The Journal of Open Source Software

NCDatasets.jl: a Julia package for manipulating netCDF data sets

Alexander Barth [1]

1 GHER, University of Liège, Liège, Belgium

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 16 January 2024
Published: unpublished

Summary

NCDatasets is a Julia package that allows users to read, create and modify netCDF files (Network Common Data Format). It is based on the Unidata netCDF library [Rew & Davis (1990); Rew2006; *NetCDF Binary Encoding Extension Standard* (2011)] which also supports reading data from remote servers using OPeNDAP (Open-source Project for a Network Data Access Protocol, https://www.opendap.org) and the Zarr file format (*Zarr Storage Specification 2.0 Community Standard*, 2022). These additional formats are also accessible to users of NCDatasets.

The aim of NCDatasets is to expose the data and metadata stored in the NetCDF file as lazy data-structures (in particular arrays and dictionaries) used in Julia. Lazy in this context means that only the requested subset of data is loaded into RAM or written to the disk. One of the design goals of NCDatasets and the netCDF library in general is being able to work with datasets which are potentially larger than the total amount of RAM in a system and to process that data per subset.

NetCDF allows users to add metadata to datasets and individual variables in form of a list of key value-pairs called attributes. The meaning of these attributes is standardized in the CF conventions (Eaton et al., 2023). While originally proposed for NetCDF files, the CF conventions are now also applied in the context of other formats like GRIB (e.g. the Julia package GRIBDatasets or the python package cfgrib).

Statement of need

NetCDF is a commonly used data format in Earth sciences (in particular oceanography, atmospheric sciences and climatology) to store model data, satellite observations and in situ observations. It is particularly well established as a format for distributing and archiving data. The Julia programming language with its native array types, just-in-time compilation and automatic function specialization based on data types are well suited for processing and analyzing large amounts of data often found in Earth sciences. Therefore, a convenient API mapping the concepts for the NetCDF format and CF convention to the corresponding equivalents of the Julia programming language is desirable. There are currently 64 registered Julia packages (as for 15 January 2024) that have NCDatasets as direct or indirect dependency (not counting for optional dependencies). For example, NCDatasets is used with satellite data (Barth et al., 2022; Doglioni et al., 2023), in situ observations (Belgacem et al., 2021; Shahzadi et al., 2021) as well as numerical ocean models (Ramadhan et al., 2020) and atmospheric models (Klöwer et al., 2023).

The Journal of Open Source Software

Installation

NCDatasets supports Julia 1.6 and later and can be installed with the Julia package manager using the following Julia commands:

```
using Pkg
Pkg.add("NCDatasets")
```

This will automatically install all dependencies and in particular the Unidata netCDF C library for which compiled binaries are currently available for Linux, FreeBSD, Mac OS and Windows thanks to the efforts of the Yggdrasil.jl project.

Features

The main objects in the netCDF data model are the dataset (typically representing a whole file), variables (named n-dimensional arrays with named dimensions), dimensions (mapping the dimension names to the corresponding length), attributes and groups (a dataset contained within a dataset). Groups can be recursively nested. Variable names must be unique within a given group, but two different groups can re-use the same name. Current features of NCDatasets include:

- Attributes, dimensions and groups are exposed to users as dictionary-like objects. Modifying them will directly modify the underlying NetCDF file as long as the file is open in write mode.
- Variables are exposed as array-like objects. Indexing these arrays with the usual Julia syntax will result in loading the corresponding subset into memory. Likewise, assigning a value to a subset will write the data to the disk.
- The netCDF C API provides several functions to query information about the various objects of the netCDF data model. It is possible to query the data and metadata of a NetCDF file in the same way that one would query an array or dictionary.
- Every time a netCDF variable is loaded the required memory is automatically allocated. Once this memory is no longer used it will be deallocated by Julia's garbage collector. For high-performance applications, the repeated allocation and deallocation can cause a significant performance overhead. For this use-case, NCDatasets provides in-place variants for loading data.
- Data stored in a contiguous ragged array representation (Eaton et al., 2023; Hassell et al., 2017) are loaded as a vector of vectors. It is typically used to load a list of in situ profiles or time series, each of different length.
- Storage parameters like compression and data chunks can be queried and defined.
- Data transformations defined via the CF conventions are applied per default (including scaling, adding an offset, conversion to the `DateTime` structure). Several calendars are standardized in the CF conventions (standard, Gregorian, proleptic Gregorian, Julian, all leap, no leap, 360 day). Where possible, dates are automatically converted to Julia's native date time type, which uses the proleptic Gregorian calendar conforming to the ISO 8601 standard. Date types are handled using the package CFTimes (originally part of NCDatasets)
- Additional functionality includes multi-file support (virtually concatenating variables of multiple NetCDF variable spanning over multiple files), a view of the variable and datasets (virtual subset without loading the whole data in memory), subset variables and dataset using coordinate values instead of indices using the package CommonDataModel (also originally part of NCDatasets).

Similar software

The Julia package NetCDF.jl from Fabian Gans and contributors is an alternative to this package which supports a more Matlab/Octave-like interface for reading and writing netCDF files while this package, NCDatasets, is more influenced by the python netCDF4 package. In the R community, the packages RNetCDF and ncdf4 fulfill a similar role.

Acknowledgements

I thank all contributors to this package, among others, George Datseris, Tristan Carion, Martijn Visser and Charles Troupin as well as Unidata for the netCDF C library and their time and efforts responding to my questions and issues. All contributors to the Yggdrasil.jl project for their effort in building the netCDF library and the required dependencies are also acknowledged.

Funding

Acknowledgment is given to the F.R.S.-FNRS (Fonds de la Recherche Scientifique de Belgique) for funding the position of Alexander Barth. This work was partly performed with funding from the Blue-Cloud 2026 project under the Horizon Europe programme, Grant Agreement No. 101094227.

References

Barth, A., Alvera-Azcárate, A., Troupin, C., & Beckers, J.-M. (2022). DINCAE 2.0: Multivariate convolutional neural network with error estimates to reconstruct sea surface temperature satellite and altimetry observations. *Geoscientific Model Development*. https://doi.org/10.5194/gmd-2021-353

Belgacem, M., Schroeder, K., Barth, A., Troupin, C., Pavoni, B., Raimbault, P., Garcia, N., Borghini, M., & Chiggiato, J. (2021). Climatological distribution of dissolved inorganic nutrients in the western mediterranean sea (1981–2017). *Earth System Science Data*, *13*(12), 5915–5949. https://doi.org/10.5194/essd-13-5915-2021

Doglioni, F., Ricker, R., Rabe, B., Barth, A., Troupin, C., & Kanzow, T. (2023). Sea surface height anomaly and geostrophic current velocity from altimetry measurements over the arctic ocean (2011–2020). *Earth System Science Data*, *15*(1), 225–263. https://doi.org/10.5194/essd-15-225-2023

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Blower, J., Caron, J., Signell, R., Bentley, P., Rappa, G., Höck, H., Pamment, A., Juckes, M., Raspaud, M., Horne, R., Whiteaker, T., Blodgett, D., Zender, C., Lee, D., … Bartholomew, S. (2023). *NetCDF Climate and Forecast (CF) Metadata Conventions v1.11*. CF Conventions Committee. http://cfconventions.org/Data/cf-conventions/cf-conventions-1.11/cf-conventions.html

Hassell, D., Gregory, J., Blower, J., Lawrence, B. N., & Taylor, K. E. (2017). A data model of the climate and forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2.1). *Geoscientific Model Development*, *10*(12), 4619–4646. https://doi.org/10.5194/gmd-10-4619-2017

Klöwer, M., Gelbrecht, M., Hotta, D., Willmert, J., Silvestri, S., Wagner, G. L., White, A., Hatfield, S., Kimpson, T., Constantinou, N. C., & Hill, C. (2023). SpeedyWeather.jl: Reinventing atmospheric general circulation models towards interactivity and extensibility. In *Journal of Open Source Software (submitted)*. The Open Journal.

NetCDF binary encoding extension standard: NetCDF classic and 64-bit offset format (OGC 10-092r3). (2011). Open Geospatial Consortium. http://www.opengis.net/doc/IS/netcdf-binary/1.0

Ramadhan, A., Wagner, G. L., Hill, C., Campin, J.-M., Churavy, V., Besard, T., Souza, A., Edelman, A., Ferrari, R., & Marshall, J. (2020). Oceananigans.jl: Fast and friendly geophysical fluid dynamics on GPUs. *Journal of Open Source Software*, *5*(53), 2018. https://doi.org/10.21105/joss.02018

Rew, R., & Davis, G. (1990). NetCDF: An interface for scientific data access. *IEEE Computer Graphics and Applications*, *10*(4), 76–82. https://doi.org/10.1109/38.56302

Shahzadi, K., Pinardi, N., Barth, A., Troupin, C., Lyubartsev, V., & Simoncelli, S. (2021). A new global ocean climatology. *Frontiers in Environmental Science*, *9*. https://doi.org/10.3389/fenvs.2021.711363

Zarr storage specification 2.0 community standard (No. 21-050r1). (2022). Open Geospatial Consortium. http://www.opengis.net/doc/CS/zarr/2.0