Earth Syst. Sci. Data Discuss., author comment AC1
https://doi.org/10.5194/essd-2021-272-AC1, 2021

Reply on RC1

Sandy P. Harrison et al.

Author 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-AC1, 2021

Thanks, Dan, for your comments on the manuscript. We can certainly improve the current database and ms based on your suggestions, but our aim was not to create the perfect database but rather to create a dataset that we can use for meaningful analyses and that is both more comprehensive than existing resources and has fewer errors. We will clarify this in a revised version of the text. Furthermore, we will revise the text so as to indicate possible issues which a user needs to be careful about.

In answer to your general points:

(1) The relationship between the RPD, Neotoma and the GCD

Our main purpose in putting these data together was to create a database that could be used for a series of planned analyses. We were driven to do this because the charcoal data coverage in Neotoma is limited, we recognised that there were problems with some of the sites we are working with in the various versions of the GCD, and there are a lot of data that we wished to use in our analyses that are currently not in any database or long-term repository. It is not our intention to create a permanent data repository for charcoal data and we agree that people generating charcoal data should ensure that they lodge their data with a long-term repository such as Neotoma. Since we are aware that there are still errors in the data set we have put together, we are reluctant to upload the RPD as a whole to Neotoma. However, we have encouraged individual data contributors to lodge their records in Neotoma or another suitable data repository. We are also happy for the data we have assembled in the RPD to be incorporated into Neotoma and/or the GCD, so that they can be used by the wider community. We welcome Jack's positive encouragement to do this, and if it appears that individual data contributors do not have the resource to do this, we will work with Neotoma on the best way forward to ensure data are not lost.

(2) Universal application of BACON age-depth modelling.

Our main purpose in putting these data together was to create a database that could be used for a series of planned analyses. For this reason, we decided to use a standard approach to age modelling and to use the latest appropriate calibrations. We recognise that this approach may not be suited to every site and that users might want to use alternative approaches for their own analyses. For this reason, we provide the information about the dates available for each site in the table "date-info". This includes all radiocarbon dates, other radiometric dates, correlative dates and core top ages as provided in the original publications or by the authors of specific records. We also indicate when the original age model excluded specific dates and why this was done. In addition, we provide the age for each sample based on the author's original age model in the "chronology" table. Thus, the RPD parallels the Neotoma structure both in terms of archiving the base data to create age models and in terms of providing an alternative chronology. We will revise the text describing the construction of the new age models to make it clear that while we are providing the models (and the uncertainties associated with them), the user can access the original age models and can also use the dates provided to construct their own age models

(3) Raw versus processed data.

We agree that the ideal is to archive raw data (count, area or mass). However, as you rightly point out, this was not available for all of the sites repatriated from the GCD. Specifically, 99.8% of the data repatriated from the GCD does not have raw data (855 out of 856 entity records). Furthermore, it was not available for all of the new sites included in the RPD. Specifically, raw data is not available for 77% of the new data we have included in the data base. During the construction of the data base, we have prioritised the inclusion of raw data (23%) or concentration data (54%) for new records wherever possible. There are some cases where we have both count and concentration data for the same records (n=24 from 9 different sites); we can remove the concentration data for all but 2 of these records for which we do not have information on sample size. In some cases, we have been able to replace influx measures with raw or concentration data for existing records taken from the GCD (n=43 sites). We may not have done this for all sites where the raw data are available, and this should certainly be a priority for future improvements to the data set. We agree that it is necessary to be careful in making analyses with the RPD data to ensure that we don't double calculate influx or concentration. We will add a caveat about this in the text.

(4) Errors in repatriated data.

The five sites that you list as containing errors were taken directly from the GCD and we apologise for not checking directly with you about these sites. We can certainly correct this information before publication of the RPD. The broader issue here of course is how many errors there might be in the rest of the data taken from various versions of the GCD. Given that our goal is to use the data for analyses, rather than to construct a permanent data base, we hope that these errors will be trapped and corrected as we go forward. However, we will correct the errors that you have pointed out in the data for Yahoo Lake, Cooley Lake, Clayoquot Lake, Rockslide Lake and Yahoo Lake. We will also check the Neotoma holdings and see if these provide additional raw count data that can be used to update the RPD records.

Response to specific suggestions:

(1) Inclusion of Neotoma IDs. We originally included the GCD ids for various sites, but this was confusing because the ids changed between versions of the GCD. We do not include the Neotoma ids for individual sites because so few of the sites are currently in Neotoma. However, we do include a field in the entity table which identifies the source of the data (i.e. whether it was from Neotoma, a specific version of the GCD, or a new contribution from one of the co-authors) and this should make it possible for users to track back and find the original data. This will also facilitate them being able e.g. to combine charcoal and other types of environmental data archived at Neotoma.

(2) Checking measurement units and changing to raw values where possible. The co-authors have already checked sites which they contributed, and we have included raw values where these are still available. We have expended a considerable effort on data checking for other sites but agree that we can and should do further checks. However, the use of the data compilation is the ideal washing machine here and we are sure that it will be easier to clean up the data as errors become apparent through use. In addition to the corrections for the five sites listed above, and checking of the Neotoma holdings, we will run a further check for measurement units for the new sites in the data base (currently 50% of the records).

(3) Analytical sample volume. We agree that it is relatively simple and that it would be useful to separate the size and the units here and will implement this. We will take the opportunity to standardise the units further e.g. to remove units that are expressed as multiples (ax100, ax1000) and to convert different weights (e.g. mg, g, kg) to a standard unit (g). We will add text to point out that these conversions have been made so that the reported data might not appear to be the same as previously published data.

(4) Separate counts, volume, concentration, influx etc. We did not do this because of the tendency for people to provide entries for all columns, which could lead to confusion if influx is recalculated using new age models. Rather than create separate columns for each type of count, we will focus on ensuring that the information given is correct and in trying to obtain raw data wherever possible.