Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2019-7-RC2, 2019

Earth System Science Data Discussions

Interactive comment on "A taxonomically harmonized and temporally standardized fossil pollen dataset from Siberia covering the last 40 ka" *by* Xianyong Cao et al.

Patrick Bartlein (Referee)

bartlein@uoregon.edu

Received and published: 18 June 2019

General comments:

This "data paper" describes the construction of a fossil-pollen data set from Siberia, that is available from Pangaea, and which will have wide application. The data are not all new, but there is considerable "value added" in placement of the data into a common taxonomic framework, and the application of a "modern" approach for establishing the chronologies of the individual records, and so I think this paper makes a useful contribution that warrants publication. I have three general comments on the paper, related to locational precision, chronologies, and relationships with other published data sets.

The individual site locations are listed in decimal longitudes and latitudes, and to two decimal places, which (roughly) yields a precision of a little over a kilometre in a N-S direction. Over a domain the size of Siberia, that sounds like a small number, but my experience with similar databases (e.g. Whitmore et al., 2005, Quat. Sci. Rev.) suggests that locational uncertainties of that order (single kilometres) creates issues in such tasks as interpolating or assigning modern climate data to core locations, or inferring elevations from a DEM. It's probably the case that the source data imposes this limitation, but if at all possible it would be good to include more precision in the locations.

I was expecting so see another table that contained information on the age assignments in individual records (age uncertainties could be worked into count and percentage tables, but I think it would be better to have all of the chronological information in one place). In addition to information on the radiocarbon ages used to build the chronologies, if other sources of information (biostratigraphic information, tephra ages, etc.) were used, that should be noted there as well. I'm not sure if the paper would be unpublishable without adding this information, but it would be a better paper if it did.

There are other databases that overlap to some extent the region documented here, in particular Binney et al. (2017, Quat. Sci. Rev.). I'm surprised by the comment that Binney et al. (2017) did not require taxonomic harmonization (because it was presenting biome reconstructions). It seems, however, that it did (see section 3.3 in the paper and Table 6 ("TaxonTaxonclean") in the Binney et al. database). I think there should be a bit more discussion about the similarity or difference between the two databases (which shouldn't be hard given the overlap among authors. . .).

Specific comments

p. 1, line 32: "transformed" sounds like, well, some kind of transformation of the data took place. Would "assigned" be a better word?

p. 2, line 3: "pollen counts" Literally counts, or were some records already expressed

as percentages? (Nevermind, explained later in the paper. . .)

p. 3, lines 15+: “homogenization of taxonomy” I think this needs to be explained a little more, because superficially it sounds like it’s a simple spreadsheet task (i.e. combining columns). The elements of a more detailed explanation should include, I think: 1) the nature of the problem (different studies used different taxon lists; there are different ways of assigning pollen type (as observed) to taxa; etc.); 2) the implications of splitting vs. lumping; 3) the “theoretical” issue of determining the target-taxon list; and 4) the practical aspects of doing the assignments. This doesn’t have to be the master tutorial for homogenization, but it should be sufficient to explain to a reader why the same record might appear different in detail in different databases.

p. 4, line 17: 100 ïĆś 10 % allows a pretty generous level of noise in the digitized data. I’m guessing that (unless the source materials were really bad) that level was not reached very often.

p. 5, line 15-15: “The presented dataset” and “this dataset” are ambiguous.

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2019-7, 2019.