NFDI4Earth

Enhancing Earth system model evaluation with data cube enabled machine learning

Breixo Soliño Fernández[1], Rémi Kazeroni[1]

[1] Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR)

14.04.2023

This work has been funded by the German Research Foundation (NFDI4Earth, DFG project no. 460036893, https://www.nfdi4earth.de/).

Pilot title	Enhancing Earth system model evaluation with data cube enabled machine learning
Project Duration	12 months
Contributors[1]	Breixo Soliño Fernández (Conceptualization, Software, Writing) Rémi Kazeroni (Conceptualization, Project administration, Validation, Writing) Veronika Eyring (Conceptualization lead, Funding acquisition, Review & editing) Fernando Iglesias-Suarez (Conceptualization, Review & editing) Axel Lauer (Conceptualization, Review & editing) Birgit Hassler (Review & editing)
DOI	10.5281/zenodo.7826038
Corresponding authors	Breixo Soliño Fernández and Remi Kazeroni
Acknowledgements	Björn Brötz (Conceptualization) Miguel D. Mahecha (Conceptualization) Markus Reichstein (Conceptualization) Jakob Runge (Conceptualization)

Abstract

Machine learning (ML) techniques represent a promising avenue to enhance climate model evaluation, better understand Earth system processes and further improve climate modeling. The application of ML techniques on multivariate climate data with high temporal and spatial frequencies may lead to several technical challenges, a major issue often being that input data can significantly exceed the memory available on computing systems. This data challenge can be circumvented by relying on cloud ready data which allows processing of data in a memory

[1] Based on the CRediT Contributor Roles Taxonomy: https://credit.niso.org/

efficient way and unlocks the application of ML methods on large input datasets. In this pilot project, the Earth System Model Evaluation Tool (ESMValTool) was extended by interfacing it with cloud-based analysis-ready data streams from the Earth system data cube infrastructure. In a second step, an ML-based analysis package was coupled to ESMValTool to demonstrate the integration of ML algorithms for climate model evaluation, with a particular focus on causal discovery applied to Arctic-midlatitude teleconnections. The main beneficiaries of this pilot are the Earth system science community, including climate model development and evaluation groups, the climate informatics community and infrastructure providers such as High Performance Computing (HPC) centers and science data providers. This pilot opens up a promising avenue towards the efficient handling of Earth system data for application of ML methods which will benefit the NFDI4Earth community.

I. Introduction

The application of ML approaches in Earth system sciences offers great opportunities to gain new insights into the climate system and improve climate models, but also raises substantial new challenges (Reichstein et al., 2019). A major technical challenge when applying ML-based algorithms to climate data is to efficiently handle large input datasets whose sizes are currently limited by the memory available on the compute nodes. In recent years, several tools and infrastructures have been developed and adopted by the Earth system science community to improve the handling and analysis of climate data, including ESMValTool and Earth system data cubes.

ESMValTool (Righi et al., 2020) is an open-source, community-developed, climate model diagnostics and evaluation software package. ESMValTool has been developed with the aim of taking model evaluation to the next level by facilitating analysis of many different model components, providing well-documented source code and scientific background of implemented diagnostics and metrics and allowing for traceability and reproducibility of results (provenance). ESMValTool was originally designed and optimized to handle the large data volume of the output from the Coupled Model Intercomparison Project (CMIP; Eyring et al., 2016). It consists of a preprocessor (ESMValCore) that performs common pre-processing operations and a diagnostic part which includes diagnostics and performance metrics for specific scientific applications. Among others, it provides diagnostics to analyze large-scale indicators of climate change (Eyring et al., 2020), emergent constraints and climate projections (Lauer et al., 2020), as well as extreme events and climate impacts (Weigel et al., 2021).

The Earth system data cube (ESDC; Mahecha et al., 2020) is a concept developed to efficiently apply user-defined functions on arbitrary dimensions of analysis-ready Earth system data cubes (e.g., location, time, variable, model ensemble member, etc.) which are provided as cloud-ready nd-arrays. To overcome the lack of interoperability between data streams and sources in Earth system science, this infrastructure provides analysis-ready data for a wide range of variables originally available at different spatial and temporal resolutions, and with various gridding approaches.

These two tools have improved the handling of Earth system models and observational data but lack support for the implementation and application of ML algorithms in a memory-efficient way. To address such limitations, this pilot project aimed at interfacing ESMValTool and the Earth system data cubes to enable model evaluation of cloud-based data. This connection offers great potential to reduce the memory footprint of the ESMValTool pre-processing chain thanks to the availability of ESDC as Zarr files which are loaded into memory only when needed. The implemented solution allows to keep large volumes of pre-processed input data on cloud storage, facilitating their usage in ML algorithms.

The integration of ML techniques in ESMValTool during this pilot project focused on causal discovery (Runge et al., 2019a). The coupling of ESMValTool with the causal discovery algorithm PCMCI (Runge et al., 2019b) has benefitted a research project aiming at better characterizing climate change in the Arctic region associated with Arctic Amplification and its further impact on midlatitude weather and climate (Galytska et al., 2022).

II. Results

In this pilot, we have extended ESMValTool in two different ways to enable usage of cloud data and the application of ML-based techniques to climate data. By coupling ESMValTool to a cloud data infrastructure, the ESDC, and to a ML-based package, PCMCI, we have laid the foundation for novel innovative workflows from cloud stores to ML-based diagnostics for model evaluation.

a) Implemented Solution

The new capability of ESMValTool to access and process the ESDC dataset is implemented as a "CMORizer script". "CMORizers" are an integral part of the ESMValTool pre-processing infrastructure and used to make a dataset compliant with the CMOR standard, the required format for input data to be used with ESMValTool. The CMOR (Climate Model Output Rewriter[2]) standard is an extension of the CF-standard[3] with additional requirements for CMIP. While usually a dataset has to be downloaded before it can be "CMORized" (i.e. adapted to the CMOR standard), the availability of the ESDC as a Zarr dataset allows the data to be CMORized directly from the cloud. ESMValTool processes data with the Python library Iris (Met Office, 2010 - 2022). Currently, Iris is not able to open Zarr datasets and therefore a work-around was implemented by opening the dataset with xarray (Hoyer and Hamman, 2017), which allows a conversion to Iris while still connected to the cloud. This conversion is not seamless, as it only supports individual variables and Iris may require metadata corrections before the conversion.

Once the data are converted to an Iris cube, the existing ESMValTool infrastructure can be used to convert the dataset to CMOR standard. This process may involve several changes depending on the original data, for example changing of units, coordinates, variable names or frequencies. In this pilot we focused on the ERA5 data stream available on ESDC because it allowed comparison with the original ERA5 dataset already supported by ESMValTool. After the CMORization process is finished, the dataset is saved to disk so that it can be used by ESMValTool.

The coupling between the causal discovery algorithm (PCMCI) and ESMValTool was performed in the framework of a study on Arctic-midlatitude teleconnections (Galytska et al., 2022), in a newly created recipe and diagnostic script. A "recipe" is a configuration file that defines the input

[2] https://cmor.llnl.gov/

[3] https://cfconventions.org/

data and preprocessing steps applied to be analyzed by ESMValTool diagnostic scripts. Diagnostic scripts are run as the last step of a recipe and are used to generate the scientific output and the final results such as plots and netCDF files. The preprocessing of data with ESMValTool is an important step to prepare data for analysis with the causal discovery analysis tool Tigramite[4]. The recipe demonstrates how ESMValTool can use an external package by using a wrapper diagnostic. Because of license incompatibilities between Tigramite (copy-left) and ESMValTool (Apache License 2.0), Tigramite could not be implemented directly into ESMValTool. Instead, ESMValTool writes the preprocessed data to disk that can then be used as input for Tigramite in a second step.

b) Data and Software availability

The work done within this pilot project was implemented in ESMValTool v2.8.0 released in March 2023[5]. ESMValTool is licensed under Apache License 2.0[6].

The latest ESMValTool release version is available as a conda-forge package[7], making it easily accessible to users of the Conda[8] or Mamba[9] package managers. ESMValTool can also be installed directly from its GitHub repository[10]. Detailed installation[11] and configuration[12] instructions are provided in the ESMValTool documentation.

The new capability of ESMValTool to access and process the ESDC dataset was implemented as a "CMORizer" script. These scripts are used to pre-process data that are not compliant to the CMOR standard before being used by ESMValTool. Typically, CMORizer scripts are run individually before running ESMValTool. Currently, ESMValTool provides CMORizer scripts for almost 90 datasets with a full list including ESDC available in the documentation[13].

The coupling of ESMValTool with Tigramite was done with a recipe and a diagnostic script which will be published in the ESMValTool v2.9.0 release. The ESMValTool documentation provides a detailed tutorial[14] for new users and developers. Currently, ESMValTool includes about 150 recipes with a full list sorted by scientific topic available online[15].

c) Innovation and FAIRness

This pilot provides a proof of concept on how to use a cloud dataset with ESMValTool. The ESDC is the first dataset in ESMValTool that is accessed and adapted to CMOR standard directly from

[4] https://jakobrunge.github.io/tigramite/
[5] https://github.com/ESMValGroup/ESMValTool/releases/tag/v2.8.0
[6] https://github.com/ESMValGroup/ESMValTool/blob/main/LICENSE
[7] https://anaconda.org/conda-forge/esmvaltool
[8] https://docs.conda.io/en/latest/index.html
[9] https://mamba.readthedocs.io/en/latest/index.html
[10] https://github.com/ESMValGroup/ESMValTool
[11] https://docs.esmvaltool.org/en/latest/quickstart/installation.html
[12] https://docs.esmvaltool.org/en/latest/quickstart/configuration.html
[13] https://docs.esmvaltool.org/en/latest/input.html#supported-datasets-for-which-a-cmorizer-script-is-available
[14] https://esmvalgroup.github.io/ESMValTool_Tutorial/
[15] https://docs.esmvaltool.org/en/latest/recipes/index.html

the cloud. This is a major advance over the standard approach where the user has to first download the data, either manually or with the help of downloading scripts.

Although the CMORized dataset still has to be saved to disk before being used, this method provides an advantage on usability and more efficient storage. Usability is improved as it merges downloading and CMORization into one step. Storage is more efficient due to accessing only a subset of variables and only storing the final product instead of having to download raw data that may contain variables not needed for a specific scientific application.

This pilot project has also helped to identify and provide insights on the challenges that need to be tackled before ESMValTool can use the cloud during the whole processing chain without requiring storing any data.

This pilot addresses the following FAIR principles. Findable, Accessible: the final product of this pilot has been released as open access software, using an open access dataset. The access to the data is automated by ESMValTool. Interoperable: The same method used within this pilot can be extended to support other Zarr-based cloud storage providers or other external ML packages. The standardization of the data allows for comparison with other datasets. Re-usable: The data reformatted to CMOR standard can be shared with users of a given HPC facility. The CMORizer scripts, recipes and diagnostics can be re-used or extended by users of ESMValTool, and they are openly available via the ESMValTool GitHub repository where all developments are discussed and conducted publicly.

III. Challenges and Gaps

The biggest challenge faced in this project was to access cloud data during run-time of ESMValTool instead of reading data from a local disk and to maintain the cloud connection during execution of an external ML package. The ESMValTool preprocessor is using the Iris library for data access and manipulation, which currently is not able to open Zarr datasets. The Iris community is aware of the desire to access Zarr datasets, but a consensus on how to achieve it has not been reached yet. One option currently under consideration to provide this missing functionality is full compatibility between xarray datasets and Iris[16]. However, the discussion is still ongoing and the implementation of that solution is beyond the scope of this pilot project.

As described in section 2a, a work-around was implemented that allows xarray to convert individual variables to Iris while still using the cloud. In principle, ESMValTool could be modified to use the same work-around to access Zarr datasets without native support on Iris. However, in practice, this would require too many significant changes to the ESMValTool infrastructure for such a temporary solution until Iris provides support for cloud data.

The current implementation as a CMORizer script was considered a good way forward: the encapsulated nature of the scripts allowed to freely explore the feasibility and requirements of this work with the possibility to include the code in a release of ESMValTool, while the functionality is similar enough for the purpose of the pilot. In case the capability of processing Zarr datasets with Iris directly will not available in the near future, the example implemented in this pilot could be used as an interim solution to add support for datasets stored on the cloud.

[16] https://github.com/SciTools/iris/issues/4994

The implementation of coupling ESMValTool with the Tigramite package also suffered from technical issues and license incompatibilities. While it could be possible to call Tigramite's functions from ESMValTool in the diagnostics script, Tigramite uses its own data format that was not compatible with the one used in ESMValTool, losing functionality provided by Iris such as lazy data. Furthermore, the distribution license used by Tigramite was found to be not compatible with ESMValTool, so the diagnostic could not be included in the ESMValTool codebase. These issues are sidestepped by writing the necessary data for Tigramite to disk.

IV. Relevance for the community and NFDI4Earth

This pilot tackles the issue of working with large data volumes that do not fit into the memory available in a single compute node, a common challenge faced by scientists working with Earth system data. Different aspects of this data challenge are addressed in this pilot: data access and retrieval, data pre-processing, and application of an ML-based algorithm. The implemented approach provides an end-to-end solution from cloud storage to scientific analysis. It relies on well-established tools in the Earth system science community, the ESDC infrastructure and ESMValTool.

Cloud storage is likely to play an important role in Earth system science and beyond in the coming years. A large collection of CMIP6 data is already available as Zarr files on cloud, and made available to the community, for instance via Amazon s3[17] and Google cloud[18]. It is of great relevance to improve support for cloud-based data in Earth systems science tools. This pilot project focused on data from the ESDC infrastructure, but the approach could easily be extended to any observational product stored on the cloud. Scientists working with cloud-based Earth observation data can use the approach proposed in this pilot to subset cloud data, reformat the data to CMOR standard and apply powerful memory-efficient pre-processing functions from ESMValTool to prepare their data for model evaluation. Reduced volume of analysis-ready data can then be downloaded and fed to ML-based algorithms or other types of diagnostic scripts developed by scientists.

The application of ML-based methods on large volumes of climate data can be computationally challenging and requires developments. In this pilot, we have demonstrated how to couple a package for ML-based analysis without the need to reimplement the package inside ESMValTool by using a diagnostic script as a wrapper to call the routines from the external package. This method could easily be used to couple other ML-based packages to ESMValTool which would open up the adoption of such packages by the Earth system science community.

The workflow developed in this pilot, from cloud data access to application of ML techniques provides a very useful stepping stone for future projects dealing with large input datasets and innovative analysis methods. It is central for research done in the Earth system science and climate informatics communities and highly relevant for data and HPC infrastructure providers. It serves as a proof-of-concept for potential applications of data cube in other sub-branches in the Earth system science and usage of cloud-based data for climate model evaluation with ESMValTool.

[17] https://registry.opendata.aws/wrf-cmip6/

[18] https://storage.googleapis.com/cmip6/

The development and achievements of this project were announced internally as part of ESMValTool monthly meetings, quickly getting interest from other ESMValTool and Iris developers. Furthermore, the capability of accessing and processing the ESDC dataset was also mentioned in the changelog of the latest ESMValTool release[19], which was advertised by the ESMValTool User Engagement Team via the ESMValTool website[20], Twitter, user mailing-list, and ESMValTool GitHub[21].

V. Future Directions

The ESMValTool community is very interested in having full cloud support available. Access to data currently relies on either shared data pools on HPC centers or on the user downloading and preprocessing their data locally. Cloud support would significantly lower the barrier of entry to new ESMValTool users by simplifying the configuration process and the storage required to use ESMValTool. This would especially benefit users without access to one of the supported HPC centers and users that are considering to use ESMValTool for their projects, who could get started quickly and focus more on their research and less on technical matters.

To be able to fully achieve this, it is necessary to add support for Zarr data to the ESMValTool preprocessor which is currently not possible with the Iris library. As mentioned above, support for Zarr is also a goal currently discussed within Iris community[22]. Once this has been achieved, the next step would be to change how the data are requested by ESMValTool, and how the tool finds that data supporting both, cloud and local sources. This can be done in multiple ways depending on how Zarr support will be implemented in Iris, so the ESMValTool community will continue to collaborate closely with the Iris developers so solutions are also well-suited for ESMValTool. Once those two issues are solved, a possible next step could be to include cloud data as an integral part of the introduction tutorials to ESMValTool.

In this pilot we have shown how ESMValTool can be used as a tool to access and process climate data to be used in packages for ML-based analyses. The data retrieval and pre-processing steps can be handled by ESMValTool directly. The coupling of external packages to ESMValTool can be achieved by a wrapper diagnostic script calling the external package from within ESMValTool. Future developments of ESMValTool would include widening of the range of scientific applications covered by the diagnostics and external analysis packages to include more scientific use cases. Preparing ESMValTool for applications to large input datasets (e.g. high-resolution models) would also require making the whole set of preprocessor functions lazy to optimize the memory footprint and further improve the performance. Some resources for starting this work have already been allocated within the German national modeling strategy project (natESM). Furthermore, more compatibility between data structures from external packages and ESMValTool is necessary to maximize the resources and avoid writing data to disk. As external packages cannot be expected to change, more compatibility between generally used data libraries is needed. The plans of full

[19] https://docs.esmvaltool.org/en/latest/changelog.html#v2-8-0

[20] https://esmvaltool.org/2023-03-29-New-release/

[21] https://github.com/ESMValGroup/Community/discussions/91

[22] https://github.com/SciTools/iris/issues/4994#issuecomment-1403705185

compatibility between xarray and Iris currently discussed could be a possible step forward to address this issue.

Future support of the implemented solution is maintained by automated testing of ESMValTool and usage of continuous integration services. The ESMValTool development team takes great care that all contents distributed within the package, including the community-provided recipes, are still working with every new release[23]. A description of the work done within this pilot has been added to the ESMValTool documentation. Instructions on how to use CMORizer scripts and recipes are also available, including an online tutorial updated after every ESMValTool release.

[23] https://docs.esmvaltool.org/en/latest/community/release_strategy/release_strategy.html#release-schedule

VI. Works Cited

Eyring, Veronika, et al. “Earth System Model Evaluation Tool (ESMValTool) v2. 0--an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP.” *Geoscientific Model Development*, vol. 13, no. 7, 2020, 3383--3438. https://gmd.copernicus.org/articles/13/3383/2020/.

Eyring, Veronika, et al. “Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization.” *Geoscientific Model Development*, vol. 9, no. 5, 2016, 1937--1958. *doi:10.5194/gmd- 9-1937-2016*.

Galytska, Evgenia, et al. “Causal model evaluation of Arctic-midlatitude teleconnections in CMIP6.” *ESS Open Archive*, October 06, 2022. *doi: 10.1002/essoar.10512569.1*.

Hoyer, Stephan, and Joe Hamman. “xarray: N-D labeled arrays and datasets in Python.” *Journal of Open Research Software*, vol. 5, no. 1, 2017. *doi:10.5334/jors.148*

Lauer, Axel, et al. “Earth System Model Evaluation Tool (ESMValTool) v2. 0--diagnostics for emergent constraints and future projections from Earth system models in CMIP.” *Geoscientific Model Development*, vol. 13, no. 9, 2020, 4205--4228. https://doi.org/10.5194/gmd-2020-60.

Mahecha, Miguel D, et al. “Earth system data cubes unravel global multivariate dynamics.” *Earth System Dynamics*, vol. 11, no. 1, 2020, 201--234.

Met Office. *Iris: A powerful, format-agnostic, and community-driven Python package for analysing and visualising Earth science data*. v3.4 ed., 2010 - 2022, doi:10.5281/zenodo.7386117, Exeter, Devon, http://scitools.org.uk/.

Reichstein, Markus, et al. “Deep learning and process understanding for data-driven Earth system science.” *Nature*, vol. 566, no. 7743, 2019, 195--204.

Righi, Mattia, et al. “Earth System model evaluation tool (ESMValTool) v2. 0--technical overview.” *Geoscientific Model Development*, vol. 13, no. 3, 2020, 1179--1199.

Runge, Jakob, et al. “Detecting and quantifying causal associations in large nonlinear time series datasets.” *Science advances*, vol. 5, no. 11, 2019b, p. eaau4996.

Runge, Jakob, et al. “Inferring causation from time series in Earth system sciences.” *Nature communications*, vol. 10, no. 1, 2019a, p. 2553.

Weigel, Katja, et al. “Earth System Model Evaluation Tool (ESMValTool) v2. 0--diagnostics for extreme events, regional and impact evaluation, and analysis of Earth system models in CMIP.” *Geoscientific Model Development*, vol. 14, no. 6, 2021, 3159--3184. *doi: 10.5194/gmd-14-3159-2021.*