NFDI4Earth

Statistical learning to assess the factors underlying environmental changes

Alexander J. Winkler[1,2], Fabian Gans[2], Miguel Mahecha[3], Ranga B. Myneni[4], Christian Kadow[5], und Markus Reichstein[2]

1 Max-Planck-Institute for Meteorology, Bundesstrasse 53, 20146 Hamburg
2 Max-Planck-Institute for Biogeochemistry, 07745 Jena
3 Remote Sensing Centre for Earth System Research, Leipzig University, 04103 Leipzig
4 Department of Earth and Environment, Boston University, Boston MA 02215, USA
5 Deutsches Klimarechenzentrum GmbH, Bundesstraße 45a, 20146 Hamburg

July 30, 2020

Abstract

The Earth system is currently undergoing profound environmental changes including but not restricted to climate, biogeochemical flows, and biodiversity. A growing body of multivariate Earth observations can now diagnose these changes. Assessing the drivers underlying these changes is a challenging task, both technically and scientifically. The objective of this pilot project is to introduce a toolkit that facilitates statistical driver attribution by combining extensive resources of spatio-temporal data (Earth System Data Lab, ESDL) and the capabilities of new tools in the Big Data geosciences (Pangeo Project and statistical learning libraries). This toolkit is designed to be universally adaptable to problems of driver attribution in Earth system sciences. To showcase the functionality of the proposed tools, it will be applied to a key question in current climate research: What are the regional drivers of the observed changes in the Earth's ecosystems and their respective roles in driving the terrestrial sink of anthropogenic carbon? The expected result of this pilot project is an integrated workflow towards the identification and regional quantification of the factors underlying observed changes in the Earth system. This knowledge will be needed above all by the modelling community to assess the plausibility of current Earth system projections; the developed workflows can be likewise applied as diagnostics for model evaluation. The envisaged toolkit will interlink the ESDL to modern statistical learning libraries and will be integrated into the Pangeo ecosystem for open-source analysis software for Earth system sciences, guaranteeing interoperability of the NFDI with international developments in the field of analyzing big n-dimensional data arrays in Earth system analysis.

I. Introduction

Multi-decadal records of field measurements as well as satellite data reveal widespread and persistent changes in the Earth system. Finding and disentangling the drivers behind the observed changes is difficult for two main reasons: First, the internal variability of the Earth system obscures the evidence of externally forced changes, and second, the observed change could be a mixed response to several external coinciding forcings (Winkler *et al.*, 2020).

A prominent example of such a driver attribution problem are the origins of terrestrial ecosystem changes across the globe. The key question in this context is whether these changes are the result of the influence of climate modes (i.e. internal variability) or of external forcings, i.e. man-made climate changes or the physiological effects of increasing atmospheric CO_2 (water use efficiency and CO_2 fertilization effect). The conventional approach to such detection and attribution problems in the Earth system sciences is the technique of Optimal Fingerprinting (Hasselmann*,* 1993), which requires a set of idealized model simulations. Various studies have shown the weaknesses and limitations of this method based on model simulations (e.g. Zhu *et al.*, 2013). In this project, we propose to make use of the comprehensive collection of Earth observation data and develop a toolkit based on statistical learning and dynamical adjustment, a recently introduced technique in the detection and attribution discipline.

The planned toolkit will combine the power of the Earth System Data Laboratory (ESDL) with that of modern libraries for statistical learning (SL, e.g. PyTorch). In doing so, it will adopt the concepts of the ESDL and the Pangeo tools, the international collaborative platform for Big Data geosciences, in such a way that it is fully compatible, but also stands independently as an ongoing development within the NFDI4Earth. This will make the tools easily accessible to the international and multi-disciplinary scientific community and detection and attribution studies can be realized rather quickly. Furthermore, the integration of state-of-the-art SL libraries into the ESDL facilitates and serves as a basis for the development of further SL-based tools within Earth system sciences.

II. Pilot description

The technological backbone of the proposed project are open-source statistical-learning frameworks, the Pangeo project and the ESDL approaches to handle Big Data and enable out-of-core computations. The ESDL (https://www.earthsystemdatalab.net/) is a concept that integrates multivariate analysis-ready data cubes (ARDCs) on common spatio-temporal resolutions and thereby enables cloud-based computations (Mahecha *et al.*, 2020). The rich collection of datasets are represented in so-called "data cubes", a scalable format in order to meet the current and future challenges of big data. In that, ESDL partly relies on tools provided and developed in the context of the Pangeo project (https://pangeo.io/), such as Xarray, a toolkit for working with labeled multi-dimensional arrays of data, or Zarr, a Python package providing an implementation of compressed, chunked, N-dimensional arrays, designed for use in parallel computing.

Pangeo aims to cultivate a Python and Julia-based ecosystem in which the next generation of open-source analysis tools for Earth system sciences can be developed, distributed, and sustained. This pilot project proposes to harness the power of both, ESDL's analysis-ready cloud-based interface, its rich collection of Earth observations alongside Pangeo's platform and tools to handle Big Data, by making use of statistical learning frameworks. There are two Python-based and open-source machine-learning libraries that together meet the requirements of this project, scikit-learn (https://scikit-learn.org/stable/) for general statistic learning purposes and PyTorch (https://pytorch.org/) for high-level deep neural networks.

The proposed "detection and attribution" toolkit will follow a two-stage approach. These two stages will be showcased on the basis of the problem outlined in the introduction: the origins of terrestrial ecosystem changes across the globe. The first stage evaluates whether the externally forced component in the response of the target variable (e.g. ecosystem productivity) can be isolated in the data at local, regional, and global scales (detection). In the second stage, individual external forcing agents (e.g. man-made climate change) are evaluated and their respective contribution to the response of the target variable estimated (attribution), again across spatial scales. Dependent on the strength of the signals, various statistical learning methods of varying complexity will be tested (see the flowchart on last page).

The development of the toolkit will rely heavily on the recently introduced technique of dynamical adjustment for driver attribution combined with statistical learning (Sippel *et al.*, 2019). Its fundamental idea is to derive an estimate of, for instance, climate-change-induced contribution to variability in a target variable, e.g. ecosystem productivity, using statistical-learning techniques (Sippel *et al.*, 2019). This estimate is then used to derive a residual time series, that is, the component of the target variable's variability that cannot be explained by climatic changes (Sippel *et al.*, 2019). The residual could contain the variability or trend explained by an external forcing, e.g. CO_2 fertilization. With this method and the comprehensive data availability in space and time, this case study aims to disentangle and quantify the drivers of the ecosystems changes solely using observations.

III. Relevance for the NFDI4Earth

With this pilot the Earth System Data Laboratory, which will be part of the TA 2Facilitate, will be tested and amended with further machine/statistical learning related functionality. While the pilot has a specific use case, driver attribution will play an important role also in many other geoscientific disciplines, such as water or solid earth research. It goes without saying that the machine/statistical learning integration into the ESDL even offers broader innovative applications, and can make the methods available to a wide range of scientists. The ESDL contains various data streams of several subsystems of the Earth, and there are no conceptual limits for adding data sets nor for the spatial scale. Further, ESDL's versatile computing interface is accompanied by a concept for effective mapping of functions in all dimensions of the contained ARDCs – this allows our framework to be seamlessly extended to other domains as well as to regional and local geoscientific research foci. In particular, this project addresses the

interoperability element of FAIR, thus facilitates the data to interoperate with applications or workflows for analysis, storage, and processing. The pilot in particular will make sure that the developments around the data cubes in the NFDI4Earth will align well with international initiatives on working with large gridded data. The primary goal here is to contribute to research tools, but given the open interface, it can easily to be interfaced with GUI and jupyter notebooks, e.g. for practitioners and undergraduate teaching, as an additional branch in NFDI.

IV. Deliverables

The deliverables of the pilot will include:

1) Toolkit interlinking the ESDL with modern statistical-learning libraries and the Pangeo ecosystem for open-source analysis software for Earth system sciences
2) Python module for domain-independent detection and attribution analysis based on statistical learning
3) Tutorial on both the detection/attribution analysis and the (more generic) use of machine/statistical learning in the ESDL-Pangeo system
4) A roadmap document: Integration of machine-learning and modelling with harmonized Earth system data. Building upon this pilot the roadmap will review best practice for using machine/statistical learning with Earth system data, and outline the link to physical modelling (model benchmarking, model development, hybrid ML-physics modelling) in the context of the ESDL, towards maximally data informed and theory-guided geoscience.

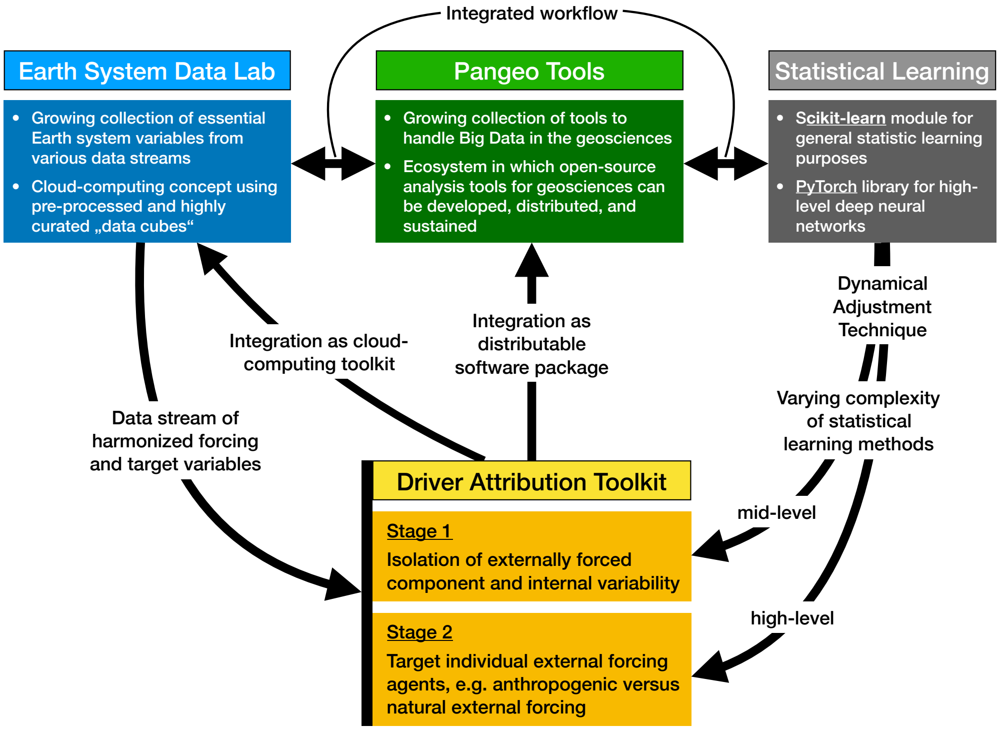
V. Work Plan & Requested funding

The work plan consist of the following four parts, the achievement of which also constitute the key milestones.

1. Integrating workflow between ESDL, Python-based statistical-learning libraries and Pangeo tools to handle Big Data
2. Developing statistical-learning framework for driver attribution based on dynamical adjustment for various levels of complexity
3. Case Study: Data-driven assessment of drivers underlying ecosystems changes
4. Testing, compiling recipes, documentation and tutorials, publishing software package

	Month											
Work plan items	1	2	3	4	5	6	7	8	9	10	11	12
Workflow integration	🟩	🟩	🟩	🟩								
Statistical learning framework				🟩	🟩	🟩	🟩					
Case study								🟩	🟩	🟩		
Testing, documentation, tutorial								🟩	🟩	🟩	🟩	🟩

The timings of the activities are summarized in the simplified Gantt chart above. The case study will serve as an important real world testcase and will help to develop comprehensive documentation/tutorial material. Thus it is in parallel for three months. All in all, funding for a one-year PostDoc full-time equivalent is requested.


Integrated workflow
Earth System Data Lab
Growing collection of essential Earth system variables from various data streams
Cloud-computing concept using pre-processed and highly curated „data cubes“
Pangeo Tools
Growing collection of tools to handle Big Data in the geosciences
Ecosystem in which open-source analysis tools for geosciences can be developed, distributed, and sustained
Statistical Learning
Scikit-learn module for general statistic learning purposes
PyTorch library for high-level deep neural networks
Dynamical Adjustment Technique
Varying complexity of statistical learning methods
Integration as cloud-computing toolkit
Integration as distributable software package
Data stream of harmonized forcing and target variables
Driver Attribution Toolkit
Stage 1
Isolation of externally forced component and internal variability
Stage 2
Target individual external forcing agents, e.g. anthropogenic versus natural external forcing
mid-level
high-level

References

Mahecha, M.D., Gans, F., …, Reichstein, M., 2020. Earth system data cubes unravel global multivariate dynamics. Earth System Dynamics 11, 201–234.

Hasselmann, K., 1993. Optimal Fingerprints for the Detection of Time-dependent Climate Change. J. Climate 6, 1957–1971.

Sippel, S., Meinshausen, N., Merrifield, A., Lehner, F., Pendergrass, A.G., Fischer, E., Knutti, R., 2019. Uncovering the Forced Climate Response from a Single Ensemble Member Using Statistical Learning. J. Climate 32, 5677–5699.

Winkler, A.J., Menyni, R., …, Brovkin, V., 2020. Slow-down of the greening trend in natural vegetation with further rise in atmospheric CO2. Earth and Space Science Open Archive. https://doi.org/10.1002/essoar.10503202.1

Zhu, Z., Piao, S., Myneni, R.B., …, Zeng, N., 2016. Greening of the Earth and its drivers. Nature Clim. Change 6, 791–795