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.)

RC3: 'Comment on essd-2025-364', Nicholas McKay, 06 Sep 2025

Citation: https://doi.org/10.5194/essd-2025-364-RC3

In this manuscript, the authors describe the rationale and methodology behind the assembly of a "Database of Databases" for paleoclimate for Common Era, and include two use cases to demonstrate the potential utility of this merged data product. The manuscript is well written and illustrated and addresses a common problem: how to use a collection of related, but not custom-built, data compilations to address a problem that could benefit from a larger collection of data.

The authors identify metadata integration, including non-overlap and terminology differences, and duplicate handling as the primary challenges in this exercise. This is consistent with my experience. The approach and methodology for identifying, handling and tracking the choices made in the deduplication process is well done, and a valuable addition to the literature.

The code and data to load in the databases, align their terminology, and remove duplicates is all available, as is the code for the use cases and figures. As always, it's great to have access to all of this to get into the details of how the authors did what they did. I thank the authors for following best practice here! Using the instructions I was able to run all the notebooks except one ("load_pages2k_vv2.ipynb") , which seemed to hang during the `pdb = cfr.ProxyDatabase().fetch('PAGES2kv2')` command.

Authors' response:

We will modify the load_pages2k notebook to load from the most recent lipdverse version, which is labeled pages2k v2.2.0 and has 647 lipid files. This includes the Palmyra (2013) update that was retrieved and used to replace the older Palmyra record, so this also serves to simplify the load_pages2k database and also remove this problem for users (which we are unable to reproduce).

Having the codebase is very helpful, however I do think that it would be a challenge for others to try to build on this approach to add additional or alternate compilations using the same design. It's certainly more of a reference with examples than it is a tool to easily create new compilations.

Authors' response: We do not share the reviewer's certainty. We hope to convince them with the requested tutorial and improved commenting. We are encouraged by the other reviewers'

enthusiasm for the detection and decision notebooks, and we also point to the applications examples (with the caveats noted by this reviewer and discussed below).

Overall, I think this is a worthwhile contribution and I think the community will find the DoD very helpful. Indeed, I expect the database itself (as opposed to the methodologies described to create) to be widely used and a starting point for many researchers keen to take a data-intensive look at many aspects of Common Era climate. And because of this expectation, I have a few suggestions to make that database both more useful, and less prone to misuse.

First, while I appreciate the need for a reduced set of metadata while integrating across datasets, there are three additional fields that I think are critically needed. Fortunately, this metadata is available in most of the original compilations so the authors would not be starting from scratch.

1. interpretation direction: A field that describes whether the variable is positively or negatively related to the interpreted variable. The need for this field is evident in the first use case, where the authors multiply the tree ring and coral d18O datasets by -1 to allow for more direct comparison with the other proxies. Although it's true that these relationships due vary by proxy and archiveType, there are many examples of variable interpretations within archive and proxy classes. Lake sediment d18O is one example where this is often variable between lakes. Critically, there are also several examples in the literature where the interpretation direction was not properly applied, and this can lead to substantially wrong conclusions. Given its importance, including explicit interpretation direction for each interpreted dataset (and interpretation) is critical. After adding these metadata, I suggest replotting the PCA results using these metadata rather than the class-specific data.

Author's response:

This is a good idea, and we will implement this. But if we are not mistaken, this metadata field only exists for the PAGES 2k database and not for the other four databases: at most about 692/4516 ~ 15%.

We propose to keep this application subsection, but revise the discussion of the results to reflect the reviewer's concerns. We would note, as opposed to composite averaging, that a PCA on standardized proxy observations in their native variable simply identifies patterns of correlation, which could be either positive or negative in sign, between records. Our analysis did so by both archive and observation type in the largest subsets (Table 3).

We acknowledge that there might be differences in sign of linear regression coefficients across, for example, lake sediment d18O records, as the reviewer states; we would appreciate citations in support of this claim. And we acknowledge that the comparison of sign across the different PCAs shown is problematic, because the sign of each EOF pattern is arbitrary.

We also believe the results may disagree within and possibly across the possible comparison subsets, because there might not be a simple linear mapping or scaling from one environmental variable to the proxy observation, especially for observations which are influenced by both moisture and temperature in different ways and proportions.

Thus in agreement with the reviewer, we would expand on the uncertainties in doing so and interpreting the results, as we have just discussed. We originally wrote:

"However, perhaps because of differences in observational networks, time resolution and/or covariance estimation interval, there appears to be little agreement between PC1 and PC2 across archive and observation subsets. Although timeseries of the mean of PC1 show, at times, some agreement for certain archives (Figure 6 B), there is no agreement regarding PC2. All this suggests careful 280 additional analysis may be needed before a multi-archive, multi-observational analysis is performed and interpreted… Although there is some degree of regional spatial agreement of EOF sign within and across archives and observations, there are also many instances of disagreement of EOF sign within and across archives and observation types. Because spatial patterns may also be sensitive to observational network and the potential for both T and M influences in this subset, we again assess that more analysis within these archives and data types is needed before we can identify large scale patterns across the multi-archive and multi-observation database."

We would expand that discussion as described. We will include the reviewer's point that as as community, we may need to go back to the original publications to add metadata on the calibration and/or interpretation of individual records, enabling use of this information in such studies, before further analysis is warranted.

2. Seasonality: Many of the datasets are interpreted to be more sensitive to climate variability during some parts of the year than others, and it's very valuable to be able to filter by interpreted seasonality to test various hypotheses.

Authors' response: we agree this would be valuable, and we will add this. But again, the metadata have not been compiled for any of these datasets except PAGES2k, which is a small subset of the dod2k.

3. Variable name: The database includes both units (what units the variable was reported in) and proxy, which describes the type of physical, chemical, and/or biological systems that imprint climatic condition onto the archive, but not the name of the variable itself. This is often similar (or identical) to the proxy, but not always. For example, chironomid data are measured as count or assemblage data, but are included in the database as a calibrated temperature variable. So the proxy is correctly "chironomid" and the units are correctly "degC", but the variable name should be "temperature" or something similar. This is particularly useful for making coherent axes labels using the metadata form the database.

Authors' response: We see the reviewer's point - in addition to the example cited, there are for instance some tree-ring width temperature reconstructions that are included in the PAGES2k

database which we would ingest as tree-ring width data without this field. We propose to add an additional dictionary term, paleodata_variableName, that can be defined as name of the variable derived from the proxy observation, which may be different from the proxy observation (see above), which is in paleodata_proxy. This may help make the codebase and database metadata more useful for future users and for when these distinctions are important in future versions of component databases.

My final major suggestion is that, to the extent possible, citation information for the original studies is provided in the manuscript and in the database. These data compilations are critical for large-scale study of the Common Era, but also tend to make it more difficult for the authors of the original study to be credited. Users will likely be unable to cite all of the studies if they're using the whole databases, but many will only use a subset and it's helpful to enable citation of the original studies whenever possible. Ideally this would be an additional field in the database.

Authors' response: As reported in Table 1 of the manuscript, this information is already included in the dod2k in the field: originalDataURL: Original data URL URL/DOI for each record. For instance, the first record of the iso2k database has originalDataURL:

'http://doi.pangaea.de/10.1594/pangaea.676709'

which points us to the full reference for citation.

However, for the fe23 dataset, originalDataURL was pointing to each flat Arizona format .rwl file, which has only the contributor name, but not the full original reference. We will modify the fe23 load notebook to instead point to the NOAA template rwl file, which also includes the original reference information as supplied to NOAA/NCEI by the scientists who deposited the data in the repository.

We recognize the reviewer's request to also potentially list the more than 4,500 individual original sources in the reference list for the manuscript. We feel this would be unwieldy but also redundant to listing the originalDataURLs for each record, and to what is listed in each of the curated component databases and their published data descriptors.

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