




# pyDARTdiags: A Python package for manipulating observation sequences and calculating observation-space diagnostics for the Data Assimilation Research Testbed (DART)

Helen Kershaw <sup>1</sup>, Marlee Smith <sup>1</sup>, Isaac Arseneau <sup>2</sup>, and Lukas Kugler <sup>3</sup>

<sup>1</sup> NSF National Center for Atmospheric Research, Boulder CO, United States  <sup>2</sup> Texas Tech University  <sup>3</sup> University of Vienna   Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

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

Editor: 

Submitted: 05 August 2025

Published: unpublished

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/))

## Summary

pyDARTdiags is a Python package for manipulating observation sequences and calculating observation-space diagnostics for the Data Assimilation Research Testbed (DART).

Data assimilation is a scientific technique that combines observations (such as measurements from weather satellites, buoys, radar, or other sensors) with predictions from numerical models to produce an improved estimate of the state of a system. It is widely used in fields like meteorology, oceanography, hydrology, and environmental science.

During assimilation, the model state is transformed into observation space by interpolating the model state to each observation location. Observation space diagnostics are used to visualize observations and model predictions (in observation space), and compare their statistical properties both before and after assimilation. Thereby, these diagnostics are key tools to model evaluation and prediction tasks.

The Data Assimilation Research Testbed (DART) ([Anderson et al., 2009](#); [UCAR/NSF NCAR/CISL/DARes, 2025](#)) is a widely used community software facility for ensemble data assimilation. DART merges diverse and complex observations into an internal data format called “observation sequence” files, which inherit metadata from each observation. Observation sequence files including the model states are also generated, wherein the model states have been transformed into the observation space for direct comparisons and analysis. DART’s observation sequence files are central to its workflow, but their format and complexity can make them challenging to manipulate and analyze outside of the DART ecosystem. To manipulate and analyze the data using open-source tools a converter/interface is required.

pyDARTdiags is a Python package created to address this challenge. It provides tools to read, manipulate, and analyze DART observation sequence files using familiar, modern Python libraries. With pyDARTdiags, users can extract data, compute diagnostics, and create visualizations, all within reproducible Python workflows that integrate seamlessly with the broader scientific ecosystem.

## Statement of need

While DART provides robust tools for data assimilation, its observation sequence files are not easily accessible for users wishing to perform custom diagnostics or integrate with Python-based workflows. Existing DART tools for calculation of observation-space diagnostics are written

in Fortran with visualization in MATLAB, which may not be freely accessible to all users. pyDARTdiags provides a single package to process and visualize observation space diagnostics.

PyDARTdiags ingests observation sequences into an ObsSequence object which contains the metadata about an observation sequence and a DataFrame containing all the data for the observations.

This provides several advantages over the existing Fortran+MATLAB DART software:

- Providing Python routines for reading and writing DART observation sequence files allows for the manipulation of observation sequences interactively via DataFrames using popular, open-source data science libraries.
- Synthesizing the manipulation, analysis, and visualization of observation sequence files into a single Python workflow improves portability and flexibility over the fractured Fortran/MATLAB workflow.
- Enabling calculation of observation-space statistics (e.g., RMSE, bias, total spread) on a DataFrame enables Data Assimilation researchers and users to write custom diagnostics based on DataFrames. By decoupling the observation sequence file format from the DataFrame-based analysis, pyDARTdiags ensures that updates to the DART file format do not disrupt user-created diagnostic routines.
- Supporting both static and interactive plotting (via Matplotlib and Plotly), facilitates the processing of observational datasets, quality control, and gaining insights about the spatial distribution of outliers or other (technical) anomalies.
- Facilitating reproducible, scriptable workflows for observation-space diagnostics enables the inclusion in Jupyter notebook workflows. A concrete use case is its integration into the the CESM Regional Ocean and Carbon Configurator with Data Assimilation and Embedding (CROCODILE) project, which is a community platform for accelerating observationally-constrained regional ocean modeling. pyDARTdiags is used for observation space diagnostics and model-to-observation comparison in the Jupyter notebook workflows for this project.

Examples for manipulating observation sequences, visualizing observational data, and generating diagnostic plots can be found in the [pyDARTdiags examples gallery](#). A detailed description of the available diagnostic statistics and available plotting functions is provided in the [user guide](#). The pyDARTdiags source code is available at <https://github.com/NCAR/pyDARTdiags>.

## Acknowledgements

We thank the DART team, Enrico Milanese (Woods Hole Oceanographic Institution), and the broader data assimilation community for their feedback and contributions.

This material is based upon work supported by the U.S. National Science Foundation under Grant No. 2311382, “Collaborative Research: Frameworks: A community platform for accelerating observationally-constrained regional oceanographic modeling”, and by the NSF National Center for Atmospheric Research, which is a major facility sponsored by the U.S. National Science Foundation under Cooperative Agreement No. 1852977.

## References

- Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., & Avellano, A. (2009). The data assimilation research testbed: A community facility. *Bulletin of the American Meteorological Society*, 90(9), 1283–1296. <https://doi.org/10.1175/2009BAMS2618.1>
- UCAR/NSF NCAR/CISL/DARes. (2025). *The data assimilation research testbed* (Version v11.11.2). <https://doi.org/10.5065/D6WQ0202>