Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2021-272-RC1, 2021

Comment on essd-2021-272

Daniel Gavin (Referee)

Referee comment on "The Reading Palaeofire database: an expanded global resource to document changes in fire regimes from sedimentary charcoal records" by Sandy P. Harrison et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-272-RC1, 2021

This data paper presents a new SQL database of sedimentary charcoal records. An important value and main motivation of this database, and of earlier versions of this database, is to understand the history of fire in the Earth system, at regional to global scales. As presented in the manuscript, there was considerable effort placed on quality checking, development of new age models, and adding site and entity metadata. The new database represents a large improvement in the breadth and depth of past databases. The authors claim several issues of data errors in the most recent database (the GCD, on paleofire.org). While paleofire.org provides a user-friendly interface to the data, I agree that metadata on chronological control is lacking.

We thus have a situation where there are multiple databases with different PIs sponsoring the data bases. To be up front regarding conflict-of-interests, I have no vested interest nor history of involvement in these databases except in their early genesis > 12 years ago when I participated in the first publications of this database. Since then, my main interest has been on site-level interpretation, or at most intercomparisons of a few sites. Thus, I am familiar with the nature of these data and I appreciate the goal of using these data in Earth-system science.

I found the manuscript to sufficiently describe the data. Modest database skills are required to read the SQL database. As described by the authors, the database is unique given the additional sites, age models, and metadata added. However, it is built from existing databases and my comments below address this issue.

My first concern is parallels with the Neotoma Paleoecology Database. The RPD has a data table structure that parallels Neotoma. All the variables in the RPD map directly to a variable in a table in Neotoma. The metadata forms in Tilia (which uploads to Neotoma) and the chronology information are much more thorough in Neotoma than in the RPD. Furthermore, many (if not most) of these records contain other data types (pollen, geochemistry) that may be archived in Neotoma. Thus, the effort at age model development would better serve the research community had the data been uploaded into Neotoma. Of course, this can still be done. As stated on the Neotoma database website: " Neotoma will enable joint analysis of multiproxy datasets to address paleoenvironmental

questions that transcend those possible with single-proxy databases." Currently, only 25 sites in Neotoma have macrocharcoal data, and 19 sites have microcharcoal data. However, many of the entities in the RPD exist in Neotoma for pollen and other data types. Neotoma is fully capable of defining all >100 types of units for measurements (area, counts, size fractions). I understand the history that lead to this situation: the GPD began many years before Neotoma was archiving charcoal data. I now think Neotoma is the proper home for sedimentary charcoal data.

However, there are limitations in the Neotoma database with respect to charcoal data for research. The API interface for Neotoma, as exists in the Neotoma package in R, can batch-download datasets. The API interface for downloading charcoal data may not be functioning for charcoal datasets. I am not very familiar with this interface. Once functional, R scripts could generate the charcoal data and various other data from the same entities (e.g., LOI, pollen of fire-sensitive taxa). Thus in the absence of this interface, stand-alone databases such as this are needed.

A second concern is the universal application of Bacon age-depth modeling. Bacon is superb when dating density is high enough to result in overlapping PDFs of the calibrated ages. In contrast, in situations when, for example, a 14,000-year record is dated by only five well-separated dates, in my experience, Bacon produces interpolations that are indistinguishable from linear interpolation. In such cases, spline fits using, for example, Clam, will not result in abrupt changes in sedimentation rate at the dated depths and not miss dates that are a little out of line. Often, original authors have placed considerable effort in the unique situations of their sediment record, and thus applying a canned approach (as described by the ageR package) may not be as desirable as simply updating the original model with INTCAL20 calibration.

This second concern provides another argument for Neotoma for dataset archiving. Neotoma has separate geochronology tables (a list of all dates measured) and chronology tables (the age controls used for a particular age model or "chron"). New "chrons" may be added to existing data sets.

Finally, a third concern is that this database does not store the raw data values at most sites, but rather some derived value. Specifically, the TYPE variable for most entities is 'concentration' or 'influx' or some other value that is very unclear to its meaning ('other'). I assume this derives from the original data collection efforts in the original databases, in which raw data may have not been archived. (By 'raw data' I mean the measured value for each sample, e.g., 'count' or 'area', or 'mass'). Many statistics depend on these values (e.g., the CharAnalysis program). It is also just good practice to archive raw values, not some derived value. For example, the new age models cannot be used to recalculate new influx rates if the data are not stored as raw values (or at least concentration data). Scanning the database, less than 20% of the entities have 'count' or 'raw count' under TYPE, while most are concentration. Furthermore, converting from concentration back to raw values (multiplying by analytical_sample_volume) is not easily done because the 'analytical sample volume' is a string field with values such as '5cm3'. I note that there was no effort to locate the raw data values (as they are available for several sites that are reported as influx), but rather to use whatever units were provided in the earlier databases. Analysis of the RPD will need to be made very carefully to prevent errors such as double-calculating influx rates or concentrations.

I examined five of my sites within the database. I found errors in each of these sites. Correct data exist on Neotoma and on my personal web page (for most):

- Yahoo Lake: 1) does not have a new age model (missing from the age_model table).

2) Data described as concentration, but the units in the sample table are influx values. 3) The sample volumes are incorrect.

- Cooley Lake: 1) data described as concentration, but the values in the sample table are counts. The data are on Neotoma and my personal web page. 2) The sample thickness values are 0 (which is impossible) or 0.01, but most samples have a 0.5 cm (0.005) thickness.
- Clayoquot Lake: data described as concentration, but the values in the sample table are counts.
- Rockslide Lake: 1) data described as influx and are indeed influx in the sample table! However, raw count data are available 2) The sample thickness values are 0 (which is impossible) or 0.01, but most samples have a 0.5 cm (0.005) thickness.
- Wentworth Lake: 1) data described as concentration, but the values in sample table are counts. 2) sample thicknesses and volumes missing.

It is interesting to surmise the sources of these errors. In my case, I provided the original raw data in spreadsheets which included measured values, concentration, and influx columns. Thus, over time, through different versions of databases, information was lost or changed. If these issues were detected in the five sites that I checked, there is reason to believe such issues exist with all the sites.

I recognize the effort placed into creating this database. I see much work was spent on adding new sites, the ageR age models, and developing the MySQL schema, and adding some new metadata fields. However, it is quite concerning those errors in old databases are perpetuated into new databases. I highly recommend Tilia and Neotoma as a means of correcting errors. The data upload process into Neotoma (performed by data stewards) involves several quality control procedures. Uploading to Neotoma is time consuming because 1) the errors in the GCD and RCD will preclude using a bulk upload method, requiring checking against original publications, and possibly contacting authors, and 2) the quality control checks in Neotoma often detect additional errors.

My recommendation for a major revision is for a large but very do-able job:

1) Check if the site and entity exist in Neotoma and provide a table that matches the entity to the Neotoma entity.

2) Check the measurement units and change values to raw measurement units where possible. Also provide correct thickness values wherever possible. The numerous coauthors should be able to help with this task.

3) Change the analytical_sample_volume to double and include a new variable for volume units or standardize all samples to cm3.

4) Have the sample table contain four columns for 1) value (e.g., count), 2) volume, 3) concentration, and 4) influx (calculated using the new age model when possible). In theory, columns 3 and 4 could be omitted and generated on the fly as needed. However, it will not be possible to have raw values for all sites, thus requiring these four columns to exist.