The Journal of Open Source Software

xeofs: Comprehensive EOF analysis in Python with xarray

Niclas Rieger[1,2,3] **and Samuel J. Levang**[4]

1 Centre de Recerca Matemàtica (CRM), Bellaterra, Spain **2** Departament de Física, Universitat Autònoma de Barcelona, Bellaterra, Spain **3** Instituto de Ciencias del Mar (ICM) - CSIC, Barcelona, Spain **4** Salient Predictions, Cambridge, MA, USA

DOI: 10.21105/joss.06060

Software

- Review
- Repository
- Archive

Editor: Samuel Forbes

Reviewers:

- @DamienIrving
- @malmans2

Submitted: 01 November 2023
Published: 02 January 2024

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

Summary

`xeofs` is a Python package tailored for the climate science community, designed to streamline advanced data analysis using dimensionality reduction techniques like Empirical Orthogonal Functions (EOF) analysis – often called Principal Component Analysis (PCA) in other domains. Integrating seamlessly with `xarray` objects (Hoyer & Hamman, 2017), `xeofs` makes it easier to analyze large, labeled, multi-dimensional datasets. By harnessing `Dask`'s capabilities (Dask Development Team, 2016), it scales computations efficiently across multiple cores or clusters, apt for extensive climate data applications.

Statement of Need

Climate science routinely deals with analyzing large, multi-dimensional datasets, whose complexity mirrors the intricate dynamics of the climate system itself. The extraction of meaningful insights from such vast datasets is challenging and often requires the application of dimensionality reduction techniques like EOF analysis (PCA outside climate science). Packages such as `scikit-learn` (Pedregosa et al., 2011) offer a range of reduction techniques, yet they often fall short of meeting the specific needs of climate scientists who work with variants of PCA (Hannachi, 2021) including ROCK-PCA (Bueso et al., 2020) and spectral, rotated PCA (Guilloteau et al., 2020).

Climate datasets are inherently multi-dimensional, usually involving time, longitude and latitude, and often include missing values representing geographical features like oceans or land. These characteristics require meticulous data transformations and tracking of missing values and dimension coordinates, which can be cumbersome and prone to error, increasing the workload, especially for smaller-scale projects. Furthermore, the size of climate datasets often necessitates out-of-memory processing.

While `xMCA` (He, 2019) and `eofs` (Dawson, 2016) have addressed some of these issues by offering analysis tools compatible with `xarray` and `Dask`, `xeofs` expands on these by including a broader range of techniques such as rotated (Kaiser, 1958), complex/Hilbert (Rasmusson et al., 1981), and extended (Weare & Nasstrom, 1982) PCA/EOF analysis. `xeofs` operates natively with `xarray` objects, preserving data labels and structure, and handles datasets with missing values adeptly. It also integrates seamlessly with `Dask` and shows improved performance in particular for larger datasets (Figure 1) due to its usage of randomized Singular Value Decomposition (SVD) (Halko et al., 2011).

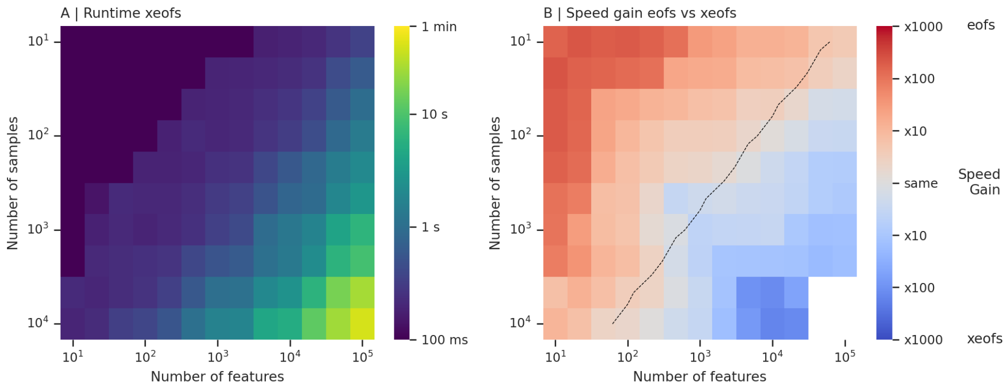

Figure 1: (A) Evaluation of xeofs computation times for processing 3D data sets of varying sizes. (B) Performance comparison between `xeofs` and `eofs` across different data set dimensions. Dashed black line indicates the contour of datasets approximately 3 MiB in size. Tests conducted [1] on an Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz, 12 threads (6 cores), with 16GB DDR4 RAM at 2667 MT/s.

Implementation

`xeofs` adopts the familiar `scikit-learn` style, delivering an intuitive interface where each method is a class with `fit`, and when applicable, `transform` and `inverse_transform` methods. It also offers flexibility by allowing users to introduce custom dimensionality reduction methods via a streamlined entry point to its internal pipeline. Additionally, the package includes a bootstrapping module for straightforward PCA model evaluation.

Available Methods

At the time of publication, `xeofs` provides the following methods:

Method	Alternative name	Reference
PCA	EOF analysis	
Rotated PCA	-	(Hendrickson & White, 1964; Kaiser, 1958)
Complex PCA	Hilbert EOF (HEOF) analysis	(Barnett, 1983; Horel, 1984; Rasmusson et al., 1981)
Complex Rotated PCA	-	(Horel, 1984)
Extended PCA	EEOF analysis / Multichannel Singular Spectrum Analysis (M-SSA)	(Broomhead & King, 1986; Weare & Nasstrom, 1982)
Optimal Persistence Analysis	OPA	(DelSole, 2001, 2006)
Geographically-Weighted PCA	GWPCA	(Harris et al., 2011)
Maximum Covariance Analysis	MCA, SVD analysis	(Bretherton et al., 1992)
Rotated MCA	-	(Cheng & Dunkerton, 1995)

[1]The script used to generate these results is available at https://github.com/nicrie/xeofs/blob/main/docs/perf/ .

Method	Alternative name	Reference
Complex MCA	Hilbert MCA/Analytical SVD	(Elipot et al., 2017)
Complex Rotated MCA	-	(Rieger et al., 2021)
Canonical Correlation Analysis	CCA	(Bretherton et al., 1992; Hotelling, 1936; Vinod, 1976)

Additionally, we are actively developing further enhancements to `xeofs`, with plans to incorporate advanced methods such as ROCK-PCA (Bueso et al., 2020) and spectral, rotated PCA (Guilloteau et al., 2020) in upcoming releases.

Acknowledgements

We express our sincere thanks to the individuals who have enhanced our software through their valuable issue reports and insightful feedback.

This work forms part of the Climate Advanced Forecasting of sub-seasonal Extremes (CAFE) project, undertaken within the Physics doctoral program at the Autonomous University of Barcelona. NR acknowledges the support of the European Union's Horizon 2020 research and innovation program, which has funded this work under the Marie Skłodowska-Curie grant (agreement No 813844).

References

Barnett, T. P. (1983). Interaction of the monsoon and pacific trade wind system at interannual time scales part i: The equatorial zone. *Monthly Weather Review*, *111*(4), 756–773. https://doi.org/10.1175/1520-0493(1983)111%3C0756:IOTMAP%3E2.0.CO;2

Bretherton, C., Smith, C., & Wallace, J. (1992). An intercomparison of methods for finding coupled patterns in climate data. *Journal of Climate*, *5*(6), 541–560. https://doi.org/10.1175/1520-0442(1992)005%3C0541:AIOMFF%3E2.0.CO;2

Broomhead, D. S., & King, G. P. (1986). Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena*, *20*(2), 217–236. https://doi.org/10.1016/0167-2789(86)90031-X

Bueso, D., Piles, M., & Camps-Valls, G. (2020). Nonlinear PCA for spatio-temporal analysis of earth observation data. *IEEE Transactions on Geoscience and Remote Sensing*, 1–12. https://doi.org/10.1109/TGRS.2020.2969813

Cheng, X., & Dunkerton, T. J. (1995). Orthogonal rotation of spatial patterns derived from singular value decomposition analysis. *Journal of Climate*, *8*(11), 2631–2643. https://doi.org/10.1175/1520-0442(1995)008%3C2631:OROSPD%3E2.0.CO;2

Dask Development Team. (2016). *Dask: Library for dynamic task scheduling*. https://dask.org

Dawson, A. (2016). *Eofs: A library for EOF analysis of meteorological, oceanographic, and climate data*. *4*(1), e14. https://doi.org/10.5334/jors.122

DelSole, T. (2001). Optimally persistent patterns in time-varying fields. *Journal of the Atmospheric Sciences*, *58*(11), 1341–1356. https://doi.org/10.1175/1520-0469(2001)058%3C1341:OPPITV%3E2.0.CO;2

DelSole, T. (2006). Low-frequency variations of surface temperature in observations and simulations. *Journal of Climate*, *19*(18), 4487–4507. https://doi.org/10.1175/JCLI3879.1

Elipot, S., Frajka-Williams, E., Hughes, C. W., Olhede, S., & Lankhorst, M. (2017). Observed basin-scale response of the north atlantic meridional overturning circulation to wind stress forcing. *Journal of Climate*, *30*(6), 2029–2054. https://doi.org/10.1175/JCLI-D-16-0664.1

Guilloteau, C., Mamalakis, A., Vulis, L., Georgiou, T. T., & Foufoula-Georgiou, E. (2020). Rotated spectral principal component analysis (rsPCA) for identifying dynamical modes of variability in climate systems. *arXiv:2004.11411 [Physics]*. https://doi.org/10.1175/JCLI-D-20-0266.1

Halko, N., Martinsson, P. G., & Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, *53*(2), 217–288. https://doi.org/10.1137/090771806

Hannachi, A. (2021). *Patterns identification and data mining in weather and climate*. Springer International Publishing. https://doi.org/10.1007/978-3-030-67073-3

Harris, P., Brunsdon, C., & Charlton, M. (2011). Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, *25*(10), 1717–1736. https://doi.org/10.1080/13658816.2011.554838

He, C. (2019). xMCA: Maximum covariance analysis in xarray for climate science. In *GitHub repository*. GitHub. https://github.com/Yefee/xMCA

Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, *17*(1), 65–70. https://doi.org/10.1111/j.2044-8317.1964.tb00244.x

Horel, J. (1984). Complex principal component analysis: Theory and examples. *Journal of Climate and Applied Meteorology*, *23*(12), 1660–1673. https://doi.org/10.1175/1520-0450(1984)023%3C1660:CPCATA%3E2.0.CO;2

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*(3), 321–377. https://doi.org/10.2307/2333955

Hoyer, S., & Hamman, J. (2017). Xarray: N-D labeled arrays and datasets in Python. *Journal of Open Research Software*, *5*(1). https://doi.org/10.5334/jors.148

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*(3), 187–200. https://doi.org/10.1007/BF02289233

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rasmusson, E. M., Arkin, P. A., Chen, W.-Y., & Jalickee, J. B. (1981). Biennial variations in surface temperature over the united states as revealed by singular decomposition. *Monthly Weather Review*, *109*(3), 587–598. https://doi.org/10.1175/1520-0493(1981)109%3C0587:BVISTO%3E2.0.CO;2

Rieger, N., Corral, Á., Olmedo, E., & Turiel, A. (2021). Lagged teleconnections of climate variables identified via complex rotated maximum covariance analysis. *Journal of Climate*, *34*(24), 9861–9878. https://doi.org/10.1175/JCLI-D-21-0244.1

Vinod, H. D. (1976). Canonical ridge and econometrics of joint production. *Journal of Econometrics*, *4*(2), 147–166. https://doi.org/10.1016/0304-4076(76)90010-5

Weare, B. C., & Nasstrom, J. S. (1982). Examples of extended empirical orthogonal function analyses. *Monthly Weather Review*, *110*(6), 481–485. https://doi.org/10.1175/1520-0493(1982)110%3C0481:EOEEOF%3E2.0.CO;2