Response to comments of reviewers

Patrick Bartlein (Referee)

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 kilometer in a NS 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 kilometer) 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.

Response: We agree with reviewer’s comment. A precision coordinate is quite important for pollen record. In the manuscript, we only present to two decimals in Appendix 1 to save space, but in our dataset uploaded to Pangaea the data are more precise. Nevertheless, we keep the precision of the coordinates as given in the original papers or database.

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.

Response: We have prepared a table including all dating data for each pollen record: see the response to the first comment mentioned by the first reviewer.

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: : :).

Response: We had already cited Binney's (2017) paper. We have carefully checked Binney's paper again recently. Their dataset includes most of the pollen records from northern Eurasia (including Siberia) together with plant macrofossil data (Binney et al., 2017). However, their dataset does not finalize the chronology standardization and the pollen data are restricted to each 1000-year time-slice. As we describe in the manuscript, our pollen dataset has finalized the temporal standardization and we present all original pollen data for each sample.

Line 66-70:

"*Binney et al. (2017) establish a pollen dataset together with a plant macrofossil dataset for northern Eurasia (excluding east Asia; and the dataset has not been made accessible yet), but the chronologies were not standardized and the pollen data restricted to 1000-year time-slices.*"

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?

Response: Done.

Line 31-33:

"*Pollen data were taxonomically harmonized, that is the original 437 taxa were assigned to 106* *combined pollen taxa.*"

p. 2, line 3: "pollen counts" Literally counts, or were some records already expressed as percentages? (Nevermind, explained later in the paper: : :)

Response: Yes, our pollen dataset includes both counted pollen and pollen percentages.

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.

Response: We have explained "homogenization of taxonomy" in more detail.

Line 135-141:

"*The pollen records were counted by different scientists that gave different pollen names to the same pollen types requiring taxonomic homogenization (from 437 original taxa to 106 combined taxa). However, this reduces the taxonomic resolution of the dataset. In cases where homogenization would have resulted in grouping pollen taxa with different growth forms (herb/shrub, tree) together, did we keep the taxa separately even though not all analysts separated them (for instance, Betula pollen is separated into Betula_shrub, Betula_tree and Betula_undiff).*"

p. 4, line 17: 100±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.

Response: Agree. In our pollen dataset, only 16 pollen records were digitized from publications, the digitized pollen percentages were re-calculated to ensure the data reflect the general pattern.

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

Response: Done.

Line 174-176:

“*The Siberian fossil pollen dataset has already been used for biome reconstruction (Tian et al., 2018), although an integration of this dataset into global or Northern Hemisphere-wide biomization research is still pending.*”