Check for updates

JGR | Machine Learning and Computation

REVIEW ARTICLE
10.1029/2024JH000496

Recommendations for Comprehensive and Independent Evaluation of Machine Learning-Based Earth System Models

P. A. Ullrich[1,2], E. A. Barnes[3], W. D. Collins[4,5], K. Dagon[6], S. Duan[1], J. Elms[7], J. Lee[1], L. R. Leung[8], D. Lu[9], M. J. Molina[10], T. A. O'Brien[4,6], and F. O. Rebassoo[1]

[1]Division of Physical and Life Sciences, Lawrence Livermore National Laboratory, Livermore, CA, USA, [2]Department of Land, Air and Water Resources, University of California, Davis, CA, USA, [3]Department of Atmospheric Science, Colorado State University, Fort Collins, CO, USA, [4]Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA, [5]Department of Earth and Planetary Science, University of California, Berkeley, CA, USA, [6]Climate and Global Dynamics Laboratory, National Center for Atmospheric Research, Boulder, CO, USA, [7]Department of Earth and Atmospheric Sciences, Indiana University Bloomington, Bloomington, IN, USA, [8]Atmospheric, Climate, and Earth Sciences, Pacific Northwest National Laboratory, Richland, WA, USA, [9]Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA, [10]Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD, USA

Abstract Machine learning (ML) is a revolutionary technology with demonstrable applications across multiple disciplines. Within the Earth science community, ML has been most visible for weather forecasting, producing forecasts that rival modern physics-based models. Given the importance of deepening our understanding and improving predictions of the Earth system on all time scales, efforts are now underway to develop Earth-system models (ESMs) capable of representing all components of the coupled Earth system (or their aggregated behavior) and their response to external changes over long timescales. Building trust in ESMs is a much more difficult problem than for weather forecast models, not least because the model must represent the alternate (e.g., future or paleoclimatic) coupled states of the system for which there are no direct observations. Given that the physical principles that enable predictions about the response of the Earth system are often not explicitly coded in these ML-based models, demonstrating the credibility of ML-based ESMs thus requires us to build evidence of their consistency with the physical system. To this end, this paper puts forward five recommendations to enhance comprehensive, standardized, and independent evaluation of ML-based ESMs to strengthen their credibility and promote their wider use.

Plain Language Summary Machine learning (ML) has the potential to be a step change in our ability to simulate the Earth system, and early efforts in this space have already yielded impressive outcomes. However, to be useful, either scientifically or in practice, we must have confidence that ML-based Earth-system models (ESMs) are producing simulations and predictions that are consistent with physical laws. Comprehensive, standardized and independent evaluation of these models provide one avenue to develop that confidence. This paper puts forward five recommendations on how the community can come together to evaluate these ML-based ESMs, with the expectation that such an effort would lead to their wider use.

Key Points:

- A multi-pronged strategy is proposed for standardized evaluation of machine learning based Earth-system models (ESMs)
- Unlike weather forecasting models, ESMs need to be credible when external factors are very different from history
- Consistency with past observations is necessary but insufficient; instead idealized test cases and emergent constraints can be used

Correspondence to:
P. A. Ullrich,
ullrich@llnl.gov

Citation:
Ullrich, P. A., Barnes, E. A., Collins, W. D., Dagon, K., Duan, S., Elms, J., et al. (2025). Recommendations for comprehensive and independent evaluation of machine learning-based Earth system models. *Journal of Geophysical Research: Machine Learning and Computation*, *2*, e2024JH000496. https://doi.org/10.1029/2024JH000496

Received 25 OCT 2024
Accepted 23 FEB 2025

Author Contributions:
Conceptualization: P. A. Ullrich, W. D. Collins, T. A. O'Brien
Funding acquisition: P. A. Ullrich
Investigation: S. Duan, J. Elms
Methodology: P. A. Ullrich, E. A. Barnes, W. D. Collins, K. Dagon, J. Lee, L. R. Leung, D. Lu, M. J. Molina, T. A. O'Brien
Project administration: P. A. Ullrich
Software: S. Duan, J. Elms, F. O. Rebassoo
Supervision: P. A. Ullrich
Visualization: S. Duan, J. Elms
Writing – original draft: P. A. Ullrich, E. A. Barnes, W. D. Collins, K. Dagon,

© 2025 The Author(s). *Journal of Geophysical Research: Machine Learning and Computation* published by Wiley Periodicals LLC on behalf of American Geophysical Union.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Earth system models (ESMs) have long been and continue to be invaluable tools for deepening our understanding of the Earth system and its response to forcing, such as greenhouse gasses, aerosols, land-use land-cover change, and solar irradiance. Scientists use ESMs to understand how the Earth system looked in the past and how it may look in the future. A wide array of end-users apply the data produced by ESMs to assess risk and inform adaptation planning. Despite their broad importance, traditional ESMs are also computationally expensive and incorporate sometimes crude assumptions about fundamental physical processes, leading to significant biases in key climatological fields (Chen et al., 2021; Christopoulos & Schneider, 2021; Moreno-Chamarro et al., 2021; Zhou et al., 2022). However, recent and rapid progress in machine learning (ML) has the potential to advance solutions to these problems. Building upon the successes of ML in other fields, ML-based weather forecast models have emerged that, once trained, can leverage hardware accelerators to produce skillful weather forecasts at a much lower time-to-solution than conventional physics-based models (e.g., Ben Bouallègue et al., 2024; Lam

J. Lee, L. R. Leung, D. Lu, M. J. Molina, T. A. O'Brien
Writing – review & editing: P. A. Ullrich, E. A. Barnes, W. D. Collins, K. Dagon, J. Lee, L. R. Leung, D. Lu, M. J. Molina, T. A. O'Brien, F. O. Rebassoo

Table 1
Descriptors Commonly Used to Describe ML-Based Weather and Climate Models

Descriptor	Summary
Forecasting	Models trained primarily on historical weather data or operational forecast archives; these systems may neglect the coupling of Earth system components or climate-relevant fields altogether, like sea-surface temperatures or irradiance, which are essential for ESMs.
Autoregressive/Roll outs	Models whose output from the current iteration becomes the input for the next iteration.
Surrogate	Models trained to predict a specific quantity (e.g., global temperature change as a function of greenhouse gas concentration).
Hybrid	Traditional physics-based weather/climate model infrastructure with one or more parameterizations/components replaced with ML.
Emulator	Models trained to emulate a single physics-based weather/climate model.
Foundation	Models pre-trained on multiple, large data sets (e.g., CMIP6 models, reanalysis data products) that can be fine-tuned to support many downstream applications.

Note. These descriptors are not meant to be mutually exclusive.

et al., 2023). These efforts have stimulated recent work on ESMs that incorporate ML components or are entirely data-driven, including a variety of hybrid models, emulators, and foundation models (see Table 1 for a definition of these terms) (Eyring et al., 2024). Despite promising advancements, building trust in ML-based ESMs for scientific discovery or applications remains an outstanding challenge. **To this end, this paper outlines a plan for comprehensive, standardized, and independent evaluation of ML-based ESMs to strengthen their credibility for wider application, based on five core recommendations.**

ESMs are widely used for various scientific and applied purposes (see Table 2 for a summary of common applications). What sets the problem of Earth system modeling apart from weather modeling is the requirement to achieve top-of-the-atmosphere energy balance in pre-industrial simulations and perform experiments out-of-sample (and sometimes extremely so, in the case of deep-time paleoclimate): ESMs need to be able to

Table 2
Applications of Earth-System Models (ESMs) and Potential Roles of ML-Based ESMs

ESM application	Role of ML-Based ESMs
Predictions of weather, subseasonal to seasonal (S2S), or subseasonal to decadal (S2D) time scales	Rapid generation of huge forecast ensembles to quantify the probability of different climate states on weather/S2S/S2D time scales. ML-based models could be more skillful than physics-based models because of direct training on observations (consequently avoiding assumptions sometimes made necessary by unresolved processes; although this is aspirational for ESMs, there is evidence of this being the case for ML-based hydrologic models, e.g., Kratzert et al., 2019).
Assessing the risk of extreme weather or low-likelihood high-impact events	Rapid generation of huge forecast or projection ensembles to better quantify the frequency and character of rare events.
Detection and attribution	Cheap generation of large climate ensembles to increase signal-to-noise ratio; using explainable AI techniques to decompose temperature or precipitation time series into components of forcing (e.g., Sweeney & Fu, 2024).
Transient climate response and equilibrium climate sensitivity	Quantifying global temperature as a result of CO_2 and/or other radiative forcing and Earth system feedbacks across climate timescales.
Process understanding and hypothesis testing	Rapid testing of ideas before running more expensive physics-based model simulations; use of auto differentiation to probe input/output relationships.
Climate change projections and uncertainty quantification	Rapid generation of huge ensembles of future climate projections to characterize uncertainties; computationally efficient quantification of the model prediction uncertainty.
Assessment of climate impacts	Computationally efficient downscaling of coarse resolution climate projections to a finer resolution for assessment of regional climate impact.
Tipping points and "what if" science	Unclear. This topic is far out of sample, and so may require physics-based modeling at a fundamental level.

29935210, 2025, 1, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000496 by Open Research - University Of East Anglia , Wiley Online Library on [14/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

29935210, 2025, 1, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000496 by Open Research - University Of East Anglia , Wiley Online Library on [14/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

simulate the coupled Earth system under forcings outside the range of the period of instrumental record (generally since industrialization). This is primarily accomplished through adherence to the physical laws that govern the Earth system, including fundamental concepts such as conservation of mass, momentum, and energy, and consistency with emergent behavior such as geostrophic balance or constraints on globally averaged precipitation. If these laws are not explicitly enforced in the physics-based or ML-based model, it is essential to demonstrate that the respective model somehow exhibits them. With only one period of record covering a short and recent span of the Earth's history available to us, we need confidence that ESMs can do more than align with these observations; they need to get the right answer for the right reasons to be useful for understanding and simulating future or past climates. Coming up with a testing strategy that goes beyond historical comparison requires us to build upon a long history of ESM evaluation and devise creative strategies to evaluate models' physical consistency.

Despite the challenges still ahead, ML-based ESMs are potentially a step-change in our ability to do meaningful science and deliver actionable information to decision-makers. Most remarkably, emulators and foundation models (see Table 1), again in large part because of their capacity to use a simplified set of operations on hardware accelerators, can provide an order-of-magnitude or greater speed-up (and lower computing power per simulated year) compared with physics-based models. This enables the construction of very large ensembles that deliver improved uncertainty quantification (e.g., Kochkov et al., 2024; Mahesh et al., 2024a, 2024b), higher signal-to-noise ratios (relevant for detection and attribution), more exhaustive exploration of low-likelihood high-impact events and the tails of statistical distributions, and the capacity to generate future climate scenarios by interpolating between available ones. Notably, and as will be discussed later, the historical record alone isn't sufficient to properly sample the space of possible weather states, and likely needs to be supplemented by training data from physics-based models.

Many in the climate community are rightly skeptical of the ability of emulators to generalize beyond the bounds of their training data sets: especially when changes in forcings are involved. Much work still remains to assess whether these models are credible for producing future projections, particularly under forcings that are far from recent history. A related question has been a central theme in ESM research over the past several decades: how do we assess the trustworthiness of an ESM's ability to operate in regimes where we do not know the correct answer? In the past, climate scientists have approached this question, in part, by various intercomparison efforts and hierarchical model evaluation approaches. We argue that such an approach is relevant for ML-based ESMs as well. This scientifically grounded effort to establish model credibility will allow ML-based ESMs to complement and expand upon the ecosystem of tools for understanding the Earth's past and future.

Motivated by the considerable excitement that has brewed around ML-based ESMs, this paper puts forward the following recommendations to the growing ML-based ESM development community for comprehensive and independent evaluation of these models:

1. Build upon the experience of the ESM community and its long history with model intercomparison while leveraging shared model code, data, and diagnostics.
2. Assemble a suite of idealized and simplified tests to evaluate the basic behavior of ML models.
3. Intercompare ML-based ESMs, physics-based regional and global models, and related statistical techniques using common metrics.
4. Develop an extensible evaluation framework useful for both ML-based and traditional models that is widely accepted by the scientific community and relevant users.
5. Identify a trusted independent party to manage regular intercomparison of ML-based and physics-based models.

The remainder of this paper walks through these recommendations, connecting with the substantial work already done in ESM evaluation. It concludes with our draft proposal for a common evaluation strategy among ML-based ESMs, aiming to inspire further developments in this space.

2. Building on the Experience of the Earth System Modeling Community

Coordinated ESM evaluation has a long and storied history, originating in the early 1980s with Robert Cess' organization of the Feedback Analysis of GCMs (general circulation models) and Intercomparison with Observations (FANGIO) project (Potter et al., 2011). This effort, using prescribed sea surface temperature experiments, led to the first multi-model constraints on climate sensitivity (which have taken decades to refine;

29935210, 2025, 1, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000496 by Open Research - University Of East Anglia , Wiley Online Library on [14/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3
Common Benchmarking Toolsets for Evaluation of Global Earth-System Models

Tools	Primary evaluation focus	Citation
PCMDI Metrics Package (PMP)	Mean climate, variability modes, extremes	Lee et al. (2024)
International Land Model Benchmarks (ILAMB)	Land surface	Collier et al. (2018)
International Ocean Model Benchmarks (IOMB)	Ocean	Fu et al. (2022)
Cyclone Metrics Package (CyMEP)	Tropical cyclones	Zarzycki et al. (2021)
ESMValTool	Mean climate, processes, variability	Eyring, Righi, et al. (2016)
Climate Variability Diagnostics Package for Large Ensembles (CVDP-LE)	Climate variability in large ensembles	Philips et al. (2020)
Model Diagnostics Task Force (MDTF)	Processes	Neelin et al. (2023)

Sherwood et al., 2020), and was the basis for the Atmospheric Model Intercomparison Project (AMIP; Gates, 1992) and subsequent Coupled Model Intercomparison Projects (CMIPs; Eyring et al., 2016b; Meehl et al., 2000, 2005; Taylor et al., 2012). The consensus from this work was that intercomparison was effective at making better climate models (Meehl et al., 1997), with a steady increase in the fidelity of "top tier" models (e.g., Ahn et al., 2022; Eyring et al., 2021; Lee et al., 2024). A wealth of products have emerged from these past efforts, including model codes, data, and diagnostics, all of which need to be considered in the ML model landscape.

There is no doubt that the ML model development community should leverage these past efforts and draw upon the deep experience of the ESM evaluation community. Perhaps the lowest hanging fruit is through direct application of the many ESM evaluation toolkits now available. Observing systems have recorded key climate variables since the late 1800s, and have been supplemented over the past four decades by comprehensive remotely sensed measurements from an array of global satellites. The summary statistics derived from these observations are an essential reference for ESM performance and have been widely vetted for use in a wide variety of model benchmarking toolkits. Benchmarking toolkits incorporate suites of metrics and diagnostics that measure and depict consistency between model simulations and observations. Metrics and diagnostics evaluate meteorological patterns, mean climatology, climate variability (including well-known modes of variability such as the El Niño-Southern Oscillation), extreme weather events, physical processes, emergent relationships, and historical trends (under changes in observed forcing). A list of some more commonly employed packages is provided in Table 3; further information on available tools in the community can be found in Dingley (2025). Use of these evaluation tools would allow ML model developers to draw on decades of evaluation experience with physics-based ESMs, and feedback from the ML community would allow developers to augment their toolkits with benchmarks appropriate for ML models. These tools are also applied with relative ease; Figure 1 depicts a portrait plot from the PCMDI Metrics Package which shows evaluation results for the Ai2 Climate Emulator (ACE; Watt-Meyer et al., 2023) as trained on the Energy Exascale Earth System Model (E3SM) v2 (Duncan et al., 2024), relative to other CMIP6 models. As an additional example, Cresswell-Clay et al. (2024) employ "Taylor diagrams" (Taylor, 2001) to evaluate modes of extratropical climate variability in the Deep Learning Earth System Model (DLESyM) as compared to ERA5. Data standardization is also important in facilitating the use of these tools, and should leverage software packages such as the Climate Model Output Rewriter (CMOR; Mauzey et al., 2024). If such tools are widely employed in the ML model development community, intercomparison would be far easier between such models. Notably, Table 4 lists several efforts underway to develop evaluation packages targeted specifically at ML-based ESMs, though these are in the early stages.

Although the application of existing ESM benchmarking packages is an essential first step, there are unavoidable challenges and limitations in this approach. The majority of the metrics and diagnostics from these packages focus on comparing against historical observations, despite the historical record only covering a period of relatively stationary climate and only one possible meteorological trajectory. For many ML-based climate models, the need to avoid testing models against the data used to train them further limits our capacity to compare them against observations (in case those observations were used as training data). Just as physics-based models may be tuned to reduce model-observation differences, ML-based models may be trained to statistically reproduce the observations without learning the underlying physics or being able to extrapolate beyond the historical training period. This concern underscores the need for more creativity in devising evaluation strategies, some of which we explore here.

10.1029/2024JH000496

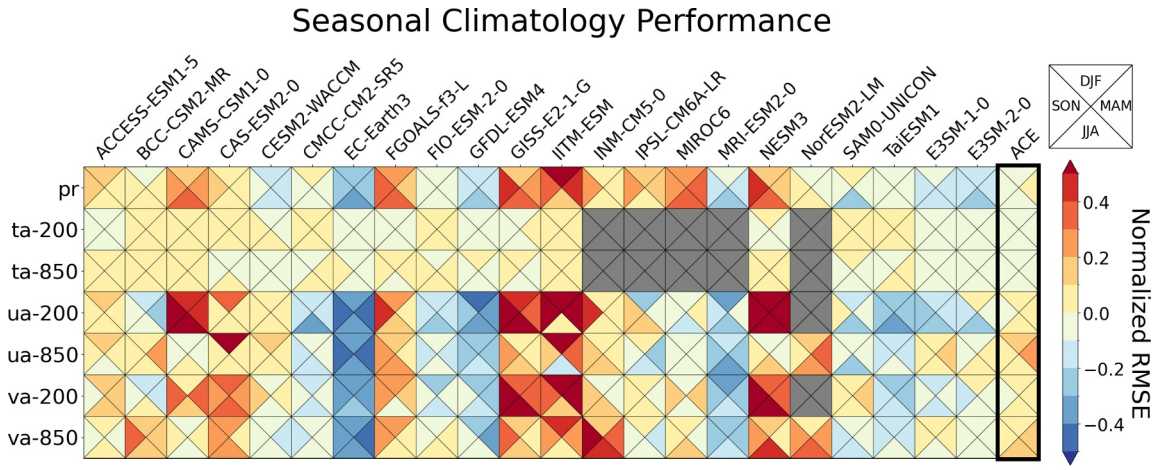

Figure 1. Tools like the PCMDI metrics package can be easily adapted to evaluate ML-based climate modeling systems. Here the Ai2 Climate Emulator (ACE; Watt-Meyer et al., 2023), trained on time-varying sea-surface temperature E3SMv2 data, is intercompared with CMIP6 models on seasonal measures of normalized root-mean-square error of precipitation (pr) compared to GPCP (Adler et al., 2018), along with 200- and 850-mb temperature (ta-200, ta-850), zonal wind speed (ua-200, ua-850) and meridional wind speed (va-200, va-850) compared to ERA5 reanalysis data (Hersbach et al., 2020). For ACE, the seasonal measures at pressure levels are interpolated from the native output vertical levels. The RMSE of each variable is normalized by the median RMSE of all models. The blue color indicates the model performs relatively better than other models, and the red color indicates the opposite. A result of +0.2 indicates an error that is 20% greater than the median RMSE across all models, whereas −0.2 indicates an error that is 20% less than the median. A detailed description of the portrait plot can be found in Lee et al. (2024).

3. Beyond Historical Evaluation

Out-of-sample projections of the climate system are generally considered credible if they are grounded in physical principles (i.e., the conservation laws that underlie all of modern physics) and their emergent properties are consistent with our understanding of the Earth system. Conservation laws in the atmosphere were discussed at length in Thuburn (2008), and it was concluded that only a few of the conserved quantities could be used effectively to measure model performance. Indeed, beyond mass (and to a lesser degree energy, with appropriate tuning, and only in coupled ESMs), no other quantity is explicitly conserved in most physics-based ESMs. With that said, mass conservation lends itself to more than global metrics of invariance. It also means that tracer concentrations can't go negative: for example, in each grid cell precipitation needs to be strictly limited by the amount of water vapor in the column. So these metrics provide a good start for evaluating models beyond their ability to reproduce the historical records. That said, these constraints are insufficient on their own: For instance, the conservation of mass and energy does not translate to any sort of useful constraint on equilibrium climate sensitivity, one of the most important quantities in climate science. Instead, some of the most scientifically and societally relevant quantities emerge from a complex series of processes and feedbacks.

Evaluation of model performance can also rely on the properties that emerge from fundamental physics. From basic atmospheric science, we know air parcel saturation limits the amount of water vapor present in a grid cell before condensation and precipitation will occur; while supersaturation can occur in the upper troposphere, it is extremely rare for it to occur in the near-surface or mid-troposphere. More complex constraints could also be included. For instance, theory indicates that globally averaged precipitation increases at a rate ∼1–2%/K while atmospheric water vapor increases at ∼7%/K (Held & Soden, 2006; Pendergrass & Hartmann, 2014). Many other

Table 4
Benchmarking Toolsets for Evaluation of Global ML-Based Earth-System Models

Tools	Primary evaluation focus	Citation
Earth-2 MIP	Weather and climate	NVIDIA (2024)
ClimateBench	Physically based ESM emulators	Watson-Parris et al. (2022)
ClimDetect	Climate change detection and attribution	Yu et al. (2024)
ClimSim-Online	Hybrid ML-physics models	Yu et al. (2023)
WeatherBench 2	Weather	Rasp et al. (2024)

29935210, 2025, 1, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000496 by Open Research - University Of East Anglia , Wiley Online Library on [14/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

emergent constraints have also been explored in the literature in recent years, developed both as a means of assessing models' consistency with the real world and as an alternate pathway to link historical observations with future change (Hall et al., 2019; Klein & Hall, 2015; Shiogama et al., 2022; Williamson et al., 2021).

Power spectra have been embraced as one approach to assessing the two-dimensional spatial structures of the simulated fields (e.g., Karlbauer et al., 2024). In multiple cases, the spectra reveal that some ML models tend to damp higher wavenumbers more strongly than physics-based models, and that the model behavior in the small-scale, damped regime may be less physically consistent (e.g., not exhibiting geostrophic balance) than physics-based models (Bonavita, 2024; Kochkov et al., 2024). These results suggest that the "effective resolution" of some ML-based weather and climate models is likely coarser than in physics-based models with similar grid spacing. This behavior may be tied to the chosen loss function and model architecture, with some recent examples seeming to avoid excessive damping (e.g., Mahesh et al., 2024a, 2024b; Price et al., 2025). Thus, metrics related to power spectra and co-spectra can help track effective versus reported grid resolution as new emulators are developed (Rasp et al., 2024).

Relationships between variables (i.e., covariances) provide another mechanism to test consistency within the model. For example, heat waves are often associated with weaker wind speeds since they are related to high-pressure ridges. Recent work by Zhang and Boos (2023) also highlighted the temporal pattern in precipitation that occurs around the hottest days of the year, where dry conditions are generally present before temperature peaks, followed by moist conditions once high temperatures drive sufficient atmospheric instability to initiate convection. Such behavior could readily inform the development of an associated metric or diagnostic.

Feature structure is another avenue for model evaluation where ML-based forecasting models have struggled (Bonavita, 2024). Feature tracking packages such as TempestExtremes (Ullrich et al., 2021) and the Toolkit for Extreme Climate Analysis (TECA; Loring et al., 2023) can be employed to track tropical cyclones, extratropical cyclones, monsoon depressions, atmospheric rivers, and other important atmospheric features. Trackers can also be employed in the ocean, such as in Mason et al. (2014), or on the land-surface, as in Liu et al. (2021). Other features, such as wildfires, marine heatwaves or sea ice extent, could also be incorporated. Composites of the environmental structure can then reveal anomalous or unphysical behavior.

4. Revisiting Simplified Tests in the Age of Machine Learning

Given the inherent complexity of physics-based ESMs, evaluation begins far before most physical processes are even incorporated into the model. Software engineering operates in much the same way—rather than only testing the software package once all the functionality has been implemented, unit testing is employed to validate each function as it is added. Applied to an ESM, hierarchical evaluation means focusing first on the simplest model that exhibits behavior consistent with the component of the Earth system being tested (e.g., dry dynamics with no topography or surface fluxes for the atmosphere), making sure physical laws are obeyed at each timestep, and gradually adding complexity. This process allows us to evaluate ESM behavior in a simpler context, where analytic or semi-analytic solutions are available. Notably, O'Loughlin et al. (2024) advocate for such a modular design as a way to develop a component-level understanding of ML models.

In the case of the atmospheric component model, testing begins with the dynamical core, which is responsible for solving the fluid flow equations in the thin spherical annulus that wraps the Earth. Major physical components are then added one at a time, beginning with moisture and moist processes, ocean-atmosphere fluxes, topography, and finally a land surface and fully coupled ocean (see below). Such a traditional ESM evaluation hierarchy for the atmospheric model component is given in Table 5.

In a manner similar to the atmospheric component model, testing of the land component can begin with the biophysical core, which is responsible for simulating terrestrial water and energy balance processes. Major physical components are then added one at a time, beginning with vegetation dynamics and photosynthesis, carbon and nitrogen cycles, topography-driven runoff, and finally human land use and land cover change (Fisher & Koven, 2020). Throughout this process, the high heterogeneity of land surfaces necessitates testing across diverse landscapes to ensure robust performance across varied environmental conditions (Byrne et al., 2024). When coupling to the atmosphere, an important consideration for land modeling is the representation of surface turbulent fluxes, and notably sub-daily fluxes, for capturing land-atmosphere feedbacks (Chaney et al., 2024).

29935210, 2025, 1, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000496 by Open Research - University Of East Anglia , Wiley Online Library on [14/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

29935210, 2025, 1, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000496 by Open Research - University Of East Anglia , Wiley Online Library on [14/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 5
The Traditional Earth-System Model Evaluation Hierarchy (Atmospheric Model Component)

Step in the Hierarchy	What is tested
2D shallow water experiments (e.g., Williamson et al., 1992)	Horizontal fluid dynamics on a sphere, barotropic energy exchange across scales
3D without moisture (e.g., Jablonowski & Williamson, 2006)	Baroclinic dynamics (e.g., baroclinic instability, one of the most fundamental atmospheric processes), baroclinic energy exchange across scales
3D with moisture and simplified physics (e.g., Reed & Jablonowski, 2012)	Isolated dynamical processes in the global system
Aquaplanet experiments (e.g., Hess et al., 1993)	Atmospheric model behavior without land-atmosphere or dynamic ocean
Fixed sea-surface temperature runs (e.g., Gates et al., 1999)	Atmosphere and land-surface behavior with tight constraints on model drift
Fully coupled simulations (e.g., Gates et al., 1995)	Fully coupled modeling system

Simplified tests for the ocean component can follow the recommendations of Bishnu et al. (2024), which outlines a suite of shallow water test cases with increasing complexity. Additional examples of simplified tests for the coupled system include testing the ML atmosphere model coupled to a hierarchy of ocean models such as slab ocean, mixed layer ocean, and full ocean models (Hsu et al., 2022).

For ML-based ESMs, such a hierarchy may seem unnecessary if the underlying model framework is kept relatively simple. However, these tests are important to demonstrate that the fundamental physical laws are respected in the model in such a way that important emergent phenomena spontaneously manifest (e.g., baroclinic Rossby waves in the atmosphere). They exercise a specific model's ability to simulate physics using initial states that are well outside the model's training data set: effectively building trust in the models by demonstrating (or failing to demonstrate) generalizability.

With that said, the traditional ESM evaluation hierarchy does not necessarily translate to ML-based models, especially tests that involve activating/deactivating/swapping specific model components that represent specific components of the Earth system. ML-based models often lack explicitly distinct components representing specific physical processes, unlike their physics-based counterparts. Consequently, deactivating particular physics (e.g., moist physics) in these ML models may not be practical through simple component deactivation. Instead, it might require retraining the entire model using data from a physics-based model where the specific process has been deactivated. However, retraining a ML-based model for a particular test (e.g., as needed for an atmosphere model running the 2D shallow water experiments) introduces new challenges with interpreting what can be learned about the more general ML-based ESM. Several of the idealized atmospheric test cases highlighted in Table 5 can be run in certain ML-based ESMs, assuming that those models allow for prescribed topography and radiative forcing. For instance, the baroclinic instability test of Jablonowski and Williamson (2006) is a dry dynamical core test case with prescribed topography and zero radiative forcing, with a well-documented numerical solution for short-period runs, suitable for a forecasting model or ESM. This test has been essential in evaluating atmospheric dynamical cores, given the importance of baroclinic instability to global weather patterns. Similarly, the tropical cyclone test of Reed and Jablonowski (2012) could be run in any ESM with at least 0.5° grid spacing, as long as topography can be removed entirely, and would be helpful for assessing the ability of a model to capture the structure of simulated tropical cyclones. Even advection could be tested under the formulation documented in the 2008 Dynamical Core Model Intercomparison Project (DCMIP) test case document for test 1-0-y (Jablonowski et al., 2008). Other such tests are documented as part of the 2008, 2012, and 2016 DCMIP test case documents (Jablonowski et al., 2008; Ullrich et al., 2012, 2016).

To demonstrate that ML-based models are capable of performing such idealized tests, despite these tests being outside the range of their training data, we show results from a moist baroclinic wave test using the Spherical Fourier Neural Operators (SFNO) model of Bonev et al. (2023). We implement the initial condition specification of Bouvier et al. (2024), which uses an analytical formula to define the initial conditions. This choice was made because other baroclinic wave tests require numerically integrating quantities in a way consistent with a model's dynamical core: this is not possible in most emulators since the numerical discretization scheme is not explicitly knowable. Results from this test are depicted in Figure 2. While the model does produce a Rossby wave, the test immediately reveals an important, albeit expected, deviation of SFNO from the behavior of physics-based models: the imprint of topography and continental geometry immediately becomes evident in many fields, despite the initial condition being zonally uniform. This expected behavior emerges because SFNO internalizes

10.1029/2024JH000496

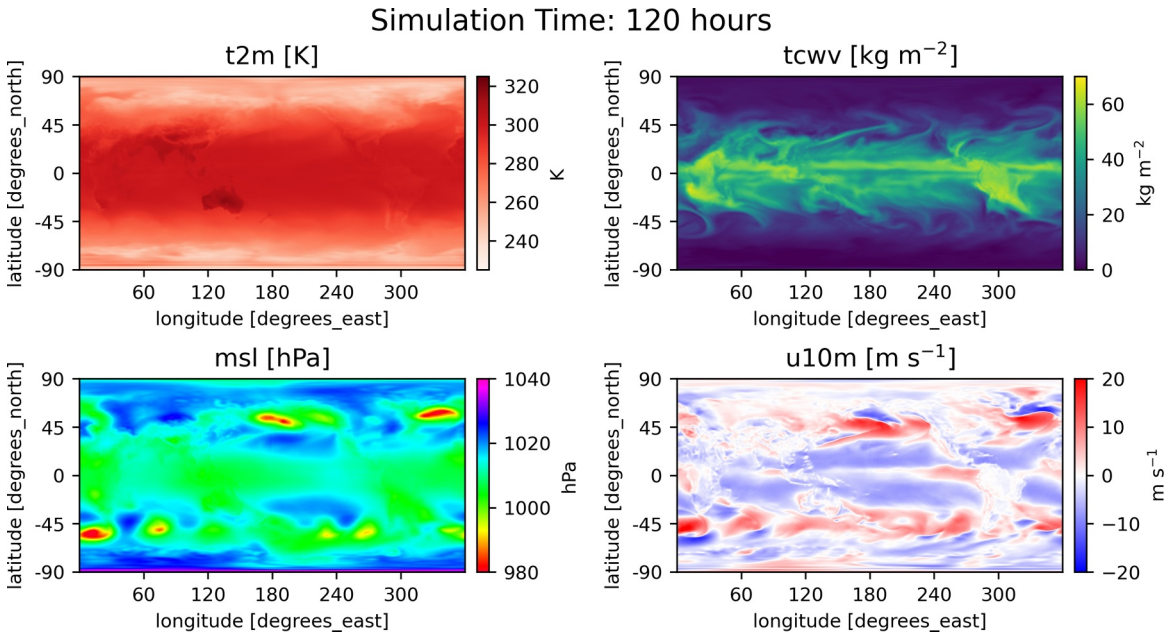

Figure 2. Snapshots of 2 m temperature (t2m), total column water vapor (tcwv), mean sea-level pressure (msl) and 10 m wind speed (u10 m) from the Bouvier et al. (2024) idealized baroclinic wave test at 120 hr after initialization, as simulated using the Spherical Fourier Neural Operators model of Bonev et al. (2023).

the impact of Earth's topography and land-sea contrast on atmospheric flow, since all the training data involved real Earth topography (as pointed out in Koldunov et al., 2024, this behavior can actually be harnessed to effectively downscale data). To the best of our knowledge, no presently available ML-based ESM supports arbitrary, prescribed topography. Nonetheless the test remains stable and the results otherwise consistent with expectations, highlighting the power in utilizing idealized tests to evaluate ML-based models.

Besides those tests commonly employed in physics-based model development, there are also tests that are unnecessary in most physics-based ESMs since the explicit numerical methods used by these ESMs encode local causality directly. However, in ML models, causality can and should be explicitly tested. To do so one could begin with a near-steady atmospheric flow and add a large local perturbation at the scale of a single grid point. The propagation of that disturbance should be limited to the acoustic wave (in the case of a hydrodynamic perturbation) or the maximum wind speed (in the case of a tracer perturbation). Such a test could reveal if the ML model allows information to propagate faster than physics would allow.

In recent work by Hakim and Masanam (2024), four idealized tests were conducted with realistic topography and a climatological mean steady state, which may be more suitable for ML-based models that do not allow modification of topography. These tests include steady tropical heating with Matsuno–Gill response, an extratropical cyclone test analogous to Jablonowski and Williamson (2006) except over the Pacific, a test of geostrophic adjustment, and a test looking at Atlantic tropical cyclogenesis. The authors used these novel experiments to conclude that Pangu-Weather (an ML-based global forecasting model; Bi et al., 2022) does likely encode physical principles since its behavior is consistent with physical understanding even for tests in which the idealized initial conditions are far outside the distribution of initial conditions in the training data. Similar tests could also be employed in the open ocean, away from the coast, in ML-based ocean models.

5. ML Models Are Part of the Broader Ecosystem

While we may not be able to definitively conclude whether ML-based ESMs are producing the right answer for the right reasons, we can evaluate if the ML-based models are within or outside the envelope of pre-existing models under alternate forcings (e.g., under projected warming). Namely, we can compare projections from ML-based models side-by-side with existing physics-based models, reduced complexity models (Nicholls et al., 2021), and statistical/dynamical downscaling methods. The toolkits discussed in Section 2 could be used for these experiments, except instead of comparing models against observations, the alternate forcing experiment

29935210, 2025, 1, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000496 by Open Research - University Of East Anglia , Wiley Online Library on [14/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

29935210, 2025, 1, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000496 by Open Research - University Of East Anglia , Wiley Online Library on [14/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

would be compared to the respective historical simulation. This approach could then be used to cover a variety of variables, regions, and experiments. One such effort in this direction is ClimateBench (Watson-Parris et al., 2022), which uses training data from the Norwegian Earth System Model (NorESM2; Seland et al., 2020) and measures models' consistency with these projections. However, a large enough ensemble of ESM projections would be needed to delineate internal variability from forced response (e.g., Lütjens et al., 2024).

Of course, care needs to be taken in case a discrepancy arises between the ML models and the other physics-based models in the ecosystem. One should not simply assume that the ML models are behaving inconsistently with physical principles, as the reasons for these differences could be far more nuanced and can include physics-based ESM biases.

Hybrid models can also be paired with their "sibling" (i.e., the ESM used to embed ML) and emulators paired with their "parent" (i.e., the physics-based ESM(s) on which they are trained), providing another route for examining consistency. Evaluations could include pairwise hybrid-to-sibling or emulator-to-parent comparisons against observational metrics, allowing us to identify similarities or differences. Alternatively, in the context of a multi-model ensemble (MME), one could ask if the projection spread of hybrid models or emulators is distinguishable from the MME of traditional physics-based ESMs? These comparisons open opportunities to reveal and understand the behaviors of emulators, hybrid models, and physics-based models beyond benchmarking using direct comparison with observations.

6. Developing a Framework for Intercomparison

There is no doubt that the significant successes in ESM development over the past four decades can be attributed to regular intercomparison and healthy competition and collaboration among modeling groups. However, frameworks for intercomparison require agreement and investment from all involved. Using our past experience with evaluating CMIP class ESMs as guidance, there is a clear need for all of the following:

- common metrics and diagnostics recipes that all groups can agree upon to measure success,
- commitment to open source code, pre-trained ML-based ESMs, data access, user documentation for each experiment, and model/data provenance from modeling groups, to ensure transparency, fairness, and reproducible science,
- a consistent output format (for when data is written to disk), compatible with widely accepted, Climate and Forecasting metadata convention standards (Hassell et al., 2017; Eaton et al., 2022, http://cfconventions.org/), and
- an independent organization responsible for validating and collating results from development groups.

The aforementioned framework would be effective at building trust in these ML-based ESM models, and making it clear that nothing is being hidden from data users. The objective would be to provide a buffer against model development groups cherry-picking metrics that show good performance, ensure that all models are evaluated on a level playing field, and support intercomparison by the broad community of ESM developers, climate scientists, and users of ESM model outputs.

With a common, prescribed metrics recipe, there are always concerns that groups will engage in "metric hacking"—namely, training the ML-based model or tuning the physics-based model to maximize their performance scores as a shortcut to building a skillful model. However, we would also argue that there should not be a single metric to summarize performance (it is not even clear if any single metric would make sense for something as complicated as an ESM). If groups are obliged to show all metrics, then optimizing among all metrics becomes nearly impossible even with sophisticated uncertainty quantification, multi-objective optimization, and model calibration methods. Regardless, such optimization would simply result in a model that is more consistent with physical principles, if these evaluation metrics are well defined. Inevitably, it is necessary to pick some quantitative measure of skill (i.e., a loss function) for training an ML system, and efforts to devise better loss functions that enable more comprehensive performance gains could be an exciting and innovative direction for research.

7. Summary

Although ML-based ESMs are in their infancy, there is no better time than now to lay out common standards for ML-based ESM evaluation and encourage common targets for intercomparison. Physics-based ESMs have been undergoing standardized intercomparison for 40 years, and in that time have made considerable progress in their

29935210, 2025, 1, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000496 by Open Research - University Of East Anglia , Wiley Online Library on [14/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 6
An Example of a Menu for ML-Based Earth-System Model Evaluation

Benchmarking packages (§2)
PCMDI Metrics Package (PMP)
• Mean climate
• ENSO metrics
• Tropical and extratropical modes of variability
• Monsoon metrics
• Extremes metrics
• Precipitation variability
• Cloud feedbacks[a]
International Land Model Benchmarks (ILAMB)[b]
International Ocean Model Benchmarks (IOMB)[c]
Cyclone Metrics Package (CyMEP) ESMValTool
Model Diagnostics Task Force (MDTF)
Sanity checks (§3)
Global mass conservation
Non-negative tracer mass
No near-surface supersaturation
Performance metrics (§3)
Zonal means
Power spectra
Covariances
Feature (e.g., TC) frequency, character and structure
Historic (1900-present) temperature trends under historic forcing
Emergent constraints (§3)
Global water vapor change (∼7%/K)
Global precipitation change (∼1–2%/K)
Ratio of OLR to Niño-3.4 index
Others?
Idealized Test Cases (§4)
Steady-state advection (DCMIP2008)[d]
Baroclinic wave test case (DCMIP2008)[d]
Mountain-induced Rossby wave (DCMIP2008)[d]
3D Rossby-Haurwitz wave (DCMIP2008)[d]
Moist baroclinic instability (DCMIP2012)[d]
Tropical cyclone test (DCMIP2012)[d]
Causality test
Steady tropical heating (H&M2024)
Extratropical cyclone development (H&M2024)
Geostrophic adjustment (H&M2024)
Atlantic hurricane development (H&M2024)

[a]AMIP and AMIP +4K experiments required. [b]Land model evaluation. [c]Ocean model evaluation. [d]Only models that allow prescription of bottom topography.

ability to simulate the Earth system with minimal intervention. By leveraging the knowledge, experience, and capabilities built up by the scientific and model evaluation community over this time, comprehensive evaluation of ML-based ESMs is achievable in a relatively short time. Our discussion of benchmarks for this new generation of models is summarized in Table 6 by no means intending to be the final say on the types of evaluation that should be performed, but a starting point for a conversation between the development and evaluation communities. The authors of this paper have already begun work to codify this menu, yielding greater insight into the capability of ML-based models to simulate the Earth system out-of-sample.

Data Availability Statement

The ACE-E3SMv2 AMIP climatology used to generate Figure 1 are available from Duan (2024). The results of the idealized baroclinic wave run with the SFNO model of Bonev et al. (2023) that were subsequently used to generate Figure 2 are available from Elms (2024).

Acknowledgments

The AMIP-type ACE simulation shown in Figure 1 was funded by LLNL-LDRD Project 22-ERD-052 PlusUp. We thank C. Bonfils, G. Pallotta and P. Caldwell for associated discussions. Works of LLNL-affiliated authors were performed under the auspices of the U.S. DOE by the Lawrence Livermore National Laboratory (LLNL) (contract no. DE-AC52-07NA27344) and their efforts were supported by the Regional and Global Model Analysis (RGMA) program area of the U.S. Department of Energy (DOE) Office of Science (SC), Biological and Environmental Research (BER) program. PNNL is operated for DOE by Battelle Memorial Institute under contract DE-AC05-76RL01830. Oak Ridge National Laboratory is operated by UT Battelle, LLC, for the DOE under contract DE-AC05-00OR22725. LBNL's contributions were supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research (BER), Regional and Global Model Analysis (RGMA) component of the Earth and Environmental System Modeling Program under Award Number DE-SC0022070 and National Science Foundation (NSF) IA 1947282. This work was also supported by the National Center for Atmospheric Research (NCAR), which is a major facility sponsored by the NSF under Cooperative Agreement No. 1852977. This work was also supported, in part, by DOE BER as part of the Program for Climate Model Diagnosis and Intercomparison Project. This research was also supported in part by the Environmental Resilience Institute, funded by Indiana University's Prepared for Environmental Change Grand Challenge initiative and in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute. This work is also supported by DOE BER RGMA under University Award Numbers DE-SC0024093 and DE-SC0023519.

References

Adler, R. F., Sapiano, M. R., Huffman, G. J., Wang, J. J., Gu, G., Bolvin, D., et al. (2018). The Global Precipitation Climatology Project (GPCP) monthly analysis (new version 2.3) and a review of 2017 global precipitation. *Atmosphere*, *9*(4), 138. https://doi.org/10.3390/atmos9040138

Ahn, M. S., Gleckler, P. J., Lee, J., Pendergrass, A. G., & Jakob, C. (2022). Benchmarking simulated precipitation variability amplitude across time scales. *Journal of Climate*, *35*(20), 6773–6796. https://doi.org/10.1175/jcli-d-21-0542.1

Ben Bouallègue, Z., Clare, M. C., Magnusson, L., Gascon, E., Maier-Gerber, M., Janoušek, M., et al. (2024). The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, *105*(6), E864–E883. https://doi.org/10.1175/bams-d-23-0162.1

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2022). Pangu-weather: A 3d high-resolution model for fast and accurate global. *Weather and Forecasting*. arXiv preprint arXiv:2211.02556.

Bishnu, S., Petersen, M. R., Quaife, B., & Schoonover, J. (2024). A verification suite of test cases for the barotropic solver of ocean models. *Journal of Advances in Modeling Earth Systems*, *16*(4), e2022MS003545. https://doi.org/10.1029/2022MS003545

Bonavita, M. (2024). On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, *51*(12), e2023GL107377. https://doi.org/10.1029/2023gl107377

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. arXiv preprint arXiv:2306.03838.

Bouvier, C., van den Broek, D., Ekblom, M., & Sinclair, V. A. (2024). Analytical and adaptable initial conditions for dry and moist baroclinic waves in the global hydrostatic model OpenIFS (CY43R3). *Geoscientific Model Development*, *17*(7), 2961–2986. https://doi.org/10.5194/gmd-17-2961-2024

Byrne, M. P., Hegerl, G. C., Scheff, J., Adam, O., Berg, A., Biasutti, M., et al. (2024). Theory and the future of land-climate science. *Nature Geoscience*, *17*(11), 1–8. https://doi.org/10.1038/s41561-024-01553-8

Chaney, N., Verhoef, A., Findell, K., Abramowitz, G., Edwards, J., Parazoo, N., et al. (2024). Back to the drawing board? Modeling land surface fluxes and their coupling with the atmosphere. *GEWEX Quarterly*, *34*(2).

Chen, D., Dai, A., & Hall, A. (2021). The convective-to-total precipitation ratio and the "drizzling" bias in climate models. *Journal of Geophysical Research: Atmospheres*, *126*(16), e2020JD034198. https://doi.org/10.1029/2020jd034198

Christopoulos, C., & Schneider, T. (2021). Assessing biases and climate implications of the diurnal precipitation cycle in climate models. *Geophysical Research Letters*, *48*(13), e2021GL093017. https://doi.org/10.1029/2021gl093017

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., et al. (2018). The International Land Model Benchmarking (ILAMB) system: Design, theory, and implementation. *Journal of Advances in Modeling Earth Systems*, *10*(11), 2731–2754. https://doi.org/10.1029/2018ms001354

Cresswell-Clay, N., Liu, B., Durran, D., Liu, A., Espinosa, Z. I., Moreno, R., & Karlbauer, M. (2024). A deep learning earth system model for stable and efficient simulation of the current climate. arXiv preprint arXiv:2409.16247.

Dingley, E. (2025). Tools for using climate data - Coupled Model Intercomparison Project. Retrieved from https://wcrp-cmip.org/tools/

Duan, S. (2024). ACE AMIP temperature climatology [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.13994330

Duncan, J. P., Wu, E., Golaz, J. C., Caldwell, P. M., Watt-Meyer, O., Clark, S. K., et al. (2024). Application of the AI2 Climate Emulator to E3SMv2's global atmosphere model, with a focus on precipitation fidelity. *Journal of Geophysical Research: Machine Learning and Computation*, *1*(3), e2024JH000136. https://doi.org/10.1029/2024jh000136

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Blower, J., et al. (2022). NetCDF Climate and Forecast (CF) meta-data conventions V1.10. Retrieved from http://cfconventions.org/Data/cf-conventions/cf-conventions-1.10/cf-conventions.html

Elms, J. (2024). Idealized baroclinic wave test results [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.13993647

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Eyring, V., Collins, W. D., Gentine, P., Barnes, E. A., Barreiro, M., Beucler, T., et al. (2024). Pushing the frontiers in climate modelling and analysis with machine learning. *Nature Climate Change*, *14*(9), 1–13. https://doi.org/10.1038/s41558-024-02095-y

Eyring, V., Gillett, N. P., Achuta Rao, K. M., Barimalala, R., Barreiro Parrillo, M., Bellouin, N., et al. (2021). Human influence on the climate system. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, et al. (Eds.), *Climate change 2021: The physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change* (pp. 423–552). Cambridge University Press. https://doi.org/10.1017/9781009157896.005

Eyring, V., Righi, M., Lauer, A., Evaldsson, M., Wenzel, S., Jones, C., et al. (2016). ESMValTool (v1. 0)–A community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP. *Geoscientific Model Development*, *9*(5), 1747–1802. https://doi.org/10.5194/gmd-9-1747-2016

29935210, 2025, 1, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000496 by Open Research - University Of East Anglia , Wiley Online Library on [14/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Fisher, R. A., & Koven, C. D. (2020). Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *Journal of Advances in Modeling Earth Systems*, *12*(4), e2018MS001453. https://doi.org/10.1029/2018ms001453

Fu, W., Moore, J. K., Primeau, F., Collier, N., Ogunro, O. O., Hoffman, F. M., & Randerson, J. T. (2022). Evaluation of ocean biogeochemistry and carbon cycling in CMIP earth system models with the international ocean model benchmarking (IOMB) software System. *Journal of Geophysical Research: Oceans*, *127*(10), e2022JC018965. https://doi.org/10.1029/2022jc018965

Gates, W. L. (1992). AN AMS continuing series: Global CHANGE--AMIP: The atmospheric model intercomparison project. *Bulletin of the American Meteorological Society*, *73*(12), 1962–1970. https://doi.org/10.1175/1520-0477(1992)073<1962:atamip>2.0.co;2

Gates, W. L., Boyle, J. S., Covey, C., Dease, C. G., Doutriaux, C. M., Drach, R. S., et al. (1999). An overview of the results of the Atmospheric Model Intercomparison Project (AMIP I). *Bulletin of the American Meteorological Society*, *80*(1), 29–56. https://doi.org/10.1175/1520-0477(1999)080<0029:aootro>2.0.co;2

Gates, W. L., Henderson-Sellers, A., Boer, G. J., Folland, C. K., Kitoh, A., McAvaney, B. J., et al. (1995). Climate models—Evaluation. *Climate Change*, *1*, 229–284.

Hakim, G. J., & Masanam, S. (2024). Dynamical tests of a deep-learning weather prediction model. *Artificial Intelligence for the Earth Systems*, *3*(3). https://doi.org/10.1175/aies-d-23-0090.1

Hall, A., Cox, P., Huntingford, C., & Klein, S. (2019). Progressing emergent constraints on future climate change. *Nature Climate Change*, *9*(4), 269–278. https://doi.org/10.1038/s41558-019-0436-6

Hassell, D., Gregory, J., Blower, J., Lawrence, B. N., & Taylor, K. E. (2017). A data model of the Climate and Forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2. 1). *Geoscientific Model Development*, *10*(12), 4619–4646. https://doi.org/10.5194/gmd-10-4619-2017

Held, I. M., & Soden, B. J. (2006). Robust responses of the hydrological cycle to global warming. *Journal of Climate*, *19*(21), 5686–5699. https://doi.org/10.1175/jcli3990.1

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hess, P. G., Battisti, D. S., & Rasch, P. J. (1993). Maintenance of the intertropical convergence zones and the large-scale tropical circulation on a water-covered earth. *Journal of the Atmospheric Sciences*, *50*(5), 691–713. https://doi.org/10.1175/1520-0469(1993)050<0691:moticz>2.0.co;2

Hsu, T. Y., Primeau, F., & Magnusdottir, G. (2022). A hierarchy of global ocean models coupled to CESM1. *Journal of Advances in Modeling Earth Systems*, *14*(8), e2021MS002979. https://doi.org/10.1029/2021ms002979

Jablonowski, C., Lauritzen, P. H., Nair, R. D., & Taylor, M. (2008). Idealized test cases for the dynamical cores of atmospheric General Circulation Models: A proposal for the NCAR ASP 2008 summer colloquium. Technical Report.

Jablonowski, C., & Williamson, D. L. (2006). A baroclinic instability test case for atmospheric model dynamical cores. *Quarterly Journal of the Royal Meteorological Society*, *132*(621C), 2943–2975. https://doi.org/10.1256/qj.06.12

Karlbauer, M., Cresswell-Clay, N., Durran, D. R., Moreno, R. A., Kurth, T., Bonev, B., et al. (2024). Advancing parsimonious deep learning weather prediction using the HEALPix mesh. *Journal of Advances in Modeling Earth Systems*, *16*(8), e2023MS004021. https://doi.org/10.1029/2023ms004021

Klein, S. A., & Hall, A. (2015). Emergent constraints for cloud feedbacks. *Current Climate Change Reports*, *1*(4), 276–287. https://doi.org/10.1007/s40641-015-0027-1

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., et al. (2024). Neural general circulation models for weather and climate. *Nature*, *632*(8027), 1–7. https://doi.org/10.1038/s41586-024-07744-y

Koldunov, N., Rackow, T., Lessig, C., Danilov, S., Cheedela, S. K., Sidorenko, D., et al. (2024). Emerging AI-based weather prediction models as downscaling tools. arXiv preprint arXiv:2406.17977.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Benchmarking a catchment-aware long short-term memory network (LSTM) for large-scale hydrological modeling. *Hydrology and Earth System Sciences Discussions*, *2019*, 1–32.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., et al. (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416–1421. https://doi.org/10.1126/science.adi2336

Lee, J., Gleckler, P. J., Ahn, M. S., Ordonez, A., Ullrich, P. A., Sperber, K. R., et al. (2024). Systematic and objective evaluation of Earth system models: PCMDI Metrics Package (PMP) version 3. *Geoscientific Model Development*, *17*(9), 3919–3948. https://doi.org/10.5194/gmd-17-3919-2024

Liu, Y., Liu, Y., Wang, W., & Zhou, H. (2021). Propagation of soil moisture droughts in a hotspot region: Spatial pattern and temporal trajectory. *Journal of Hydrology*, *593*, 125906. https://doi.org/10.1016/j.jhydrol.2020.125906

Loring, B., Elbashandy, A., O'Brien, T. A., Dufek, A., Krishnan, H., & Keen, N. (2023). *LBL-EESA/TECA: TECA 6.0.0 (6.0.0)*. Zenodo. https://doi.org/10.5281/zenodo.10129872

Lütjens, B., Ferrari, R., Watson-Parris, D., & Selin, N. (2024). The impact of internal variability on benchmarking deep learning climate emulators. arXiv preprint arXiv:2408.05288.

Mahesh, A., Collins, W., Bonev, B., Brenowitz, N., Cohen, Y., Elms, J., et al. (2024a). Huge ensembles part I: Design of ensemble weather forecasts using spherical Fourier neural operators. arXiv preprint arXiv:2408.03100.

Mahesh, A., Collins, W., Bonev, B., Brenowitz, N., Cohen, Y., Harrington, P., et al. (2024b). Huge ensembles part II: Properties of a huge ensemble of hindcasts generated with spherical fourier neural operators. arXiv preprint arXiv:2408.01581.

Mason, E., Pascual, A., & McWilliams, J. C. (2014). A new sea surface height–based code for oceanic mesoscale eddy tracking. *Journal of Atmospheric and Oceanic Technology*, *31*(5), 1181–1188. https://doi.org/10.1175/jtech-d-14-00019.1

Mauzey, C., Doutriaux, C., Nadeau, D., Taylor, K. E., Durack, P. J., Betts, E., et al. (2024). *PCMDI/CMOR: CMOR v3.9.0 (v3.9.0)*. Zenodo. https://doi.org/10.5281/zenodo.13388462

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., & Stouffer, R. J. (1997). Intercomparison makes for a better climate model. *Eos, Transactions American Geophysical Union*, *78*(41), 445–451. https://doi.org/10.1029/97eo00276

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., & Stouffer, R. J. (2000). The coupled model intercomparison project (CMIP). *Bulletin of the American Meteorological Society*, *81*(2), 313–318. https://doi.org/10.1175/1520-0477(2000)081<0313:tcmipc>2.3.co;2

Meehl, G. A., Covey, C., McAvaney, B., Latif, M., & Stouffer, R. J. (2005). Overview of the coupled model intercomparison project. *Bulletin of the American Meteorological Society*, *86*(1), 89–93.

Moreno-Chamarro, E., Caron, L. P., Loosveldt Tomas, S., Gutjahr, O., Moine, M. P., Putrasahan, D., et al. (2021). Impact of increased resolution on long-standing biases in HighResMIP-PRIMAVERA climate models. *Geoscientific Model Development Discussions*, *2021*, 1–33.

29935210, 2025, 1, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000496 by Open Research - University Of East Anglia , Wiley Online Library on [14/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Neelin, J. D., Krasting, J. P., Radhakrishnan, A., Liptak, J., Jackson, T., Ming, Y., et al. (2023). Process-oriented diagnostics: Principles, practice, community development, and common standards. *Bulletin of the American Meteorological Society*, *104*(8), E1452–E1468. https://doi.org/10.1175/bams-d-21-0268.1

Nicholls, Z., Meinshausen, M., Lewis, J., Corradi, M. R., Dorheim, K., Gasser, T., et al. (2021). Reduced complexity model intercomparison project phase 2: Synthesizing earth system knowledge for probabilistic climate projections. *Earth's Future*, *9*(6), e2020EF001900. https://doi.org/10.1029/2020ef001900

NVIDIA. (2024). Earth-2 MIP (Beta). Retrieved from https://github.com/NVIDIA/earth2mip

O'Loughlin, R., Li, D., & O'Brien, T. (2024). *Moving beyond post-hoc XAI: Lessons learned from dynamical climate modeling* (pp. 1–24). EGUsphere.

Pendergrass, A. G., & Hartmann, D. L. (2014). The atmospheric energy constraint on global-mean precipitation change. *Journal of Climate*, *27*(2), 757–768. https://doi.org/10.1175/jcli-d-13-00163.1

Phillips, A. S., Deser, C., Fasullo, J., Schneider, D. P., & Simpson, I. R. (2020). Assessing climate variability and change in model large ensembles: A user's guide to the "climate variability diagnostics package for large ensembles". 1001. Retrieved from https://opensky.ucar.edu/islandora/object/manuscripts

Potter, G. L., Bader, D. C., Riches, M., Bamzai, A., & Joseph, R. (2011). Celebrating two decades of the program for climate model diagnosis and intercomparison. *Bulletin of the American Meteorological Society*, *92*(5), 629–631. https://doi.org/10.1175/2011bams3018.1

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., et al. (2025). Probabilistic weather forecasting with machine learning. *Nature*, *637*(8044), 84–90. https://doi.org/10.1038/s41586-024-08252-9

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., et al. (2024). WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, *16*(6), e2023MS004019. https://doi.org/10.1029/2023ms004019

Reed, K. A., & Jablonowski, C. (2012). Idealized tropical cyclone simulations of intermediate complexity: A test case for AGCMs. *Journal of Advances in Modeling Earth Systems*, *4*(2). https://doi.org/10.1029/2011ms000099

Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., et al. (2020). Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations. *Geoscientific Model Development*, *13*(12), 6165–6200. https://doi.org/10.5194/gmd-13-6165-2020

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., et al. (2020). An assessment of Earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics*, *58*(4), e2019RG000678. https://doi.org/10.1029/2019rg000678

Shiogama, H., Watanabe, M., Kim, H., & Hirota, N. (2022). Emergent constraints on future precipitation changes. *Nature*, *602*(7898), 612–616. https://doi.org/10.1038/s41586-021-04310-8

Sweeney, A., & Fu, Q. (2024). Interannual variability of zonal mean temperature, water vapor, and clouds in the tropical tropopause layer. *Journal of Geophysical Research: Atmospheres*, *129*(3), e2023JD039002. https://doi.org/10.1029/2023jd039002

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, *106*(D7), 7183–7192. https://doi.org/10.1029/2000jd900719

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/bams-d-11-00094.1

Thuburn, J. (2008). Some conservation issues for the dynamical cores of NWP and climate models. *Journal of Computational Physics*, *227*(7), 3715–3730. https://doi.org/10.1016/j.jcp.2006.08.016

Ullrich, P. A., Jablonowski, C., Kent, J., Lauritzen, P. H., Nair, R. D., & Taylor, M. A. (2012). Dynamical core model intercomparison project (DCMIP) test case document. *DCMIP Summer School*, *83*.

Ullrich, P. A., Jablonowski, C., Reed, K. A., Zarzycki, C., Lauritzen, P. H., Nair, R. D., et al. (2016). Dynamical Core Model Intercomparison Project (DCMIP2016) test case document. https://doi.org/10.5281/zenodo.1298671

Ullrich, P. A., Zarzycki, C. M., McClenny, E. E., Pinheiro, M. C., Stansfield, A. M., & Reed, K. A. (2021). TempestExtremes v2.1: A community framework for feature detection, tracking, and analysis in large datasets. *Geoscientific Model Development*, *14*(8), 5023–5048. https://doi.org/10.5194/gmd-14-5023-2021

Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls, G., et al. (2022). ClimateBench v1. 0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2021MS002954. https://doi.org/10.1029/2021ms002954

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

Williamson, D. L., Drake, J. B., Hack, J. J., Jakob, R., & Swarztrauber, P. N. (1992). A standard test set for numerical approximations to the shallow water equations in spherical geometry. *Journal of Computational Physics*, *102*(1), 211–224. https://doi.org/10.1016/s0021-9991(05)80016-6

Williamson, M. S., Thackeray, C. W., Cox, P. M., Hall, A., Huntingford, C., & Nijsse, F. J. (2021). Emergent constraints on climate sensitivities. *Reviews of Modern Physics*, *93*(2), 025004. https://doi.org/10.1103/revmodphys.93.025004

Yu, S., Hannah, W. M., Peng, L., Bhouri, M. A., Gupta, R., Lin, J., et al. (2023). ClimSim: An open large-scale dataset for training high-resolution physics emulators in hybrid multi-scale climate simulators. arXiv preprint arXiv:2306.08754.

Yu, S., White, B. L., Bhiwandiwalla, A., Hinck, M., Olson, M. L., Nguyen, T., & Lal, V. (2024). ClimDetect: A benchmark dataset for climate change detection and attribution. arXiv preprint arXiv:2408.15993.

Zarzycki, C. M., Ullrich, P. A., & Reed, K. A. (2021). Metrics for evaluating tropical cyclones in climate data. *Journal of Applied Meteorology and Climatology*, *60*(5), 643–660. https://doi.org/10.1175/jamc-d-20-0149.1

Zhang, Y., & Boos, W. R. (2023). An upper bound for extreme temperatures over midlatitude land. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(12), e2215278120. https://doi.org/10.1073/pnas.2215278120

Zhou, W., Leung, L. R., & Lu, J. (2022). Linking large-scale double-ITCZ bias to local-scale drizzling bias in climate models. *Journal of Climate*, *35*(24), 4365–4379. https://doi.org/10.1175/jcli-d-22-0336.1

29935210, 2025, 1, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024JH000496 by Open Research - University Of East Anglia , Wiley Online Library on [14/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License