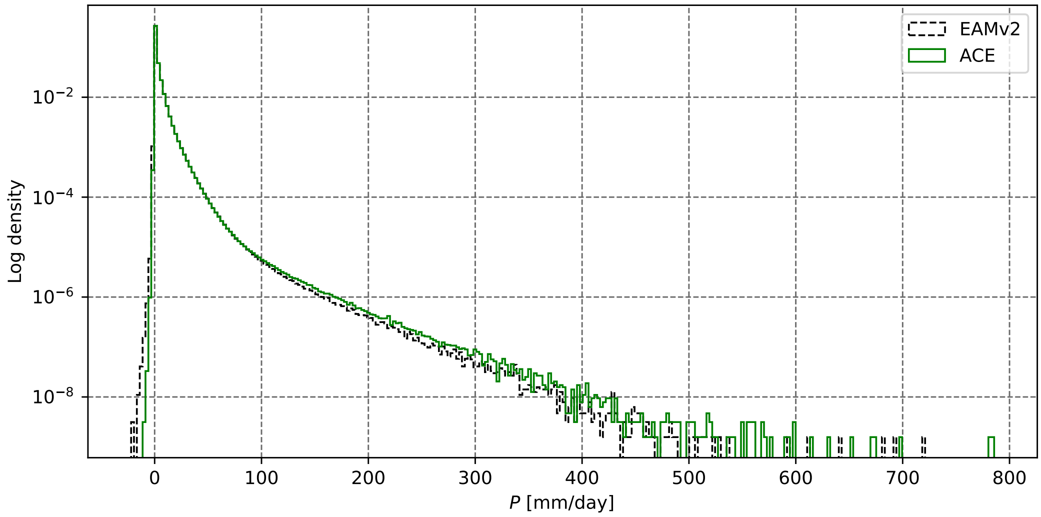

Figure 2. Frequency distribution of daily mean precipitation across all grid points over 10 years.

3.2 Precipitation density and spatial bias patterns

Establishing the precipitation extremes possible under various forcing scenarios is an important task for any climate model. Changes in the spatial distribution of time-mean precipitation under a range of possible future climate scenarios also inform many aspects of water-resource planning. Below, we examine ACE's ability to match EAMv2 in terms of (1) the frequency distribution of precipitation and (2) patterns of spatial bias in time-mean precipitation and strongly associated top-of-atmosphere fluxes.

Figure 2 shows the frequency distribution of daily precipitation in EAMv2 (black, dashed line) and ACE, including all grid points, over the 10 year validation period. Note that both the target and generated precipitation fields have a small number of negative values due to the spherical harmonic transform round-trip applied to the data, an important data preprocessing step that removes polar artifacts as explained in Watt-Meyer et al. (2023). Overall, we see that ACE captures EAMv2's precipitation distribution well, including at the extreme upper quantiles. ACE's ability to capture precipitation extremes is an encouraging sign of the usefulness of deep learning GCM emulation for downstream climate science tasks.

Figure 3 shows 10 year time-mean spatial bias patterns of precipitation and two highly correlated fields: top-of-atmosphere upward short- and longwave radiative fluxes. The left column labeled "EAMv2 vs. observation" displays the bias patterns observed when comparing the EAMv2 simulation temporal mean over the validation years 54–63 to historical observations. The observed precipitation comes from the Global Precipitation Climatology Project (GPCP) (Huffman et al., 2023) version 3.2 and corresponds to the period 1983–2021. The observed fluxes are from Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) (Loeb et al., 2018) version 4.1, over the period 2001–2018. In the right column, the corresponding validation target emulation outputs from ACE, initialized from the first timepoint of year 54, are compared against EAMv2. This way we can get a sense of the magnitude of ACE's emulation biases relative to EAMv2's observational biases.

The time-mean precipitation biases of ACE vs. EAMv2 range from -2.5 to 3.7 mm/day depending on location. The global spatial RMSE of time-mean precipitation is a remark-

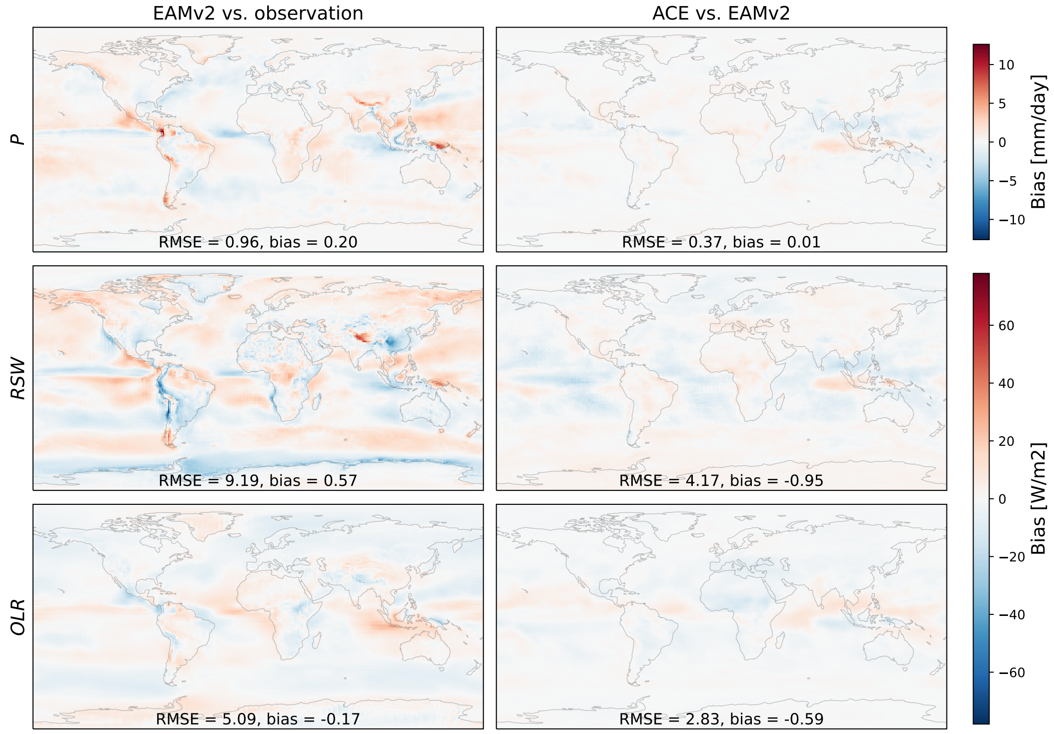

Figure 3. Temporal average of biases for surface precipitation rate (top row), outgoing top-of-atmosphere shortwave (RSW, middle row) and longwave (OLR, bottom row) radiative fluxes. The right column shows the mean spatial distribution of ACE biases vs. EAMv2, comparing the generated 6-hourly outputs to the corresponding simulation targets for the same timestep. The left column compares EAMv2 to the observed temporal mean (from GPCP for precipitation and CERES-EBAF for radiation; see main text.)

ably small 0.37 mm/day, which is comparable to the value of 0.46 reported in Watt-Meyer et al. (2023). EAMv2 observational biases lie between -6.5 and 12.6 mm/day (Figure 3) with a RMSE of 0.96 mm/day. Thus ACE emulates EAMv2 precipitation patterns much better than EAMv2 can simulate them.

OLR biases follow an expected inverse relationship with precipitation biases, a good sign of ACE's ability to emulate the radiative effects of precipitating cloud systems with cold cloud tops. Their spatial pattern RMSE is only 2.8 W/m^2, with a global-mean bias of -0.59 W/m^2. ACE's shortwave biases are larger, with a spatial pattern RMSE of 4.2 W/m^2 and a global-mean bias of -0.95 W/m^2. They are not just associated with deep precipitating cloud systems, but also 'dim' subtropical trade cumulus regimes, 'bright' Southern Ocean clouds, and excessive reflected shortwave radiation over Antarctica. As with precipitation, these emulation biases are small in comparison to EAMv2's observational biases. See Table S1 for additional summary metrics.

3.3 Tracking tropical precipitation and the MJO

Most tropical precipitation falls from organized deep convective systems, including tropical cyclones, the Madden Julian Oscillation (MJO), and diverse convectively-coupled waves. Thus it is important that global atmospheric models accurately represent the space-

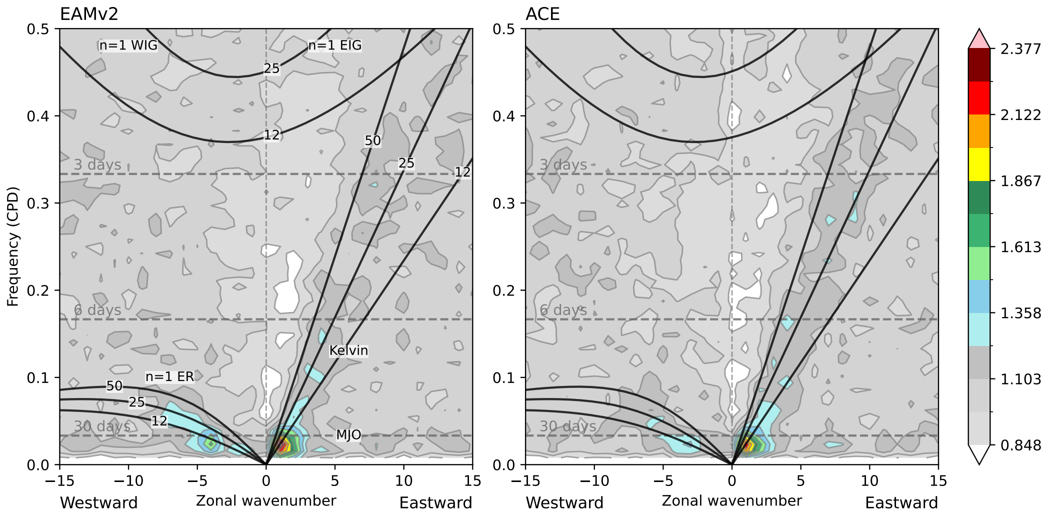

Figure 4. Normalized symmetric component of the wavenumber-frequency spectrum of daily-mean precipitation over a 10 year period for (left) withheld EAMv2 simulation output and (right) corresponding outputs from ACE. As with Figure 17 of Golaz et al. (2022), we label prominent wave types in the left panel and plot shallow water dispersion curves for equivalent depths 12, 25, and 50 m as solid black lines. ER = equatorial Rossby; EIG = eastward inertia-gravity; WIG = westward inertia-gravity.

time organization of tropical precipitation, and that an emulator of such a model replicates the organization of its tropical precipitation.

The wavenumber-frequency spectrum (Wheeler & Kiladis, 1999) of daily-mean precipitation meridionally averaged over 15°S-15°N is a widely used diagnostic of the large-scale organization of tropical precipitation. In Figure 4, we plot the normalized symmetric component of this wavenumber-frequency spectrum over the 10 year validation period for the target EAMv2 simulation data and the corresponding outputs from ACE. EAMv2's spectrum is the appropriate ground truth against which to evaluate ACE, and the emulator broadly captures EAMv2's precipitation variability.

Some minor discrepancies include slightly reduced power in the MJO and the equatorial Rossby wave, the latter also peaking at a lower wavenumber in ACE compared to EAMv2. Figure S2 provides a closer look at these features. As noted by Golaz et al. (2022), compared to satellite retrievals of the historical period, EAMv2's spectrum has weaker normalized spectral power in the wavenumber-frequency bands corresponding to the MJO and the equatorial Rossby wave and severely underestimates precipitation variability associated with Kelvin and westward inertia-gravity waves. By construction, a perfect emulator should inherit these biases.

The Madden-Julian Oscillation (MJO) is a convectively-coupled Earth-spanning atmospheric oscillation that is characterized by a large eastward-propagating band of anomalous precipitation in the tropics (Madden & Julian, 1971; Zhang, 2005). It is the most regular and predictable sub-seasonal oscillation of the Earth's atmosphere and affects many aspects of tropical and extratropical weather (Waliser et al., 2009; Zhang et al., 2020). Thus, a good emulator of an atmospheric model should replicate the statistical characteristics of its MJO.

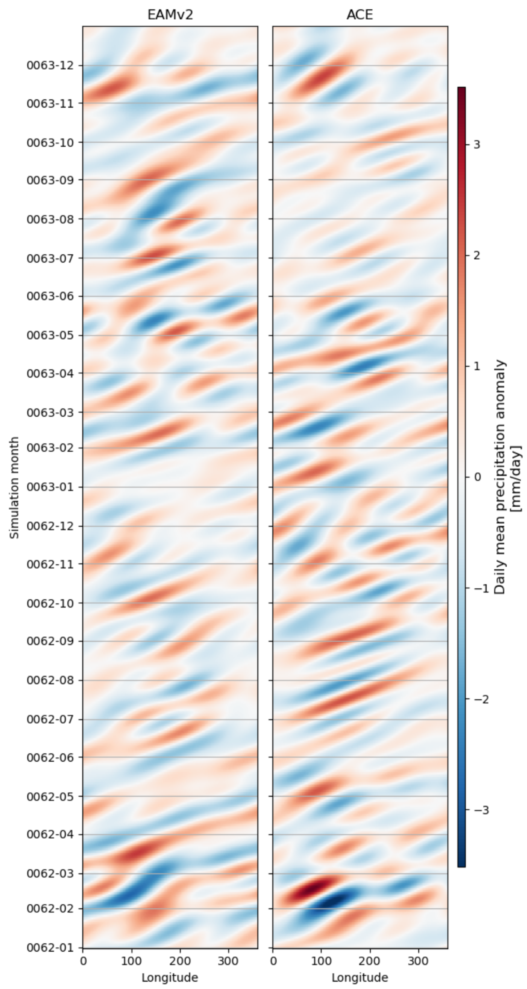

Figure 5. Hovmöller diagrams of daily mean tropical-mean precipitation over two typical years, bandpassed to retain 20-100 day periods. Both EAMv2 and ACE show patterns of eastward propagating tropical precipitation anomalies that last around 30 to 90 days.

Figure 4 suggests that ACE captures key statistical characteristics of EAMv2's simulated MJO. This skill is more directly verified by isolating the MJO frequency band with a 20-100 day bandpass filter to daily- and meridional-mean (15°S-15°N) tropical precipitation anomalies. Figure 5 shows longitude-time Hovmöller diagrams of a typical two year segment from ACE and EAMv2 simulations of the 10-year validation period. The bandpass filter drives the roughly 50-day period of the features. It is nevertheless impressive that ACE (right panel) accurately captures the amplitude and eastward propagation of the MJO spatiotemporal evolution simulated by EAMv2 (left panel).

4 Conclusions

With approximately the same training and testing protocol, ACE emulates EAMv2 with excellent skill similar to the FV3GFS model on which ACE was originally trained, as measured using 10-year time-mean climatological biases of geographically varying fields such as precipitation, near-surface and upper-tropospheric temperature and precipitable water. This suggests that ACE could easily be trained to also emulate other global atmosphere models.

ACE emulates diverse characteristics of EAMv2-simulated precipitation encouragingly well. The emulator nearly matches the EAMv2 frequency distribution of daily precipitation out to its extreme-precipitation tail. A Wheeler-Kiladis spectral analysis of tropical convectively coupled waves also shows good consistency between ACE and EAMv2, including in the simulated Madden-Julian Oscillation. That is, ACE captures the space-time organization of precipitation simulated by EAMv2.

These results were obtained for the important special case of annually-repeating climatological sea-surface temperatures. It remains to be seen how ACE will fare when faced with more realistic time-varying forcing or observational data. Over the longer term, we envision integrating future versions of ACE with other conventional or machine-learned Earth system components, such as a dynamical ocean, as part of the E3SM ecosystem and other climate and earth system models. This would enable coupled climate simulations or simulation ensembles with greatly reduced computational cost. We also envision using ACE to emulate finer-grid global atmosphere models, such as DOE's SCREAM (Caldwell et al., 2021), using ML to affordably translate the enhanced fidelity of such models into more reliable centennial climate simulations.

Open Research

Data Availability Statement

ACE model weights (2.5 GB) and the EAMv2 10-year validation set (165 GB) are available to download over HTTP from the E3SM project's NERSC science gateway at `https://portal.nersc.gov/archive/home/projects/e3sm/www/e3smv2-fme-dataset`. Documentation, inference code, and an example configuration for running ACE are available in the following repository: `https://github.com/ai2cm/ace` (Watt-Meyer et al., 2023).

Acknowledgments

This research was funded by Laboratory Directed Research and Development (LDRD 22-ERD-052) at Lawrence Livermore National Laboratory. It was initiated when James Duncan was a 2023 summer intern at AI2. The Energy Exascale Earth System Model (E3SM) project is funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research. It used computational resources of the E3SM project and NERSC, a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, using NERSC award BER-ERCAP0024832. EAMv2 simulations were performed using a high-performance computing cluster (Chrysalis) provided by the BER Earth System Modeling program and operated by the Laboratory Computing Resource Center at Argonne National Laboratory. ACE data preprocessing, training, and inference runs used NERSC's Perlmutter system. In addition, we thank James Benedict and Walter Hannah of the E3SM project for helpful conversations on tropical variability in EAMv2 and for sharing their tropical diagnostics code. Caldwell and Golaz's work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

Ben-Bouallegue, Z., Weyn, J. A., Clare, M. C. A., Dramsch, J., Dueben, P., & Chantry, M. (2023). Improving medium-range ensemble weather forecasts with hierarchical ensemble transformers. doi: 10.48550/arXiv.2303.17195

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*, 533–

538. doi: 10.1038/s41586-023-06185-3

Bonev, B., Kurth, T., Hundt, C., Pathak, J., Baust, M., Kashinath, K., & Anandkumar, A. (2023). Spherical Fourier neural operators: Learning stable dynamics on the sphere. doi: 10.48550/arXiv.2306.03838

Caldwell, P. M., Terai, C. R., Hillman, B., Keen, N. D., Bogenschutz, P., Lin, W., ... Zender, C. S. (2021). Convection-permitting simulations with the E3SM global atmosphere model. *Journal of Advances in Modeling Earth Systems*, *13*, e2021MS002544. doi: 10.1029/2021MS002544

Golaz, J.-C., Van Roekel, L. P., Zheng, X., Roberts, A. F., Wolfe, J. D., Lin, W., ... Bader, D. C. (2022). The DOE E3SM Model Version 2: Overview of the physical model and initial model evaluation. *Journal of Advances in Modeling Earth Systems*, *14*, e2022MS003156. doi: 10.1029/2022MS003156

Hannah, W. M., Bradley, A. M., Guba, O., Tang, Q., Golaz, J.-C., & Wolfe, J. (2021). Separating physics and dynamics grids for improved computational efficiency in spectral element earth system models. *Journal of Advances in Modeling Earth Systems*, *13*, e2020MS002419. doi: 10.1029/2020MS002419

Huffman, G. J., Adler, R. F., Behrangi, A., Bolvin, D. T., Nelkin, E. J., Gu, G., & Ehsani, M. R. (2023). The new Version 3.2 Global Precipitation Climatology Project (GPCP) monthly and daily precipitation products. *Journal of Climate*, *36*, 7635–7655. doi: 10.1175/JCLI-D-23-0123.1

IPCC. (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press. doi: 10.1017/CBO9781107415324

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. doi: 10.48550/arXiv.1412.6980

Kucharski, F., Molteni, F., King, M. P., Farneti, R., Kang, I.-S., & Feudale, L. (2013). On the need of intermediate complexity general circulation models: A "SPEEDY" example. *Bulletin of the American Meteorological Society*, *94*, 25–30. doi: 10.1175/BAMS-D-11-00238.1

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., ... Battaglia, P. (2023). GraphCast: Learning skillful medium-range global weather forecasting. doi: 10.48550/arXiv.2212.12794

Loeb, N. G., Doelling, D. R., Wang, H., Su, W., Nguyen, C., Corbett, J. G., ... Kato, S. (2018). Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) Top-of-Atmosphere (TOA) Edition-4.0 Data Product. *Journal of Climate*, *31*, 895–918. doi: 10.1175/JCLI-D-17-0208.1

Madden, R. A., & Julian, P. R. (1971). Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific. *Journal of the Atmospheric Sciences*, *28*, 702–708. doi: 10.1175/1520-0469(1971)028⟨0702:DOADOI⟩2.0.CO;2

Meinshausen, M., Raper, S. C. B., & Wigley, T. M. L. (2011). Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, MAGICC6 – Part 1: Model description and calibration. *Atmospheric Chemistry and Physics*, *11*, 1417–1456. doi: 10.5194/acp-11-1417-2011

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). ClimaX: A foundation model for weather and climate. doi: 10.48550/arXiv.2301.10343

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., ... Anandkumar, A. (2022). FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. doi: 10.48550/arXiv.2202.11214

Waliser, D., Sperber, K., Hendon, H., Kim, D., Wheeler, M., Weickmann, K., ... Stern, W. (2009). MJO simulation diagnostics. *Journal of Climate*, *22*, 3006-3030. doi: 10.1175/2008JCLI2731.1

Watt-Meyer, O., Dresdner, G., McGibbon, J., Duncan, J., Henn, B., Clark, S., & Perkins, W. A. (2023). Inference code for the AI2 Climate Emulator [Software]. doi: 10.5281/zenodo.10463348

Watt-Meyer, O., Dresdner, G., McGibbon, J., Clark, S. K., Henn, B., Duncan, J., ... Bretherton, C. S. (2023). ACE: A fast, skillful learned global atmospheric model for climate prediction. doi: 10.48550/arXiv.2310.02074

Wheeler, M., & Kiladis, G. N. (1999). Convectively coupled equatorial waves: Analysis of clouds and temperature in the wavenumber–frequency domain. *Journal of the Atmospheric Sciences*, *56*, 374–399. doi: 10.1175/1520-0469(1999)056⟨0374:CCEWAO⟩2.0.CO;2

Zhang, C. (2005). Madden-Julian Oscillation. *Reviews of Geophysics*, *43*. doi: 10.1029/2004RG000158

Zhang, C., Adames, A. F., Khouider, B., Wang, B., & Yang, D. (2020). Four theories of the Madden-Julian Oscillation. *Reviews of Geophysics*, *58*, e2019RG000685. doi: 10.1029/2019RG000685

Application of the AI2 Climate Emulator to E3SMv2's global atmosphere model, with a focus on precipitation fidelity

James P. C. Duncan[1], Elynn Wu[2], Jean-Christophe Golaz[3], Peter M. Caldwell[3], Oliver Watt-Meyer [2], Spencer K. Clark[2], Jeremy McGibbon[2], Gideon Dresdner[2], Karthik Kashinath[4], Boris Bonev[4], Michael S. Pritchard[4], and Christopher S. Bretherton[2]

[1]University of California, Berkeley
[2]Allen Institute for Artificial Intelligence (AI2), Seattle, USA
[3]Lawrence Livermore National Laboratory, Livermore, California, USA
[4]NVIDIA, Santa Clara, California, USA

Key Points:

- The ACE weather-climate emulator yields an accurate climate when trained on EAMv2, E3SMv2's global atmosphere model.
- Time-mean biases vs. EAMv2 in diverse atmospheric fields are similar to those seen before for ACE applied to the FV3GFS atmospheric model.
- ACE captures the space-time organization of EAMv2 precipitation well, with a much smaller time-mean bias than EAMv2's observational bias.

Corresponding author: James Duncan, `jpduncan@berkeley.edu`

Abstract

Can the current successes of global machine learning-based weather simulators be generalized beyond two-week forecasts to stable and accurate multiyear runs? The recently developed AI2 Climate Emulator (ACE) suggests this is feasible, based upon 10-year simulations trained on a realistic global atmosphere model using a grid spacing of approximately 110 km and forced by a repeating annual cycle of sea-surface temperature. Here we show that ACE, without modification, can be trained to emulate another major atmospheric model, EAMv2, run at a comparable grid spacing for at least ten years with similarly small climate biases. ACE accurately reproduces EAMv2's frequency distribution of daily-mean precipitation, its time-mean spatial pattern of precipitation, and its space-time structure of tropical precipitation, including the Madden-Julian Oscillation. Moreover, ACE's climate biases with respect to EAMv2 are substantially smaller than EAMv2's own biases compared to the observed historical average surface precipitation rate and top-of-atmosphere radiative fluxes.

Plain Language Summary

Traditional methods to predict the weather use mathematical models of the Earth's atmosphere that are costly to run. However, "data-driven" weather prediction methods, which learn to predict future weather directly from data on past weather, have come to match or even beat traditional methods and do so with much less running cost. In contrast to weather prediction where the goal is to predict the weather in the near future, in *climate modeling* the goal is to study the Earth's long-term weather trends under different possible future scenarios for many years into the future. Until the introduction of the AI2 Climate Emulator (ACE), a recent data-driven method for climate modeling, no data-driven method could match traditional climate models. In this work we test ACE's climate modeling skills and find that it is able to faithfully mimic a traditional model of the climate when looking at patterns of rainfall around the globe and in the tropics. With ACE, we can study the potential future of Earth's climate under many more scenarios and with much lower cost than ever before.

1 Introduction

In recent years, the field of numerical weather prediction has undergone a significant transformation, with researchers and institutions worldwide embracing machine learning (ML) based techniques to make weather forecasts (Pathak et al., 2022; Lam et al., 2023; Bi et al., 2023; Ben-Bouallegue et al., 2023). Notably, the European Centre for Medium-Range Weather Forecasts (ECMWF) unveiled an Artificial Intelligence based Forecasting System (AIFS) as a new companion to their physics-based numerical weather prediction model (IFS). The shift from solely physics-based numerical weather prediction to integrating ML-based systems has sparked considerable excitement within the scientific community. While most studies have focused on short to medium-range weather forecasts (up to 14 days), the AI2 Climate Emulator (ACE) has demonstrated the ability to emulate an existing global atmosphere model, FV3GFS, at climate timescales (Watt-Meyer et al., 2023) by accurately simulating weather variability and deriving climate from the statistics of the simulated weather, as do conventional global climate models. For this reason we call ACE a weather-climate emulator, to distinguish it from much simpler surrogate models that bypass weather simulation. Such models can instead be based on global or large-scale budget equations, e.g. the Model for the Assessment of Greenhouse-Gas Induced Climate Change (MAGICC) (Meinshausen et al., 2011) used in IPCC assessment reports (e.g. Sec. 8.8.2 of IPCC (2013)), in which a few parameters are tuned to give the same climate sensitivity, ocean heat uptake, and other salient global properties as a target global climate model. Alternatively, ML-based surrogate models such as ClimaX (Nguyen et al., 2023) directly predict monthly climate evolution.

ACE approximately conserves mass and moisture, and accurately predicts the climatology of key variables throughout the depth of the atmosphere. ACE can make a decade-long simulation in one hour of wall clock time of one A100 GPU, making it 100 times faster and more energy-efficient than FV3GFS run at a similar grid spacing.

Inspired by the achievements of ACE, in this paper we investigate its generalizability to emulating a different global atmosphere model, the E3SM Atmosphere Model version 2 (EAMv2). EAMv2 is the atmospheric component of the U.S. Department of Energy's Energy Exascale Earth System Model version 2 (E3SMv2) (Golaz et al., 2022). As configured for this study, EAMv2 fluid dynamics uses a grid spacing of approximately 110 km, like the FV3GFS implementation used for ACE. While FV3GFS is based on a finite-volume dynamical core with 64 vertical layers, EAMv2 uses a spectral-element approach with 72 layers while other processes use a finite-volume grid that divides each element into 2×2 cells of equal size, giving a horizontal resolution of 165 km (Hannah et al., 2021). The physical parameterizations of EAMv2 are also substantially different than those of FV3GFS.

We also analyze the emulation of precipitation in more detail than Watt-Meyer et al. (2023), including its time-mean geographic distribution, its frequency distribution of daily variability, and its organization in the tropics. A final goal of this work is to bring awareness of ACE and ML-based climate emulation into the traditional climate modeling literature.

2 Data and Methods

2.1 EAMv2 Dataset

Our training data is derived from 6-hourly outputs of a 73-year simulation of EAMv2, a model described in detail in Section 2.1 of Golaz et al. (2022). The simulation is configured to run with the "F2010" component set[1], forcing the model with perpetual 2010 greenhouse gas concentrations and emissions of aerosols and precursors, along with an annually repeating cycle of sea surface temperature and sea ice derived from the observed 2005-2014 average. The initial 11 years are discarded as spinup because the EAMv2 stratosphere is equilibrating; the following 42 years are used for training; the subsequent 10 years are used for validation; and the final 10 years are reserved for evaluating EAMv2's internal decadal variability. This simulation is performed on the E3SM Chrysalis cluster, achieving 24 simulated years per day using 30 nodes. See Text S2 for a comparison of the computational efficiencies of EAMv2 and ACE.

We make several other design choices following ACE (Watt-Meyer et al., 2023). First, we perform a conservative regridding from the native EAMv2 output to a 1° Gaussian grid to ensure compatibility with the underlying Spherical Fourier Neural Operator (SFNO) architecture (Bonev et al., 2023). Second, we filter the data with a spherical harmonic transform (SHT) round-trip to help eliminate artifacts in the high latitudes. Third, to reduce the emulator's memory footprint, we coarsen the vertical model-level coordinate from the native 72 down to 8 layers. For more details see Table S2.

2.2 ACE Training Overview

As described by Watt-Meyer et al. (2023), ACE is a modified version of NVIDIA's open-source FourCastNet global atmospheric emulator (Pathak et al., 2022) that employs the SFNO architecture for efficient spatial information exchange (Bonev et al., 2023). Much as traditional physics-based numerical models of atmospheric dynamics recursively step forward the atmospheric state X_t at time t, ACE is trained to autoregressively gener-

[1] `https://acme-climate.atlassian.net/wiki/spaces/DOC/pages/961250902/F2010C5-CMIP6-LR`

ate predictions of the atmospheric state at time $t + \delta t$: $\hat{X}_{t+\delta t}$. We use $\delta t = 6$ hours and minimize the average "one-step" loss over a random batch $\mathcal{B}$ of initial condition times t:

$$\frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \frac{\|\hat{X}_{t+\delta t} - X_{t+\delta t}\|_2}{\|X_{t+\delta t}\|_2}$$

Whereas FourCastNet uses identical input and output variables and trains a separate model to predict diagnostic variables (Pathak et al., 2022), ACE uses a set of prognostic variables which are both inputs and outputs, a set of specified forcing input variables such as insolation and sea surface skin temperature which are exogenous to the dynamical system, and a set of diagnostic variables which are incorporated in the training loss but are output-only. This and a variety of other improvements enable ACE, unlike past weather emulators, to produce stable, skillful, more interpretable multiyear emulations of the target model. For more details see Table S3, Watt-Meyer et al. (2023), and Bonev et al. (2023).

3 Results

Watt-Meyer et al. (2023) provide a holistic evaluation of ACE's physical consistency when trained on 100 years of FV3GFS simulation outputs in terms of physical budgets and time- and global-mean biases and pattern errors.

Section 3.1 shows a similar analysis of ACE's global- and time-mean absolute bias and root mean square error (RMSE) metrics on EAMv2. This analysis shows that ACE produces a similarly high-quality emulation of the climatology of EAMv2 as for FV3GFS, demonstrating that ACE's training methodology generalizes across reference models of comparable grid resolution with different dynamical cores and physical parameterizations. In the remainder of Section 3, we present some key metrics of how well ACE emulates EAMv2's precipitation variability over the 10 year validation period, a topic not documented in detail by Watt-Meyer et al. (2023).

3.1 Global- and time-mean biases and RMSE

In Figure 1, we compare ACE's climatological skill to that of an unseen EAMv2 reference dataset, years 64–73 of the EAMv2 simulation run. Both ACE and the reference are evaluated against the validation target years 54–63. The reference values give a 'noise floor' estimate, computed as the difference of time means from a single pair of ten-year segments of the reference simulation. Different pairs of ten-year periods would give different estimates for each output, with a scatter of positive-definite RMSEs and zero-centered biases. For every output variable, we compute global-mean bias and spatial RMSE as in Watt-Meyer et al. (2023) equations (6) and (7), respectively. Figure 1 also includes the previously reported values for ACE trained and evaluated on FV3GFS simulation outputs.

ACE's time-mean RMSEs are comparable to the estimated noise floors for the reference set, falling within a factor of two for many important lower-tropospheric fields and within the same order of magnitude in all but a handful of cases. Global- and time-mean biases are also quite small in real terms and fall within one to two orders of magnitude of the single-pair estimates of the EAMv2 reference dataset biases, with some noted exceptions such as surface pressure (top row in Figure 1). Global-mean surface pressure is the sum of dry air mass (which should be conserved) and a much lesser water mass (which is exchanged with the underlying ocean and land surface). In EAMv2, the 10-year mean of this quantity is tightly constrained, varying little between different decadal samples (i.e. small absolute bias in Figure 1). The current version of ACE does not enforce exact global conservation equations for dry air and water and this causes larger temporal drifts in global mean surface pressure when emulating both EAMv2 and FV3GFS. Nev-

Figure 1. Global- and time-mean absolute bias (left panel) and RMSE (right panel) metrics for all output variables, averaged over the 10 year validation period. From top to bottom, prognostic variables are listed first with diagnostic variables starting with RSW. Metrics computed on ACE EAMv2 outputs ("ACE-EAMv2") are compared against: equivalent metrics for the "ACE-FV3GFS" model of (Watt-Meyer et al., 2023) with respect to the 10-year FV3GFS validation set; the best-case scenario EAMv2 metrics ("Reference"), as in Figure 3. Metrics are plotted with log scaling and units are given on the right margin for clarity.

ertheless, ACE produces a realistic time-mean map of surface pressure (not shown). With a 10 year global-time-mean of −11 Pa the magnitude of ACE's surface pressure bias is only around 0.01% of the typical surface pressure on Earth.

Overall, we find that with 42 years of training data, ACE is able to learn a representation of EAMv2 in terms of these metrics that is of similarly high quality to the results obtained for FV3GFS using 100 years of training data. In what follows, we analyze the frequency distribution of daily precipitation and time-mean spatial bias patterns of precipitation together with highly correlated top-of-atmosphere radiative fluxes. Then we examine the spectrum and temporal evolution of tropical precipitation variability between 15°S and 15°N.

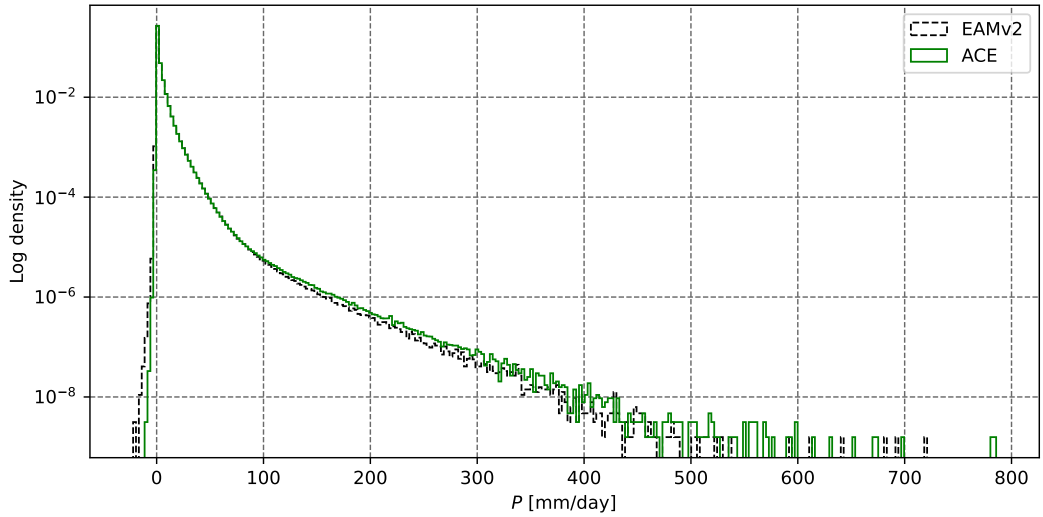

Figure 2. Frequency distribution of daily mean precipitation across all grid points over 10 years.

3.2 Precipitation density and spatial bias patterns

Establishing the precipitation extremes possible under various forcing scenarios is an important task for any climate model. Changes in the spatial distribution of time-mean precipitation under a range of possible future climate scenarios also inform many aspects of water-resource planning. Below, we examine ACE's ability to match EAMv2 in terms of (1) the frequency distribution of precipitation and (2) patterns of spatial bias in time-mean precipitation and strongly associated top-of-atmosphere fluxes.

Figure 2 shows the frequency distribution of daily precipitation in EAMv2 (black, dashed line) and ACE, including all grid points, over the 10 year validation period. Note that both the target and generated precipitation fields have a small number of negative values due to the spherical harmonic transform round-trip applied to the data, an important data preprocessing step that removes polar artifacts as explained in Watt-Meyer et al. (2023). Overall, we see that ACE captures EAMv2's precipitation distribution well, including at the extreme upper quantiles. ACE's ability to capture precipitation extremes is an encouraging sign of the usefulness of deep learning GCM emulation for downstream climate science tasks.

Figure 3 shows 10 year time-mean spatial bias patterns of precipitation and two highly correlated fields: top-of-atmosphere upward short- and longwave radiative fluxes. The left column labeled "EAMv2 vs. observation" displays the bias patterns observed when comparing the EAMv2 simulation temporal mean over the validation years 54–63 to historical observations. The observed precipitation comes from the Global Precipitation Climatology Project (GPCP) (Huffman et al., 2023) version 3.2 and corresponds to the period 1983–2021. The observed fluxes are from Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) (Loeb et al., 2018) version 4.1, over the period 2001–2018. In the right column, the corresponding validation target emulation outputs from ACE, initialized from the first timepoint of year 54, are compared against EAMv2. This way we can get a sense of the magnitude of ACE's emulation biases relative to EAMv2's observational biases.

The time-mean precipitation biases of ACE vs. EAMv2 range from −2.5 to 3.7 mm/day depending on location. The global spatial RMSE of time-mean precipitation is a remark-

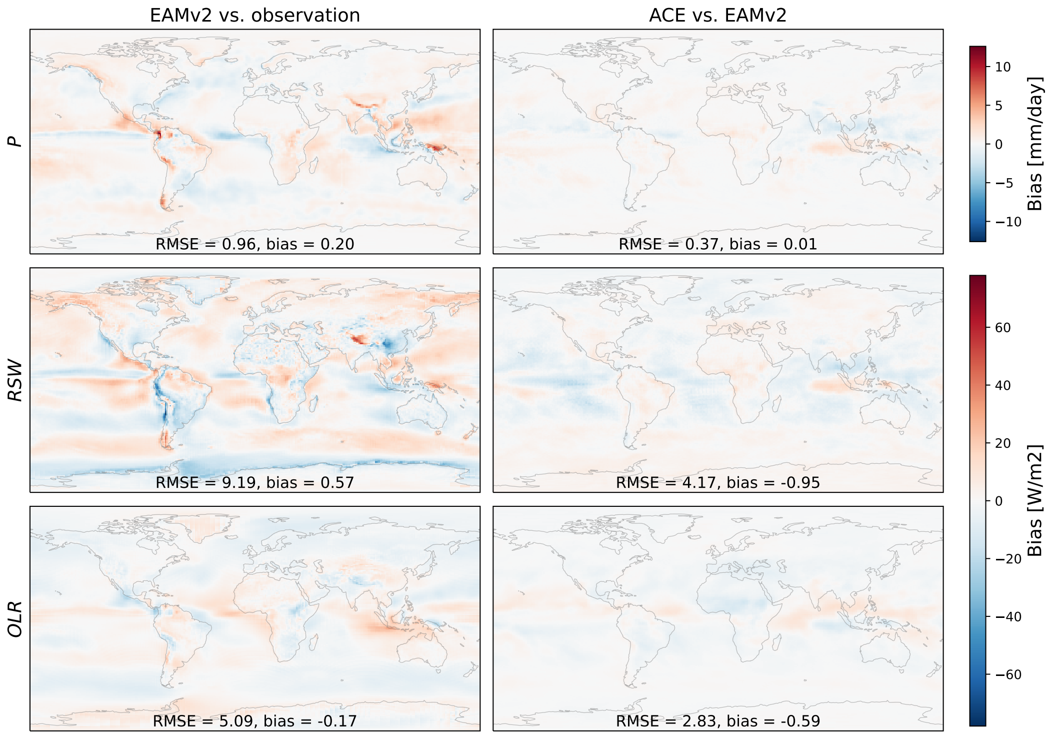

Figure 3. Temporal average of biases for surface precipitation rate (top row), outgoing top-of-atmosphere shortwave (RSW, middle row) and longwave (OLR, bottom row) radiative fluxes. The right column shows the mean spatial distribution of ACE biases vs. EAMv2, comparing the generated 6-hourly outputs to the corresponding simulation targets for the same timestep. The left column compares EAMv2 to the observed temporal mean (from GPCP for precipitation and CERES-EBAF for radiation; see main text.)

ably small 0.37 mm/day, which is comparable to the value of 0.46 reported in Watt-Meyer et al. (2023). EAMv2 observational biases lie between -6.5 and 12.6 mm/day (Figure 3) with a RMSE of 0.96 mm/day. Thus ACE emulates EAMv2 precipitation patterns much better than EAMv2 can simulate them.

OLR biases follow an expected inverse relationship with precipitation biases, a good sign of ACE's ability to emulate the radiative effects of precipitating cloud systems with cold cloud tops. Their spatial pattern RMSE is only 2.8 W/m^2, with a global-mean bias of -0.59 W/m^2. ACE's shortwave biases are larger, with a spatial pattern RMSE of 4.2 W/m^2 and a global-mean bias of -0.95 W/m^2. They are not just associated with deep precipitating cloud systems, but also 'dim' subtropical trade cumulus regimes, 'bright' Southern Ocean clouds, and excessive reflected shortwave radiation over Antarctica. As with precipitation, these emulation biases are small in comparison to EAMv2's observational biases. See Table S1 for additional summary metrics.

3.3 Tracking tropical precipitation and the MJO

Most tropical precipitation falls from organized deep convective systems, including tropical cyclones, the Madden Julian Oscillation (MJO), and diverse convectively-coupled waves. Thus it is important that global atmospheric models accurately represent the space-

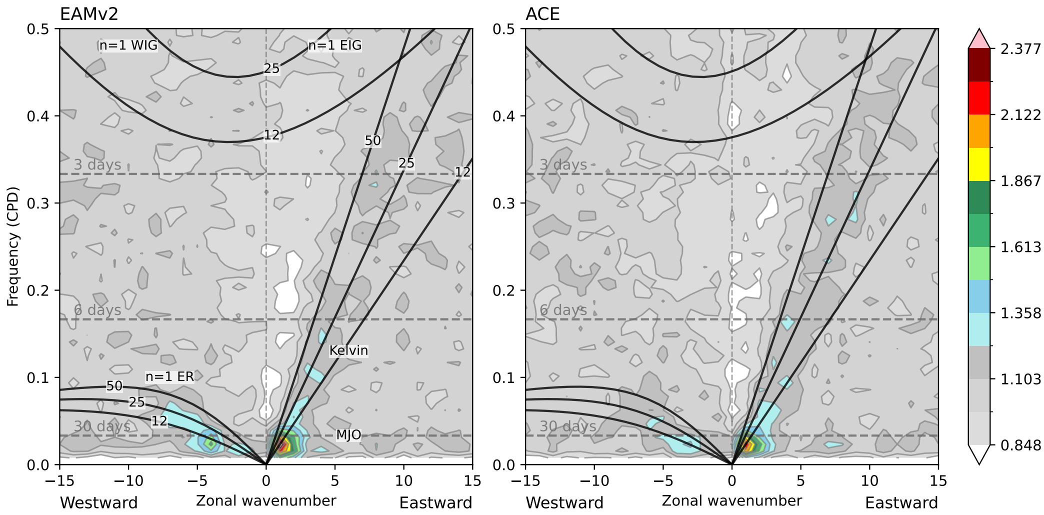

Figure 4. Normalized symmetric component of the wavenumber-frequency spectrum of daily-mean precipitation over a 10 year period for (left) withheld EAMv2 simulation output and (right) corresponding outputs from ACE. As with Figure 17 of Golaz et al. (2022), we label prominent wave types in the left panel and plot shallow water dispersion curves for equivalent depths 12, 25, and 50 m as solid black lines. ER = equatorial Rossby; EIG = eastward inertia-gravity; WIG = westward inertia-gravity.

time organization of tropical precipitation, and that an emulator of such a model replicates the organization of its tropical precipitation.

The wavenumber-frequency spectrum (Wheeler & Kiladis, 1999) of daily-mean precipitation meridionally averaged over 15°S-15°N is a widely used diagnostic of the large-scale organization of tropical precipitation. In Figure 4, we plot the normalized symmetric component of this wavenumber-frequency spectrum over the 10 year validation period for the target EAMv2 simulation data and the corresponding outputs from ACE. EAMv2's spectrum is the appropriate ground truth against which to evaluate ACE, and the emulator broadly captures EAMv2's precipitation variability.

Some minor discrepancies include slightly reduced power in the MJO and the equatorial Rossby wave, the latter also peaking at a lower wavenumber in ACE compared to EAMv2. Figure S2 provides a closer look at these features. As noted by Golaz et al. (2022), compared to satellite retrievals of the historical period, EAMv2's spectrum has weaker normalized spectral power in the wavenumber-frequency bands corresponding to the MJO and the equatorial Rossby wave and severely underestimates precipitation variability associated with Kelvin and westward inertia-gravity waves. By construction, a perfect emulator should inherit these biases.

The Madden-Julian Oscillation (MJO) is a convectively-coupled Earth-spanning atmospheric oscillation that is characterized by a large eastward-propagating band of anomalous precipitation in the tropics (Madden & Julian, 1971; Zhang, 2005). It is the most regular and predictable sub-seasonal oscillation of the Earth's atmosphere and affects many aspects of tropical and extratropical weather (Waliser et al., 2009; Zhang et al., 2020). Thus, a good emulator of an atmospheric model should replicate the statistical characteristics of its MJO.

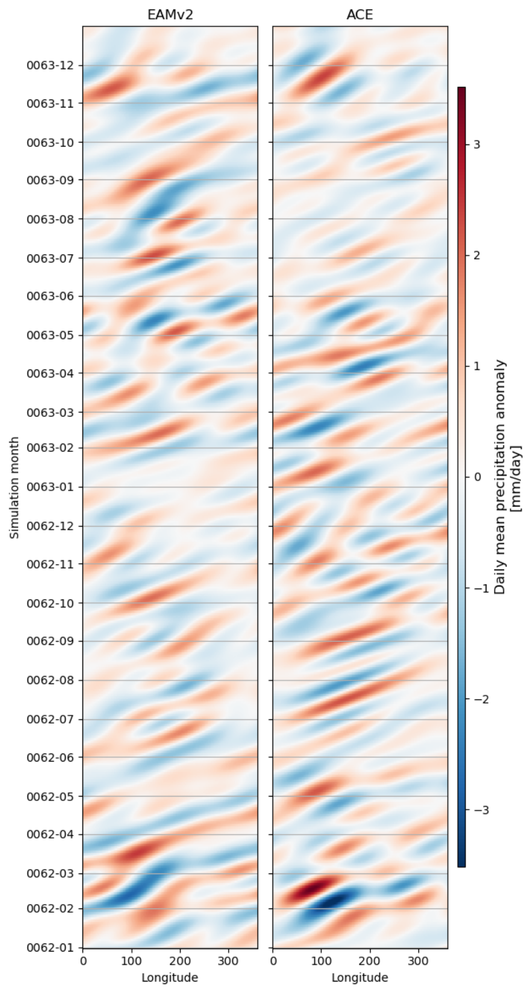

Figure 5. Hovmöller diagrams of daily mean tropical-mean precipitation over two typical years, bandpassed to retain 20-100 day periods. Both EAMv2 and ACE show patterns of eastward propagating tropical precipitation anomalies that last around 30 to 90 days.

Figure 4 suggests that ACE captures key statistical characteristics of EAMv2's simulated MJO. This skill is more directly verified by isolating the MJO frequency band with a 20-100 day bandpass filter to daily- and meridional-mean (15°S-15°N) tropical precipitation anomalies. Figure 5 shows longitude-time Hovmöller diagrams of a typical two year segment from ACE and EAMv2 simulations of the 10-year validation period. The bandpass filter drives the roughly 50-day period of the features. It is nevertheless impressive that ACE (right panel) accurately captures the amplitude and eastward propagation of the MJO spatiotemporal evolution simulated by EAMv2 (left panel).

4 Conclusions

With approximately the same training and testing protocol, ACE emulates EAMv2 with excellent skill similar to the FV3GFS model on which ACE was originally trained, as measured using 10-year time-mean climatological biases of geographically varying fields such as precipitation, near-surface and upper-tropospheric temperature and precipitable water. This suggests that ACE could easily be trained to also emulate other global atmosphere models.

ACE emulates diverse characteristics of EAMv2-simulated precipitation encouragingly well. The emulator nearly matches the EAMv2 frequency distribution of daily precipitation out to its extreme-precipitation tail. A Wheeler-Kiladis spectral analysis of tropical convectively coupled waves also shows good consistency between ACE and EAMv2, including in the simulated Madden-Julian Oscillation. That is, ACE captures the space-time organization of precipitation simulated by EAMv2.

These results were obtained for the important special case of annually-repeating climatological sea-surface temperatures. It remains to be seen how ACE will fare when faced with more realistic time-varying forcing or observational data. Over the longer term, we envision integrating future versions of ACE with other conventional or machine-learned Earth system components, such as a dynamical ocean, as part of the E3SM ecosystem and other climate and earth system models. This would enable coupled climate simulations or simulation ensembles with greatly reduced computational cost. We also envision using ACE to emulate finer-grid global atmosphere models, such as DOE's SCREAM (Caldwell et al., 2021), using ML to affordably translate the enhanced fidelity of such models into more reliable centennial climate simulations.

Open Research

Data Availability Statement

ACE model weights (2.5 GB) and the EAMv2 10-year validation set (165 GB) are available to download over HTTP from the E3SM project's NERSC science gateway at `https://portal.nersc.gov/archive/home/projects/e3sm/www/e3smv2-fme-dataset`. Documentation, inference code, and an example configuration for running ACE are available in the following repository: `https://github.com/ai2cm/ace` (Watt-Meyer et al., 2023).

Acknowledgments

This research was funded by Laboratory Directed Research and Development (LDRD 22-ERD-052) at Lawrence Livermore National Laboratory. It was initiated when James Duncan was a 2023 summer intern at AI2. The Energy Exascale Earth System Model (E3SM) project is funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research. It used computational resources of the E3SM project and NERSC, a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, using NERSC award BER-ERCAP0024832. EAMv2 simulations were performed using a high-performance computing cluster (Chrysalis) provided by the BER Earth System Modeling program and operated by the Laboratory Computing Resource Center at Argonne National Laboratory. ACE data preprocessing, training, and inference runs used NERSC's Perlmutter system. In addition, we thank James Benedict and Walter Hannah of the E3SM project for helpful conversations on tropical variability in EAMv2 and for sharing their tropical diagnostics code. Caldwell and Golaz's work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

Ben-Bouallegue, Z., Weyn, J. A., Clare, M. C. A., Dramsch, J., Dueben, P., & Chantry, M. (2023). Improving medium-range ensemble weather forecasts with hierarchical ensemble transformers. doi: 10.48550/arXiv.2303.17195

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*, 533–

538. doi: 10.1038/s41586-023-06185-3

Bonev, B., Kurth, T., Hundt, C., Pathak, J., Baust, M., Kashinath, K., & Anandkumar, A. (2023). Spherical Fourier neural operators: Learning stable dynamics on the sphere. doi: 10.48550/arXiv.2306.03838

Caldwell, P. M., Terai, C. R., Hillman, B., Keen, N. D., Bogenschutz, P., Lin, W., … Zender, C. S. (2021). Convection-permitting simulations with the E3SM global atmosphere model. *Journal of Advances in Modeling Earth Systems*, *13*, e2021MS002544. doi: 10.1029/2021MS002544

Golaz, J.-C., Van Roekel, L. P., Zheng, X., Roberts, A. F., Wolfe, J. D., Lin, W., … Bader, D. C. (2022). The DOE E3SM Model Version 2: Overview of the physical model and initial model evaluation. *Journal of Advances in Modeling Earth Systems*, *14*, e2022MS003156. doi: 10.1029/2022MS003156

Hannah, W. M., Bradley, A. M., Guba, O., Tang, Q., Golaz, J.-C., & Wolfe, J. (2021). Separating physics and dynamics grids for improved computational efficiency in spectral element earth system models. *Journal of Advances in Modeling Earth Systems*, *13*, e2020MS002419. doi: 10.1029/2020MS002419

Huffman, G. J., Adler, R. F., Behrangi, A., Bolvin, D. T., Nelkin, E. J., Gu, G., & Ehsani, M. R. (2023). The new Version 3.2 Global Precipitation Climatology Project (GPCP) monthly and daily precipitation products. *Journal of Climate*, *36*, 7635–7655. doi: 10.1175/JCLI-D-23-0123.1

IPCC. (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press. doi: 10.1017/CBO9781107415324

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. doi: 10.48550/arXiv.1412.6980

Kucharski, F., Molteni, F., King, M. P., Farneti, R., Kang, I.-S., & Feudale, L. (2013). On the need of intermediate complexity general circulation models: A “SPEEDY” example. *Bulletin of the American Meteorological Society*, *94*, 25–30. doi: 10.1175/BAMS-D-11-00238.1

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., … Battaglia, P. (2023). GraphCast: Learning skillful medium-range global weather forecasting. doi: 10.48550/arXiv.2212.12794

Loeb, N. G., Doelling, D. R., Wang, H., Su, W., Nguyen, C., Corbett, J. G., … Kato, S. (2018). Clouds and the Earth’s Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) Top-of-Atmosphere (TOA) Edition-4.0 Data Product. *Journal of Climate*, *31*, 895–918. doi: 10.1175/JCLI-D-17-0208.1

Madden, R. A., & Julian, P. R. (1971). Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific. *Journal of the Atmospheric Sciences*, *28*, 702–708. doi: 10.1175/1520-0469(1971)028⟨0702:DOADOI⟩2.0.CO;2

Meinshausen, M., Raper, S. C. B., & Wigley, T. M. L. (2011). Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, MAGICC6 – Part 1: Model description and calibration. *Atmospheric Chemistry and Physics*, *11*, 1417–1456. doi: 10.5194/acp-11-1417-2011

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). ClimaX: A foundation model for weather and climate. doi: 10.48550/arXiv.2301.10343

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., … Anandkumar, A. (2022). FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. doi: 10.48550/arXiv.2202.11214

Waliser, D., Sperber, K., Hendon, H., Kim, D., Wheeler, M., Weickmann, K., … Stern, W. (2009). MJO simulation diagnostics. *Journal of Climate*, *22*, 3006-3030. doi: 10.1175/2008JCLI2731.1

Watt-Meyer, O., Dresdner, G., McGibbon, J., Duncan, J., Henn, B., Clark, S., & Perkins, W. A. (2023). Inference code for the AI2 Climate Emulator [Software]. doi: 10.5281/zenodo.10463348

Watt-Meyer, O., Dresdner, G., McGibbon, J., Clark, S. K., Henn, B., Duncan, J., ... Bretherton, C. S. (2023). ACE: A fast, skillful learned global atmospheric model for climate prediction. doi: 10.48550/arXiv.2310.02074

Wheeler, M., & Kiladis, G. N. (1999). Convectively coupled equatorial waves: Analysis of clouds and temperature in the wavenumber–frequency domain. *Journal of the Atmospheric Sciences*, *56*, 374–399. doi: 10.1175/1520-0469(1999)056⟨0374:CCEWAO⟩2.0.CO;2

Zhang, C. (2005). Madden-Julian Oscillation. *Reviews of Geophysics*, *43*. doi: 10.1029/2004RG000158

Zhang, C., Adames, A. F., Khouider, B., Wang, B., & Yang, D. (2020). Four theories of the Madden-Julian Oscillation. *Reviews of Geophysics*, *58*, e2019RG000685. doi: 10.1029/2019RG000685

Supporting Information for "Application of the AI2 Climate Emulator to E3SMv2's global atmosphere model, with a focus on precipitation fidelity"

James P. C. Duncan[1], Elynn Wu[2], Jean-Christophe Golaz[3], Peter M. Caldwell[3], Oliver Watt-Meyer [2], Spencer K. Clark[2], Jeremy McGibbon[2], Gideon Dresdner[2], Karthik Kashinath[4], Boris Bonev[4], Michael S. Pritchard[4], and Christopher S. Bretherton[2]

[1]University of California, Berkeley

[2]Allen Institute for Artificial Intelligence (AI2), Seattle, USA

[3]Lawrence Livermore National Laboratory, Livermore, California, USA

[4]NVIDIA, Santa Clara, California, USA

Contents of this file

Introduction

In this Supporting Information, we give additional metrics related to ACE's climatological skill and supplementary figures which provide additional perspectives on the figures of the main text. We also provide further details on the computational efficiency of ACE, the vertical coarsening of raw EAMv2 simulations outputs, and the optimization hyperparameters employed during ACE training.

Text S1. Another perspective on ACE's emulation biases

Figure S1 compares ACE's emulation biases to EAMv2's internal variability. The left column labeled "EAMv2 reference vs. EAMv2" displays the bias patterns observed when comparing EAMv2 to itself, which serves as an 'oracle' emulator with the highest climate skill possible in terms of faithfulness to the original simulation, given natural variability due to weather fluctuations. These biases are computed by comparing the unseen reference set, years 64–73 of the EAMv2 simulation run, against the validation target years 54–63. The column labeled "ACE vs. EAMv2" visualizes the same data as the right column of Figure 3 of the main text. Table S1 provides additional bias and RMSE metrics for these variables when evaluating ACE and EAMv2 internally (i.e., against EAMv2 simulation outputs) as in Figure S1 and against historical observations as in the left column of Figure 3.

Text S2. Computational efficiency of ACE

We carried out the 73 year EAMv2 simulation on the Chrysalis supercomputer at Argonne National Laboratory, which is a dedicated E3SM machine[1]. Using 30 CPU nodes on Chrysalis, each of which has 2× 32-core AMD EPYC 7532 CPUs, the simulation achieved 24 simulated years per day, or about 10 seconds per simulation day. After training, we ran

ACE inference using a single NVIDIA A100 40 GB GPU on Lawrence Berkeley National Laboratory's Perlmutter supercomputer with a wall clock time of 1 second per simulation day, an approximate 10x speedup. The discrepancy with the 100x speedup found in Watt-Meyer et al. (2023) is explained by the much larger number of cores used for the EAMv2 simulation compared to the FV3GFS simulation, which used a total of 96 cores across two higher-efficiency AMD EPYC 7H12 CPUs. We estimate the energy consumption of 1 second on 1 A100 GPU at maximum power consumption of 400 W is 0.4 kJ, while 10 seconds on 60 total EPYC 7532 CPUs at 200 W is approximately 120 kJ. This amounts to an approximate 300x energy savings when using ACE as a surrogate for EAMv2.

References

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. doi: 10.48550/arXiv.1412.6980

Kucharski, F., Molteni, F., King, M. P., Farneti, R., Kang, I.-S., & Feudale, L. (2013). On the need of intermediate complexity general circulation models: A "SPEEDY" example. *Bulletin of the American Meteorological Society*, *94*, 25–30. doi: 10.1175/BAMS-D-11-00238.1

Watt-Meyer, O., Dresdner, G., McGibbon, J., Clark, S. K., Henn, B., Duncan, J., ... Bretherton, C. S. (2023). ACE: A fast, skillful learned global atmospheric model for climate prediction. doi: 10.48550/arXiv.2310.02074

Notes

1. `https://climatemodeling.science.energy.gov/news/chrysalis-ready-emerge-e3sm-v2-runs`

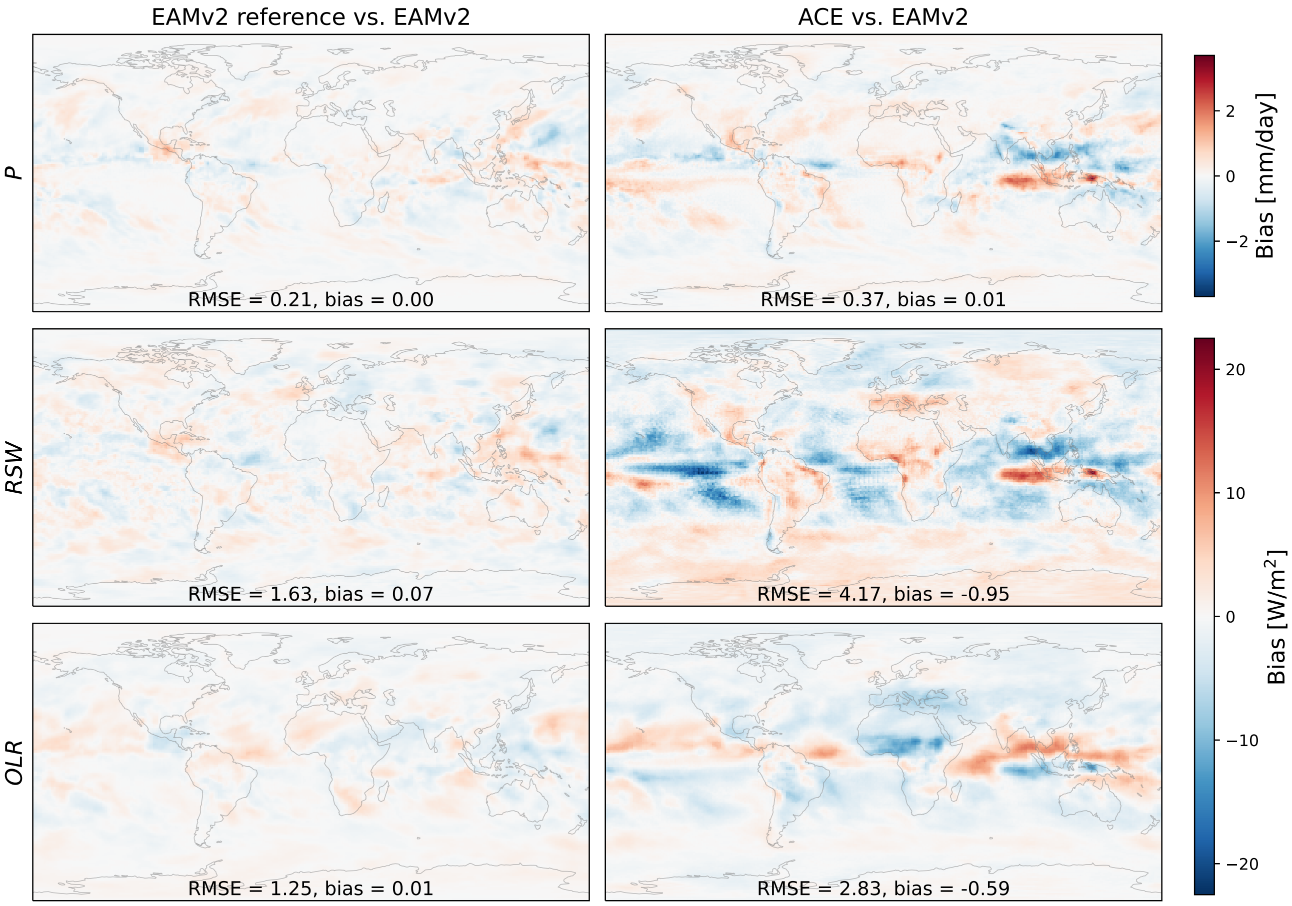

Figure S1. Time average biases (*predicted* - *target*) for precipitation (top row) and top-of-atmosphere outgoing shortwave (RSW, middle row) and longwave (OLR, bottom row) radiative fluxes. The right column ("ACE vs. EAMv2") shows the mean spatial distribution of ACE biases, comparing the generated 6-hourly outputs to the corresponding targets for the same timestep. The left column ("EAMv2 reference vs. EAMv2") compares EAMv2 to itself by recalculating biases using the final 10 years of the simulation set in the place of the *predicted* data, giving a best-case scenario reference.

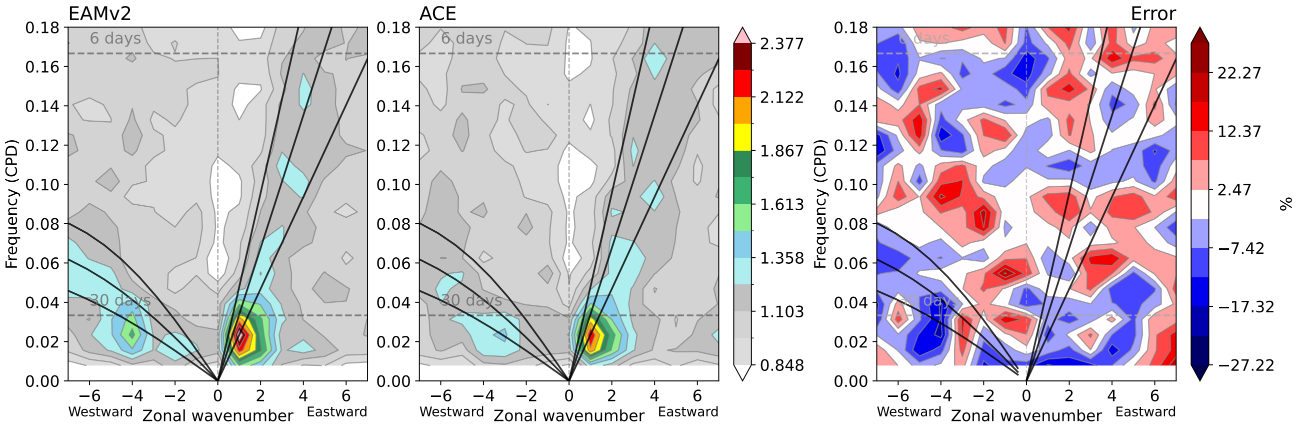

Figure S2. Same as Figure 4 of the main text but zoomed in for a closer look at the tropical spectra between wavenumbers -6 and 6 and frequencies smaller than 0.18. In addition, the third panel displays relative errors within this region, calculated as: $100 \times \frac{\text{predicted power}-\text{target power}}{\text{target power}}\%$.

Table S1. ACE and E3SMv2 biases and RMSEs with respect to various references. ACE$_{int}$: ACE compared against EAMv2 outputs over the 10 year validation period. EAMv2$_{int}$: EAMv2 outputs over the 10 year reference period compared against EAMv2 outputs over the 10 year validation period. ACE$_{obs}$: ACE compared against historical observations. EAMv2$_{obs}$: EAMv2 outputs over the 10 year validation period compared against historical observations.

	ACE$_{int}$		EAMv2$_{int}$		ACE$_{obs}$		EAMv2$_{obs}$	
Variable	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
P [mm/day]	5.7e-3	0.37	1.6e-3	0.21	0.20	0.93	0.20	0.96
RSW [W/m^2]	-0.95	4.17	6.7e-2	1.63	-0.38	8.87	0.57	9.19
OLR [W/m^2]	-0.59	2.83	8.5e-3	1.25	-0.77	5.64	-0.17	5.09

Table S2. EAMv2 vertical interface coordinates that were used for vertical coarsening of the raw 3D outputs, reducing the number of vertical levels from 72 to 8 for computational tractability. As in Watt-Meyer et al. (2023), we chose the 9 vertical interfaces listed below that best align with those of the SPEEDY model (Kucharski et al., 2013), in sigma coordinates, assuming a constant reference surface pressure of $p_8^{ref} = 1000$ hPa. The coarsened interfaces are indexed starting from the top of the atmosphere by k from 0 to 8, while the corresponding original EAMv2 interfaces are indexed by I_k. In each grid column, the terrain-following interfacial pressures $p_k = a_k + b_k p_s$ are computed from the hybrid coordinates a_k and b_k and the surface pressure p_s. The original model levels are vertically integrated by mass in order to preserve the total dry air and moisture budget, using the true surface pressure at each point in space and time. For further details, see Watt-Meyer et al. (2023).

k	a_k [Pa]	b_k [unitless]	I_k	p_k^{ref} [hPa]
0	10.0	0.0	0	0.1
1	4943.694	0.0	19	49.4
2	13913.118	0.0	30	139
3	16254.503	0.10464	38	267
4	12435.282	0.31152	44	436
5	8945.939	0.50053	48	590
6	5115.018	0.70804	53	759
7	2027.536	0.87529	61	896
8	0.0	1.0	72	1000

Table S3. Following Watt-Meyer et al. (2023), we employ the Adam optimizer (Kingma & Ba, 2017) with a cosine annealing learning rate schedule decaying to zero by the end of training and use an exponential moving average of the model parameters across training steps. We conducted a thorough hyperparameter search across 29 combinations of batch size, initial learning rate, and number of epochs, arriving at the final choice of hyperparameters based upon a comparison of 10-year time-mean validation metrics, multiyear stability, and visual artifacts. See Watt-Meyer et al. (2023) for additional details on training and SFNO architectural hyperparameters.

Name	Value
Initial learning rate	3×10^{-4}
Number of epochs	50
Batch size	8