Eos

Climate Model Benchmarking: Building Trust and Advancing Science

A critical look at how well climate models work, how to successfully use them, and the anticipated challenges for their next-generation development.

By **Birgit Hassler**, **Forrest Hoffman**, **Ranjini Swaminathan,** and **Beth Dingley**
24 August 2026

Hurricane Edouard in September 2014. Credit: NASA

Editors' Vox is a blog from AGU's Publications Department.

As climate models become more advanced and complex, it is increasingly important to evaluate model performance with respect to observations. A **new article** in ***Reviews of Geophysics*** explores climate model evaluation and benchmarking efforts in the scientific community. Here, we asked the authors to give an overview of climate modeling, how scientists evaluate models, and what challenges remain.

When did scientists first start developing climate models?

While today's climate models are comprised of millions of lines of code and need to be run on large supercomputers, climate modeling started over a century ago. Lewis Fry Richardson published his 1922 book 'Weather Predication by Numerical Process', outlining his idea to forecast the weather using differential equations. It took him six weeks to derive an eight-hour forecast by hand. It wasn't until 1950 when the first computerized weather forecast was run at Princeton University, led by Jule G. Charney. Six years later, Norman Phillips published a paper titled '**The general circulation of the atmosphere: a numerical experiment**', in which he details the first General Circulation Model (GCM) of the atmosphere.

Climate modeling, as a distinct field of study from weather forecasting, developed significantly throughout the 1960s, with several seminal studies being published which still underpin climate research today, culminating in the first results from a coupled atmosphere-ocean GCM, in which the atmosphere, ice, and ocean interact with each other for the first time (Manabe et al., 1975 **Part 1** and **Part 2**).

What's the difference between physical climate models and Earth System Models (ESMs)?

> *GCMs are predecessors of ESMs, which have been expanded to include more processes describing the climate system.*

The difference between physical ("physics-based") climate models (GCMs) and Earth System Models (ESMs) can be very clearly determined: GCMs are predecessors of ESMs, which have been expanded to include more processes describing the climate system. GCMs focus strictly on the physical processes and exchanges of energy and matter between the atmosphere, oceans, land surface, and sea ice. They therefore provide the means to understand the physics of the system.

ESMs are based on more comprehensive model configurations. In addition to the physical representation of the climate system, they also include processes such as interactive atmospheric chemistry, biogeochemical cycles, and marine/terrestrial ecosystems, and carbon exchange between Earth system components. These additional processes help with the realistic representation of the climate system and increase the confidence in future climate projections, but they come with increased costs in resources and time. Which configuration of climate models are used to prepare a simulation is therefore still very much dependent on the planned use of the simulation.

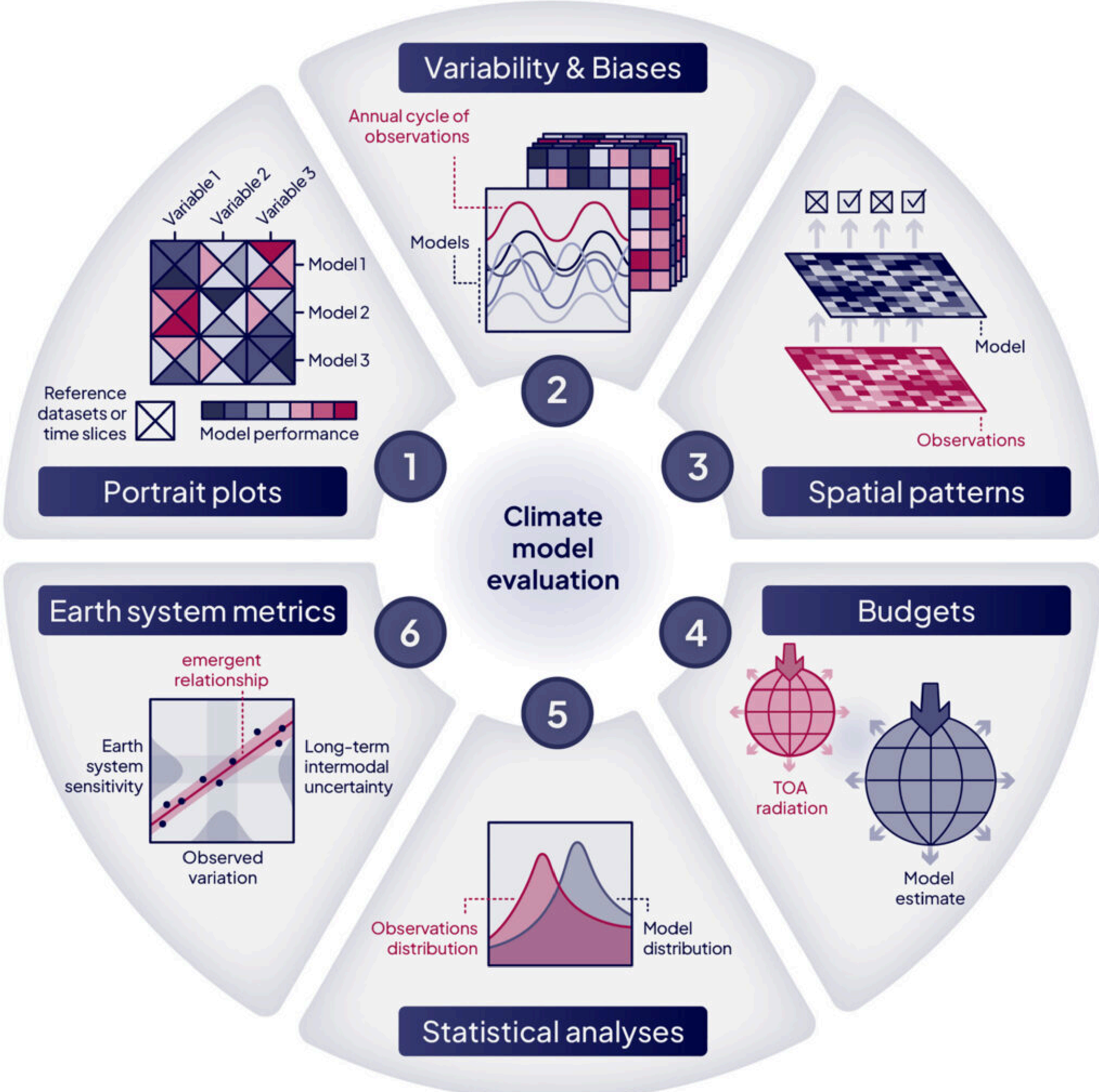
Why is it important to evaluate and benchmark climate models?

Climate models are among the most important and useful scientific mechanisms to study the Earth system, and the past and future climate. Every generation of climate models brings new advances in representing processes in different components of the Earth system such as the atmosphere, oceans and land and their interactions. State-of-the-art models also include the representation of the carbon cycle across these different domains, allowing us to understand the role of carbon cycling in determining the Earth's climate. Climate model evaluation with a wide variety of diagnostics is how we develop confidence in the fidelity with which models represent processes that affect future climate projections (IPCC AR 4). A diverse set of tests and diagnostics can help understand differences between models, differences caused by changes over model generations, and identify priorities for future scientific development.

A diverse set of tests and diagnostics can help understand differences between models, differences caused by changes over model generations, and identify priorities for future scientific development.

What are some of the different ways that scientists evaluate climate model simulations?

Climate model evaluation is described as the process of comparing model simulations against observational or reanalyzes datasets. Benchmarking is the process where model simulations are evaluated with observations, reanalysis data, or with other models often resulting in a statement made about the "goodness" of the simulation or model based on a predetermined set of standards or criteria. One or more models can be compared against the same set of observations in evaluation. Some of the more basic diagnostics used in model evaluation are the comparison of climatological (average over a period of 30 years or so) means, inter-annual variability and biases often expressed as time series. Maps or spatial patterns comparing models and observations, statistical analyses such as distributions of a specific variable and portrait plots for multi-model evaluation are other approaches commonly used to evaluate models. More complex diagnostics such as water, energy or carbon budgets and cross-domain interactions (e.g. how does temperature vary with precipitation) are also used for better process-based evaluation.



The evaluation of climate model simulations can be grouped in six general approaches. Most approaches require observations and can be applied on a global or regional scale. Credit: [Hassler et al. [2026]](), Figure 2

What are some specific model biases that have been improved due to routine assessments?

Systematic model evaluation assures scientists are frequently reminded of biases that have not been eliminated or that often recur, and it helps modeling centers identify process representations to prioritize in future model development.

Biases in cloud and water vapor process representations have been reduced over the past 20 years due to the growth of detailed observations and improvements in cloud vertical distribution and parameter tuning.

The East Asian summer monsoon (EASM), which carries moisture from the Indian and Pacific Oceans to East Asia, exhibits intense interannual variability that results in severe droughts and floods. Several multi-model benchmarking studies have focused on model reproductions of the EASM, and they identified a weakened western North Pacific anticyclone as a primary cause of biases in the western Pacific subtropical high and the Meiyu-Baiu-Changma rainband.

Early efforts in evaluating land carbon cycle models identified persistent biases in the timing of the seasonal variation of vegetation growth, especially with respect to satellite-derived leaf area index (LAI), which were tied to an underestimate of carbohydrate pools carried from one growing season to the next. Land models substantially overestimated aboveground live biomass in the Amazon Basin compared to estimates from satellite observations, which researchers attributed primarily to low autotrophic respiration and excessive allocation of net primary production to wood. Improvements in model representation of vegetation processes significantly reduced these long-standing biases that have been routinely assessed through comparison with observational data for at least three decades.

What are some of the remaining biases in climate models where additional research and development efforts are needed?

“

Identifying persistent biases and tracking their reductions are crucial to strengthening the utility of projections from climate and Earth system models.

A variety of biases exhibited by many climate models have been difficult to reduce or completely eliminate despite frequent model evaluation and bias characterization. Many of these have led scientists to perform extensive model tuning exercises or apply bias removal techniques to reduce the impacts of such biases. Examples include both positive and negative regional biases in precipitation, the double ITCZ (intertropical convergence zone), warming biases in the tropical troposphere, and low Arctic Sea ice sensitivity to global warming compared to observations. Identifying such persistent biases and tracking their reductions are crucial to strengthening the utility of projections from climate and Earth system models. Creating a wealth of additional model performance metrics and enhancing collections of observational data sets are required to understand the sources of these long-standing model biases.

—Birgit Hassler (**birgit.hassler@dlr.de**, **0000-0003-2724-709X**), Deutsches Zentrum für Luft- und Raumfahrt (DLR), Germany; Forrest Hoffman (**0000-0001-5802-4134**), Oak Ridge National Laboratory, United States; Ranjini Swaminathan (**0000-0001-5853-2673**), University of Reading, United Kingdom; and Beth Dingley (**0000-0002-9831-9671**), European Space Agency, United Kingdom

13 CLIMATE ACTION

The research reported here supports **Sustainable Development Goal 13**. AGU is committed to supporting the United Nations 2030 Agenda for Sustainable Development, which provides a shared blueprint for peace and prosperity for people and the planet, now and into the future.

Editor's Note: It is the policy of AGU Publications to invite the authors of articles published in Reviews of Geophysics to write a summary for Eos Editors' Vox.

Citation: Hassler, B., F. Hoffman, R. Swaminathan, and B. Dingley (2026), Climate model benchmarking: building trust and advancing science, *Eos, 107,* https://doi.org/10.1029/2026EO265032. Published on 24 August 2026.

What do you think?

0 Responses

Upvote Funny Love Surprised Angry Sad

0 Comments

1 Login ▾

G

Start the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Share

Best Newest Oldest

Be the first to comment.

Subscribe **Privacy** **Do Not Sell My Data**