Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2019-62-AC1, 2019

Earth System Dynamics
Discussions

Interactive comment on "Earth system data cubes unravel global multivariate dynamics" *by* Miguel D. Mahecha et al.

Miguel D. Mahecha et al.

mmahecha@bgc-jena.mpg.de

Received and published: 18 December 2019

The comments of the reviewer are repeated here in bold font; *our answers are given in italics.*

The authors present a new data cube approach, called Earth System Data Cube, where multiple spatiotemporal data streams are treated as one singular, very high-dimensional data stream. This paper is of high quality and clearly outlines the authors' reasoning of implementing the ESDC in the way it is implemented and the advantages it brings to Earth System dynamics studies. The ESDC datacube approach brings different data streams to a common grid, which is beneficial for specific scientific Earth System studies. Common operations, as e.g.

Printer-friendly version

Discussion paper

resampling and bringing the data to the same spatial and temporal resolution, are just taken care of and users can focus more on science. Yet, the unified grid from the beginning might be a limitation for other application areas. *We thank the reviewer for the very positive comments and for sharing her/his concerns w.r.t. the technical approach chosen here. Re-gridding the data to a common format was essentially necessary to start thinking in "cubes" where common axes are indeed needed. Otherwise the formalism, but more importantly, the implementation would have been substantially more complicated, yet not impossible. But we agree that this can be regarded as suboptimal for certain applications. We will, in the revised version, discuss the potential drawback of a pre gridded data set and also differentiate better between the "concept" of the ESDL and the "implementation". In fact, the latter can be extended to deal with data of different grids and then work on the mismatch on the fly.*

Some comments and areas for minor revision:

What is not so clear for me so far is the overall implementation strategy of the ESDC. I understand that I can now use the ESDC with the datasets as outlined in the paper appendix. Is there the option that users can extend the ESDC data list? *Indeed, as described by in the paper there is the possibility to 1) use the ESDL with the current data, 2) add more variables to an existing cube, or 3) use the code implementation with own data.*

Or will it be similar to the OpenDataCube concept where multiple implementations of the ESDC concept can be set up with different data? *As we describe we have already several cubes (see L211 and L215) implemented and more can follow (L298). See also our discussion on future perspectives section 5.3 L540.*

Where is the current ESDC data hosted? *We quote from our paper, page 13; table 2: "The cubes are currently hosted on the Object Storage Service by the Open Telecom Cloud under https://obs.eu-de.otc.t-systems.com/obs-esdc-v2.0.0/ "*

And how long does the data preparation take? *There is no generic answer to this*

question as it depends on the native format of the data that should be ingested. If the data are already in suitable NetCDF, it is essentially a very fast conversation to the currently supported zarr-Format. It also depends on the size and other aspects.

I suggest to bring in these additional aspects. *We thank you for the suggestion and will focus on making the points on "own data" more clear. Given that we have had alreday addressed the aspects, we simply will try to describing them more prominently in the text.*

What are the bottlenecks in gathering the data? I understand that in order to implement the ESDC, data has to be moved from the respective data repositories. Is this correct? Would be good to elaborate on this aspect as well. *Yes, we need a common repository - or at least a common format that serves the data in a joint data model as said in L64. For details on the storage we had written Section 3.3 (L 260).*

It is without doubt that Julia is a powerful and efficient programming language. However, the fact that the ESDC package has been developed in Julia could be a restrictive factor in the uptake, as I argue that most potential users of the ESDC use Python or R and might be less motivated to take up a new programming language. Are there plans to extend it to Python or R? *The reviewer is right that many users would prefer Python and R over Julia. But please note that in fact the ESDL is perfectly working from python. As we write in L. 258 and 279, we have a running Python interface that can be inspected here https://cablab.readthedocs.io/en/latest/. Working with R is more complicated as there is no suitable implementation of the zarr-format available yet. In response to your comment and other requests, one of the co-authors of this paper (Guido Kraemer) is working on an implementation for reading the zarr format in R (https://github.com/gdkrmr/zarr-R), this can serve as a basis for a future implementation of the data cube framework in R. In particular we aim to use data cube concepts as developed e.g. by Edzer Pebesma and Marius Appel (see commentary uploaded by them to this discussion) in the stars framework (https://github.com/r-spatial/stars) to talk with our data-cubes. In response to the comment we will mention the Python*

implementation more prominently in the revised version and also point the users to the R developments in progress.

The authors argue that the ESDL is closest to the Climate Data Store. In my current understanding of both data systems, I would disagree. Do the authors talk about the Climate Data Store, which is primarily a data dissemination system, or the Climate Data Store toolbox, which is the processing editor on top of the CDS? It would be necessary to elaborate more on what aspects both systems are close and why the authors come to this assessment. It is also important that the authors differentiate between the CDS and the CDS toolbox. *Thank you for the advice to be more precise on this topic. We will elaborate this in the revision. Our point is that the CDS indeed does offer also data analytic access via jupyter notebooks that allow the user to map UDFs as we do on an arbitrary set of of data stored there.*

Some additional typing errors discovered: ... *Thank you for spotting these errors - we will work on these in the revisions.*

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2019-62, 2019.