

ESMBenchmarkViz: A Python Toolkit for Interactive Visualization of Earth System Model Evaluation and Benchmarking

Jiwoo Lee ¹, Kristin Y. Chang ¹, Peter Gleckler ¹, and Paul Ullrich ^{1,2}

¹ Lawrence Livermore National Laboratory, United States ² University of California, Davis, United States

DOI: [10.21105/joss.10644](https://doi.org/10.21105/joss.10644)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Mengqi Zhao](#)  

Reviewers:

- [@shenyulu](#)
- [@carbform](#)

Submitted: 01 December 2025

Published: 09 September 2026

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

ESMBenchmarkViz is a Python toolkit designed to generate interactive graphs for visualizing statistics and metrics from the evaluation of climate and Earth system models. This toolkit enables researchers to perform more straightforward benchmarking and intercomparison of models.

State of the Field

Earth System Models (ESMs) are essential for understanding and predicting the complex interactions within the Earth's climate system. These models integrate components such as the atmosphere, oceans, land surface, and biosphere to address fundamental and applied questions about Earth system processes, feedbacks, and the system's response to external forcing. As ESMs grow in complexity, the need for effective evaluation and benchmarking methods to ensure their reliability and accuracy becomes increasingly important ([Hassler et al., 2026](#)).

Evaluating ESMs involves comparing model outputs against observational data and other models to assess performance. This process is vital for identifying model strengths and weaknesses, guiding improvements, and deepening our understanding of climate processes.

Statement of need

ESM evaluations (particularly comprehensive evaluations) often generate large volumes of data, as shown in Lee et al. (2019), Lee et al. (2021), Planton et al. (2021), Ahn et al. (2022), and Lee et al. (2024), which can be challenging to interpret and communicate effectively. Wrangling this data to produce an effective, polished visualization is a cumbersome process that often requires constructing data and label lists, calculating plot coordinates, and making manual adjustments for final elements like color bar position or label orientation.

To address these challenges, we have developed ESMBenchmarkViz, a modern Python library specifically designed for efficient interactive visualization of statistical performance metrics from ESM evaluation and intercomparison. This library leverages the power of Bokeh ([Bokeh Development Team, 2018](#)) to provide researchers and practitioners with user-friendly interactive tools for dynamic exploration of complex datasets. By enabling real-time manipulation of visualizations, ESMBenchmarkViz facilitates deeper insights into model performance and inter-model comparisons, making it easier to identify outliers and persistent biases.

In this document, we describe the core functionalities of ESMBenchmarkViz and demonstrate example applications. Our aim is to empower climate scientists with enhanced visualization

capabilities, contributing to more robust ESM evaluations and benchmarking, and ultimately supporting informed decision-making for climate policy.

Software Design

ESMBenchmarkViz was developed from the interactive visualization dashboard of the [Program for Climate Model Diagnosis and Intercomparison \(PCMDI\) Metrics Package](#) (Lee et al., 2024), which displays diverse evaluation metrics for ESMs alongside diagnostic information on the [PCMDI website](#). We refer to these diagnostics as “dive-down information,” as they enable users to investigate metrics in greater detail.

ESMBenchmarkViz is structured as a modular Python library that provides high-level plotting functions tailored to common Earth system model (ESM) evaluation workflows. The core design goal is to minimize boilerplate for end users while retaining flexibility for advanced customization. To achieve this, the package separates three main concerns: (1) data preparation and validation, (2) plot configuration (layout, styling, and interactivity), and (3) rendering via the Bokeh backend. Each supported plot type (Taylor diagram, portrait plot, scatter plot) is implemented as a self-contained module exposing a concise, stable API, while sharing common utilities for color mapping, tooltips, legends, and export.

A key design trade-off was choosing Bokeh as the primary visualization engine rather than building on top of more general plotting wrappers. Bokeh offers native support for interactive HTML output, hover tools, and linked selections, which aligns with the “dive-down” paradigm that motivated this toolkit. Although other visualization libraries (e.g., Matplotlib, Plotly, Holoviews) provide overlapping capabilities, they either lack specialized implementations of ESM-oriented diagnostics (such as Taylor diagrams and portrait plots with standardized climate-centric defaults) or require substantial custom code to reproduce similar interactive features.

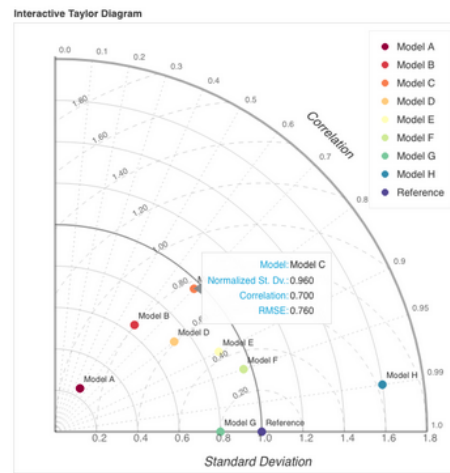
In developing ESMBenchmarkViz, we considered contributing directly to existing packages, particularly the [PCMDI Metrics Package](#) and general-purpose plotting libraries. We elected to build a standalone but interoperable toolkit for two reasons. First, we aimed for an architecture that is agnostic to any particular metrics engine, so that it can be reused across different ESM evaluation pipelines beyond PCMDI. Second, keeping visualization logic in a dedicated library allows us to iterate on interaction patterns, layout presets, and performance optimizations without constraining upstream packages or imposing additional dependencies on users whose workflows do not require interactive plots. At the same time, ESMBenchmarkViz is designed to integrate smoothly with existing climate analysis ecosystems through simple data structures (e.g., pure Python lists, Pandas DataFrames, numpy or xarray outputs) and clear [APIs](#).

Key Features

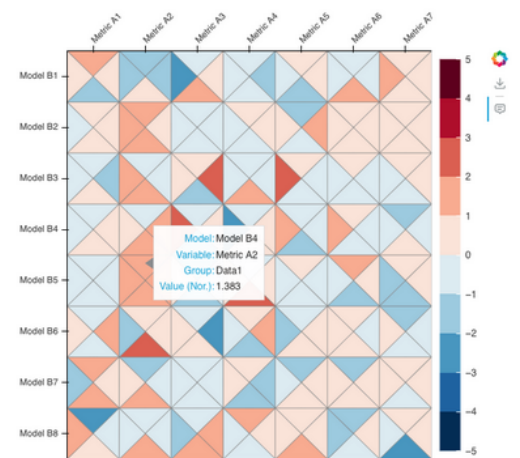
ESMBenchmarkViz provides reusable functions to generate a suite of interactive plots customized for the evaluation, intercomparison, and benchmarking of ESMs. The toolkit is developed in Python 3 and built on top of the Bokeh library for interactive visualization. API reference documentation and interactive demo Jupyter notebooks are available for each type of plot.

With singular functions for each plot, the library integrates seamlessly with existing data analysis workflows and promotes reproducibility in climate research. Users can interact with data by zooming, filtering, and hovering for detailed tooltips or displaying additional details as a sidenote, enhancing the communication of findings.

(a) Taylor Diagram



(b) Portrait Plot



(c) Scatter plot (with enabling the side dive-down-image viewer)

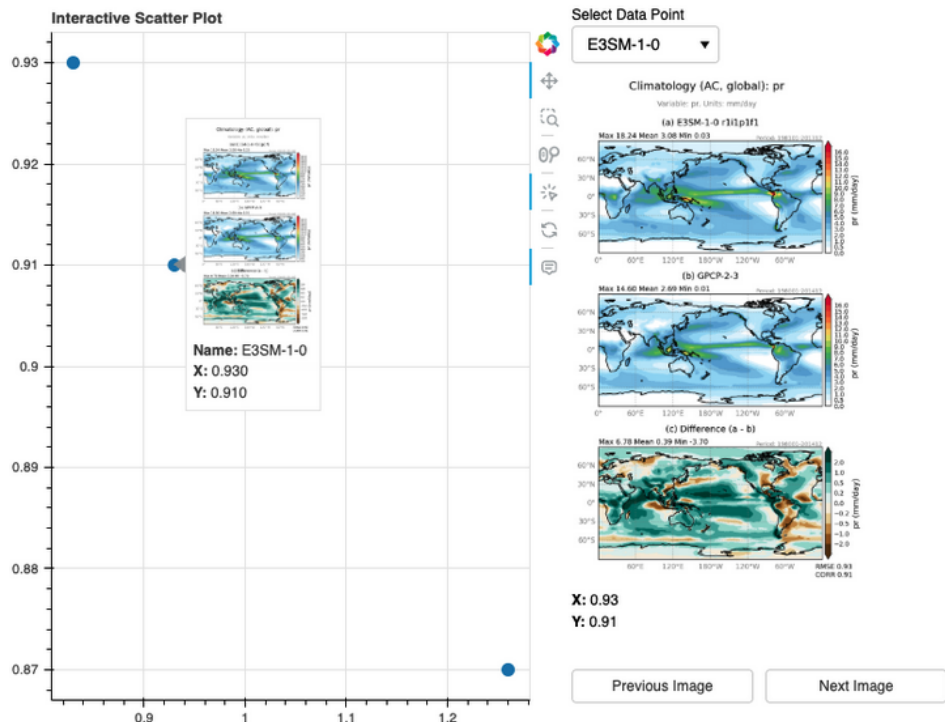


Figure 1: Demonstration of the core features: (a) Taylor Diagram (Taylor, 2001), (b) Portrait Plot (Gleckler et al., 2008), and (c) scatter plot with the side dive-down image viewer option activated. A user's mouse cursor hovering over a specific data point (i.e., a specific ESM and its metrics) interactively shows a tooltip that includes detailed information, with the capability of clicking it to open the associated "dive-down" image. Images can also be included in the tooltips or in the side viewer.

The toolkit offers convenient APIs to generate and save the following types of graphs in both static and interactive modes: two specialized plots used in ESM evaluation—Taylor Diagram (Taylor, 2001) (Fig. 1a) and Portrait Plot (Gleckler et al., 2008) (Fig. 1b)—as well as the widely used scatter plot (Fig. 1c). These graph types were selected for their utility in ESM evaluation and benchmarking, and because few tools currently provide such capabilities.

Taylor Diagram

The Taylor Diagram ([Taylor, 2001](#)), shown as Fig. 1a, provides a concise graphical summary of how well patterns simulated by a model match observations. It simultaneously displays three statistics—spatial pattern correlation, standard deviation, and root-mean-square error—making it especially useful for comparing multiple models or datasets against a reference.

Portrait Plot

The Portrait Plot ([Gleckler et al., 2008](#)), shown as Fig. 1b, presents a matrix-like visualization that summarizes model performance across multiple variables, metrics, or regions. It enables quick identification of patterns, strengths, and weaknesses by displaying performance scores as colored cells, facilitating comprehensive intercomparison among models. This type of plot has been actively used for various climate model evaluation studies, as shown in [Lee et al. \(2019\)](#), [Lee et al. \(2021\)](#), and [Ahn et al. \(2022\)](#). The [PCMDI Metrics Package](#) ([Lee et al., 2024](#)) team had developed a precursor to this package to present evaluation results from hundreds of simulations in an efficient way as shown on the [PCMDI website](#).

Scatter Plot

The Scatter Plot (Fig. 1c) displays the relationship between two variables, allowing users to visually assess correlations, trends, and outliers in model evaluation data. It is a flexible and widely used tool for exploring and communicating the distribution and association of key metrics. Although it is a very widely applied type of plot across disciplines, we have included it in the package for its synergy with tooltips and images together, as shown in Fig. 1c.

There are more types of plots planned for the future advancement of the package, for example, Parallel Coordinate Plots in the way it has been used for ESM evaluations ([Lee et al., 2024](#)).

Research Impact statement

ESMBenchmarkViz directly addresses a recurrent bottleneck in climate and Earth system model evaluation: turning large, heterogeneous collections of performance metrics into clear and interactive visual summaries. By encapsulating complex plots (Taylor diagrams, portrait plots, and enhanced scatter plots) into reusable functions, the toolkit lowers the barrier to producing publication-quality figures and interactive dashboards that help researchers explore and communicate model behavior.

The library is already integrated into the workflow of the [PCMDI Metrics Package](#), where it has been used to visualize evaluation results for large ensembles of climate simulations. This integration demonstrates its practical utility in real-world ESM assessment and supports ongoing activities within the PCMDI. The ability to link summary metrics with “dive-down” diagnostic images enables more efficient identification of systematic biases and facilitates targeted model development discussions between modeling centers and diagnostic teams.

From a community perspective, ESMBenchmarkViz is positioned as a general-purpose, domain-aware visualization layer. The public documentation site, API reference, and example Jupyter notebooks are designed to support adoption by researchers who may not be visualization experts. The package is open source, versioned on GitHub, and distributed for Linux and macOS, with plans for conda-forge availability to further streamline installation. Automated tests and continuous integration (as appropriate for the project’s maturity) help preserve stability as new features and plot types are added. Together, these elements indicate that ESMBenchmarkViz provides not only novel capabilities but also the infrastructure required for reproducible research and sustainable community use in climate model evaluation and benchmarking.

Documentation

The ESMBenchmarkViz [documentation](#) includes the [public API list](#) and a Jupyter Notebook [Gallery](#) that demonstrates usage of the package.

Distribution

ESMBenchmarkViz is available for Linux and MacOS via the conda-forge channel on Anaconda or pip, following the [installation instructions](#). We host all development activity at the [GitHub Repository](#).

Acknowledgements

This work was performed under the auspices of the U.S. [Department of Energy](#) by Lawrence Livermore National Laboratory ([LLNL](#)) under Contract DE-AC52-07NA27344. The efforts of the authors were supported by the Regional and Global Model Analysis ([RGMA](#)) program of the United States [Department of Energy's Office of Science](#). The authors thank Charles Doutriaux (LLNL) and William Hill (formerly LLNL) for their contributions to the earlier stage of the precursor of this work, which enabled initial proof of concept that later evolved to this project.

AI Usage Disclosure

The authors used generative AI tools, including LLNL LivChat (based on the gpt-5.1 model) and ChatGPT, in the development of both this manuscript and the ESMBenchmarkViz software.

For the manuscript, AI tools were used to improve the initial draft. All AI-generated text was subsequently reviewed, edited, and integrated by the human authors, who are responsible for the final content and its accuracy.

For the software, AI tools were used to assist with drafting portions of the initial source code, such as boilerplate structures, refactoring suggestions, and example usage patterns. Any AI-generated code was inspected, tested, and modified as needed by the authors before inclusion in the repository. Architectural decisions, algorithmic choices, and domain-specific logic for ESMBenchmarkViz were made by the authors, and the overall design and maintenance of the project remain under human supervision.

No AI tools were given access to proprietary, non-public data in the course of this work.

References

- 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>
- Bokeh Development Team. (2018). *Bokeh: Python library for interactive visualization*. <https://bokeh.pydata.org/en/latest/>
- Gleckler, P. J., Taylor, K. E., & Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres*, 113(D6). <https://doi.org/10.1029/2007JD008972>
- Hassler, B., Hoffman, F. M., Beadling, R., Blockley, E., Huang, B., Lee, J., Lembo, V., Lewis, J., Lu, J., Madaus, L., Malinina, E., Medeiros, B., Pokam, W., Scoccimarro, E., & Swaminathan, R. (2026). Systematic benchmarking of climate models: Methodologies,

- applications, and new directions. *Reviews of Geophysics*, 64(1), e2025RG000891. <https://doi.org/10.1029/2025RG000891>
- Lee, J., Gleckler, P. J., Ahn, M.-S., Ordonez, A., Ullrich, P. A., Sperber, K. R., Taylor, K. E., Planton, Y. Y., Guilyardi, E., Durack, P., Bonfils, C., Zelinka, M. D., Chao, L.-W., Dong, B., Doutriaux, C., Zhang, C., Vo, T., Boutte, J., Wehner, M. F., ... Krasting, J. (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>
- Lee, J., Sperber, K. R., Gleckler, P. J., Taylor, K. E., & Bonfils, C. (2021). Benchmarking performance changes in the simulation of extratropical modes of variability across CMIP generations. *Journal of Climate*, 34(17), 6945–6969. <https://doi.org/10.1175/JCLI-D-20-0832.1>
- Lee, J., Sperber, K., Gleckler, P., Bonfils, C., & Taylor, K. (2019). Quantifying the agreement between observed and simulated extratropical modes of interannual variability. *Climate Dynamics*, 52, 4057–4089. <https://doi.org/10.1007/s00382-018-4355-4>
- Planton, Y. Y., Guilyardi, E., Wittenberg, A. T., Lee, J., Gleckler, P. J., Bayr, T., McGregor, S., McPhaden, M. J., Power, S., Roehrig, R., Vialard, J., & Voldoire, A. (2021). Evaluating climate models with the CLIVAR 2020 ENSO metrics package. *Bulletin of the American Meteorological Society*, 102(2), E193–E217. <https://doi.org/10.1175/BAMS-D-19-0337.1>
- Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, 106(D7), 7183–7192. <https://doi.org/10.1029/2000JD900719>