Authors' Response:

Thank you to all three reviewers for the insightful remarks and constructive criticisms of the submitted manuscript, database, and codebase. We are pleased to describe plans for revising the manuscript as follows (reviewer remarks in blue; our response in black.)

RC2: 'Comment on essd-2025-364', Julien Emile-Geay, 21 Aug 2025

Citation: https://doi.org/10.5194/essd-2025-364-RC2

Review of "A database of databases for Common Era paleoclimate applications" by Evans et al

Summary:

The article presents an attempt at synthesizing paleoclimate proxy records across 5 different databases with partial overlap. A detailed procedure for identifying and removing duplicates is described, and two applications of analysis on the unified database are shown. This careful work will be suitable for publication after minor revisions.

Scientific Comments:

Given the scope of the journal, my comments will focus on the data and associated code.

1) Since this is, in part, an attempt at standardization, I would like to point out that many (3/5) of the constituent databases have recently been updated on lipdverse.org, and now use terminology that espouses the community-sourced LinkedEarth ontology (https://linked.earth/ontology/), which itself uses a number of controlled vocabularies (https://lipdverse.org/vocabulary/). Some of these vocabularies have been aligned to relevant terms in the NCEI PaST Thesaurus (https://www.ncei.noaa.gov/products/paleoclimatology/paleoenvironmental-standard-terms-thesaurus). To the extent possible, it would be good to align the terms used in this study (cf Table 1) to those standards, and refer to them in the text so readers are more aware of them.

Authors' response:

We have mapped our compact common dictionary, which was originally based on the PAGES2k (2017) terminology, to match the first level of the LiPDverse vocabulary (https://lipdverse.org/vocabulary/) as follows:

- Terms which we have changed to match LiPDverse:
  - 'climateInterpretation_variable' -> interpretation_variable
  - 'climateInterpretation_variableDetail' -> interpretation_variableDetail
  - 'climateInterpretation_direction' -> interpretation_direction (new in v2.0)
  - 'climateInterpretation_seasonality' -> interpretation_seasonality (new in v2.0)

    - 'paleoData_variableName' (new in dod2k v2.0: see below, in response to Nick McKay's comments: name of the variable derived from the proxy observation, which may be different from the proxy observation)
- Terms which are already in agreement with the LiPDverse terminology:
    - archiveType, paleoData_units
- Terms which have no clear match in the LiPDverse first level we have left the same.

However based on your suggestion we have also changed the terminology for the entries of the archiveType and paleoData_proxy (e.g. we renamed 'tree' to 'Wood' and 'd2H' to 'dD' to comply with LiPDverse terminology).

We hope these changes will make the usage of the database and the code more convenient for users.

With regard to the versions of the curated databases, we will update the versions ingested into dod2k as follows:

- PAGES2k: we propose to use the lipd serialization version 2.2.0 that is on lipdverse.
- Iso2k: update from iso2k v1.0.1 to iso2k v1.1.2 using lipdverse current version.
- Breitenmoser et al (2014): no update is available.
- CoralHydro2k: update from v1.0.0 to v1.0.1 using lipdverse current version.
- Sisal: no update, as we are using v3 (2024), and lipdverse is at v2.1.1, although as of this writing the directories at https://lipdverse.org/SISAL-LiPD/current_version/ are empty.

2) While the database itself will undeniably be useful for some applications, I believe the associated workflows are of greater value still. In particular, the workflow to identify and remove duplicates addresses a recurring issue in this line of work, and to my knowledge it is the first published instance of such a workflow being described in detail, and shared in code form.

Unfortunately, there is no universal standard for sharing workflows. It is very helpful that the authors made notebooks and auxiliary Python modules available through GitHub, but the notebooks are still a little rough around the edges (cf a lot of commented out old code) and lack a narrative. I would like to invite the authors to organize their cleaned-out notebooks as a JupyterBook, and share it through a gallery like PaleoBooks: https://linked.earth/PaleoBooks/. I believe the work will have greater visibility there, and will have more enduring value to the community.

Authors' response: We will clean up the notebooks and create a tutorial to bridge from the Quickstart Guide to the notebook workflow. We are using mkdocs (https://github.com/mkdocs/mkdocs) as mentioned in the response to reviewer 1 who made a similar request. We will discuss with linked.earth whether a PaleoBooks submission would be a useful contribution.

Editorial Comments:

The paper is well written, though I have a handful of suggestions.

> there is inconsistent terminology throughout the manuscript, sometimes referring to PAGES2K, or PAGES2k. The proper nomenclature is PAGES 2k (lowercase, with space).

Authors' response:

We certainly can agree that a consistent nomenclature is good practice. The PAGES 2k Network conforms to the suggested nomenclature when referring to the research network. The reviewer's own lead authored data descriptor refers consistently to PAGES2k (no space), as do Iso2k and Coralhydro2k, when referring to databases themselves produced by that network, and we will do the same here.

> L117: complimentary —> should be "complementary"

Authors' response:

We will make this revision. Thank you, here and elsewhere, for pointing out these errors.

> Table 2: it looks like the authors loaded the constituent databases from various static files. For most up to date information on PAGES 2k, Iso2k, and CoralHydro2k, it is recommended that they download the latest from lipdverse.org, as many updates were made this summer.

Authors' response:

Thank you. We were not aware of updates to the lipdverse versions, having sourced from the NCEI and other repositories; we will make these revisions (see also notes earlier). We will modify the load notebooks to ask whether to download from source before loading from the local copy.

> L167: "The evidently true duplicate records …". How many such duplicates were found, and what fraction of the total number does that represent?

Authors' response:

This is given in the submitted manuscript, l. 198, first results sentence: "DoD2k v1 consists of 4516 records (4841 before duplicate screening)." We will update this result along with the planned updating of component databases.

L238: “the sensor model in PRYSM (Dee et al., 2015),” —> It should be noted that this is the sensor model introduced by Partin et al (2013), http://dx.doi.org/10.1130/G34718.1.

Authors’ response:

Thank you, this is an important original citation to make, which we will add where indicated.

Section 4: it is odd to put results in the discussion. I recommend renaming Section 4 “Applications” and having a very short Section 5 called “discussion” or “outlook” that incorporates what is currently in Section 4.3. Conclusions could be merged in there too.

Authors’ response:

We will create a new Section 5 from Section 4.3 to clearly separate Results from Discussion.

L310: references should be parenthetical (\citep{}, not \citet{}).
L327: “ filtering by dictionary terms we have employed” —> filtering by THE dictionary terms we have employed (missing THE).

Authors’ response:

We will make these revisions.

In surveying S_analysis.ipynb, if appears that the authors re-invented the wheel in how they implemented something as straightforward as linear regression. I recommend that they use statsmodels (https://www.statsmodels.org/stable/), as it will provide all the information the authors need via model summaries, and lead to a more lightweight notebook, less prone to errors since statsmodels has been more thoroughly vetted.

Authors’ response:

We will update the speleothem analysis notebook for use of the statsmodels package, and update the results in the manuscript accordingly.

Julien Emile-Geay

Michael N. Evans
Lucie J. Lücke
Kevin J. Fan
Feng Zhu