https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

Open Access Earth System Science Data Discussions

A database of databases for Common Era paleoclimate applications

Michael N. Evans[1,2], Lucie J. Lücke[2], Kevin J. Fan[3], and Feng Zhu[4]

[1]Department of Geology and Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA
[2]School of Geosciences, University of Edinburgh, Edinburgh, UK
[3]Computer Science and Geology, University of Maryland, College Park, MD, USA
[4]NSF National Center for Atmospheric Research, Boulder, CO USA

Correspondence: Michael N. Evans (mnevans@umd.edu)

Abstract. We present a database of curated databases (DoD2k version 1) developed for Common Era (1-2000 A.D.) paleoclimate research. The DoD2k leverages existing community efforts, many of which arise from the PAGES (Past Global Changes) 2k working group, and the codebase developed by the paleoclimate data informatics communities over the past decade. Using a common, compact set of terms for metadata and data management, we merge five existing curated databases. These individual curated databases represent a range of approaches, from single archive-single observation to multiarchive-multiobservation collections, and span a total of 14 archives, 49 data types, and 4613 records within the Common Era. We then use a multistage algorithm to remove duplicates, checking against a common set of metadata and comparison metrics. We illustrate the value of the DoD2k with two applications. In the first, we extract the moisture and temperature subset of records and perform an empirical orthogonal function (EOF) analysis on the resulting multi-archive, multi-observation dataset. In the second, we show that calcite speleothem oxygen isotopic composition is consistent with proxy system simulations. DoD2k may also be useful for paleoclimatic detection and attribution analysis using proxy system modeling, data assimilation, and deep learning for the development and testing of improved proxy system models.

1 Introduction

The climate is changing, forced primarily by human-caused increases in greenhouse gas concentrations, aerosols and land use change, toward a warmer and more moisture-inequitable state, in which extreme events are more likely, and more extreme, than observed during the 20th century (Arias et al., 2021). Superimposed on that are other causes of climate variation and change, for instance, arising from volcanic activity, solar and orbital variations, as well as the tendency of the climate to wander without any forcing at all: the internal, unforced variability. If the former is thought to be well understood, the latter is not: how the climate system, broadly defined as the coupled ocean, atmosphere, land surface, land and sea ice, biota, and solid earth integrates and responds to such natural forcings, may take tens to thousands of years to be fully realized (Miller et al., 2012; McGregor et al., 2015; Abram et al., 2016; Gebbie and Huybers, 2019). How then, to define both the spatial imprint and the amplitude of the climate change that arises from such forcings, and distinguish it from the human-driven forcings? The answers are important: first, for defining the equilibrium and transient climate change in response to a unit of forcing, over time and in different parts of the world (Forster et al., 2021); second, for detecting and projecting the impacts of both anthropogenic and natural climate

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

forcing over past and future decades and centuries (Fox-Kemper et al., 2021; Marvel et al., 2019b, a). For such goals we need realistically forced paleoclimate simulations and observations (Neukom et al., 2019a).

The development of the observational target for such work is the focus of the present contribution, in particular for the so-called Common Era (1-2000 CE), for which observations from paleoclimatic archives are most dense and diverse, and permit an approximate 10-fold increase in the time interval of study relative to the historical record. More specifically, we would desire the most dense and random sampling in space and time, of all possible observations, imprinted with a diverse set of climatic information, and resolving timescales of variation from subannual to multicentennial with similar observational temporal resolution, and with well-characterized chronological uncertainty. The natural starting point for such an effort would be public repositories of individual paleoenvironmental datasets and databases, such as at the National Center for Environmental Information (https://www.ncei.noaa.gov/products/paleoclimatology) and PANGAEA (https://pangaea.de). However, this is impractical for multiple reasons, including some nonuniformity in dataset submissions and metadata, changes in repository submission templates and requirements over time, and the presence of multiple versions of datasets in repositories whose prime directive is preservation and availability (Anderson et al., 2019).

An alternate foundation for the development of such a database is in existing databases, many of them compiled by years-long efforts by PAGES (Past Global Changes; www.pastglobalchanges.org) Working Groups. PAGES databases are the result of leveraging community-level specialist expertise that is difficult for any single research group to assemble or maintain. The work of many individuals in multiple groups has enabled the development of publicly available observational datasets, the metadata that describes them, and most recently, the open semantic formalisms (Emile-Geay and Eshleman, 2013) and codebases (McKay and Emile-Geay, 2016) that enable their re-use. However, each such database, although rich in metadata, metadata uniformity, error checking and quality control, is generally assembled for a specific purpose. For example, the PAGES 2k Consortium (Consortium, 2013; Emile-Geay et al., 2017a) originally planned development of a multiarchive database (wood, coral, ice, documents, lake and marine sediments) of many different temperature-sensitive observations in these archives for the purpose of global mean and spatially resolved surface temperature reconstructions (Neukom et al., 2019a, b). The SISAL Working Group (Kaushal et al., 2024a) developed a single (speleothem) archive of multiple observations (e.g. δ^{18}O, Mg/Ca) made in that particular archive. The Iso2k Working Group (Konecky et al., 2020a) produced a multiarchive (marine sediment, lake sediment, marine carbonate, speleothem carbonate, wood, ice) database of solely δ^{18}O and δD observations in those archives, agnostic of climatic interpretation. For facilitating the repurposing of these databases for other scientific goals, such as the reconstruction of hydroclimatic variability (Falster et al., 2023) and the detection and attribution of climate change in both moisture and temperature via paleoclimatic data modeling (Franke et al., 2022a), we might need to combine multiple existing databases.

There are multiple challenges to creating a unified database of databases for Common Era paleoclimate applications. These include differences in the metadata and terminology for describing datasets across databases but within even the same proxy observation types and biogeochemical archival materials; differences between databases of the required metadata, sampling resolution, age model development, time resolution, level of replication, descriptions of observational uncertainty, and interpretational notes; and the problem of duplicate detection across combined databases (Anderson et al., 2019; Tardif et al.,

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

2019; Steiger et al., 2022). Unfortunately, differences in metadata and terminology for defining properties across different curated databases, as well as differences in database terminology, structure, management, make merging databases and cleaning them for duplicates difficult. For instance, PAGES2K (Emile-Geay et al., 2017a) has 173 dictionary terms, and as it happens, SISALv3 (Kaushal et al., 2024a) has 173 unique dictionary terms linking its 21 constituent csv files into a database. However, these are not the same 173 dictionary terms as for PAGES2k. Although there is common overlap in metadata, such as site identification name, they have different keys (`'paleoData_TSid'` in PAGES2K, `'site_id'` in SISALv3).

Here we describe an approach to resolving these challenges with a flexible and open framework, implemented as Python functions, scripts and Jupyter notebooks, in which a standard set of metadata by which to merge existing databases could be specified (McKay and Emile-Geay, 2016), and in which duplicates across merged datasets could be identified and removed. The framework is extensible and can incorporate new databases or updates to existing ones. It provides methodical and comprehensive testing for duplicate records and can be easily attached to paleoclimatic analysis and reconstruction toolsets (Zhu et al., 2023, 2024b).

The existing Common Era paleoclimate datasets to be merged are briefly described in section 2.1. The data, approach and the code base for merging are described in section 2.2. The resulting Database of Common Era paleoclimate Databases (hereinafter, DoD2k) is described in section 3, and some applications are illustrated in section 4. Conclusions and an outlook for future development are in section 5.

2 Data and Codebase

2.1 Data

As a target, we assemble five such databases of Common Era paleoclimate records.

- Breitenmoser et al. (2014a); Franke et al. (2022b) (hereinafter, fe23): restandardization of tree-ring width (TRW) chronologies for comparison with their 20th century simulation using the VS-Lite data model (Tolwinski-Ward et al., 2011), with data available within the interval 850-2000 CE and with climatic interpretations re-estimated as published in a subsequent study (Franke et al., 2022a);

- Emile-Geay et al. (2017a, b): PAGES2k (hereinafter, p2k): multiproxy, multiarchive compilation of temperature-sensitive proxies for the Common Era, updated for records from Palmyra Atoll (Dee et al., 2020);

- Konecky et al. (2020a); Konecky and McKay (2020): Iso2k (hereinafter, iso2k): multiarchive compilation of oxygen and deuterium isotopic records extending through the Common Era and into previous periods;

- Walter et al. (2020, 2022): CoralHydro2k (hereinafter, ch2k): single archive, multiproxy compilation of records from coral carbonates, within the Common Era;

- Kaushal et al. (2024a, c): SISAL2k (hereinafter, sisal): single archive, multiproxy compilation of cave carbonate records, extending from the present through the Common Era and into previous periods.

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

Open Access Earth System Science Data Discussions

2.2 Workflow and Codebase

An overview of the workflow for creating DoD2k is shown in Figure 1. DoD2k as well as the codebase for its production are supplied as a GitHub repository (https://www.github.com/lluecke/dod2k). The codebase is written in Python 3, and includes a series of jupyter notebooks and Python scripts, which read in the original databases, concatenate to a large database, perform a duplicate search and eliminate duplicates, output the product (DoD2k), and provide some summary plots (Figure 1). Thus, we here present not only DoD2k, but also the opportunity for users to modify the source code based on their specific requirements, including but not limited to, add data, add variables, and make their own expert decisions regarding the potential duplicate candidates. In the following paragraphs, we present an overview of the DoD2k database and the associated Python utilities.

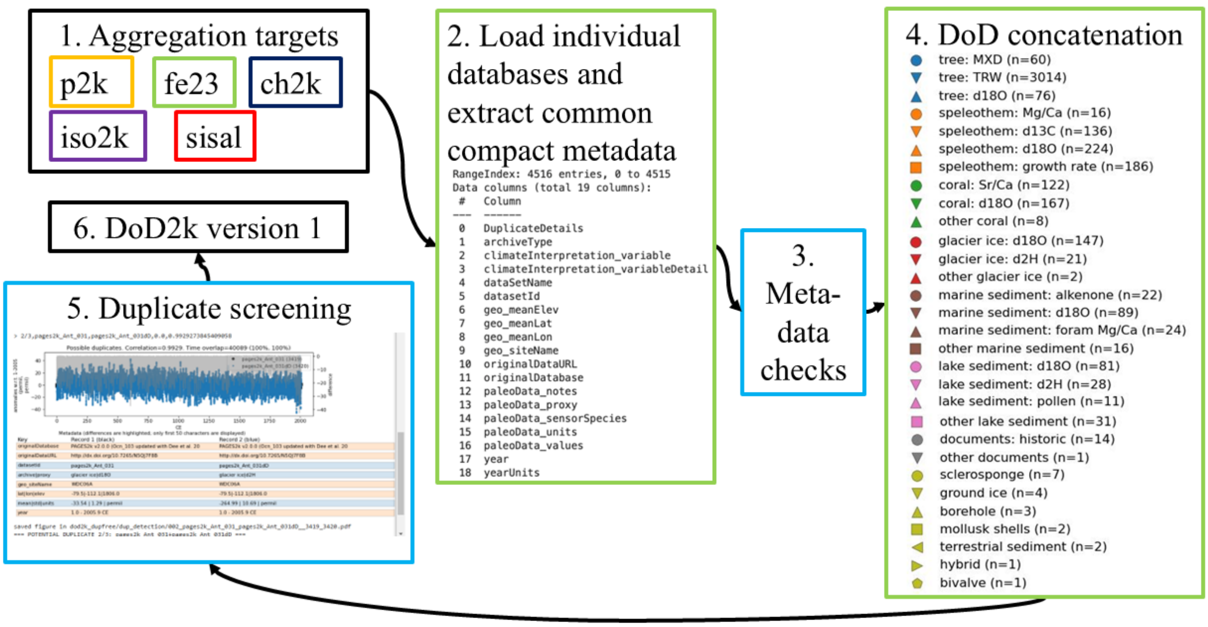

Figure 1. Schematic overview of the DoD2k workflow. From upper left: starting with the databases to be aggregated, we load each dataset as a compact common subset of the metadata and data using a standard set of dictionary terms, if necessary translating from the original terms. We then concatenate the data, perform basic metadata checks, and check for duplicates (optionally with operator review and temporal compositing) before creating the DoD2k. All operator choices are journaled and the notebooks may be commented for subsequent review and for reproducibility of processing and traceability back to original databases and their entries (Bush et al., 2020).

The codebase consists of two key parts. The first includes loading the original databases and their concatenation to a common database (referred to as the 'load notebooks'). The second involves the duplicate detection process. In addition, we supply several notebooks for data visualization purposes, and two notebooks for applications described in section 4. An overview of the notebooks and scripts associated with the key parts of the DoD2k development process (Fig. 1) is given in Table 2.

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

Open Access Earth System Science Data Discussions

2.3 Environment

A virtual environment for running the Python functions, scripts and Jupyter notebooks built within a Jupyterhub installation (`https://tljh.jupyter.org/en/`) can be found at the aforementioned github repository in the file `cfr-env.yml`. To reproduce the environment, users just need to type the following command: `conda env create -f cfr-env.yml -n cfr-env`.

2.4 Metadata fields

To assemble a database of databases, we identified a set of 16 metadata fields (Table 1) which satisfy the following criteria across all the individual databases. Apart from a few fields specific to DoD2k, the nomenclature of the fields was largely adapted from Emile-Geay et al. (2017a) and the PAGES2k v2.0.0 temperature-sensitive database. Criteria were that the metadata is commonly used within the community, and the majority of the original databases have non-missing entries within the field. For a description of how the individual metadata parameters were collected, we refer back to the original databases and their development by teams of specialist researchers. Note, however, that this set of fields is by no means exhaustive, and if desired by the community or by an individual user, may be expanded. For example, the p2k database has 173 metadata fields. However, half of these fields are missing for 85% of entries, making it more difficult for users to identify and extract relevant metadata. We have tried to keep the majority of fields populated, and only `climateInterpretation_variableDetails` and `duplicateDetails` have a significant number of missing entries in DoD2k. However, both these fields are complimentary with the only purpose to provide extra information where needed.

2.5 Load notebooks

The load notebooks read the five original databases (ch2k, fe23, iso2k, p2k and sisal), extract a number of shared variables, and concatenate the databases. Each notebook follows the structure:

- set up environment, load source data
- potentially process data according to database provider (may use code supplied by authors of the original database) to obtain a `pandas` dataframe
- identify and extract the relevant variables
- convert metadata and data to the correct format
- save output as a “compact” `pandas` dataframe with standardized metadata

The output of each load notebook is a standardized “compact” set of 18 metadata and data fields (Table 1). Output is saved in pickle (.pkl) format as well as a series of comma separated value files. A unique `datasetId` is pulled from the identifier used in the component curated dataset. Each additional metadata field contains either information from the original dataset as published, or has been added for the purpose of improving data identification. In the case of the former, some information has

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

Key	Variable	Description & Comments
`archiveType`	Proxy archive type	E.g. tree, glacier ice, speleothem, coral, ...
`climateInterpretation_variable`	Climate interpretation variable	temperature or moisture sensitive (or both or neither)
`climateInterpretation_variableDetail`	Details on climate interpretation variable	Data available for Iso2k and PAGES2k.
`dataSetName`	Record name	May be identical to dataset ID or site name. Non-unique.
`datasetId`	Unique Record Identifier	Uses original ID, modified to make values unique, refers to original database.
`geo_meanElev`	Mean elevation	(m asl)
`geo_meanLat`	Mean Latitude	(deg N)
`geo_meanLon`	Mean Longitude	(deg E)
`geo_siteName`	Site name	Site names may differ for duplicate records.
`originalDatabase`	Original database	PAGES2k, FE23, CoralHydro2k, Iso2k, SISAL
`originalDataURL`	Original data URL	URL/DOI for each record
`paleoData_notes`	Notes	Notes from original authors (if applicable)
`paleoData_proxy`	Proxy measurement	E.g. MXD, d18O, d13C, Sr/Ca, pollen
`paleoData_sensorSpecies`	Proxy archive species	Available only for biological proxies.
`paleoData_units`	Units of proxy data	e.g. mm, mmol/mol, ‰
`paleoData_values`	Proxy data	Data time series for each record
`year`	Year	Time coordinate for `paleoData_values`. Year≥1 CE
`yearUnits`	Year units	All records transformed to CE
`DuplicateDetails`	*Notes on duplicates*	*Applicable only for duplicates. Saves associated duplicates, decisions and operator comments.*

Table 1. Metadata and data parameters of DoD2k, by common compact dictionary key. Note that the Key *DuplicateDetails* is added as a result of the duplicate screening process (section 2.6).

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

notebook	description	input/output
load notebooks		
`load_pages2k.ipynb`	Load source data, extract key variables (see common dictionary), where necessary format metadata, standardise entries and convert data, create standard 'compact' dataframe, save as csv.	In: `PAGES2k_v2.0.0ts.pklz` (Emile-Geay et al. (2017a)), utilizes *cfr* (Zhu et al., 2024b) to update Palmyra record.
`load_fe23.ipynb`		In: `fe23.nc` (Breitenmoser et al., 2014b; Franke et al., 2022a).
`load_sisal.ipynb`		In: `sisalv3_database_mysql_csv.zip` (Kaushal et al. (2024a))
`load_ch2k.ipynb`		In: `CoralHydro2k1_0_1.zip` (Walter et al., 2020)
`load_iso2k.ipynb`		In: `iso2k1_0_1.RData` (Konecky et al., 2020b) Out: `iso2k_compact_PARAMETER.csv`
duplicate screening		
`dup_detection.ipynb`	Loop through the dataframe, flag candidate duplicate pairs and produce summary figures of candidates.	In: `dod2k_compact_PARAMETER.csv` (original data matrix) Out: `dup_detection_candidates_dod2k.csv` (list of potential duplicate candidates)
`dup_decision.ipynb`	Loop through the candidates, create summary figures. Automatically reject fully identical records. Operator to make decision otherwise. Decisions are saved as a summary csv.	In: `dup_detection_candidates_dod2k.csv` Out: `dup_decisions_dod2k_INITIALS_YY-MM-DD.csv` (spreadsheet of decisions for each candidate pair)
`dup_removal.ipynb`	Apply the decisions made in the previous step. Produce a duplicate free dataframe, save as series of csv files.	In: `dup_decisions_dod2k_INITIALS_YY-MM-DD.csv` Out: `dod2k_INITIALS_YY-MM-DD_dup_free_PARAMETER.csv` (duplicate free data matrix)

Table 2. Notebooks for the creation of DoD2k and in-/output data (PARAMETER: paleoData_values, year, metadata; INITIALS: Operator's initials as specified in the notebook; YY-MM-DD: date).

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

been homogenized to a standard set of values across the databases, e.g. in case of the `archiveType`: 'tree' to identify all tree-ring type proxies, or for `climateInterpretation_variable`: 'T' to identify all temperature sensitive proxies. Thus, some of the information may slightly differ from the original authors in terms of nomenclature. However, this is user customizable via modification of the load notebooks. The main purpose is to allow user friendly filtering for individual values, and thus easy access to e.g. a database for an individual archive type. We have also transformed chronological assignments such that they are all provided in units of years CE, and we have restricted the data to the Common Era, 1-2000 CE. The resulting homogenized original databases are subsequently assembled into DoD2k, and are ready for duplicate screening.

2.6 Duplicate screening

The duplicate detection process consists of three separate steps, each implemented as a separate notebook. This process uses the output of any of the load notebooks and may also be applied to the assembled database. The first step, duplicate detection (`dup_detection.ipynb`), identifies the potential duplicate candidates. Duplicate detection begins with a simple threshold correlation coefficient test between any two records with `paleoData_values` x and y. We choose $\mathrm{corr}(x,y) > 0.98$, which we consider sufficient to find exact duplicates, but also allows detection of records which, for example, might differ by only a few points at the beginning or end of the observed time interval, have a limited number of missing values, or contain a different number of retained significant digits. However, this criterion is not sufficient on its own (as discovered in analysis of the Steiger et al. (2022) database) to identify duplicates that might arise from standardization choices, compositing, truncation, and metadata differences that might differ across databases. We therefore add multiple additional diagnostics, metadata comparison, and operator screening of candidates. Here, any two records with `paleoData_values` x and y, and their associated z-scores z_x and z_y, are scanned according to the following necessary criteria:

1. data type: archive type (`archiveType`) and proxy type (`paleoData_proxy`) must be identical,
2. site: there must be an overlap in the site name (`geo_siteName`), e.g. at least one shared expression,
3. location: distance between the point coordinates (`geo_meanLat` and `geo_meanLon`) ≤ 8 km,
4. correlation: $[\mathrm{corr}(x,y) > 0.9 \vee \mathrm{corr}(z_x,z_y) > 0.9] \wedge [\mathrm{RMSE}(x,y) < 0.1 \vee \mathrm{RMSE}(z_x,z_y) < 0.1]$
5. overlap: time intersect (`year`) at least ≥ 10 points, unless one of the records is overall shorter than that, in which case the other criteria are sufficient,
6. URL: the data identifier (`originalDataURL`) are identical if both candidates originate from the same database (`originalDatabase`), otherwise the other criteria are sufficient.

The so-detected candidate pairs are flagged, and their summary figures as well as a list of the candidate pairs are saved. In addition to the 18 fields previously described, the assembled and duplicate-screened database DoD2k contains the field `duplicateDetails` (Table 1). This field is populated during the duplicate screening process and includes information on the detected duplicates and any choices made by the user, such as the selection of one record or the other of the candidates, and whether a composite record has been created (see below).

The second step, the decision process (`dup_decision.ipynb`), uses the output generated in the previous step to perform a decision process on each candidate pair. In order to make this process more operator-friendly, we have implemented a mixture of automated and user-operated decisions. For candidate pairs which are evidently true duplicates, no user input is necessary. Instead, a record is removed, based on a user-specified database hierarchy (see next paragraph). The evidently true duplicate records satisfy the following criteria:

7. `geo_meanLat` and `geo_meanLon` agree up to a decimal place,
8. `geo_meanElev` agrees up to 1 unit,
9. `archiveType`, `paleoData_proxy`, `originalDataURL` and `geo_siteName` are identical.

These user-specific criteria were empirically determined during the decision process and are thus tailored to our purposes. They were based on a number of true duplicates, where the geographical location differed slightly due to different authors' precision in the specification of coordinates.

We also specify a hierarchy for adoption of duplicate records from databases, with the assumption that newer databases were improvements over older ones, i.e. sisal > ch2k > iso2k > fe23 > p2k (updated with the Dee et al. (2020) record). In addition, no user input is needed either for cases in which one of the candidates is evidently a recollection/update of the other record. Here, the `geo_siteName` is scanned for the keywords 'update' and 'recollection', and to determine if the location criterion (see #3 in the list above) is satisfied.

For any remaining candidate pair, the notebook operator is required to make a manual decision based on a summary figure of the duplicates (Fig. 1, step 5). The manual decision encompasses the following choices: keep record #1 and remove record #2 (and vice versa), keep both records, remove both records or composite both records. The latter choice is useful for individual records which were either composited by the original data producer, or represent a site-local concentrated data collection which represents paleoclimatic information from a particular site over time, and for which treatment as unique records is less valuable (Mann et al., 1999; Cobb et al., 2013). In the case of manual decisions, the operator is encouraged to leave a comment regarding the decision made for each candidate pair, in order to ensure that the decision process remains in retrospect comprehensible and reproducible. This note is later saved in the field `duplicateDetail` (Table 1).

The final step of the duplicate screening process consists in the removal of the duplicates (`dup_removal.ipynb`). Here, duplicate records are removed or composited from the database according to the operator's decisions. For the compositing process, the duplicate pair is standardized to z-scores and averaged over their intersecting time period, and their metadata is joined. After the removal and compositing is finished, the duplicate screened database is saved as comma separated value files, respectively: `year`, `paleoData_values`, `metadata` (17 fields) and a README file containing information about the duplicate screening process, such as operator name and correspondence details, date and operator's comments. Because the decision process is dependent on the operator's choices, the operator details should always be provided alongside the database to ensure transparency and traceability.

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

Open Access Earth System Science Data Discussions

3 Results

3.1 Spatial and temporal coverage

DoD2k v1 consists of 4516 records (4841 before duplicate screening). This can be compared to p2k (692 records), fe23 (2754 records), iso2k (596 records), ch2k (272 records), and sisal (546 records). As is true for the parent databases, the spatial distribution (Figure 2) is heterogeneous and represents mid and high northern latitude terrestrial regions more densely than the oceans, Saharan Africa, subtropical Eurasia, the tropical oceans, the Pacific Ocean and the Southern Ocean.

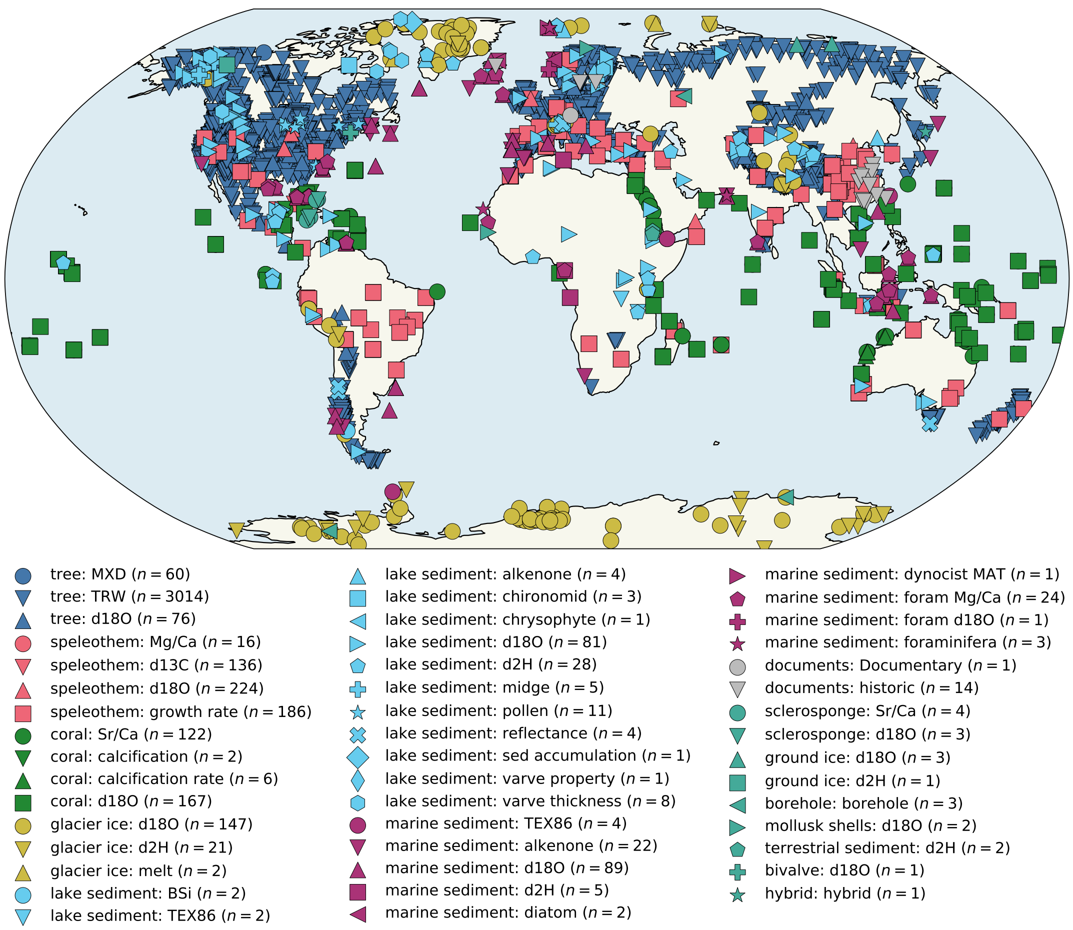

Figure 2. Spatial distribution of all available proxy records after execution of load, concatenation, duplicate detection and operator supervised decision and removal notebooks. Symbols indicate archive types and legend captions indicate observation types.

Although the total number of records available are more than 6 times those available in p2k, p2k patterns in temporal availability persist (Emile-Geay et al., 2017a, Figure 3). The total DoD2k record number is highly dominated by the wood (tree) archive, in which the vast majority of observations are tree-ring width (TRW), with 3014 TRW records compared to 60

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

maximum latewood density records (MXD) and 76 δ^{18}O isotope records. Tree/wood archives in particular represent the most dominant record type after around 1500 CE, when the number of available tree samples rapidly increases. Before this, other archive types are more abundant, especially speleothems, followed by marine and lake sediments and glacier ice. Speleothem records include 16 Mg/Ca, 136 δ^{13}C, 224 δ^{18}O isotope and 186 growth rate type proxies. Speleothems and marine and lake sediments, in addition to a wider range of observation types included in DoD2k relative to p2k, supply data records on longer timescales, often reaching even further back than the Common Era, and therefore offer a relatively constant coverage over the Common Era, relative to that afforded by p2k (Figure 3). Although the total number of coral records is also relatively high, these, like records from trees, drop out relatively quickly when moving back in time, and exist only in the most recent millennium. Other rarer archive types include documents and sclerosponges as well as boreholes, hybrid, bivalve, ground ice, mollusk shells, and terrestrial sediments.

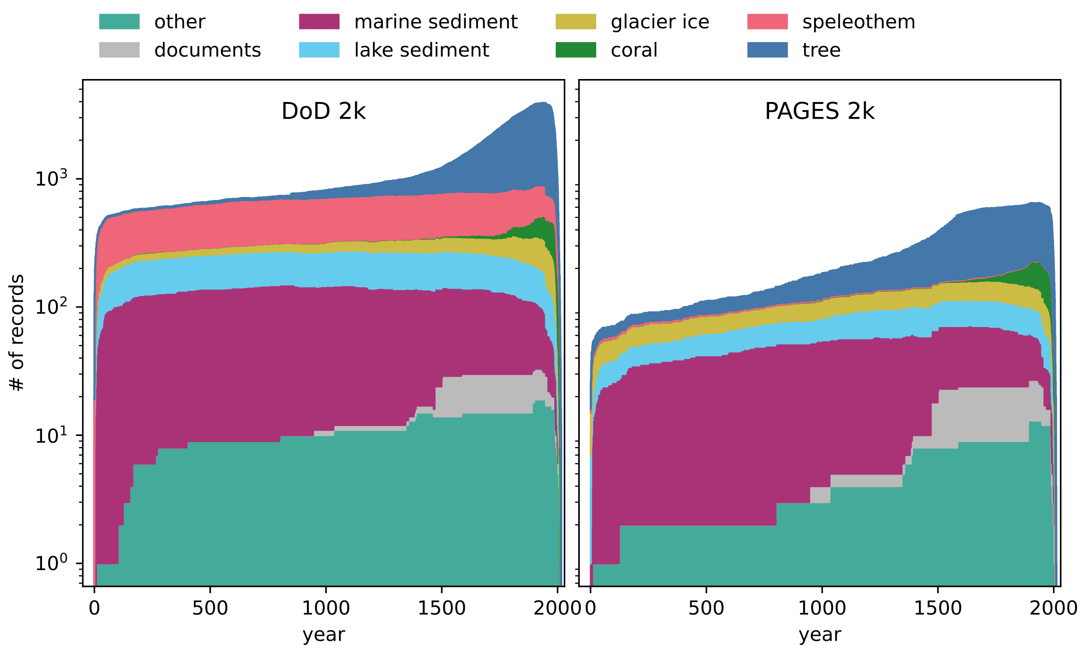

Figure 3. Temporal distribution of the DoD2k records by archive type. Records from boreholes, hybrid, bivalve, ground ice, mollusk shells, and terrestrial sediment archives, including all such observations made in those archives, as shown in Fig. 2, are grouped for plotting as "other" for simplicity of presentation. For comparison, right panel shows record availability by archive type from the p2k T-sensitive database (Emile-Geay et al., 2017a). Note the log scale for total numbers over time.

3.2 Applications

The development of DoD2k via open-source Python functions and scripts, and as Jupyter notebooks, not only allows for future growth of the aggregate database, but also extraction of subsets of data for specific paleoclimatic analysis. Here we

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

illustrate this with two applications, one utilizing `climateInterpretation_variable` to filter the data, and the second, utilizing `archiveType` and `paleoData_proxy` to provide a target for data modeling (Evans et al., 2013).

3.2.1 Hydroclimate record selection

For the analysis of specific climate conditions, the database can be filtered for `climateInterpretation_variable`. Around two thirds of DoD2k records have non-missing entries. In the Jupyter notebook entitled `df_filter.ipynb`, we have coded these from the original entries to be listed in DoD2k as 'temperature' (T), 'moisture' (M), 'temperature+moisture' (T+M) and 'NOT temperature NOT moisture' (∼T∼M). For hydroclimatic reconstruction purposes (Kaushal et al., 2022), we can filter the database for M-only or T-only records, but find that only a small subset of records are available and suitable for PCA analysis from a small subset of archives and observation types (see notebooks `M_analysis.ipynb`, `M_analysis.ipynb`, results not shown here).

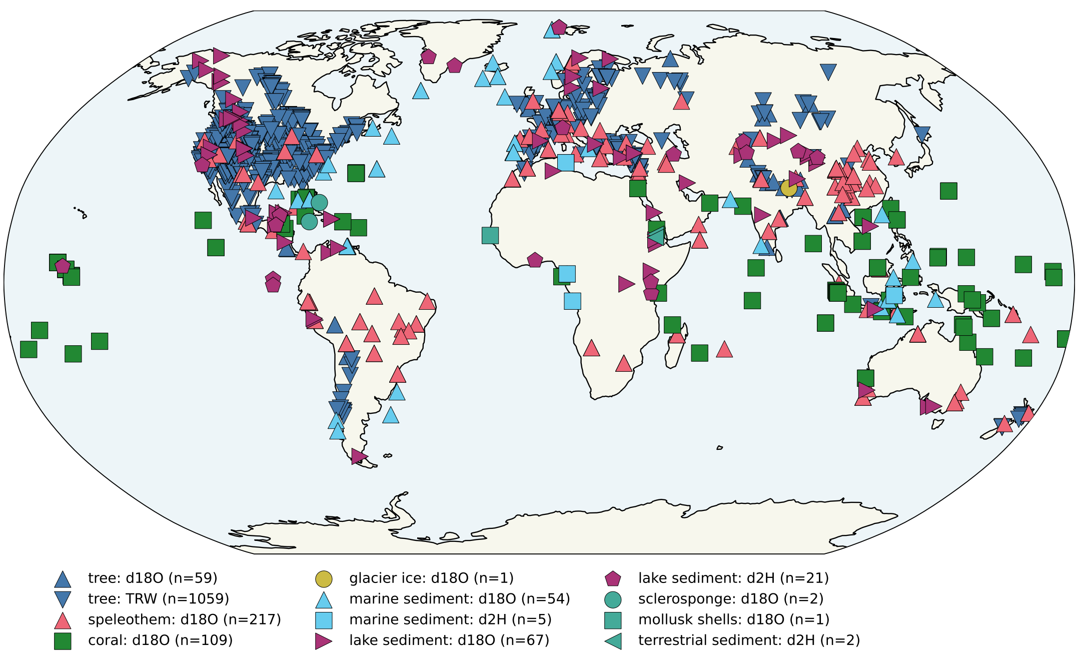

Figure 4. Spatial distribution of the available proxy records with `climateInterpretation` of moisture (M) or moisture and temperature sensitive (MT). These data are not drawn from p2k by definition (see text), but are from fe23 (1054 records), iso2k (266 records), ch2k (92 records), and sisal (179 records), and 6 DoD2k-constructed composite standardized records.

For M and T+M records, resulting in around 1597 hydroclimate sensitive records, a subset we designate as 'MT', data are available primarily in the Northern Hemisphere, but also in the tropics and at lower latitudes in the Southern hemisphere (Figure 4). Out of these records, 92 originate from ch2k, 179 from sisal, 266 from iso2k, 1054 from fe23, and 6 were standardized composites formed from these four databases during the operator duplicate decision process. Although the possibility that some p2k records are temperature *but also* moisture sensitive cannot be excluded, as noted by Emile-Geay et al. (2017a), none

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

of these records are taken from the PAGES2k temperature-sensitive database, by its own definition, unless those records were tagged as T+M in the other component databases.

3.2.2 Speleothem δ^{18}O record modeling

We provide a second application for data modeling within a specific archive and observation type, the oxygen isotopic composition (δ^{18}O) of cave carbonates (speleothems). We first filter DoD2k for only speleothem δ^{18}O records. We then simulate δ^{18}O of speleothem calcite using the sensor model in PRYSM (Dee et al., 2015), which takes as inputs observed δ^{18}O of precipitation, surface temperature, and mean transit time τ between the surface and the speleothem drip, assuming a well-mixed aquifer source. The archive model is a simple rescaling of the temperature-mediated dripwater δ^{18}O to account for the difference in scales for the precipitation source (SMOW reference) to the calcite observation (PDB reference). We compare results across the spatial gradients in climatological environmental inputs to assess the null of no difference between simulated and observed speleothem calcite δ^{18}O ($\delta^{18}O_{cs}$ and $\delta^{18}O_c$, respectively). Thus in some ways, this is an extension of the approach taken by Okazaki and Yoshimura (2019) to speleothem calcite δ^{18}O observations, but without the intermediate step of using an isotope-enabled atmospheric model nudged to observed sea surface temperatures; it is similarly an approach simplified relative to that of Bühler et al. (2021), who compared past millennium (850-1850 CE) climate model simulations of $\delta^{18}O_p$ to air temperature-corrected $\delta^{18}O_c$ observations from 85 calcite speleothem records from 71 sites, using the PRYSM speleothem sensor model to estimate transit times from surface conditions to cave calcification.

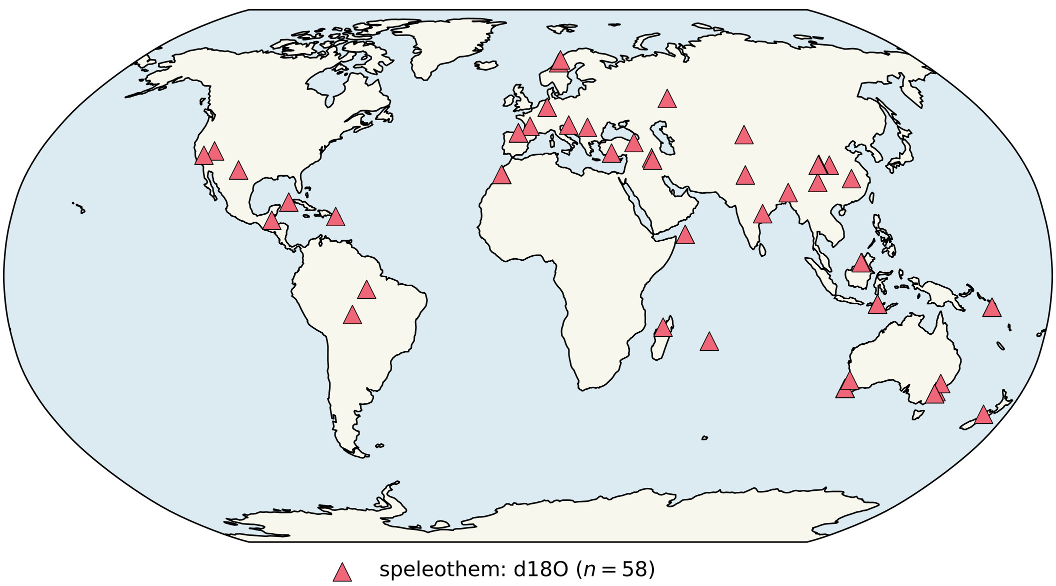

Figure 5. Map of locations for which speleothem calcite δ^{18}O observations are available from the DoD2k within the period 1960-2005 CE.

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

We perform this exercise in the Jupyter notebook `S_analysis.ipynb`. When we filter DoD2k for `paleoData_proxy` and `archiveType`, we find that the `paleoData_values` come primarily from the `SISALv3` database but also the `PAGES2k` and `Iso2k` compilations. Within the comparison period with simulations, 1960-2005 (see below), there are 107 such records available (Figure 5) from all continents except Antarctica.

For speleothem calcite δ^{18}O modeling, we import the `psm` and `cfr` python packages (Dee et al., 2015; Zhu et al., 2024b). We use as inputs the gridded mean annual precipitation amount-weighted climatologically averaged and interpolated terrestrial δ^{18}O product of Bowen et al. (2005), with estimates taken from the product grid point nearest to the observations (Fig. 5). As this product is based on precipitation amount and δ^{18}O for 1960-2005, we use this time interval for averaging of other environmental inputs and for the estimation of climatological mean δ^{18}O of observed speleothem calcite. For surface temperature we use similar nearest gridpoint estimates from the 0.5^o x 0.5^o CRUTS v4.08 surface temperature product (Harris et al., 2020), and average to the same time interval. Because the time interval is short and we average over time, we specify a mean transit time τ of 1 year and we specify constant $\delta^{18}O_p$ for the period of simulation, then average the resulting $\delta^{18}O_{cs}$ over time. As might be expected the temporal averages of $\delta^{18}O_{cs}$ are not sensitive to specification of τ.

4 Discussion

4.1 Analysis of the multiarchive, multi-observation MT subset

We perform principal component analysis (PCA) in the Jupyter notebook `MT_analysis_v9.3.ipynb` to assess the extent to which there are large scale patterns with and across observational archives and observational types, for the major classes of archives and observations in this subset of data. We perform separate PCA on the following six archive and observation subsets of the complete dataset: tree TRW; tree δ^{18}O; coral δ^{18}O; speleothem δ^{18}O; lake sediment δ^{18}O and δ^{2}H; and marine sediment δ^{18}O. Because of the potential for PCA to be influenced by large changes in record availability over time, for each subset we subjectively identify a smaller set of records with a common temporal resolution and a relatively constant data availability over some time period determined by the balance between record length and data availability. For the archive/data subsets described above, the record numbers and temporal availability are given in Table 3. For instance, we gathered TRW records from tree archives at annual time resolution, and found PCA could be performed with 71 records all available over the time interval 1400-1963 for the covariance estimation, and for a time interval spanning 1000-2000, from a 71 record subset. All identified subsets are standardized prior to calculation of their covariance (correlation) matrices and covariance eigendecompositions.

Results of the PCA are shown in Figures 6 and 7, respectively. All PCA have a leading pattern which explains from about 30-75% of the variance, with cumulative explained variance from 55-90% for PCs 1 and 2 (Figure 6 A). 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 additional analysis may be needed before a multi-archive, multi-observational analysis is performed and interpreted.

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

Archive	Observation	resolution (yrs)	min. record length (yr)	covariance timespan (CE)	total timespan (CE)	N
Tree	TRW	1	600	1400-1963	1000-2000	71
Tree	δ^{18}O	2	100	1901-1969	1700-2000	23
Coral	δ^{18}O	1	90	1905-1990	1750-2000	28
Speleothem	δ^{18}O	11	500	1431-1563	650-1950	21
Lake	δ^{18}O + δ^{2}H	55	100	630-1620	300-1800	18
Marine	δ^{18}O	100	100	100-1300	100-1800	14

Table 3. Parameters on the PCA for MT sensitivity subset archive and observation types. Remaining columns are the time resolution at which data were averaged to provide a common resolution across available records, the period of overlap for all records included for covariance estimation, the complete time interval covered by the data subset, and the number of available records entered into each PCA.

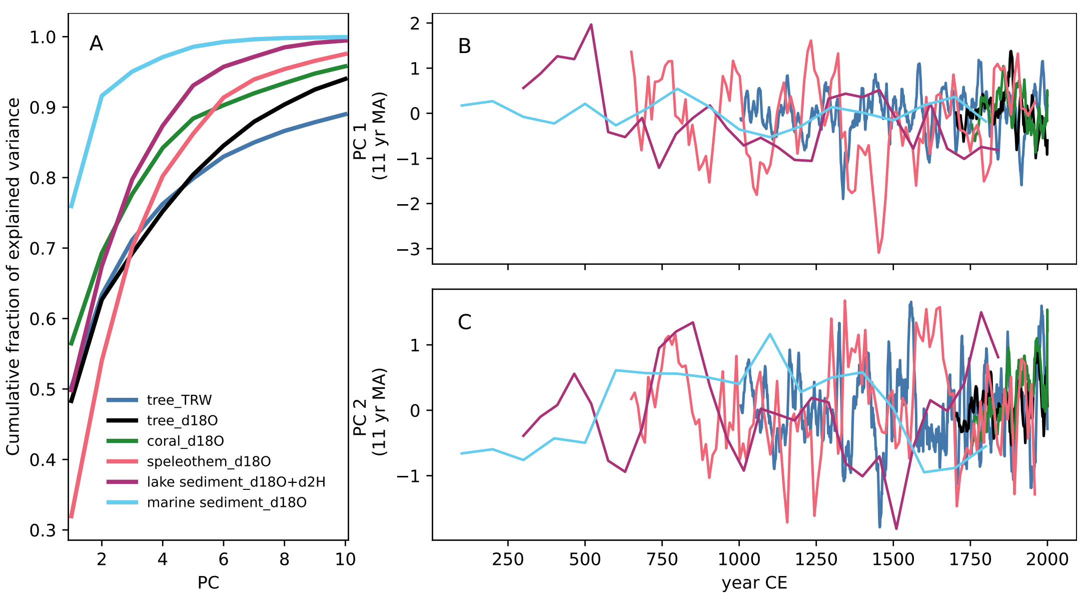

Figure 6. Results of the PCA on tree TRW, tree δ^{18}O, coral δ^{18}O, speleothem δ^{18}O, lake sediment δ^{18}O + δ^{2}H, and marine sediment δ^{18}O. A. Cumulative fraction of explained variance as the sum of normalized eigenvalues corresponding to PCs 1-10. Note: for PCA with dimension > 10, only first 10 cumulative fractional eigenvalues are shown. B, C. 1st and 2nd time series expansion for all 6 PCA. All series with resolution higher than 11 years per timepoint (see Table 3) have been averaged to 11 years per timepoint plotted. PCs of tree δ^{18}O and coral δ^{18}O have been multiplied by -1 for the influence of warming on isotopic composition in these archives (Barbour et al., 2004; Konecky et al., 2020a; Walter et al., 2020).

Figure 7 shows the EOF1 and EOF2 loadings from the 6 independent PCA as a function of their map location. 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

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

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.

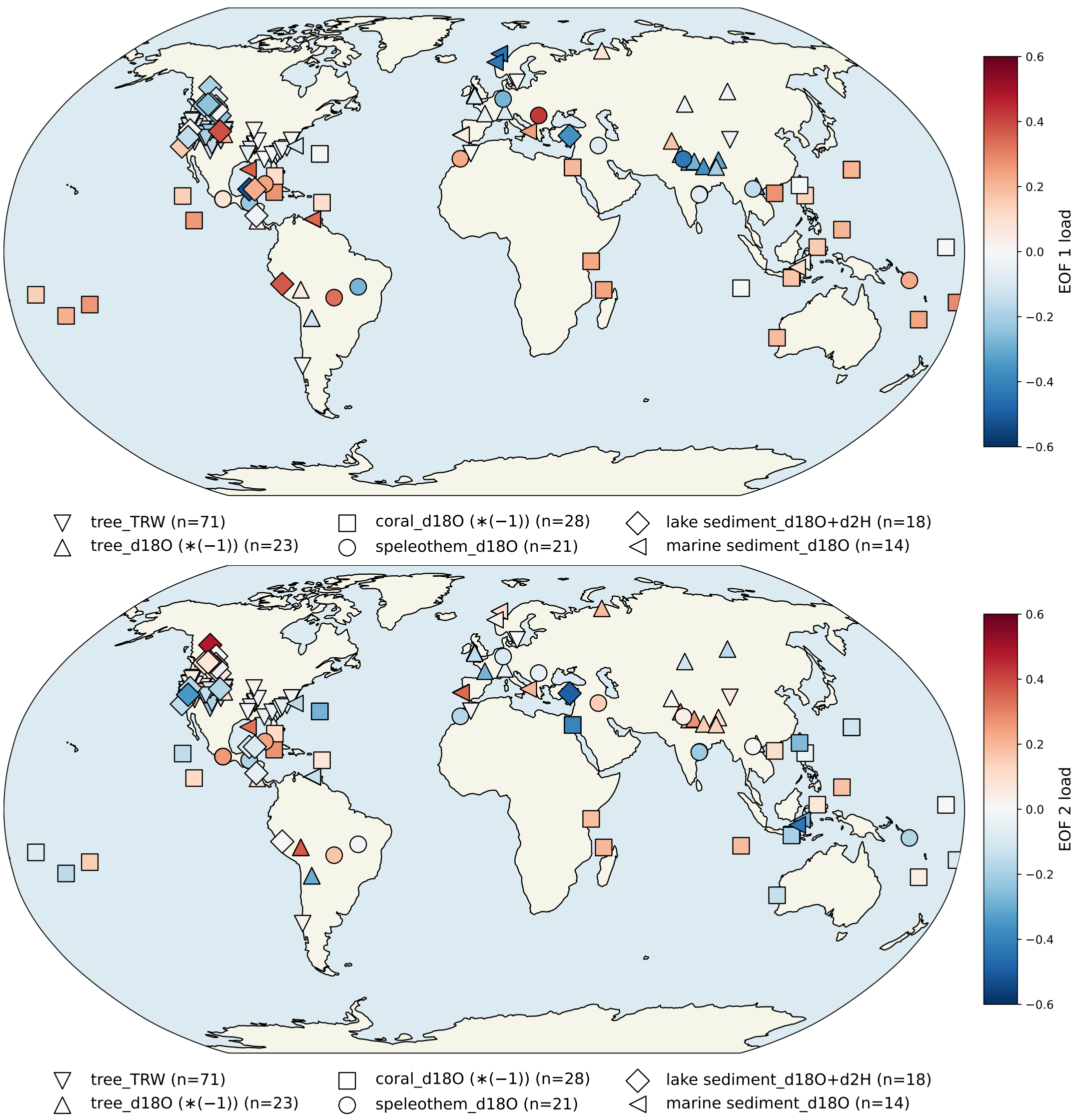

Figure 7. Top: EOF1 loadings across the 6 PCA, corresponding to the first PC shown in Fig. 6. Bottom: as in top panel, except for EOF2 loadings. As in Fig. 6, EOF loadings for tree δ^{18}O and coral δ^{18}O have been multiplied by -1.

4.2 Speleothem proxy system modeling

We develop and analyze results of the speleothem modeling application in notebook `Sanalysis.ipynb`. We find an agreement between observed ($\delta^{18}\mathrm{O}_{c,o}$) and simulated $\delta^{18}\mathrm{O}_{c,s}$) speleothem calcite isotopic composition that is significantly different from

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

zero (Table 4). This result arises primarily from the regression of calcite δ^{18}O on precipitation δ^{18}O, and to a much lesser extent, on the dependence of calcite δ^{18}O on temperature. The regression of $\delta^{18}O_{c,o}$ on $\delta^{18}O_{c,s}$ has a slope of 0.93+/-0.12 ($p<0.001$). The slope is not different from 1:1 by t-test (t=0.56; p=0.58). The mean difference $\delta^{18}O_{c,s} - \delta^{18}O_{c,o}$ is significantly less than zero by t-test (-1.33‰) with p=0.09. That value of -0.93‰ is about 10% of the range in observed calcite δ^{18}O. This mean difference might arise from either the estimated environmental controls or the translation of surface into karst conditions, or the archive model effects, or all of these factors together. Overall, these results demonstrate that simulated and observed mean calcite δ^{18}O are broadly consistent across a spatial gradient, but with no reduction of true variance across the spatial gradient (Dee et al., 2015; Hu et al., 2017; Okazaki and Yoshimura, 2019; Bühler et al., 2021).

Statistical Model	Intercept(+/-1SE)	Slope(+/-1SE)	R^2	RMSE	p_{int}	p_{slope}
$\delta^{18}O_{c,o}$ on T	-7.32(6.92)‰	0.09(0.39)‰/^{o}C	0.10	22.9‰	0.81	0.29
$\delta^{18}O_{c,o}$ on $\delta^{18}O_p$	-2.75(0.75)‰	0.45(0.10)‰/‰	0.54	2.66‰	<0.001	<0.001
$\delta^{18}O_{c,s}$ on $\delta^{18}O_{c,o}$	-1.33(0.77)‰	0.93(0.12)‰/‰	0.56	2.1‰	0.09	<0.0001

Table 4. Regression diagnostics for regression of observed calcite oxygen isotopic composition ($\delta^{18}O_c$) on T, on mean annual amount-weighted oxygen isotopic composition ($\delta^{18}O_p$), and for regression of simulated ($\delta^{18}O_{cs}$) on observed calcite oxygen isotopic composition ($\delta^{18}O_c$). In all cases, N = 58 and df = 57.

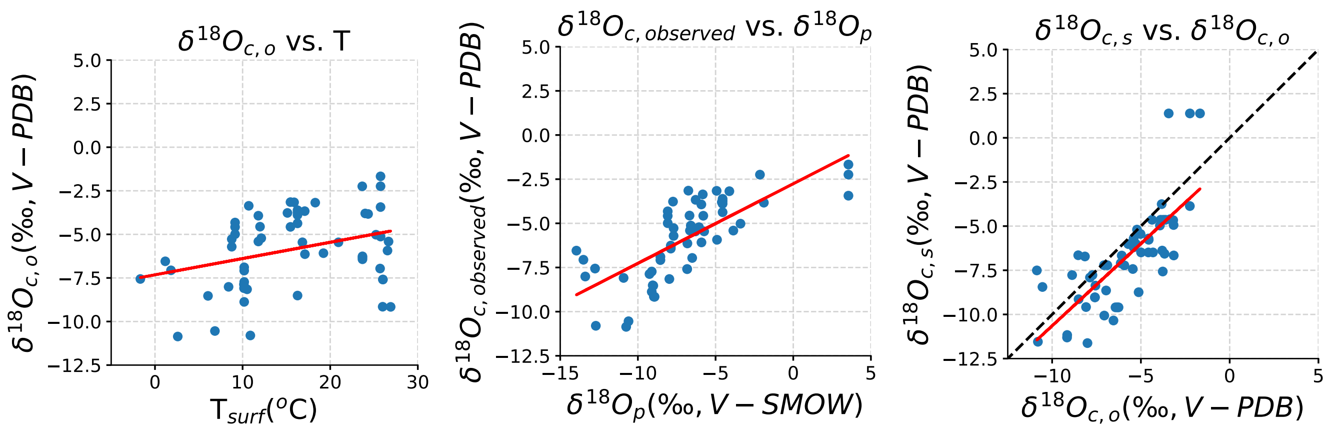

Figure 8. Left: regression of observed calcite oxygen isotopic composition ($\delta^{18}O_c$) on T, Middle: regression of $\delta^{18}O_c$ on mean annual amount-weighted oxygen isotopic composition ($\delta^{18}O_p$); Right: regression of simulated ($\delta^{18}O_{cs}$) on $\delta^{18}O_c$. Dashed line is 1:1.

If including all mineralogies (calcite, aragonite, mixed) from the SISALv3 database, we would obtain a slope of regression of $\delta^{18}O_{c,s}$ on $\delta^{18}O_{c,o}$ significantly smaller than unity, and simulated $\delta^{18}O_{c,s}$ lower than $\delta^{18}O_{c,o}$ (results not shown). This demonstrates the importance of including that metadata in the original database (Kaushal et al., 2024a) and retaining it in the DoD2k as `paleoData_notes`. If confirmed with downcore time series and spectral comparisons, this result might be used

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

to either improve the sensor model, by including differential temperature fractionation associated with aragonite vs calcite (Bühler et al., 2021), or construct a more realistic, location-specific archive model (Dee et al., 2015; Bühler et al., 2021).

4.3 Outlook

We have produced the DoD2k as a starting point into the detection and attribution of climate variability over the Common Era, using the paleoclimate observations directly, in conjunction with proxy system models attached to paleoclimate process modeling (Franke et al., 2022a). Because the DoD2k is not limited to any one subset of observations, archives, or paleoclimatic interpretations, a feature but also a limitation of its 5 component datasets, it produces a unique target that spans more of the Common Era with greater spatial and temporal coverage than would otherwise have been possible. The common dictionary structure can be used to explicitly represent a proxy system model structure Evans et al. (2013); Dee et al. (2015); Dolman and Laepple (2018) across multiple component datasets. With minor modification of the load notebooks provided, it is also possible to extend the database with new and emerging curations. An example is the documentary database DOCU-CLIM (Brönniman and Bergdorf, 2022; Burgdorf et al., 2024), which compiles 622 records of phenological and climatological observations over the 15th-19th centuries, including many records from underobserved regions of Africa (their Fig. 6; this manuscript, Fig. 2), at seasonal to annual resolution. The data are provided in a format that should be straightforward to adapt into the DoD2k v1 compact common dictionary. Proxy system models exist for the most important of the Dod2k archives and observations, making it possible to perform a multivariate, spatially and temporally resolved fingerprinting exercise (Evans et al., 2024), and to create paleoclimate data assimilation products using available tools (Zhu et al., 2024a). Challenges for these exercises will be to adequately sample observational, proxy system model and climate process model uncertainty, and to aggregate results across the diverse spectrum of paleoclimatic estimates, resolutions, timescales and differential sensitivities to environmental forcing (Evans et al., 2013; Tardif et al., 2019).

Additional challenges that should be carefully considered by adopters and modifiers of the DoD2k begin with the choice and definition of compact dictionary terms by which the aggregation across datasets is performed. We begin with 17 dictionary terms, primarily following the lead of the PAGES2k T sensitive compilation (Emile-Geay et al., 2017a), but this could and should be expanded, contracted or modified to fit other purposes. For instance, the development of a moisture and temperature subset (Section 3.2.1) requires revising original expert `climateInterpretation_variable` metadata into either moisture or temperature keywords for filtering. There may be cases in which filtering by dictionary terms we have employed result in ambiguous results, for instance in the case when a desired observation type is found in multiple archive types but with different proxy models and climatic interpretations (for instance: Mg/Ca in speleothems and foraminifera; δ^{18}O in wood, marine carbonates, lacustrine carbonates, ice cores). In the case of the SISALv3 database integration and speleothem analysis (Section 3.2.2), there are many other metadata fields than might be useful for successfully leveraging all the information included in that archive (Kaushal et al., 2024a) but which we have neglected in our compact dictionary. More generally, user-supervised choices about selecting and adapting dictionary terms, and managing candidate duplicates may vary according to the application of the DoD2k.

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

5 Conclusions

We have developed a compilation of publicly available notebooks and functions, which transform five expert-curated, community sourced Common Era paleoclimate datasets into a database of databases (DoD2k). We demonstrate the utility of the DoD2k with two applications which permit analysis of the DoD2k by paleoclimatic interpretation and by a small subset of archive and observation type. Challenges to this approach include the need to produce a common set of metadata categories, here implemented as dictionary terms, by which to aggregate the component databases within specific groupings by metadata characteristic. The notebooks, functions, and database are available on GitHub. Being user-customizable, they can be used to produce specific DoD2k versions for particular applications.

6 Code and data availability

Jupyter notebooks, Python functions, and scripts are on GitHub, at https://github.com/lluecke/dod2k, with DOI: https://doi.org/10.5281/zenodo.15676255.

7 Data availability

Databases used to construct the DoD2k are available online (Emile-Geay et al., 2017a; Walter et al., 2020; Konecky et al., 2020b; Franke et al., 2022a; Kaushal et al., 2024b). The DoD2k database (Evans et al., 2025) is available on GitHub at https://github.com/lluecke/dod2k, and at the NOAA/NCEI World Data Service for Paleoclimatology (https://www.ncei.noaa.gov/access/paleo-search/study/41981) with DOI: https://doi.org/10.25921/sptp-g618.

Author contributions. Michael N. Evans (MNE) conceived the database of databases, designed the duplicate detection algorithm, developed the plans for applications, wrote the first draft of the manuscript, and revised functions and notebooks. Lucie J. Luecke (LJL) implemented the database aggregation, duplicate detection and underlying python functions for analysis and visualization and drafted section 2.2 with input and minor code modifications from MNE; she wrote the moisture/temperature sensitive data analysis notebook. Kevin J. Fan (KJF), MNE and and Feng Zhu (FZ) wrote the speleothem analysis application notebook and developed visualization of the results, and MNE wrote the associated subsection of section 4 with input from KJF and FZ. FZ gave important suggestions and direction to initial stages of the manuscript drafting. MNE, LJL, KJF and FZ all contributed to revision and development of the submitted manuscript.

Competing interests. The authors declare no competing interests.

Acknowledgements. MNE is grateful to the PReSTo https://paleopresto.com/about.html data informatics working group for a February 2023 meeting in which the need for this work became evident, and to Julien Emile-Geay in particular who suggested using existing curated

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

databases instead of raw records from the NOAA/NCEI/WDS repository for Paleoclimatology. The wisdom of this becoming clearer during some windy and wet bicycling and subsequent sipping on Islay. Dylan Jones (University of Edinburgh) undertook the initial efforts to identify a compact set of common dictionary terms across databases. MNE thanks the University of Edinburgh Geosciences community and in particular Gabi Hegerl, Andrew Schurer, Simon Tett, Patrick Meir and their research groups for support and constructive feedback. All the authors thank the individuals and working groups who painstakingly created the individual datasets and made them publicly available. We thank the individuals and working groups organized and supported by PAGES for then producing the component databases without which this work would have been impossible. This work was supported by a Royal Society of London/Wolfson Visiting Fellowship grant Award \RSWVF\R1\221018 to MNE, which partly supported LJL, US NSF/P4CLIMATE Award AGS2303530 to MNE, which supported KJF and FZ, and by the University of Maryland, College Park and the University of Edinburgh, School of Geosciences, for funding and hosting MNE during a sabbatical year visit.

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

References

Abram, N. J., McGregor, H. V., Tierney, J. E., Evans, M. N., McKay, N. P., Kaufman, D. S., and the PAGES2k Consortium: Early onset of industrial-era warming across the oceans and continents, Nature, 536(7617), 411–418, doi: 10.1038/nature19082, 2016.

Anderson, D. M., Tardif, R., Horlick, K., Erb, M. P., Hakim, G. J., Noone, D., Perkins, W. A., and Steig, E.: Additions to the Last Millennium Reanalysis Multi-Proxy Database, Data Sci. J., 18, 2–11, https://doi.org/10.5334/dsj-2019-002, 2019.

Arias, P., Bellouin, N., Coppola, E., Jones, R., Krinner, G., Marotzke, J., Naik, V., Palmer, M., Plattner, G.-K., Rogelj, J., Rojas, M., Sillmann, J., Storelvmo, T., Thorne, P., Trewin, B., Achuta Rao, K., Adhikary, B., Allan, R., Armour, K., Bala, G., Barimalala, R., Berger, S., Canadell, J., Cassou, C., Cherchi, A., Collins, W., Collins, W., Connors, S., Corti, S., Cruz, F., Dentener, F., Dereczynski, C., Di Luca, A., Diongue Niang, A., Doblas-Reyes, F., Dosio, A., Douville, H., Engelbrecht, F., Eyring, V., Fischer, E., Forster, P., Fox-Kemper, B., Fuglestvedt, J., Fyfe, J., Gillett, N., Goldfarb, L., Gorodetskaya, I., Gutierrez, J., Hamdi, R., Hawkins, E., Hewitt, H., Hope, P., Islam, A., Jones, C., Kaufman, D., Kopp, R., Kosaka, Y., Kossin, J., Krakovska, S., Lee, J.-Y., Li, J., Mauritsen, T., Maycock, T., Meinshausen, M., Min, S.-K., Monteiro, P., Ngo-Duc, T., Otto, F., Pinto, I., Pirani, A., Raghavan, K., Ranasinghe, R., Ruane, A., Ruiz, L., Sallée, J.-B., Samset, B., Sathyendranath, S., Seneviratne, S., Sörensson, A., Szopa, S., Takayabu, I., Tréguier, A.-M., van den Hurk, B., Vautard, R., von Schuckmann, K., Zaehle, S., Zhang, X., and Zickfeld, K.: Technical Summary, p. 33-144, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896.002, 2021.

Barbour, M. M., Roden, J. S., Farquhar, G. D., and Ehleringer, J. R.: Expressing leaf water and cellulose oxygen isotope ratios as enrichment above source water reveals evidence of a Péclet effect, Oecologia, 138, 426–435, doi: 10.1007/s00442-003-1449-3, 2004.

Bowen, G. J., Wassenaar, L. I., and Hobson, K. A.: Global application of stable hydrogen and oxygen isotopes to wildlife forensics, Oecologia, 143, 337–348, https://doi.org/10.1007/s00442-004-1813-y, 2005.

Breitenmoser, P., Brönnimann, S., and Frank, S.: Forward modeling of tree-ring width and comparison with a global network of tree-ring chronologies, Clim. Past., 10, 437–449, doi:10.5194/cp-10-437-2014, 2014a.

Breitenmoser, P., Brönnimann, S., and Frank, D.: Forward modelling of tree-ring width and comparison with a global network of tree-ring chronologies, Climate of the Past, 10, 437–449, https://doi.org/10.5194/cp-10-437-2014, 2014b.

Brönniman, S. and Bergdorf, A.-M.: DOCU-CLIM, https://boris-portal.unibe.ch/entities/product/f7e92d0d-ab6f-4d4c-a692-ee7287a74d63/full, https://doi.org/10.48620/167, accessed: 2025-06-06, 2022.

Bühler, J. C., Roesch, C., Kirschner, M., Sime, L., Holloway, M. D., and Rehfeld, K.: Comparison of the oxygen isotope signatures in speleothem records and iHadCM3 model simulations for the last millennium, Clim. Past, 17, 985–1004, https://doi.org/10.5194/cp-17-985-2021, 2021.

Burgdorf, A. M., Brónnimann, S., Adamson, G., Amano, T., Aono, Y., Barriopedro, D., Bullón, T., Camenisch, C., Camuffo, D., Daux, V., del Rosario Prieto, M., Dobrovolný, P., Gallego, D., García-Herrera, R., Gergis, J., Grab, S., Hannaford, M. J., Holopainen, J., Kelso, C., Kern, Z., Kiss, A., Lin, E. K.-H., Loader, N. J., Možný, M., Nash, D., Nicholson, S. E., Pfister, C., Rodrigo, F. S., Rutishauser, T., Sharma, S., Takács, K., Vargas, E. T., and Vega, I.: DOCU-CLIM: A global documentary climate dataset for climate reconstructions, Sci. Data, 10, https://doi.org/10.1038/s41597-023-02303-y, article 402, 2024.

Bush, R., Dutton, A., Evans, M., Loft, R., and Schmidt, G. A.: Perspectives on Data Reproducibility and Replicability in Paleoclimate and Climate Science, Harvard Data Science Review, 2(4), doi: 10.1162/99608f92.00cd8f85, 2020.

Cobb, K. M., Westphal, N., Sayani, H., Watson, J. T., Di Lorenzo, E., Cheng, H., Edwards, R. L., and Charles, C. D.: Highly variable El Niño-Southern Oscillation throughout the Holocene, Science, 339, 67–70, doi: 10.1126/science.1228246, 2013.

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

Consortium, P. K.: Continental-scale temperature variability during the past two millennia, Nat. Geosci., 6, 339–346, doi: 10.1038/ngeo1797, 2013.

Dee, S. G., Emile-Geay, J., Evans, M., Allam, A., Stei, E. J., and Thompson, D. M.: PRYSM: An open-source framework for PRoxY System Modeling, with applications to oxygen-isotope systems, J. Adv. Mod. Earth Sys., 07, doi: doi:10.1002/2015MS000447, 2015.

Dee, S. G., Cobb, K. M., Emile-Geay, J. E., Ault, T. R., Edwards, R. L., Cheng, H., and Charles, C. D.: No consistent ENSO response to volcanic forcing over the last millennium, Science, 367(6485), 1477–1481, doi: 10.1126/science.aax200, 2020.

Dolman, A. and Laepple, T.: Sedproxy: a forward model for sediment-archived climate proxies, Clim. Past, 14, 1851–1868, https://doi.org/10.5194/cp-14-91851-2018, 2018.

Emile-Geay, J. and Eshleman, J. A.: Toward a semantic web of paleoclimatology, Geochem., Geophys., Geosyst., 14, 457–469, https://doi.org/10.1002/ggge.200672, 2013.

Emile-Geay, J., McKay, N. P., Kaufman, D. S., von Gunten, L., Wang, J., Anchukaitis, K., Abram, N., Addison, J. A., Curran, M. A. J., Evans, M. N., Henley, B. J., Hao, Z., Martrat, B., McGregor, H. V., Neukom, R., Pederson, G. T., Stenni, B., Thirumalai, K., Werner, J. P., Xu, C., Divine, D. V., Dixon, B. C., Gergis, J., Mundo, I. A., Nakatsuka, T., Phipps, S. J., Routson, C. C., Steig, E. J., Tierney, J. E., Tyler, J. E., Allen, K. J., Bertler, N. A. N., Bjórklund, J., Chase, B. M., Chen, M.-T., Cook, E. R., De Jong, R., DeLong, K. L., Dixon, D. A., Ekaykin, A. A., Ersek, V., Filipsson, H. L., Francus, P., Freund, M. B., Frezzotti, M., Gaire, N. P., Gajewski, K., Ge, Q., Goosse, H., Gornostaeva, A., Grosjean, M., Horiuchi, K., Hormes, A., Husum, K., Isaksson, E., Kandasamy, S., Kawamura, K., Kilbourne, K. H., Koć, N., Leduc, G., Linderholm, H. W., Lorrey, A. M., Mikhalenko, V., Mortyn, P. G., Motoyama, H., Moy, A. D., Mulvaney, R., Munz, P. M., Nash, D. J., Oerter, H., Opel, T., Orsi, A. J., Ovchinnikov, D. V., Porter, T. J., Roop, H. A., Saenger, C., Sano, M., Sauchyn, D., Saunders, K. M., Seidenkrantz, M.-S., Severi, M., Shao, X., Sicre, M.-A., Sigl, M., Sinclair, K., St. George, R. S., St. Jacques, J. M., Thamban, M., Kuwar Thapa, U., Thomas, E. R., Turney, C., Uemura, R., Viau, A. E., Vladimirova, D. O., Wahl, E. R., White, J. W. C., Yu, Z., Zinke, J., and the PAGES2k Consortium: A global multiproxy database for temperature reconstructions of the Common Era, Sci. Data, 4, article no. 170088, doi: https://doi.org/10.1038/sdata.2017.88, 2017a.

Emile-Geay, J., McKay, N. P., Kaufman, D. S., von Gunten, L., Wang, J., Anchukaitis, K. J., Abram, N. J., Addison, J. A., Curran, M. A. J., Evans, M. N., Henley, B. J., Hao, Z., Martrat, B., McGregor, H. V., Neukom, R., Pederson, G. T., Stenni, B., Thirumalai, K., Werner, J. P., Xu, C., Divine, D. V., Dixon, B., Gergis, J., Mundo, I. A., Nakatsuka, T., Phipps, S. J., Routson, C., Steig, E. J., Tierney, J. E., Tyler, J. J., Allen, K. J., Bertler, N. A. N., Björklund, J., Chase, B. M., Chen, M. T., Cook, E. R., de Jong, R., DeLong, K. L., Dixon, D. A., Ekaykin, A. A., Ersek, V., Filipsson, H. L., Francus, P., Freund, M. B., Frezzotti, M., Gaire, N., Gajewski, K., Ge, Q., Goosse, H., Gornostaeva, A., Grosjean, M., Horiuchi, K., Hormes, A., Husum, K., Isaksson, E., Kandasamy, K., Kawamura, K., Kilbourne, K. H., Ko̧ N., Leduc, G., Linderholm, H. W., Lorrey, A., Mikhalenko, V. N., Mortyn, P. G., Motoyama, H., Moy, A., Mulvaney, R., Munz, P. M., Nash, D., Oerter, H., Opel, T., Orsi, A. J., Ovchinnikov, D. V., Porter, T. J., Roop, H. A., Saenger, C. P., Sano, M., Sauchyn, D. J., Saunders, K. M., Seidenkrantz, M. S., Severi, M., Shao, X., Sicre, M. A., Sigl, M., Sinclair, K., St. George, S., St. Jacques, J. M., Meloth, T., Thapa, U. K., Thomas, E. R., Turney, C. S. M., Uemura, R., Viau, A. E., Vladimirova, D. O., Wahl, E. R., White, J. W. C., Yu, Z., and Zinke, J.: NOAA/WDS Paleoclimatology - PAGES2k Global 2,000 Year Multiproxy Database, https://www.ncei.noaa.gov/access/paleo-search/study/21171, https://doi.org/10.6084/m9.figshare.c.3285353.v2, https://doi.org/10.25921/ycr3-7588, accessed: 2023-06-15, 2017b.

Evans, M., Franke, J., Schurer, A. P., Zhu, F., Lücke, L., Jones, D., Valdes, P. J., and Hegerl, G. C.: Climate change detection and attribution using paleoclimate observations directly, meeting website at: http://www.meteo.fr/cic/meetings/2024/IMSC/, meeting agenda at: http://www.meteo.fr/cic/meetings/2024/IMSC/IMSC_2024_final_detailed_agenda.pdf, slides published at: https://pacificclimate.org/∼IMSC/3-Wednesday/2-Orals/S07/Evans.pdf, 2024.

Evans, M. N., Tolwinski-Ward, S. E., Thompson, D. M., and Anchukaitis, K. J.: Applications of proxy system modeling in high resolution paleoclimatology, Quat. Sci. Rev., 76, 16–28, doi: 10.1016/j.quascirev.2013.05.024, 2013.

Evans, M. N., Lücke, L. J., Fan, K. J., and Zhu, F.: DoD2k:A database of curated databases for Common Era Paleoclimate Applications, https://www.ncei.noaa.gov/access/paleo-search/study/41981, https://www.github.com/lluecke/dod2k/, https://doi.org/10.25921/sptp-g618, accessed: 2025-06-17, 2025.

Falster, G., Konecky, B., Coats, S., and Stevenson, S.: Forced changes in the Pacific Walker circulation over the past millennium, Nature, 622, 93–100, doi: 10.1038/s41586-023-06447-0, 2023.

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., Lunt, D., Mauritsen, T., Palmer, M., Watanabe, M., Wild, M., and Zhang, H.: The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity, p. 923–1054, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896.009, 2021.

Fox-Kemper, B., Hewitt, H., Xiao, C., Aðalgeirsdóttir, G., Drijfhout, S., Edwards, T., Golledge, N., Hemer, M., Kopp, R., Krinner, G., Mix, A., Notz, D., Nowicki, S., Nurhati, I., Ruiz, L., Sallée, J.-B., Slangen, A., and Yu, Y.: Ocean, Cryosphere and Sea Level Change, p. 1211–1362, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896.011, 2021.

Franke, J., Evans, M. N., Schurer, A., and Hegerl, G. C.: Climate change detection and attribution using observed and simulated tree-ring width, Clim. Past., 18, 2583–2597, doi:10.5194/cp-18-2583-2022, 2022a.

Franke, J., Evans, M. N., Schurer, A. P., and Hegerl, G. C.: NOAA/WDS Paleoclimatology - Northern Hemisphere Observed and Simulated Tree Ring Width Data 1401-2000 CE, https://doi.org/10.25921/8HPF-A451, https://doi.org/10.25921/8HPF-A451, accessed: 2024-11-01, 2022b.

Gebbie, G. and Huybers, P.: The Little Ice Age and 20th-century deep Pacific cooling, Science, 363, 70–74, 2019.

Harris, I., Osborn, T. J., Jones, P. D., and Lister, D.: Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset, Sci. Data, 7, 109, doi: 10.1038/s41597-020-0453-3, 2020.

Hu, J., Emile-Geay, J., and Partin, J.: Correlation-based interpretations of paleoclimate data – where statistics meet past climates, Earth and Planet. Sci. Lett., 459, 362–371, https://doi.org/10.1016/j.epsl.2016.11.048, 2017.

Kaushal, N., Atwood, A., Bothe, O., Eggleston, S., Falster, G., Henley, B. J., Jones, M., Jonkers, L., Martrat, B., McGregor, H. V., Orsi, A., Phipps, S. J., and Sayani, H.: Phase 4 of the PAGES 2k Network: Hydroclimate of the Common Era, PAGES Magazine, 30, 52, doi: 10.22498/pages.30.1.52, 2022.

Kaushal, N., Lechleitner, F. A., Wilhelm, M., Azennoud, K., Bühler, J. C., Braun, K., Ait Brahim, Y., Baker, A., Burstyn, Y., Comas-Bru, L., Fohlmeister, J., Goldsmith, Y., Harrison, S. P., Hatvani, I. G., Rehfeld, K., Ritzau, M., Skiba, V., Stoll, H. M., Szűcs, J. G., Tanos, P., Treble, P. C., Azevedo, V., Baker, J. L., Borsato, A., Chawchai, S., Columbu, A., Endres, L., Hu, J., Kern, Z., Kimbrough, A., Koç, K., Markowska, M., Martrat, B., Ahmad, S. M., Nehme, C., Novello, V. P., Pérez-Mejías, C., Ruan, J., Sekhon, N., Sinha, N., Tadros, C. V., Tiger, B. H., Warken, S., Wolf, A., Zhang, H., and SISAL Working Group members: SISALv3: a global speleothem stable isotope and trace element database, Earth Sys. Sci. Dat., 16, 1933—-1963, https://doi.org/10.5194/essd-16-1933-2024, 2024a.

Kaushal, N., Lechleitner, F. A., Wilhelm, M., and Members of the SISAL Working Group: SISALv3: Speleothem Isotopes Synthesis and AnaLysis database version 3.0, https://doi.org/10.5287/ORA-2NANWP4RK, https://doi.org/10.5287/ORA-2NANWP4RK, 2024b.

Kaushal, N., Lechleitner, F. A., Wilhelm, M., and Members of the SISAL Working Group: SISALv3: Speleothem Isotopes Synthesis and AnaLysis database version 3.0, https://doi.org/10.5287/ORA-2NANWP4RK, accessed: 2023-06-03, 2024c.

Konecky, B. L. and McKay, N. P.: NOAA/WDS Paleoclimatology - Iso2k Database Global Common Era Paleo-d18O and d2H Records, https://doi.org/10.25921/57j8-vs18, https://doi.org/10.25921/57J8-VS18, accessed: 2024-11-01, 2020.

Konecky, B. L., McKay, N. P., Churakova (Sidorova), O. P., Comas-Bru, L., Dassié, E. P., DeLong, K. L., Falster, G. M., Fischer, M. J., Jones, M. D., Jonkers, L., Kaufman, D. S., Leduc, G., Managave, S. R., Martrat, B., Opel, T., Orsi, A. J., Partin, J. W., Sayani, H. R., Thomas, E. K., Thompson, D. M., Tyler, J. J., Abram, N. J., Atwood, A. R., Cartapanis, O., Conroy, J. L., Curran, M. A., Dee, S. G., Deininger, M., Divine, D. V., Kern, Z., Porter, T. J., Stevenson, S. L., von Gunten, L., and Iso2k Project Members: The Iso2k database: a global compilation of paleo-δ^{18}O and δ^{2}H records to aid understanding of Common Era climate, Earth Sys. Sci. Dat., 12, 2261—-2288, https://doi.org/10.5194/essd-12-2261-2020, 2020a.

Konecky, B. L., McKay, N. P., Churakova (Sidorova), O. V., Comas-Bru, L., Dassié, Émilie P.., DeLong, K. L., Falster, G. M., Fischer, M. J., Jones, M. D., Jonkers, L., Kaufman, D. S., Leduc, G., Managave, S. R., Martrat, B., Opel, T., Orsi, A. J., Partin, J. W., Sayani, H. R., Thomas, E. K., Thompson, D. M., Tyler, J. J., Abram, N. J., Atwood, A. R., Cartapanis, O., Conroy, J. L., Curran, M. A., Dee, S. G., Deininger, M., Divine, D. V., Kern, Z., Porter, T. J., Stevenson, S. L., and von Gunten, L.: The Iso2k database: a global compilation of paleo-δ^{18}O and δ^{2}H records to aid understanding of Common Era climate, Earth System Science Data, 12, 2261–2288, https://doi.org/10.5194/essd-12-2261-2020, 2020b.

Mann, M. E., Bradley, R. S., and Hughes, M. K.: Northern hemisphere temperatures during the past millennium: Inferences, uncertainties and limitations, Geophys. Res. Lett., 26, 759–762, 1999.

Marvel, K., Biasutti, M., Bonfils, C., Durack, P. J., Kushnir, Y., and Cook, B. I.: Observed and Projected Changes to the Precipitation Annual Cycle, J. Clim., 30, 4983–4995, https://doi.org/10.1175/JCLI-D-16-0572.1, 2019a.

Marvel, K., Cook, B. I., Bonfils, C. J. W., Taylor, K. E., Smerdon, J. E., and Williams, A. P.: Twentieth-century hydroclimate changes consistent with human influence, Nature, 569, 59–65, https://doi.org/10.1038/s41586-019-1149-8, 2019b.

McGregor, H. V., Evans, M. N., Goosse, H., Leduc, G., Martrat, B., Addison, J. A., Mortyn, P. G., Oppo, D. W., Seidenkrantz, M.-S., Sicre, M.-A., Phipps, S. J., Selvaraj, K., Thirumalai, K., Filipsson, H. L., and Ersek, V.: Robust global ocean cooling trend for the pre-industial Common Era, Nat. Geosci., doi: 10.1038/ngeo2510, 2015.

McKay, N. P. and Emile-Geay, J.: The Linked Paleo Data framework–a common tongue for paleoclimatology, Climate of the Past, 12, 1093–1100, doi:10.5194/cp-12-1093-2016, 2016.

Miller, G. H., Geirsdóttir, A., Zhong, Y., Larsen, D. J., Otto-Bliesner, B. L., Holland, M. M., Bailey, D. A., Refsnider, K. A., Lehman, S. J., Southon, J. R., Anderson, C., Bj ornsson, H., and T., T.: Abrupt onset of the Little Ice Age triggered by volcanismand sustained by sea-ice/ocean feedbacks, Geophys. Res. Lett., 39, L02 708, https://doi.org/10.1029/2011GL050168, 2012.

Neukom, R., Barboza, L. A., Erb, M. P., Shi, F., Emile-Geay, J., Evans, M. N., Franke, J., Kaufman, D. S., Lücke, L., Rehfeld, K., Schurer, A., Zhu, F., Brönnimann, S., Hakim, G. J., Henley, B. J., Ljungqvist, F. C., McKay, N., Valler, V., and von Gunten, L.: Consistent multidecadal variability in global temperature reconstructions and simulations over the Common Era, Nat. Geosci., 12, 643–649, doi: 10.1038/s41561-019-0400-0, 2019a.

Neukom, R., Steiger, N., Gómez-Navarro, J. J., Wang, J., and Werner, J.: No evidence for globally coherent warm and cold periods over the preindustrial Common Era, Nature, 571, 550–554, doi: 10.1038/s41586-019-1401-2, 2019b.

Okazaki, A. and Yoshimura, K.: Global Evaluation of Proxy System Models for Stable Water Isotopes With Realistic Atmospheric Forcing, J. Geophys. Res. - Atm., 124, 8972–8993, https://doi.org/10.1029/2018JD029463, 2019.

Steiger, N., Horlick, K., Tardif, R., Erb, M., Emile-Geay, J., Steig, E., and Hakim, G.: A global collection of paleoclimate proxy time series over the Common Era: Nov 2022 Update, https://zenodo.org/records/1189006 accessed 11 February 2025, 2022.

https://doi.org/10.5194/essd-2025-364
Preprint. Discussion started: 1 July 2025

Open Access Earth System Science Data Discussions

Tardif, R., Hakim, G. J., Perkins, W. A., Horlick, K. A., Erb, M. P., Emile-Geay, J., Anderson, D. M., Steig, E. J., and Noone, D.: Last Millennium Reanalysis with an expanded proxy database and seasonal proxy modeling, Climate of the Past, 15, 1251–1273, doi:10.5194/cp-15-1251-2019, 2019.

Tolwinski-Ward, S. E., Evans, M. N., Hughes, M. K., and Anchukaitis, K. J.: An efficient forward model of the climatic controls on interannual variation in tree-ring width, Clim. Dyn., 36, 2419–2439, doi: 10.1007/s00382-010-0945-5, 2011.

Walter, R. M., Sayani, H. R., Felis, T., Cobb, K. M., Abram, N. J., Arzey, A. K., Atwood, A. R., Brenner, L. D., Dassié, E. P., DeLong, K. L., Ellis, B., Emile-Geay, J., Fischer, M. J., Goodkin, N. F., Hargreaves, J. A., Kilbourne, K. H., Krawczyk, H., McKay, N. P., Moore, A. L., Murty, S. A., Ong, M. R., Ramos, R. D., Reed, E. V., Samanta, D., Sanchez, S. C., Zinke, J., and PAGES CoralHydro2k Project Members: The CoralHydro2k database: a global, actively curated compilation of coral δ^{18}O and Sr/Ca proxy records of tropical ocean hydrology and temperature for the Common Era, Earth Sys. Sci. Dat., 15, 2081—-2116, https://doi.org/10.5194/essd-15-2081-2023, 2020.

Walter, R. M., Sayani, H. R., Felis, T., Cobb, K. M., Abram, N. J., Arzey, A. K., Atwood, A., Brenner, L. D., Dassié, E. P., DeLong, K. L., Ellis, B., Fischer, M. J., Goodkin, N. F., Hargreaves, J. A., Kilbourne, K. H., Krawczyk, H. A., McKay, N. P., Murty, S. A., Ramos, R. D., Reed, E. V., Samanta, D., Sanchez, S. C., Zinke, J., and PAGES CoralHydro2k Project Members: NOAA/WDS Paleoclimatology - CoralHydro2k Database (Common Era coral d18O and Sr/Ca data compilation), https://doi.org/10.25921/yp94-v135, https://doi.org/10.25921/YP94-V135, accessed: 2024-05-24, 2022.

Zhu, F., Emile-Geay, J., Anchukaitis, K. J., McKay, N. P., Stevenson, S., and Meng, Z.: A Pseudoproxy Emulation of the PAGES 2k Database Using a Hierarchy of Proxy System Models, Scientific Data, 10, 624, https://doi.org/10.1038/s41597-023-02489-1, 2023.

Zhu, F., Emile-Geay, J., Hakim, G. J., Guillot, D., Khideř, D., Tardif, R., and Perkins, W. A.: `cfr` (v2024.1.26): a Python package for climate field reconstr uction, Geoscientific Model Development, 17, 3409–3431, https://doi.org/10.5194/gmd-17-3409-2024, 2024a.

Zhu, F., Emile-Geay, J., Hakim, G. J., Guillot, D., Khider, D., Tardif, R., and Perkins, W. A.: cfr (v2024.1.26): a Python package for climate field reconstruction, Geoscientific Model Development, 17, 3409–3431, https://doi.org/10.5194/gmd-17-3409-2024, 2024b.