The Journal of Open Source Software

scores: A Python package for verifying and evaluating models and predictions with xarray and pandas

Tennessee Leeuwenburg [1¶], **Nicholas Loveday** [1], **Elizabeth E. Ebert**[1], **Harrison Cook** [1], **Mohammadreza Khanarmuei** [1], **Robert J. Taggart** [1], **Nikeeth Ramanathan** [1], **Maree Carroll** [1], **Stephanie Chong** [2], **Aidan Griffiths**[3], **and John Sharples**[1]

1 Bureau of Meteorology, Australia **2** Independent Contributor, Australia **3** Work undertaken while at the Bureau of Meteorology, Australia ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished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).

Summary

scores is a Python package containing mathematical functions for the verification, evaluation and optimisation of forecasts, predictions or models. It primarily supports the geoscience communities; in particular, the meteorological, climatological and oceanographic communities. In addition to supporting the Earth system science communities, it also has wide potential application in machine learning and other domains such as economics.

scores not only includes common scores (e.g. Mean Absolute Error), it also includes novel scores not commonly found elsewhere (e.g. FIxed Risk Multicategorical (FIRM) score, Flip-Flop Index), complex scores (e.g. threshold-weighted continuous ranked probability score), and statistical tests (such as the Diebold Mariano test). It also contains isotonic regression which is becoming an increasingly important tool in forecast verification and can be used to generate stable reliability diagrams. Additionally, it provides pre-processing tools for preparing data for scores in a variety of formats including cumulative distribution functions (CDF). At the time of writing, scores includes over 50 metrics, statistical techniques and data processing tools.

All of the scores and statistical techniques in this package have undergone a thorough scientific and software review. Every score has a companion Jupyter Notebook tutorial that demonstrates its use in practice.

scores primarily supports xarray datatypes for Earth system data, allowing it to work with NetCDF4, HDF5, Zarr and GRIB data sources among others. scores uses Dask for scaling and performance. It has expanding support for pandas.

The software repository can be found at https://github.com/nci/scores/.

DRAFT

The Journal of Open Source Software

Statement of Need

The purpose of this software is (a) to mathematically verify and validate models and predictions and (b) to foster research into new scores and metrics.

Key Benefits of scores

In order to meet the needs of researchers and other users, `scores` provides the following key benefits.

Data Handling

- Works with n-dimensional data (e.g., geospatial, vertical and temporal dimensions) for both point-based and gridded data. `scores` can effectively handle the dimensionality, data size and data structures commonly used for:
  - gridded Earth system data (e.g. numerical weather prediction models)
  - tabular, point, latitude/longitude or site-based data (e.g. forecasts for specific locations).
- Handles missing data, masking of data and weighting of results.
- Supports `xarray` (Hoyer & Hamman, 2017) datatypes, and works with NetCDF4 (Unidata, 2024), HDF5 (The HDF Group & Koziol, 2020), Zarr (Miles et al., 2020) and GRIB (World Meteorological Organization, 2024) data sources among others.

Usability

- A companion Jupyter Notebook (Jupyter Team, 2024) tutorial for each metric and statistical test that demonstrates its use in practice.
- Novel scores not commonly found elsewhere (e.g. FIRM (Taggart et al., 2022), Flip-Flop Index (Griffiths et al., 2019, 2021)).
- All scores and statistical techniques have undergone a thorough scientific and software review.
- An area specifically to hold emerging scores which are still undergoing research and development. This provides a clear mechanism for people to share, access and collaborate on new scores, and be able to easily re-use versioned implementations of those scores.

Compatability

- Highly modular - provides its own implementations, avoids extensive dependencies and offers a consistent API.
- Easy to integrate and use in a wide variety of environments. It has been used on workstations, servers and in high performance computing (supercomputing) environments.
- Maintains 100% automated test coverage.
- Uses Dask (Dask Development Team, 2016) for scaling and performance.
- Expanding support for pandas (McKinney, 2010; The pandas development team, 2024).

Metrics, Statistical Techniques and Data Processing Tools Included in scores

At the time of writing, `scores` includes over 50 metrics, statistical techniques and data processing tools. For an up to date list, please see the `scores` documentation.

The ongoing development roadmap includes the addition of more metrics, tools, and statistical tests.

DRAFT

Table 1: A **curated selection** of the metrics, tools and statistical tests currently included in `scores`

	Description	A Selection of the Functions Included in scores
Continuous	Scores for evaluating single-valued continuous forecasts.	Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Additive Bias, Multiplicative Bias, Pearson's Correlation Coefficient, Flip-Flop Index (Griffiths et al., 2019, 2021), Quantile Loss, Murphy Score (Ehm et al., 2016).
Probability	Scores for evaluating forecasts that are expressed as predictive distributions, ensembles, and probabilities of binary events.	Brier Score (Brier, 1950), Continuous Ranked Probability Score (CRPS) for Cumulative Distribution Functions (CDFs) (including threshold-weighting, see Gneiting & Ranjan (2011)), CRPS for ensembles (Ferro, 2013; Gneiting & Raftery, 2007), Receiver Operating Characteristic (ROC), Isotonic Regression (reliability diagrams) (Dimitriadis et al., 2021).
Categorical	Scores for evaluating forecasts of categories.	Probability of Detection (POD), Probability of False Detection (POFD), False Alarm Ratio (FAR), Success Ratio, Accuracy, Peirce's Skill Score (Peirce, 1884), Critical Success Index (CSI), Gilbert Skill Score (Gilbert, 1884), Heidke Skill Score, Odds Ratio, Odds Ratio Skill Score, F1 Score, Symmetric Extremal Dependence Index (Ferro & Stephenson, 2011), FIxed Risk Multicategorical (FIRM) Score (Taggart et al., 2022).
Spatial	Scores that take into account spatial structure.	Fractions Skill Score (Roberts & Lean, 2008).
Statistical Tests	Tools to conduct statistical tests and generate confidence intervals.	Diebold-Mariano (Diebold & Mariano, 1995) with both the Harvey et al. (1997) and Hering & Genton (2011) modifications.
Processing Tools	Tools to pre-process data.	Data matching, discretisation, cumulative density function manipulation.

DRAFT

Use in Academic Work

In 2015, the Australian Bureau of Meteorology began developing a new verification system called Jive, which became operational in 2022. For a description of Jive see Loveday, Griffiths, et al. (2024). The Jive verification metrics have been used to support several publications (Foley & Loveday, 2020; Griffiths et al., 2017; Taggart, 2022a, 2022b, 2022c). `scores` has arisen from the Jive verification system and was created to modularise the Jive verification functions and make them available as an open source package.

`scores` has been used to explore user-focused approaches to evaluating probabilistic and categorical forecasts (Loveday, Taggart, et al., 2024).

Related Software Packages

There are multiple open source verification packages in a range of languages. Below is a comparison of `scores` to other open source Python verification packages. None of these include all of the metrics implemented in `scores` (and vice versa).

`xskillscore` (Bell et al., 2021) provides many but not all of the same functions as `scores` and does not have direct support for pandas. The Jupyter Notebook tutorials in `scores` cover a wider array of metrics.

`climpred` (Brady & Spring, 2021) uses `xskillscore` combined with data handling functionality, and is focused on ensemble forecasts for climate and weather. `climpred` makes some design choices related to data structure (specifically associated with climate modelling) which may not generalise effectively to broader use cases. Releasing `scores` separately allows the differing design philosophies to be considered by the community.

`METplus` (Brown et al., 2021) is a substantial verification system used by weather and climate model developers. `METplus` includes a database and a visualisation system, with Python and shell script wrappers to use the `MET` package for the calculation of scores. `MET` is implemented in C++ rather than Python. `METplus` is used as a system rather than providing a modular Python API.

`Verif` (Nipen et al., 2023) is a command line tool for generating verification plots whereas `scores` provides a Python API for generating numerical scores.

`Pysteps` (Imhoff et al., 2023; Pulkkinen et al., 2019) is a package for short-term ensemble prediction systems, and includes a significant verification submodule with many useful verification scores. `PySteps` does not provide a standalone verification API.

`PyForecastTools` (Morley & Burrell, 2020) is a Python package for model and forecast verification which supports `dmarray` rather than `xarray` data structures and does not include Jupyter Notebook tutorials.

Acknowledgements

We would like to thank Jason West and Robert Johnson from the Bureau of Meteorology for their feedback on an earlier version of this manuscript.

We would like to thank and acknowledge the National Computational Infrastructure (nci.org.au) for hosting the `scores` repository within their GitHub organisation.

References

Bell, R., Spring, A., Brady, R., Huang, A., Squire, D., Blackwood, Z., Sitter, M. C., & Chegini., T. (2021). *xarray-contrib/xskillscore: Metrics for verifying forecasts*. Zenodo. https://doi.org/10.5281/zenodo.5173153

Brady, R. X., & Spring, A. (2021). climpred: Verification of weather and climate forecasts. *Journal of Open Source Software*, *6*(59), 2781. https://doi.org/10.21105/joss.02781

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078%3C0001:vofeit%3E2.0.co;2

Brown, B., Jensen, T., Gotway, J. H., Bullock, R., Gilleland, E., Fowler, T., Newman, K., Adriaansen, D., Blank, L., Burek, T., & others. (2021). The Model Evaluation Tools (MET): More than a decade of community-supported forecast verification. *Bulletin of the American Meteorological Society*, *102*(4), E782–E807. https://doi.org/10.1175/bams-d-19-0093.1

DRAFT

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

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. https://doi.org/10.3386/t0169

Dimitriadis, T., Gneiting, T., & Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, *118*(8), e2016191118. https://doi.org/10.1073/pnas.2016191118

Ehm, W., Gneiting, T., Jordan, A., & Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *78*(3), 505–562. https://doi.org/10.1111/rssb.12154

Ferro, C. A. T. (2013). Fair scores for ensemble forecasts: Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, *140*(683), 1917–1923. https://doi.org/10.1002/qj.2270

Ferro, C. A. T., & Stephenson, D. B. (2011). Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting*, *26*(5), 699–713. https://doi.org/10.1175/WAF-D-10-05030.1

Foley, M., & Loveday, N. (2020). Comparison of single-valued forecasts in a user-oriented framework. *Weather and Forecasting*, *35*(3), 1067–1080. https://doi.org/10.1175/waf-d-19-0248.1

Gilbert, G. K. (1884). Finley's tornado predictions. *American Meteorological Journal*, *1*(5), 166–172.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, *29*(3), 411–422. https://doi.org/10.1198/jbes.2010.08110

Griffiths, D., Foley, M., Ioannou, I., & Leeuwenburg, T. (2019). Flip-Flop Index: Quantifying revision stability for fixed-event forecasts. *Meteorological Applications*, *26*(1), 30–35. https://doi.org/10.1002/met.1732

Griffiths, D., Jack, H., Foley, M., Ioannou, I., & Liu, M. (2017). *Advice for automation of forecasts: A framework*. Bureau of Meteorology. https://doi.org/10.22499/4.0021

Griffiths, D., Loveday, N., Price, B., Foley, M., & McKelvie, A. (2021). Circular Flip-Flop Index: Quantifying revision stability of forecasts of direction. *Journal of Southern Hemisphere Earth Systems Science*, *71*(3), 266–271. https://doi.org/10.1071/es21010

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*(2), 281–291. https://doi.org/10.1016/S0169-2070(96)00719-4

Hering, A. S., & Genton, M. G. (2011). Comparing spatial predictions. *Technometrics*, *53*(4), 414–425. https://doi.org/10.1198/tech.2011.10136

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

Imhoff, R. O., De Cruz, L., Dewettinck, W., Brauer, C. C., Uijlenhoet, R., Heeringen, K.-J. van, Velasco-Forero, C., Nerini, D., Van Ginderachter, M., & Weerts, A. H. (2023). Scale-dependent blending of ensemble rainfall nowcasts and numerical weather prediction in the open-source pysteps library. *Quarterly Journal of the Royal Meteorological Society*,

DRAFT

149(753), 1335–1364. https://doi.org/10.1002/qj.4461

Jupyter Team. (2024). *Jupyter interactive notebook*. GitHub. https://github.com/jupyter/notebook

Loveday, N., Griffiths, D., Leeuwenburg, T., Taggart, R., Pagano, T. C., Cheng, G., Plastow, K., Ebert, E., Templeton, C., Carroll, M., Khanarmuei, M., & Nagpal, I. (2024). *The Jive verification system and its transformative impact on weather forecasting operations*. arXiv. https://doi.org/10.48550/arXiv.2404.18429

Loveday, N., Taggart, R., & Khanarmuei, M. (2024). A user-focused approach to evaluating probabilistic and categorical forecasts. *Weather and Forecasting*. https://doi.org/10.1175/waf-d-23-0201.1

McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Miles, A., Kirkham, J., Durant, M., Bourbeau, J., Onalan, T., Hamman, J., Patel, Z., shikharsg, Rocklin, M., dussin, raphael, Schut, V., Andrade, E. S. de, Abernathey, R., Noyes, C., sbalmer, bot, pyup.io, Tran, T., Saalfeld, S., Swaney, J., … Banihirwe, A. (2020). *Zarr-developers/zarr-python: v2.4.0* (Version v2.4.0). Zenodo. https://doi.org/10.5281/zenodo.3773450

Morley, S., & Burrell, A. (2020). *Drsteve/PyForecastTools: Version 1.1.1* (Version v1.1.1). Zenodo. https://doi.org/10.5281/zenodo.3764117

Nipen, T. N., Stull, R. B., Lussana, C., & Seierstad, I. A. (2023). Verif: A weather-prediction verification tool for effective product development. *Bulletin of the American Meteorological Society*, *104*(9), E1610–E1618. https://doi.org/10.1175/bams-d-22-0253.1

Peirce, C. S. (1884). The numerical measure of the success of predictions. *Science*, *ns-4*(93), 453–454. https://doi.org/10.1126/science.ns-4.93.453.b

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., & Foresti, L. (2019). Pysteps: An open-source Python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development*, *12*(10), 4185–4219. https://doi.org/10.5194/gmd-12-4185-2019

Roberts, N. M., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 78–97. https://doi.org/10.1175/2007MWR2123.1

Taggart, R. (2022a). *Assessing calibration when predictive distributions have discontinuities*. http://www.bom.gov.au/research/publications/researchreports/BRR-064.pdf

Taggart, R. (2022b). Evaluation of point forecasts for extreme events using consistent scoring functions. *Quarterly Journal of the Royal Meteorological Society*, *148*(742), 306–320. https://doi.org/10.1002/qj.4206

Taggart, R. (2022c). Point forecasting and forecast evaluation with generalized huber loss. *Electronic Journal of Statistics*, *16*(1), 201–231. https://doi.org/10.1214/21-ejs1957

Taggart, R., Loveday, N., & Griffiths, D. (2022). A scoring framework for tiered warnings and multicategorical forecasts based on fixed risk measures. *Quarterly Journal of the Royal Meteorological Society*, *148*(744), 1389–1406. https://doi.org/10.1002/qj.4266

The HDF Group, & Koziol, Q. (2020). *HDF5-version 1.12.0*. https://doi.org/10.11578/dc.20180330.1

The pandas development team. (2024). *Pandas-dev/pandas: pandas* (Version v2.2.2). Zenodo. https://doi.org/10.5281/zenodo.10957263

DRAFT

Unidata. (2024). *Network common data form (NetCDF)*. UCAR/Unidata Program Center. https://doi.org/10.5065/D6H70CW6

World Meteorological Organization. (2024). *WMO no. 306 FM 92 GRIB (edition 2)*. World Meteorological Organization. https://codes.wmo.int/grib2

DRAFT