Enhancing Earth system model evaluation with data cube enabled machine learning

Veronika Eyring[1,2], Björn Brötz[1], Fernando Iglesias-Suarez[1], Axel Lauer[1], Miguel D. Mahecha[3], Markus Reichstein[4,5], and Jakob Runge[6]

[1]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen
[2]University of Bremen, Institute of Environmental Physics (IUP), Bremen
[3]Remote Sensing Centre for Earth System Research, Universität Leipzig, Leipzig
[4]Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena
[5]Michael-Stifel-Center Jena for Data-driven and Simulation Science, Jena
[6]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Datenwissenschaften, Jena

Date of submission: 30 July 2020

Abstract

The application of machine learning methods to scientific questions in Earth system sciences is opening new ways to explore and better understand Earth system processes such as dynamical interdependencies between modes of climate variability or extreme events. One major technical challenge of machine learning applications in Earth system science is to efficiently handle the required large volume of input data. The objective of the pilot is to enable the use of cloud ready data and machine learning methods (here we focus on causal discovery) for routine Earth system model evaluation. This will be done by enhancing the well-established Earth System Model Evaluation Tool (ESMValTool) with the novel Earth System Data Lab's (ESDL) "data cube" concept and infrastructure. This interface will also enable the efficient integration of machine learning algorithms that is currently unfeasible due to memory limitations. The result of the pilot will be the release of an enhanced ESMValTool version that includes the data cube concept and innovative machine learning diagnostics. The stakeholders of this pilot are (1) the Earth system science community including groups participating in the Coupled Model Intercomparison Project (CMIP) and ESMValTool development, (2) the climate informatics community, and (3) technology and infrastructure groups such as HPC centers and Earth science data providers. This pilot will contribute to improve efficient handling of Earth system data which is of key importance for application of machine learning methods by the NFDI4Earth community.

I. Introduction

The application of machine learning (ML) techniques offers great potential to overcome some of the existing limitations in Earth system sciences, but also raises substantial new challenges (Reichstein et al., 2019). One major technical challenge of ML-based applications in Earth system sciences is to efficiently handle the required large volume of input data. Two tools already exist that facilitate the handling of Earth system model and observational data, yet without the implementation of ML techniques so far.

The Earth System Model Evaluation Tool (ESMValTool, https://www.esmvaltool.org/) is a diagnostics and performance metrics tool developed by an international consortium with more than 80 institutions. The goal of this internationally well-recognized effort is to improve comprehensive and routine evaluation of Earth system models (ESMs) participating in the World Climate Research Programme (WCRP) Coupled Model Intercomparison Project (CMIP, Eyring et al., 2016) with observations. It has undergone rapid development since the first release in 2016 and is now a well-tested tool that provides end-to-end provenance tracking to ensure reproducibility of the results. It consists of an easy-to-install, well documented Python package providing the core functionalities that performs common pre-processing operations (Righi et al. 2020) and a diagnostic part 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 impacts (Weigel et al., 2020).

The Earth System Data Lab (ESDL, earthsystemdatalab.net; Mahecha et al. 2020) is a concept developed to efficiently apply user defined functions on arbitrary dimensions of an Earth system data cube (space, time, variable, model ensemble member etc.). The idea is that such user defined functions talk efficiently to analysis ready data cubes (ARDCs), which are provided as cloud-ready nd-arrays. In order to enable researchers to efficiently work with a plethora of data streams, a wide range of climate and in particular land-surface parameters are typically curated as ARDCs and the research can focus on the data-analytics only. Novel developments shall allow ingesting arbitrary gridded data to an existing instance of the ESDL and thereby guarantee interoperability amongst data streams.

An interface of these two tools will improve efficient handling of Earth system data with ML techniques, which is of key importance for application of ML methods by the NFDI4Earth community.

II. Pilot description

In this pilot an interface will be created between the ESMValTool and the ESDL data cube concept and infrastructure to enhance ESM evaluation and analysis (Eyring et al., 2019) with ML. This integration of existing ML tools, focusing in this pilot project on causal discovery (Runge et al, 2019a), will allow addressing data intensive scientific problems such as identifying dynamical interdependencies between the major modes of climate variability or extreme events. In this pilot we will exemplary integrate the PCMCI causal discovery algorithm (Runge et al. 2019b). The PCMCI algorithm has already been successfully applied for the evaluation of ESMs (Nowack et al., 2020) and has now reached a level of maturity that makes it suitable for further use by the wider community and for different applications. This will also facilitate future work on linking causal discovery with deep learning that can potentially optimize for example predictive models by revealing “true” linkages (Javed et al., 2020).

The causal discovery algorithms are applied to spatio-temporal features in multivariate Earth system data. The data-challenge in this situation is that the input data is far larger than the available memory of any machine and requires efficient out-of-core computations. As a work-around, a number of tedious pre-processing/transformation steps such as reformatting and reshaping, are required to make the input data available to the ML algorithms. Relying on ARDCs provided through the ESDL concept and infrastructure is one promising approach to

address this challenge. A significant step towards such end would be applying data-parallel training methods where different nodes train the same model on different subsets of the data. Due to the I/O bound situation the approach of zarrdata arrays is to “trade compute for I/O” by loading heavily compressed data chunks into memory and unpacking the data there. Once the ESMValTool has been extended to allow for operations on a data cube, a new set of ML-based diagnostics would become possible, which are currently out of reach due to technical limitations. The ESMValTool as a well-established instrument for Earth system model evaluation could then be used as a benchmark and proof-of-concept for the data cube concept.

The technological backbone of this pilot consists of the software components ESMValTool and the implementation of the data cube concept in Julia, ESDL.jl, the infrastructure components of HPC file systems (computation and parallelization), as well as cloud storage of ARDC provided by the ESDL. The pilot also takes advantage of already available data from Earth system models and Earth observational data hosted at the DKRZ. The pilot makes use of all standards that are used in the current ESMValTool and ESDL implementation such as the W3C provenance standard. The development will be carried out as an open source project in a publicly available development environment, following state of the art software engineering standards such as code quality control and automated testing. The ESDL package is based on zarrdata arrays and offers flexible selection of storage backends. This allows for interoperability between the ESMValTool and multiple cloud storages, which is currently not possible with the ESMValTool. Exemplary results of the analyses carried out with the ESMValTool will be provided as data cubes to an object store (e.g. at DKRZ or Amazon S3). The proposed way of this pilot to address the data challenge described above is to integrate the concept of the “data cube” into the ESMValTool to allow for new ML diagnostics that would otherwise not be possible because of memory limitations. In turn, this also enables the ESMValTool to act in general on data cube cloud storages. The innovation compared to the status quo is therefore to make ML-based applications feasible from the well-established ESMValTool and to provide cloud-ready output.

III. Relevance for the NFDI4Earth

What are expected users and stakeholders and how do they benefit? The nature of Earth system model evaluation is interdisciplinary as it brings together the climate modelling and observational community that will both benefit from the improved accessibility through the newly developed interfaces. The pilot is also central for activities that use ML techniques in climate informatics and the software and infrastructure that will be delivered are in the focus of high performance computing centers (HPC) and infrastructure providers. It will be relevant for scientists and engineers working in those communities and therefore for NFDI4Earth. Given that the developments in this pilot will contribute to the open source software ESMValTool the adaption of the proposed enhancements by the Earth system model evaluation community is very likely. Earth system data providers who offer their data products in cloud environments rely on the ability of their data users to handle cloud based Earth system data with their tools. So the contribution of this project will be relevant for a diverse set of communities in NFDI4Earth.

What is the potential for other sub-branches in the Earth System Sciences? The data cube approach addresses a core problem of all Earth system scientists working on large data

volumes that are difficult to fit into the memory of a computing element (GPU or CPU). The integration of the data cube concept into the ESMValTool can serve as a showcase and proof-of-concept for potential applications of the data cube in other sub-branches of Earth system sciences. The application of the ESMValTool within the context of ML-diagnostics to data available from the ESDL cloud store will contribute to enhancing the data quality and availability at the ESDL that will then be used by scientists from other branches of Earth system science.

What elements of FAIR are particularly addressed? *Findable, Accessible:* All products will be released as open access datasets and software. This will be a vital element of scientific assessment for Earth system data analysis beyond this pilot as the volume of the data is becoming too prohibitive to perform systematic analysis without having software in place that can cope with it. *Interoperable:* the integration of the data cube based on zarrdata arrays allows access to a multitude of different storage backends which makes the ESMValTool with data cube enhancement interoperable with different cloud storages. Furthermore, exemplary ESMValTool output will be also saved as data cubes for cloud usage. *Re-usable*: the ESMValTool cloud-ready data can be re-used by other stakeholders.

What aspects of the research data life cycle are particularly addressed? The developments in this pilot will address in particular the easy access and re-usability of Earth system data in its mature stage of its life cycle as cloud-ready data.

Are there particular contributions that help the NFDI4Earth to engage with? The pilot will contribute the infrastructure for a model evaluation approach with innovative ML methods included accessible through the NFDI4Earth one-stop shop. Also the enhancement of the existing platforms ESDL and ESMValTool can be used to further engage the NFDI4Earth community.

IV. Deliverables

Technical operability of the pilot:

- **Software and developed interface:** ESMValTool-ESDL-interface as part of a new open-source software release that includes ML-diagnostics (here causal discovery) [Month 12]

Roadmap document for the community:

- “Enhancing Earth system model evaluation with data cube enabled machine learning” [Month 12]

V. Work Plan & Requested funding

- Milestone plan for the planned one-year implementation phase:
  - **Milestone 1**: connection of the ESMValTool to the data stream of the ESDL cloud store [Month 6]
  - **Milestone 2:** application of ML-based diagnostics to the data stream within the framework of the ESMValTool [Month 12]
  - **Milestone 3:** exemplary ESMValTool output as data cube for cloud usage [Month12]
- Funding request: 1 FTE for this project.

References

Eyring, V. et al., Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937-1958, doi:10.5194/gmd-9-1937-2016, 2016.

Eyring, V. et al., Taking climate model evaluation to the next level. Nature Climate Change, 9(2): 102-110, 2019.

Eyring, V. 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, Geosci. Model Dev., 13, 3383–3438, https://doi.org/10.5194/gmd-13-3383-2020, 2020.

Javed, K. et al., Learning Causal Models Online, 181–184, http://arxiv.org/abs/2006.07461, 2020.

Lauer, A. et al., Earth System Model Evaluation Tool (ESMValTool) v2.0 – diagnostics for emergent constraints and future projections from Earth system models in CMIP, Geosci. Model Dev., https://doi.org/10.5194/gmd-2020-60, in press, 2020.

Mahecha, M.D. et al., Earth system data cubes unravel global multivariate dynamics. Earth System Dynamics, 11, 201–234, doi:10.5194/esd-11-201-2020, 2020.

Nowack, P., Runge, J., Eyring, V., and Haigh, J.D., Causal networks for climate model evaluation and constrained projections. Nat Commun 11, 1415, https://doi.org/10.1038/s41467-020-15195-y, 2020.

Reichstein, M. et al., Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743): 195-204, 2019.

Righi, M. et al., Earth System Model Evaluation Tool (ESMValTool) v2.0 – technical overview, Geosci. Model Dev., 13, 1179-1199, 2020.

Runge, J. et al., Inferring causation from time series in Earth system sciences. Nature Communications, 10(1), 1–13, 2019a.

Runge, J. et al., Detecting and quantifying causal associations in large nonlinear time series datasets. Sci Adv, 5(11): eaau4996, 2019b.

Weigel, K. 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, Geosci. Model Dev., submitted, 2020.