Clim. Past Discuss.,
https://doi.org/10.5194/cp-2017-157-SC6, 2018
© Author(s) 2018. This work is distributed under
the Creative Commons Attribution 4.0 License.

Climate of the Past Discussions

Interactive comment on "Technical Note: Open-paleo-data implementation pilot – The PAGES 2k special issue" *by* Darrell Kaufman and PAGES 2k special-issue editorial team

O. Bothe

ol.bothe@gmail.com

Received and published: 1 March 2018

The discussion on open-data is essential not only in paleoclimatology but in climate-science more generally. A number of criticisms, allegations, and attacks against climate scientists based and base on a real, perceived, or simply asserted lack of openness.

Reproducibility and validation of results by others relies on clear communication of methods and, often, on the availability of code and (input and output) data.

Let me start with a disclaimer: I am part of the Pages 2k coordinators team. As such

I was shortly involved in the initial discussions of the data-policy for the Special Issue and involved myself recently in discussions about the comments on the Technical Note among the SI editors, the data policy team, and the Pages 2k coordinators. Furthermore, I am not a data-producer (as defined so far in the discussion) but a data-user. If I became a data-producer, it would likely be of model-simulation-output.

I am one of those persons reluctant to publish their data because of, in short, losing control of the data, which however entails various only vaguely conscious and vaguely formulated concerns.

In the context of the discussion on the Technical Note, I would like to very shortly comment on a number of points.

There is an ideal world where every manuscript is open access, the methods used in the manuscript are openly accessible, i.e., the code for the analyses is available, and the input data of the manuscript as well as the output data are public as long as they may be input for new studies.

This does not describe the current state of scientific publishing and I think it is unlikely that it will become the state of things anytime soon. But some relevant issues originate from this idea.

In this ideal, the idea of publishing the data would define our workflows to some extent. That is, when a data set approaches the state in which it could be published, the workflow would include a step of putting the data into its publishable format, preparing the meta-data, and so on. Thus, if we aim at reproducible research, this data-preparation step has to become an essential part of our workflows, and it's current absence - at least in my workflow - is one of the major hindrances in being truly open.

If this becomes part of the workflow, publishing the data is little more than pushing a few buttons. Then researchers can decide themselves when it is the correct point in

time to publish.

This ideal world assumes it is desirable to make data publicly available and thus to lose control. If this is the case, projects and grant proposals include the publishing of the data as an integral part and are structured in a way that it is done without harming the project or the researchers. This harm-prevention implies that the ideal world includes a mechanism to provide the meta-data of a dataset to the public but to still withhold the data for a certain period of time.

As a side-note, I, personally, would be more afraid of scoops where someone does similar analyses and comes to similar conclusions based on a different data-set than that someone uses my data.

Another point is, which output data we really consider to be essential. Someone once stated, every data that is visualized in a manuscript should be available to reproduce the plots. My personal feeling is this is excessive. Is the output of a principal component analysis on an available input data set essential? Is a transformation of an input data set essential? Is the combined transformation of more than one input data sets essential?

This prompts the related question of whether it is better to provide the code or the output data, or whether it is indeed necessary to provide both.

Thus what is essential for reproducing an analysis and essential for subsequent studies?

On a slightly different note, I am not sure to what extend the discussions on the data-policy also considered the output of climate simulations. While many of these are routinely archived at relevant data centers, there are many other simulations which are not archived in an easily accessible manner. They may even only be stored on a collection of external hard drives. Is it enough to put

the subsets used in a manuscript in a repository or should a larger part of the output be publicly available. More generally, how do we deal with large data sets.

Adding to this, if this was a review of the Technical Note, I would recommend to discuss in (even) more detail the following three topics, (i) the concerns on the procedure of the data team as expressed in the discussion so far, (ii) recommendations on work flows and tools supporting the publication of new or previously unavailable data sets, and (iii) specifics for various different data types, e.g., proxies, proxy collections, model simulations, or ensembles of results.

As a general comment, I think there is little doubt that we have to come to a point where we not only encourage making data public but where we as a community take steps to ensure that data is made public. The community as a whole and each individual researcher will likely benefit from the openness of their colleagues more than from withholding their own data. The additional steps in workflows require additional efforts, but I assume these are not bigger than learning to use a new software or hardware tool.

The discussion so far shows from my point of view that while it is - as Kaufman (2018, https://doi.org/10.5194/cp-2017-157-SC4) puts it - indeed time to commit to making data open, there is work to do. We need to be more specific about the how and what and we have to provide mechanisms of harm-prevention not only but especially for early career researchers.

Interactive comment on Clim. Past Discuss., https://doi.org/10.5194/cp-2017-157, 2017.