GC34A-04 ClimateBench2.0: Probabilistic Climate Model Scoring

Wednesday, 17 December 2025

16:45 - 16:55

New Orleans Theater C (NOLA CC)

Author will be Presenting:

In-person

Abstract

Despite their central role in climate science and policy, Earth system models (ESMs) remain difficult to compare in any rigorous or transparent way. Most existing evaluations either emphasize specific processes or rely on qualitative assessments across diverse metrics, making it nearly impossible to rank models by their predictive skill. ClimateBench2.0 introduces a probabilistic scoring framework that focuses instead on what matters most: a model's ability to accurately simulate the historical climate and project future multi-decadal change.

The benchmark leverages high-quality observations from the satellite era (1980–present), with a particular focus on present-day metrics such as top-of-atmosphere (TOA) energy balance, seasonal cycle fidelity, and variability in clouds, aerosols, precipitation, and ocean heat uptake for which observational constraints are strongest. Paleoclimate reconstructions (LGM, LIG, Mid-Holocene) are incorporated as out-of-distribution tests to evaluate models beyond the narrow window of recent data. Scoring is based on robust probabilistic metrics such as CRPS and Brier scores, designed to assess ensemble skill and uncertainty quantification.

Crucially, statistical performance alone is not sufficient. ClimateBench2.0 will also introduce a dedicated Physical Consistency category, evaluating properties such as global energy balance closure, conservation of water and carbon, and realistic land-ocean-atmosphere energy exchanges. These physical integrity checks are essential for trusting a model's out-of-distribution predictions - especially under strong forcings not seen in the historical record.

By combining empirical benchmarks with physically grounded constraints, ClimateBench2.0 transforms evaluation into a reproducible, quantitative, and outcome-driven ranking framework. It applies across model types, from physical to hybrid to ML-based, and integrates with existing efforts (e.g., CMIP, Obs4MIPs) to ensure transparency and broad adoption.

First Author

W

Duncan Watson-Parris

University of California San Diego

Authors

B

Venkatramani Balaji

B

Kevin W Bowman

Schmidt Futures

NASA Jet Propulsion Laboratory

Christopher Stephen Bretherton
Allen Institute for AI

Peter Martin Caldwell
Lawrence Livermore National Laboratory

Will Chapman
NSF National Center for Atmospheric Research

William Drew Collins
Berkeley Lab and UC Berkeley

Gregory Elsaesser
Columbia University/NASA GISS

Veronika Eyring
German Aerospace Center DLR Oberpfaffenhofen

Pierre Gentine
Columbia University

Stephan Hoyer
Google

Ralph F Keeling
Univ California San Diego

Nikolay Koldunov
Alfred Wegener Institute Helmholtz-Center for Polar and Marine Research Bremerhaven

David M Lawrence
NSF National Center for Atmospheric Research

Christian Lessig
Otto-von-Guericke Universität

J David Neelin
University of California, Los Angeles

Mike S Pritchard
University California Irvine

Sarah G Purkey
Scripps Institution of Oceanography, University of California San Diego

Gavin A Schmidt
NASA/GISS

Tapio Schneider
California Institute of Technology

Michael Schulz
Norwegian Meteorological Institute

Isla Simpson
National Center for Atmospheric Research

Tiffany Shaw
University of Chicago

Graeme L Stephens
Jet Propulsion Laboratory

Joao Teixeira
Jet Propulsion Laboratory, California Institute of Technology

Willa Tobin
UC San Diego

Andrew Williams
University of Oxford

Laure Zanna
University of Oxford

Rose Yu
University of California San Diego

View Related