Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2022-172-RC1, 2022

Comment on essd-2022-172

Oliver Bothe (Referee)

Referee comment on "The CoralHydro2k database: a global, actively curated compilation of coral $\delta^{18}O$ and Sr□/□Ca proxy records of tropical ocean hydrology and temperature for the Common Era" by Rachel M. Walter et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-172-RC1, 2022

Dear authors, dear editor,

Summary

The PAGES CoralHydro2k Project team presents their database of paired coral δ18O and Sr/Ca proxy records in the manuscript in question. This is a major contribution for our understanding of tropical climatology over the Common Era and particularly the past 200 years.

I have to emphasize that I cannot comment on the quality of the manuscript with respect to questions specific to coral research but I review it mainly with respect to general paleoclimatological and data publication aspects.

I have a number of minor comments and questions on the manuscript, which, however, are not critical. I have two larger - though still not major - notes on the manuscript assets and the presentation.

From my point of view the manuscript can be quickly published.

Recommendation

I recommend publication after minor revisions.

Comments:

Larger notes::

1.
a. As far as I can see the document assets are not yet available at the already prepared persistent location (https://doi.org/10.25921/yp94-v135, last accessed 19. August 2022). This makes it hard to assess database version 1.0, which supposedly is described in the manuscript. I would prefer to have access to version 1.0 and not to have to fall back to version 0.5.4. Indeed, I don't think reviewers are clearly enough pointed to version 0.5.4 in the editorial interface but that may be more a comment towards Copernicus and not the authors. Version 0.5.4 is apparently a smaller subset of version 1.0. Therefore there is an obvious discrepancy between the data as described and the data that can be reviewed. I trust the authors and the editorial team at Copernicus that they ensure that (i) the data will be available at publication at the given location and (ii) that version 1.0 will be as accessible with the available tools as version 0.5.4.
Version 0.5.4 is available at the Lipdverse (https://lipdverse.org/CoralHydro2k/current_version/, last access 19. August 2022).

b. Going from the description in the manuscript, I am not sure what I should find at the given DOI. Neither can I verify that what is described at NOAA as downloadable data is identical with the manuscript description. NOAA NCEI states that there is a description of the file, an example file that may be the Matlab script mentioned in the manuscript but may also be something else, a pickle-file, a zip-file for the LiPD-files, an RData-file, a Matlab-data-file and a NCEI Direct Download, which at the moment does not point to a download but to the Google-form for submission of new records.

2.
I am not convinced that the technicalities of accessing the data are sufficiently presented in the manuscript (its section 4) and in the supposedly accompanying Matlab-script. This is not least the case because I am uncertain how commonly the LiPD format is now used by colleagues.
The Matlab-script is, as far as I can find, not available so far and may or may not be included as a potential item at the database's NOAA NCEI access page.

I would welcome it if the authors provide more detailed documentation that quickly guides the potential user through loading the database, filtering it, and plotting an example series or even redoing one or more of the plots in the manuscript. I quote from the review guidelines: "The authors should point to suitable software or services for simple visualization and analysis, keeping in mind that neither the reviewer nor the casual "reader" will install or pay for it." While the authors point to the LiPD format and the associated tools and while the serializations can be accessed without any knowledge about LiPD, such a simple walk through may increase the later reuse and utilization of the database. An example of what I have in mind could be Nick McKay's (one of the manuscript's co-authors) tutorial for the geochronR package (e.g., https://nickmckay.github.io/GeoChronR/articles/Introduction.html, last accessed 19. August 2022).

Sidenotes:
I did not test the access to the Matlab version. I did test the access to the Python and R serializations. I did test the access to the LiPD-files from within R. I only had a slightly more detailed look from R, which suggests that the data is accessible as described.
I did not check the consistency of the appendix table.

Minor

Page 2, line 57ff: I am not sure this sentence is relevant for the topic of the manuscript. If it is from the authors' point of view, I nevertheless wonder if they really mean aspects of large-scale hydrology being tied to large-scale dynamics or if they mean more generally aspects of hydrology.

Page 3, line 89ff: The authors mention SISAL later in the paper, but I think the database is also relevant here.

Page 3, line 106ff: The paragraph includes the phrases "active curation" and "opportunities for future data collection". While both are indeed mentioned later, the phrasing here suggests more prominence for both than eventually realized.

Page 5, Line 153: I am surprised - and apparently didn't pay attention to Iso2k - by using only two digits of the publication year. In a sense it probably is a realistic perspective on the longevity of any data today but the philosophy may result in conflicts at some point.

Tables generally: The authors clarify the meaning of "standardized" fields in the manuscript text, but I am not sure that the reader will get what is meant from the table captions alone.

Table 2: This is minor but I think it may become important if more databases use comparable structures. The CoreID-variable has the fieldname "paleoData_ch2kCoreCode". May this ID better have a fieldname that is more directly interoperable with other LiPD IDs as it is of the same structure as an Iso2k ID - if I understand it correctly. What I mean is, if "paleoData_coreCode" or "paleoData_code" are potentially better fieldnames and other databases may, then, want to use the exact same fieldname.

Table 2: Similarly to the previous comment I wonder if "geo_secondarySiteName" is standard nomenclature for comparable types of data.

Table 2: The authors call the paleoData_TSid a LiPD ID but it is also in the serializations. So, I am not sure whether LiPD in the manuscript refers to the data container or file format or "vehicle" as the original paper calls it, or to the framework of structuring the data. Maybe it is not so much a LiPD ID but a "time series" or "record" ID.

Table 2: I am a bit confused by the connection between the fieldname "paleoData_hasUncertainty" and the variable Error TSid. First, the fieldname for me suggests a logic variable or flag but not a TSid. For an ID, I would rather expect a fieldname like "paleoData_errorTSid" in agreement with the "paleoData_TSid". Second, I am of two minds if I agree that there should be a difference between Error TSid and TSid. Both are in the end TSids, both are generated the same way presumably. They serve different functions. I suggest to the authors to consider if Error TSid may better also be named TSid - but I myself tend right now to a "no".

Table 4: I personally would welcome standardization and quality control on the "Original data source" in an upcoming update on the database as well as inclusion of a persistent identifier (PID) for this Original source as an additional entry in the publication metadata. That is, an entry "Original data source PID" with variable "originalDataPID". Maybe that is even a major shortcoming of the database in its current state.

Table 6: As a reader and a potential user - who likely would not get into the documentation first - I wonder if the entries "calibration_dataset" and "calibration_datasetRange" are clear enough transporting what they are or whether users may expect something different.

Section 3 generally but starting with section 3.2: Regarding the given significant correlations: a correlation of 0.13 may be significant but what really can we expect to learn from such a weak relation between proxy and variable of interest. Extending on that, are these very weak correlations significantly different from zero. To be clear, I do not expect the authors to answer by extra analysis but I think it should be commented on shortly.

Section 3.2: again on the correlations: The authors state that a higher percentage of records is correlated significantly for bimonthly data than for annual data. How much of this potentially is due to seasonal signals? And if so, what does this imply for subsequent reconstructions, if the, e.g., annual cycle peaks dominate the correlation skill? Or has this basically no repercussions at all?
The authors address this point in parts on page 18 in line 317. Thus, there also applies my question: is the seasonal cycle correlation a feature in records or may it even negatively affect reconstructions of interannual climate variability?

Figure 5: I think it may be helpful to add more information on the filtering also to the caption. However, I also understand if, then, the caption becomes too lengthy.

Page 17, line 296: I am not sure that the authors use the term “mode of variability” in its commonly understood meaning here. Modes of variability usually - in my understanding - do not refer to frequency bands but to large scale features of climate variations.

Page 18, line 303: Sentence: “Conversely …” Looking at Figure 5, my impression is that this statement is not correct in its absoluteness, but the authors have done more analysis than looking at the Figure, so my eyeballing may be wrong.

Figures 6 and 7: Maybe the captions could benefit from some more details.

Page 20, line 366ff: The references for the sentence on “vital effects” are quite old. As I am not a coral-person I am curious: have there not been any updates on this topic?

Page 21, line 380: I am surprised that the sentence singles out the impact of calibration. Isn’t it more the impact of each step in the workflow that requires more work? Indeed this made me wonder if the coral community could do - or maybe they even already did it or are in the process of doing it - something like the tree ring community did for Büntgen et al. (2021, 10.1038/s41467-021-23627-6)?

Page 21, line 399: I am not sure that “LiPD serialization” is clearly understandable, and that it is clear that the author’s view on their data is that they provide (a) the database as in LiPD-formats and (b) a number of serializations of the database to serve different languages.
In addition the following paragraph and list could be understood as meaning that these are the only possibilities to subset the data but - again unless I am mistaken - this list is not comprehensive.

Page 21, line 398: Are D and TS correctly described as “variables” - not least as variable means something different in the data.

Page 22, line 401: The authors write, the database can be searched. Naively one may assume that there are specialised tools for the database. It may help to clarify that, unless I am mistaken, the “searchability” basically means to use a coding language to reorganize the data.

Page 22, line 418: It would be helpful if the MATLAB script was available already. It would also be a great service to the community, if further scripts or notebooks for other languages are provided in the future.

Page 22, line 425ff: “It is anticipated” is a rather weak statement. Does NOAA NCEI allow for such a change log and is CoralHydro2k striving to provide it?

Page 22, line 430: “If only a subset …”: I disagree and I welcome if the authors change their message here. If any subset of the database is used each member of this subset should be referenced. Similarly, if any record is singled out, these records should be explicitly referenced. This ideally includes citations to a relevant publication and the record/dataset.

Page 23, line 432: I recommend that the authors also include a persistent identifier (PID) to the original public archive to foster FAIRness, reproducibility, provenance, and a culture of giving credit where credit is due.

Page 23, line 437: “improving the skill of future climate projections”. I agree but I think this statement would benefit from a reference - or if the point is supposed to be made above already, then a reference and more emphasis are necessary there.

Appendix table: I suggest that the authors include further information in this table: (a) a persistent identifier (PID), e.g., a DOI, for each record, (b) a data citation for each record, and (c) the DOIs for the publication. (c) may be unnecessary assuming the reference list is complete and (a) may also be obsolete if (b) is fulfilled and all data citations are in the reference list and include such a PID.
Beyond that I did not check the consistency of this table.

Page 37, line 487: Acknowledgements: As former PAGES 2k coordinator I am unsure if CoralHydro2k received funding from PAGES within their Data Stewardship Scholarship, if so, I think this should be acknowledged and put into the funding information.

Software: If there is code to write or access the data structures, sharing it publically may foster wider adaptation of the CoralHydro2k database. This could also be referenced including “data”/code citations.

Finally, not so much a comment on the manuscript but on the chosen data format. As an R-user I still would welcome it if all the LiPD tools were available from CRAN and not only from Github. If I recall correctly, the LiPD-crew is pursuing this goal but I thought I might emphasize my wish once more here.

Technical

Page 1, Line 39ff: I am not sure that the sentence "Most coral-based …" is clear on first reading. Maybe consider clarifying.

Page 3, Line 85ff: Again, I am not sure if the sentence "Whereas …" is clear for the reader. If I understand it correctly the main point is the contrast between success at sites and the limited assessment of larger scale signals. I think restructuring the sentence may clarify the point.

Page 4, Line 112: "is" and "make up". In a sense phase 3 is gone and PAGES 2k is now in phase 4 - and CoralHydro2k is still part of it. I wonder if it may be an idea to rephrase this to be more aligned to the current status. However, it isn't wrong as written, so may also stand.

Pages 4, Line 121: Do Google Suite, Slack, and Zoom require references?

Page 4, Line 114 and line 126: The authors mention the project goals in line 126 but I think they better fit in line 114.

Section 2.2: Is a reference to the FAIR principles already needed here?

Tables in general: I do wonder if the clarity of the manuscript and the understanding of the tables would benefit from slightly more worded/detailed captions.

Tables again: As nothing in Table 2 is italicized as far as I can see, I invite the authors to check the italicization in all tables.

Metadata field names: Most fieldnames are structured as "word1_word2Word3" but the publication metadata in table 4 is simply written English. I think this should be aligned between different tables and if some changes have to happen to the data files, this should also be done.

Table 2: the description for paleoData_variableName has "will be" and "will have" and I wonder if the tense is correct.

Figure 1: The Figure would become even clearer if there was a bit more white space between panels a and b but that certainly is a very minor point.

Page 17, line 286: The authors write of “significant discrepancies”. Is this a tested significance or simply a figure of writing? If it is the latter, I suggest replacing the word.

Page 22, line 415: I am not convinced that using LiPD follows from being guided by the FAIR principles.

Page 22, line 420: I ask the authors to check that the form is correctly labeled as such on the repository website.