Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2019-223-RC1, 2020

Interactive comment on "Integrating palaeoclimate time series with rich metadata for uncertainty modelling: strategy and documentation of the PALMOD 130k marine palaeoclimate data synthesis" *by* Lukas Jonkers et al.

Anonymous Referee #1

Received and published: 29 January 2020

The authors present a compilation of paleoclimate data from marine sediment cores covering the past 130 kyr. They give a clear account of their data acquisition strategy, which focussed on cores with d18O measured on benthic foraminifera and radiocarbon dates so that a robust common chronology could be constructed for the entire compilation. Where other paleoclimate data were available for the same core, these data were included in the compilation. They pay close attention to including meta data required to analyse the data further.

Parameter and metadata names have been harmonised and the original naming is

preserved so that these can be traced in the original publications. The data are all well referenced with DOIs and citations. New depth-age models have been constructed for all sites using BACON and the published chronologies preserved for reference.

In all this represents a very well researched and harmonised dataset with rich and useful metadata that does not exist elsewhere. The data are supplied in a variety of formats, as R data objects, NetCDF, and in the LiPD format which is itself a set of zipped plain text (csv) files containing the data in a highly structured JSON format (http://wiki.linked.earth/Linked_Paleo_Data).

However, the structure of the data within the R objects makes it very difficult to search for and extract subsets of data. For example, all the records of variable "planktonic.d18O", or all the records in a certain geographical region. In the NetCDF and LiPD formats the data are also structured in a similar way, although there may be tools available to help work with the LiPD data.

Use of this data compilation would be greatly enhanced if the data were restructured into a set of "partially normalized tables" in a "star schema" so that queries can be made in an SQL-like way by joining tables and using select and filter type statements. See Brian McGill's 3rd Commandment here (https://dynamicecology.wordpress.com/2016/08/22/ten-commandments-for-good-data-management/). A key table in this format would for example be the "ParameterListWithRefs.csv" table linked to in the Data availability statement of this manuscript but not found within the data objects.

No specific database software needs to be used; these could be plain text files that could be read in by many data analysis software. This is not a "big" dataset so the structure does not need to optimise storage or retrieval efficiency.

I'm not suggesting that reformatting the data in this way would be trivial for the authors, but the data in their current format are well structured and so it should be possible to write code to do it – and this should be much easier for the authors than for someone coming to it fresh.

Minor comment: l. 486 - In the section “recommendations for data archiving” “Include metadata” I would also recommend including information about the size of the sample on which the parameter was measured, e.g. number of foraminifera, mass of sample, total peak area for Alkenones. As this can be very useful when assessing the uncertainty of the value.

Text errors:

l. 147 “were” -> “where”

l. 161 “more of data”

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2019-223, 2020.