Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2019-62-AC2, 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: 19 December 2019

Comments of the reviewer are pasted here in bold font; *our answers are given in italics.*

In this article Mahecha et al. present the concept of data cubes to handle the growing body of Earth system data and introduce the computing interface "Earth system data lab" (ESDL) as a cloud-based solution. In the introduction the authors describe different data sources and variables of the Earth system, the hurdles of using the data, and illustrate the data cube approach as solution to the problem. Then, they delve into the concept and definition of data cubes, provide a generic description and mathematical formulations including how to apply

customized operations on data cubes. In this context, Mahecha et al. explain the detailed implementation of the data cube approach in the ESDL project and depict its representation and processing of the various data streams. The authors showcase three example studies to demonstrate the functioning and usefulness of the ESDL. The ESDL is novel and unique in its approach to focus on the fusion of global multivariate data streams and thus enables an simultaneous exploration of many facets of the Earth system. Therein, the ESDL is well equipped to face the challenges in the upcoming era of machine learning. Therefore, the ESDL and this descriptive article is an important contribution to the Earth system sciences and possibly to a wider community. The manuscript is well structured and written. However, I have a few minor points of criticism that need consideration, before I can recommend this manuscript for publication. *We thank the reviewer for the precise summary of the ESDL and enthusiasm with respect to the potential of this approach. We will address all points in the remainder of the review in due detail.*

1 General Comments:

1.1 A large part of this article deals with the mathematical formulation and technical implementation of data cubes. However, you completely miss out on the technical description of the processing of the various data streams and how you treat uncertainty in the ESDL. For example, how do you treat the provided uncertainty estimates / quality flags of the individual data products? How does this affect the remapping / resampling algorithms? Is the ESDL capable to take error propagation into account? Could you provide a flow-chart to illustrate the procedure of how you incorporate data streams? *The reviewer is right that this paper has been written with an emphasis on the conceptual aspects of the "data cube idea". Many concepts for data cube have been proposed in the last few years in the geo community and many are still under active development. However, we feel that we are still lacking an overarching vision of how data cubes can empower Earth system sciences.*

Regarding the specific question on the uncertainty: We have to admit that we have had many discussions on how to consider uncertainty and propagate it to the ESDL. We found, however, that each data product comes with its own uncertainty, some e.g. with flags, others with confidence bands, and other entirely without information on the uncertainty. Hence, we couldn't come up with a unified view on an "uncertainty dimension" in the ESDL framework that could give due credit to all of them. The procedure to incorporate a data stream is extensively described in the documentation of the ESLD: `https://cablab.readthedocs.io/en/latest/esdc_prod.html` *- which will be mentioned in the revisions of the paper. Regarding the last point on adding a flowchart we agree that his is a good idea and will add it in a revision version of the paper.*

1.2 The second case study on the intrinsic dimension(s) of land surface variables clearly demonstrates the usefulness of the ESDL and thereby supports the title of the manuscript. Also the first case study corroborates the statement that multivariate dynamics can be better studied with data cubes, less convincingly though. However, I am not convinced that the third study really supports the need for the multivariate approach in ESDL. Here, you basically analyze only two variables (ecosystem respiration and temperature), which one could easily do with any other tool not based on data cubes. Can you make more clear why this case study supports the claim in the title? *Thank you for this important comment. We understand that addressing a two-variable problem is not that convincing. The rationale for this use-case is actually to show that there is no limit to address problems beyond "data exploration" in the ESDL. To put it in other terms: this example was also chosen for its simplicity (although it is not trivial), but can serve as an example for implementing a parameterizing more complex models. We will clarify and discuss this in our revision.*

1.3 The ESDL really lives on the various data streams. Many researchers have their specific datasets which they would want to analyze alongside the data streams provided in the ESDL environment. How do you enable the usage of ex-

ternal datasets? What are the disk usage constraints for each individual user? Is it possible to stream data (e.g. in Zarr format) from external data storage? If incorporating own datasets constitutes a complicated endeavor, researchers might be hesitant to use ESDL. *This concern is one of the most often raised questions in the various user consultation meetings we had so far and was also raised by reviewer 1. To give a brief answer: Yes, it is indeed possible to add "own" data sets any pre-curated data cube. One can, for instance, read any xarray data set - as long as it shares common axis with the existing cube. In the experimental platform that was set up now we have, of course, disk usage constraints, but the paper describes a generic system that is independent of the concrete jupyter hub running right now. We also note that the implementation in Julia and the Python based xarray allow for reading in additional NetCDF files and concatenating them with an existing cube (again - under the assumption of shared axes). We can also read additional cubes into memory.*

2 Specific comments:

2.1 L36: You cannot only analyze the state, but also the change of the system using ESDL, right? *Yes indeed! For an example study in this direction we refer to a paper in discussion by our co-author Guido Kraemer et al. https://www.biogeosciences-discuss.net/bg-2019-307/bg-2019-307.pdf i.e. figures 5 and 6 therein. But we will mention this now also in this manuscript more prominently.*

2.2 L35: You start the paragraph claiming that we are well prepared in terms of data availability, but here you say there are access barriers. Maybe the term "availability" is not accurate here. A huge amount of data are collected, but they are not necessarily available for science. So, you could start this paragraph saying that we are well-prepared in terms of data collection. *Thank you for this sharp observation! We will change this accordingly.*

2.3 L70: I suggest to remove 'we believe ... **' and phrase the sentence: Due to its** .... **interface, the ESDL is well-suited** .... *We agree and will change the text*

accordingly.

2.4 L86: This sentence ("However, ...") is somewhat complicated to grasp. I guess you want to motivate why "variable" should be treated as an additional dimension. Please revise this sentence. *We agree and will change the text accordingly.*

2.5 L90: Here, the subscript of Y denotes the different variables k and the superscript denotes the different domains j - this is not consistent with the definition of Y in LL85-87. *Thank you for spotting this! The first author apologizes for sloppiness to his co-authors as this remark has been stated internally before. We will change the text accordingly.*

2.6 L98: If the dimensions for only one grid point are dropped, do you not lose information? For example, a point measurement with certain lat and lon coordinates and a timestamp cannot be represented without losing the coordinate information, right? *Well in fact it can. We have chosen this notation for the sake of simplicity and because this is how most programs of scientific computing deal with such phenomena. But one can always define the mapping such that the collapsed dimension retains a length of 1. The risk we see is that very long workflows become intractable as there is no "simplification" in the dimensionality. But it is a bit irritating at times We will add a remark in the text to clarify that both ways are thinkbale but that we prefer this approach here.*

2.7 L100: Here you use a math symbol / notation (large `x`) to describe a cartesian product, which I have rarely seen before. Please mention that large `x` refers to cartesian product. *We will add a footnote clarifying the meaning of this symbol.*

2.8 L105: Please define NA - does it refer to "not available", thus missing data? *Yes, will clarify this in the text accordingly. See also response to comment by Apel and Pebesma.*

2.9 L111: In the modelling community data cubes can be used to represent large ensembles, thus I suggest to also list "ensemble member" as another relevant dimension. *Please note that we have this already in the paper cf. lines 532 and 539. But we agree that it should be listed in L111 as well.*

2.10 L214: If I may suggest to contact the maintainers of the Integrated Climate Data Center (ICDC, `https://icdc.cen.uni-hamburg.de/daten.html`) at the University of Hamburg. They do a very good job in collecting, processing (e.g. remapping, quality assurance), and maintaining/updating data of any kind relevant for Climate / Earth system sciences - provided in netCDF, which can easily be converted to Zarr. I assume one could join forces and build an even more comprehensive ESDL. *Thank you for the hint. Indeed we have had various points of contact with Hamburg and very much hope that future collaborations can emerge in the direction you suggest here.*

2.11 L219: ' ... given of their ... " does not sound correct. I suggest to omit 'of' and write "... been ingested given their recurrent ... ". *We agree and will change the text accordingly.*

2.12 LL237-247: This part explains some functions of the ESDL.jl toolbox. At this point, this information is not necessarily important for the reader. Maybe it is enough to refer to ESDL.jl documentation and the case studies which are accompanied with code - as you do it for the python implementation of ESDL. *We note that this request to remove this part stands in opposition to the comment posted by Prof. Pebesma and Dr. Appel who actually requested us to extend this part. We think this is an editorial decision and look forward to the opinion of the handling editor to proceed accordingly. We agree, however, that software developments can change rapidly so that it could be better having their description separated from the scientific concepts.*

2.13 L290: 'an univariate time series' *We agree and will change the text accordingly.*

2.14 L291: I suggest the following corrections and revisions: "If one stored the same data cube with complete time series contained in one chunk, read operations could perform much faster." *We agree and will change the text accordingly.*

2.15 L303: Delete first occurrence of 'behaviour'. I would also omit 'time', since it is redundant in the combination with 'long-term', i.e. 'long-term system behaviour in time' *We agree and will change the text accordingly.*

2.16 L317: Please capitalize "northern" or use lowercase consistently for all occurrences of "northern" and "southern hemisphere". *We agree and will change the text accordingly.*

2.17 L320: Please capitalize "Fig" or use lowercase consistently for all occurrences. *We agree and will change the text accordingly.*

2.18 L321: "Southern Hemisphere" in singular. *We agree and will change the text accordingly.*

2.19 L348: I suggest using present tense when describing what you did in your study and using past tense when describing what others did before you, thus 'In our application, we follow this approach ...'. Please use tenses consistently across the paper. *We agree and will change the text accordingly.*

2.20 L352: Where do you introduce all the acronyms? Please provide the written-out terms here or refer to the table in the Appendix. *All acronyms were introduced in section 3.1. I.e. way before L352 and some again at the beginning of Section 4.*

2.21 L353: I suggest to delete ' the latter two", since it is not needed. *We agree and will change the text accordingly.*

2.22 L361: I suggest to use the term "seasonal cycle" in singular. 2.23 *We agree and will change the text accordingly.*

L374: You use 'essentially', so delete 'only' in "... driven essentially by solar

forcing only ...". *We agree and will change the text accordingly.*

2.24 L377: I recommend to use 'complex' instead of 'complicated'. *We respectfully disagree as the notion of complexity is not trivial and would require additional analytics that are beyond the scope of this paper.*

2.25 LL377-379: This sentence ('Zooming...') is somehow complicated to read and grasp. Maybe you can split the sentence and provide more information why it is important/interesting to focus on the northern regions of South America. *We agree and will change the text accordingly, i.e. we will provide a rationale for our interest in this region.*

2.26 L382: Please use 'land surface' or 'land-surface' consistently throughout the paper. *We agree and will change the text accordingly.*

2.27 L392: Here you put Reco in italic letters and earlier (e.g. L352) you don't. Please use math notation consistently and follow the conventions explained in the Copernicus LaTeX template. *We actually were actually unsure about this, as we have not written variable names in italic, but here need it as mathematical symbol. We will go to the Copernicus style guide and change the text accordingly.*

2.28 L402: Eq. (15) is incomplete. The minimal output dimensions are 'para' and 'time'. Where is the time term in the equation? Also, correct typo 'par' to 'para'. *Thank you for spotting this! This is inherited from an earlier version of the paper where we had a more complex approach. We will change the text accordingly.*

2.29 L405: Please check for consistent usage of 'high-latitude, high-dimensional, high-resolution, etc.', so, with or without hyphen. *We agree and will change the text accordingly.*

2.30 L418: Correct typo 'supporting materials'. *We agree and will change the text accordingly.*

2.31 L426: Please check for consistent usage of 'semi-arid' versus 'semiarid'

(L428). *We agree and will change the text accordingly.*

2.32 L461: Where are these simplifications described in detail? *We agree and will change the text accordingly.*

2.33 L464: Can 'would be' replaced by 'is'? *We prefer to keep this wording as is as the statement is of a speculative nature.*

2.34 LL465-469: Please consider my comment again w.r.t. to contacting the ICDC maintainers (Comment 2.10)! *Yes, indeed!*

2.35 L510: Delete one 'several'. *We agree and will change the text accordingly.*

2.36 L512: Maybe better '... , with no claim to completeness, ...' or '..., without claiming completeness, ...'. *We agree and will change the text accordingly.*

2.37 L525: Please explain shortly 'kaggle' or provide an URL / reference. *We agree and will change the text accordingly.*

2.38 L530: Delete one 'the' . *We agree and will change the text accordingly.*

2.39 L535: The straightforward implementation of ESDL to handle / analyze CMIP multi-model but also the emerging grand ensembles of several hundreds of simulations is a key strength in the ESDL approach, in my opinion. I suggest to promote this aspect more strongly and include respective keywords, such as 'multi-model ensemble', 'large' or 'grand ensemble' in the abstract. *We cannot agree more and will promote it in the abstract accordingly. Please note that in fact, co-author Fabian Gans has implemented a prototype that can be explored online as shown in this gist: https://gist.github.com/meggart/2d544be2c1368f8774d0a21ea4633985. However, this is still work in progress in collaboration with the Pangeo community so we did not include it in this paper.*

2.40 L557: Delete 'on'. *We agree and will change the text accordingly.*

2.41 L557: "Dimension" in plural. *We agree and will change the text accordingly.*

2.42 L594: The last sentence of the conclusion section is somewhat cumbersome. Can you boil down or split this sentence? As a reader, I expect the last sentence of the paper to be a strong and precise statement. *We agree and will have to think about it and will change the text accordingly.*

2.43 Figure 2: This is certainly an appealing visualization, however, it does not convey much information. Maybe this is also the reason why do not reference this figure anywhere in the text. I suggest to reconsider if this figure is really needed or if the url to the animation or providing the animation as ESD asset accompanying the article is sufficient. If the figure is needed, than I suggest to make some modifications as illustrated in Figure 1 and add a legend so that the figure is understandable without studying the details in the caption. *We admit that it was a mistake not to reference the figure properly in the text. However, we respectfully disagree with the statement that this figures does not convey information and hence would like to ask the editor for permission to keep this figure without further modification. The rationale is the following: This figure is a new variant of an earlier figure published here https://figshare.com/articles/Earth_Data_Cube/4822930 which has been used widely. For instance we found many copies of it at the last EGU conference posters (actually without proper citation). Hence, we believe that it is a very suitable means to convey the fundamental idea of the paper in a conceptual manner. If we would go for a technical illustration as you suggest it here with axes labels or colorbars, units etc. it would lose the character of a conceptual figure Hence, we would like to suggest to keep it and explain its purpose more accurately + referencing it properly in the text.*

2.44 Figure 3: The actual values represented by the gray color-scale are not visible at all and are thus redundant in the current visualization. I suggest to include labels for actual values in the polar plot, as described here: https://matplotlib.org/gallery/pie_and_polar_charts/polar legend.htmlsphx-glr-gallery-pie-and-polar-charts-polar-legend-py *We had this*

discussion as well among coauthors. We had a version with labels as you suggested, but these were not readable in print either. Hence, we prefer the current version as the colorbar gives exactly the range and allows us to annotate the usints properly.

2.45 Figure 4: Please explain why there are no data for the Arctic. *We will explain this with data fractionations.*

2.46 Figure 5: Better vertically center the y-label. *We believe that the esthetics is up to the authors.*

2.47 Figure 6: Please correct typos: ‘ the latter reduces ... and leads...” *We agree and will change the text accordingly.*

As a final sentence we would like to thank the reviewer again for the very detailed feedback that will greatly improve the quality of the manuscript! We will acknowledge this is the revised paper as well.

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