The Journal of Open Source Software

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]

1 Lawrence Livermore National Lab, Livermore, USA

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 01 December 2025
Published: unpublished

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.

Statement of Need

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.

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. However, ESM evaluations (particularly comprehensive evaluations) often generate large volumes of data (e.g., (J. Lee et al., 2019), (Jiwoo Lee et al., 2021), (J. Lee et al., 2024); (Ahn et al., 2022); (Planton et al., 2021)), 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.

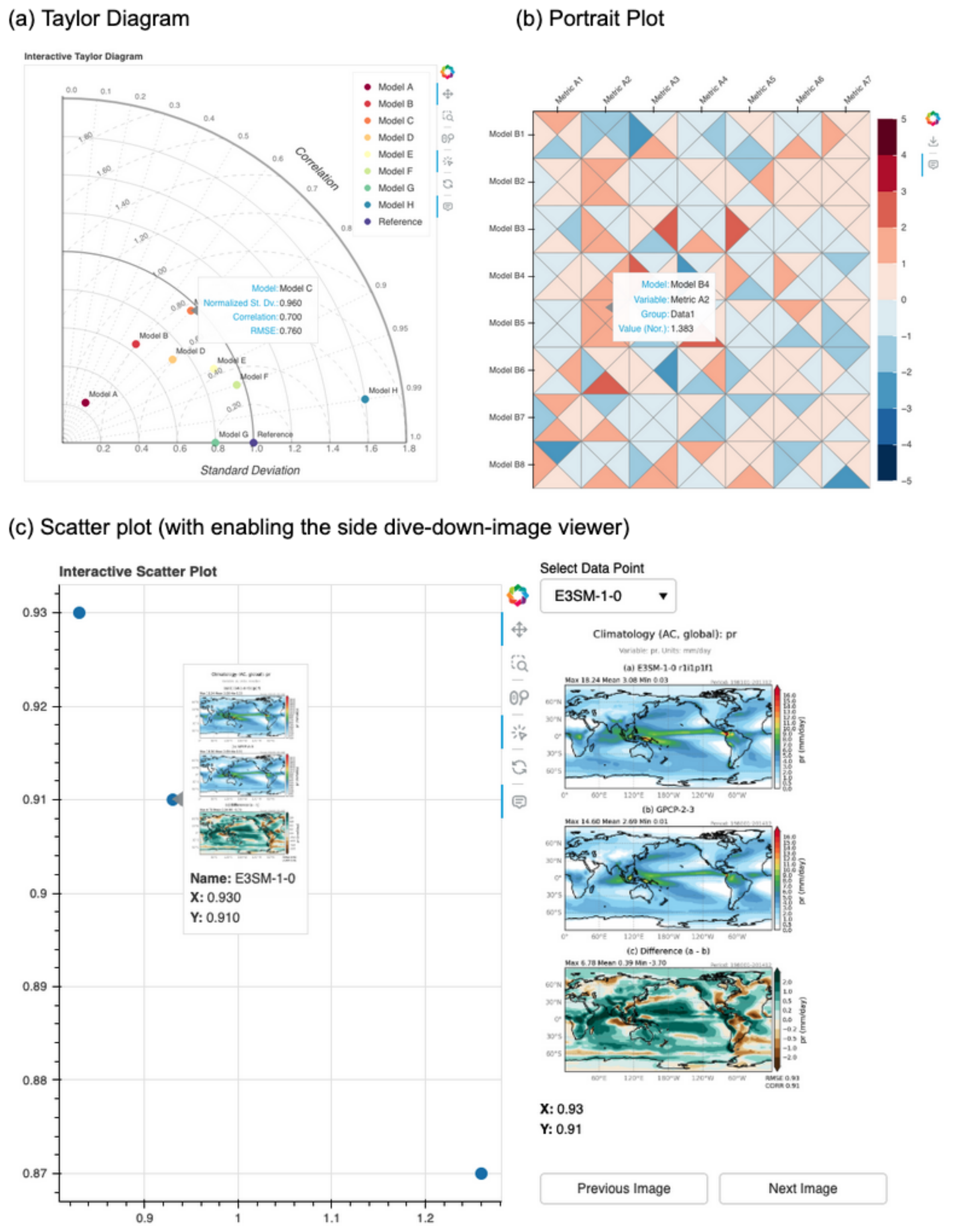
The development of ESMBenchmarkViz originated from the interactive visualization dashboard of the Program for Climate Model Diagnosis and Intercomparison (PCMDI) Metrics Package (Jiwoo Lee et al., 2023) (J. Lee et al., 2024), showing diverse evaluation metrics for ESMs along with diagnostic information (https://pcmdi.llnl.gov/research/metrics/). We refer to these diagnostics as "dive-down information," as they enable users to investigate metrics in greater detail.

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.

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. Users' mouse cursor hovering over for a specific data point (i.e., a specific ESM model and for 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 be also included to the tooltips or to 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); 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); 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 (e.g., (J. Lee et al., 2019), (Jiwoo Lee et al., 2021), (Ahn et al., 2022)). The PCMDI Metrics Package ((Jiwoo Lee et al., 2023), (J. Lee et al., 2024)) Team had developed a precursor version of this package to present evaluation output from hundreds of simulations in an efficient way (https://pcmdi.llnl.gov/metrics/).

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 interdisciplinary, we have included it to the package for its synergy with tooltips and images accompanying 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 ((J. Lee et al., 2024)).

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, following the installation instructions. We host all development activity at the GitHub Repository. We plan to set up a conda-forge channel on Anaconda for an easier installation.

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.

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/https://doi.org/10.1029/2007JD008972

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, Jiwoo, Gleckler, P., Ordonez, A., Ahn, M.-S., Ullrich, P., Vo, T., Boutte, J., Doutriaux, C., Durack, P., Shaheen, Z., Muryanto, L., Painter, J., & Krasting, J. (2023). *PCMDI/pcmdi_metrics: PMP version 3.1.2* (Version v3.1.2). Zenodo. https://doi.org/10.5281/zenodo.10236521

Lee, Jiwoo, Sperber, K. R., Gleckler, P. J., Taylor, K. E., & Bonfils, C. J. W. (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