The Journal of Open Source Software

CrocoLakeTools: A Python package to convert ocean observations to the parquet format

Enrico Milanese[1], **David Nicholson**[1], **Gaël Forget**[2], and **Susan Wijffels**[1]

1 Woods Hole Oceanographic Institution, United States of America **2** Massachusetts Institute of Technology, United States of America

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Investigations of the ocean state are possible thanks to the ever growing number of measurements performed with multiple instruments by different research missions. The vast and variegated efforts have brought the community to define data storage conventions (e.g. CF-netCDF) and to assemble collections of datasets (e.g. the World Ocean Database). Yet, accessing these datasets often requires the usage of multiple tools, is inefficient over the cloud, and presents an overall high entrance barrier in terms of knowledge and time required to effectively access these resources. CrocoLakeTools is a Python package that address those shortcomings by providing workflows to convert several datasets from their original format to a uniform parquet dataset with a shared schema.

Statement of need

`CrocoLakeTools` is a Python package to build workflows that convert ocean observations from different formats (e.g. netCDF, CSV) to parquet. CrocoLakeTools take advantage of Python's well-established and growing ecosystem of open tools: it uses `dask`'s parallel computing capabilities to convert multiple files at once and to handle larger-than-memory data. `dask` is already well-integrated with `xarray` and `pandas`, two widely used Python libraries for the treatment of array and tabular data, respectively, and with `pyarrow`, the API to the Apache Arrow library which is used to generate the parquet dataset.

Parquet is a data storage format for big tidy data which presents several advantages: it is language agnostic (it can be accessed with Python, Matlab, Julia and web developement technologies); it offers faster reading performances than other tabular formats such as CSV; it is optimized for cloud systems storage and operations; it is widespread in the data science community, leading to a multitude of freely accessible tools and educational material.

`CrocoLakeTools` was developed with the goal of building and serving CrocoLake, a regularly refreshed database of oceanographic observations that are pre-filtered to contain only quality-controlled measurements. `CrocoLakeTools` was designed to be used by researchers and data scientists and engineers in oceanograhy. `CrocoLake` was designed to be accessed by the wider oceanographic community.

Code architecture

Converters

The core task of CrocoLakeTools is to take one or more files from a dataset and convert them to parquet, ensuring that CrocoLake's schema is followed. This is achieved through the methods contained in the `Converter` class and its subclasses. While the conversion of all datasets requires some general functionality (e.g. renaming the original variables to the final schema), each conversion requires specific tools for the specific dataset (e.g. the map used to rename the variable). CrocoLakeTools then hosts a converter for each dataset that implement the specific needs of that datasets and inherits from `Converter`, which contains the shared methods.

Workflow

The first step in the workflow is to retrieve the original files. This is generally left to the user, although we provide methods to download Argo data and we wish to be able in the future to provide this type of tools for other datasets as well.

[TD: ADD STEPS]

Accessing CrocoLake

The `examples` folder contains examples for how to access parquet datasets with several programming languages: Python, Matlab, Julia. We hope to include soon R too. Separate open repositories contains more examples for each language, see: CrocoLake-Python, CrocoLake-Matlab, CrocoLake-Julia.

Example

[TD ADD FIGURES]

Documentation

Documentation is available at [TD add link]. It describes the CrocoLake dataset, thesub-datasets that it is made, and how they are obtained. It provides references to the original files origins for download before converting them.

Citation

If you use CrocoLakeTools and/or CrocoLake, do not limit yourself to citing this manuscript. Remember to cite the datasets that you have used as indicated in the documentation. For example, if your work relies on Argo measurements, acknowledge Argo (Wong et al., 2020). This is important for each product to track their impact.

Acknowledgements

We acknowledge funding from [NSF CSSI CROCODILE details]

References

Wong, A. P., Wijffels, S. E., Riser, S. C., Pouliquen, S., Hosoda, S., Roemmich, D., Gilson, J., Johnson, G. C., Martini, K., Murphy, D. J., & others. (2020). Argo data 1999–2019: Two million temperature-salinity profiles and subsurface velocity observations from a global array of profiling floats. *Frontiers in Marine Science*, *7*, 700.