NFDI4Earth

Propagating complex uncertainties in data cubes

Nils Weitzel[1], Kira Rehfeld[1,2]

[1]Department of Geosciences, University of Tübingen, Tübingen, Germany

[2]Department of Physics, University of Tübingen, Tübingen, Germany

Date of submission: 31.03.2023

Abstract

Complex uncertainty structures, which are analytically intractable, occur across the Earth System Sciences when large datasets from numerical simulations and measurements are compared, for example in climatology, hydrology, and remote sensing. In paleoclimatology, proxy system models map Earth System Model (ESM) output onto proxy measurements through a process chain with multiple sources of autocorrelated and non-Gaussian uncertainties. In this pilot, we aim at integrating in-situ measurements and methods for propagating uncertainties with Monte Carlo techniques into a data cube architecture using paleoclimate proxy data as example. This will expand the availability of computationally efficient operations on data cubes from processing ESM output to model-proxy comparison. We will develop a software package and demonstrate the methods by quantifying the consistency of simulated forest cover changes over the last 25,000 years with a global database of pollen records. Our pilot will allow the use of more sophisticated frameworks for inferring past environmental changes from ESMs and global proxy databases. By employing consistent and efficient methods for all processing steps in model-proxy comparison, the software package will improve interoperability and reusability of data analysis workflows in paleoclimatology. It can be further extended to other applications in Earth System Sciences where measurement operators are subject to complex uncertainties.

I. Introduction

Comparing numerical simulations with measurements often involves operators that map model output to the measurement space. In many Earth System Science disciplines such as climatology (Kennedy et al. 2019), hydrology (McMillan et al. 2018), and remote sensing (Merchant et al. 2017), operators need to propagate complex uncertainty structures, which are analytically intractable. In paleoclimatology, so-called proxy system models (PSMs) map Earth System Model (ESM) output onto proxy measurements from natural climate archives such as ice and sediment cores. PSMs consist of process chains with multiple sources of autocorrelated and non-Gaussian uncertainties (Dolman and Laepple 2018).

Over time, the complexity and simulation periods of ESMs have increased strongly, which creates enormous amounts of data available for analysis. This led to the development of efficient processing methods on data cubes for gridded ESM output (e.g. ESMValTool[1], MetPy[2], xCDAT[3]). Meanwhile, proxy data is stored in tabular or database formats, with larger datasets of proxy records being compiled in recent years (e.g. Comas-Bru et al. 2019). A suite of processing tools for proxy data exists, but they rarely perform efficient computations on large datasets. This makes it currently not feasible to employ computationally efficient workflows, which leverage the progress in ESMs, global proxy databases, and PSMs, for climate process understanding and model validation. To overcome this issue, efficient processing of ESM output needs to be combined with Monte Carlo methods for uncertainty propagation, and in-situ measurements such as paleoclimate proxies need to be incorporated into climate data cubes. This pilot addresses these challenges. It will enable tackling new research questions in paleoclimatology such as evaluating the ability of ESMs to simulate climate variability patterns across spatial and temporal scales.

The pilot contributes to more efficient, interoperable, and reusable model-proxy comparison workflows that combine all mappings between ESM output and proxy measurements. A modular approach will facilitate the integration of many existing PSM functions. Furthermore, the concepts of this pilot can be applied to measurement operators with complex uncertainty structures in other Earth System Science disciplines. In the long-term, results from our pilot can help to include measurement operators for in-situ measurements in user-friendly multi-language platforms like Earth System Data Lab (Mahecha et al. 2020).

II. Pilot description

The pilot integrates proxy data and efficient Monte Carlo based uncertainty propagation into the climate data cube concept. This will facilitate creating efficient workflows for large and heterogeneous datasets with complex uncertainty structures. The main application area are comparisons between ESM output and proxy measurements. We will use data cubes, which are a data representation designed to efficiently perform operations that require only small chunks of large datasets (Mahecha et al. 2020). They are already separately employed for processing ESM output and Monte Carlo based inference. Data cubes are suitable for paleoclimate applications because operators rarely involve more than two dimensions of a data cube at once. This permits lazy loading and efficient parallelization. We will demonstrate the applicability of data cube workflows in paleoclimatology by evaluating simulated forest cover changes over the last 25,000 years against a global database of pollen records (Adam et al. 2021). This period includes the transition from the Last Glacial Period into the Holocene, and is therefore highly valuable for understanding large amplitude environmental changes.

The envisaged software package will be written in Python and builds on the data cube implementation xarray (Hoyer and Hamman 2017). We employ Python because it is increasingly used for processing ESM output, and it is the most-used language for paleoclimate data analysis besides R. Xarray is a flexible data cube implementation that is interoperable with other data cube implementations (e.g. iris data cubes[4]) and supports netCDF files for storing large datasets. We plan to build on uncertainty quantification standards from the widely-used Arviz package (Kumar et al. 2019), which also uses xarray data structures. Proxy databases have so far not been integrated into data cube architectures. We will explore different data structures to find a solution that permits efficient processing of all typical steps in PSMs.

The data standards of our software package will assure compatibility and interoperability with Python tools for processing ESM output and proxy data that are suitable for lazy loading and efficient parallelization. Operators will follow data cube concepts and uncertainty propagation will use ARVIZ standards, similar to advanced Bayesian inference methods (e.g. PyMC[5], PyStan[6]).

III. *Relevance for the NFDI4Earth*

The pilot primarily addresses the data processing and analysis aspects of the research data life cycle by enhancing the computational efficiency of paleoclimate model-proxy comparison. Our results will help to reuse research data products by facilitating reproducible and adaptable data workflows. We expect that our results will be mainly used by scientists to explore research questions which require the use of long simulations with comprehensive ESMs, large proxy data compilations, and rigorous uncertainty quantification. Furthermore, scientists from the wider NFDI4Earth community and other Earth System Science disciplines, in which uncertainty quantification for large datasets are important, could integrate our concepts into their workflows. In particular, the pilot is a step towards an interface between validating ESMs for the instrumental (e.g. with ESMValTool) and paleoclimate periods. Our vision also emphasizes the benefit of using file formats in data repositories that are compatible with big data concepts, even if the individual datasets are comparatively small. Finally, our framework will reduce the barriers for students to work with different types of paleoclimate data.

To advertise our pilot, we will present our software package at the EGU24 conference. PI1 (Nils Weitzel) will promote the pilot through his activities as a fellow of the NFDI4Earth Academy. The PIs will advertise the results in national (PalMod, natESM) and international (PMIP, PAGES) projects. Our proposed software solutions will be findable and accessible for users through open access publication and extensive documentation. Using harmonized data structures will make ESM output and proxy data more accessible, in particular to users who are not working with the respective data types regularly. This can foster new collaborations between proxy experts and climate modelers. Interoperability and reusability will be ensured by providing wrapper functions to integrate existing Python and R tools (e.g. pyleoclim[7], BACON[8], sedproxy[9]).

IV. *Deliverables*

We work towards four deliverables. PI1 will present a preliminary version of the software package at the EGU24 conference (D1, month 8). The finalized software package will be published on github (D2, month 10). We will submit a manuscript describing the software package and a prototype application comparing simulated forest cover with pollen data to an open access journal (D3, month 12). Finally, a roadmap containing a project evaluation and an assessment of future prospects will be published (D4, month 12). The intended title is “Towards model-to-measurement data cubes with rigorous uncertainty propagation”.

V. *Work Plan & Requested funding*

The work is separated into three tasks (see Gantt Chart below). Task 1 is carried out by both PIs. Tasks 2 and 3 are carried out mainly by PI1 with support from a student assistant. In task 1, we define standards for incorporating uncertainty propagation methods and proxy measurements in data cubes (Milestone M1). Task 2 develops the software package. First, we implement data structures for ESM output, proxy data, and intermediate products based on xarray objects (Milestone M2, month 5). Then, we add PSM operators with rigorous propagation of uncertainties (Milestone M3, month 9), before publishing the package in month 10. Task 3 implements the prototype application. We request in total EUR 71,163.36, which are split into a 0.75 FTE position for PI1 Nils Weitzel (EUR 60,075 following DFG Personnel Rates for 2023), a student assistant (EUR 9588,36), and presenting the pilot at EGU24 in Vienna (EUR 1,500.00).

Task	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
1) Define standards	🟩	M1										
2) Software package			🟩	🟩	M2	🟩	🟩	D1	M3	D2		
3) Implement prototype									🟩	🟩	🟩	D3,4

References

Adam et al.: Identifying Global-Scale Patterns of Vegetation Change During the Last Deglaciation From Paleoclimate Networks, Paleoceanog and Paleoclimatol, 36, doi: 10.1029/2021PA004265, 2021.

Comas-Bru et al.: SISALv2: a comprehensive speleothem isotope database with multiple age–depth models, Earth Syst. Sci. Data, 12, doi: 10.5194/essd-12-2579-2020, 2020.

Dolman and Laepple: Sedproxy: a forward model for sediment-archived climate proxies, Clim. Past, 14, doi: 10.5194/cp-14-1851-2018, 2018.

Hoyer and Hamman: xarray: N-D labeled Arrays and Datasets in Python, J. Open Source Softw., 5, doi: 10.5334/jors.148, 2017.

Kennedy et al.: An ensemble data set of sea surface temperature change from 1850: The Met Office Hadley Centre HadSST.4.0.0.0 data set, J. Geophys. Res. Atmos., 124, doi: 10.1029/2018JD029867, 2019.

Kumar et al.: ArviZ a unified library for exploratory analysis of Bayesian models in Python. J. Open Source Softw., 4, doi: 10.21105/joss.01143, 2019.

Mahecha et al.: Earth system data cubes unravel global multivariate dynamics, Earth Syst. Dynam., 11, doi: 10.5194/esd-11-201-2020, 2020.

McMillan et al.: Hydrological data uncertainty and its implications, WIREs Water, 5, doi: 10.1002/wat2.1319, 2018.

Merchant et al.: Uncertainty information in climate data records from Earth observation, Earth Syst. Sci. Data, 9, doi: 10.5194/essd-9-511-2017, 2017.

[1] esmvaltool.org, [2] unidata.ucar.edu/software/metpy, [3] github.com/xCDAT, [4] scitools-iris.readthedocs.io, [5] pymc.io, [6] mc-stan.org/users/interfaces/pystan, [7] pypi.org/project/pyleoclim, [8] cran.r-project.org/web/packages/rbacon, [9] github.com/EarthSystemDiagnostics/sedproxy