DATA SCIENCE JOURNAL

Anderson, DM, et al. 2019. Additions to the Last Millennium Reanalysis Multi-Proxy Database. *Data Science Journal*, 18: 2, pp. 1–11. DOI: https://doi.org/10.5334/dsj-2019-002

DATA ARTICLE

Additions to the Last Millennium Reanalysis Multi-Proxy Database

David M. Anderson[1], Robert Tardif[2], Kaleb Horlick[3], Michael P. Erb[4], Gregory J. Hakim[2], David Noone[3], Walter A. Perkins[4] and Eric Steig[5]

[1] NOAA Paleoclimatology Program (retired), Boulder, US

[2] Department of Atmospheric Sciences, University of Washington, Seattle, US

[3] College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, US

[4] School of Earth and Sustainability, Northern Arizona University, Flagstaff, US

[5] Department of Earth and Space Sciences, University of Washington, Seattle, US

Corresponding author: David M. Anderson (davidmanderson2@gmail.com)

Progress in paleoclimatology increasingly occurs via data syntheses. We describe additions to a collection prepared for use in paleoclimate state estimation, specifically the Last Millennium Reanalysis (LMR). The 2290 additional series include 2152 tree ring chronologies and 138 other series. They supplement the collection used previously and together form a database titled LMRdb 1.0.0. The additional data draws from lake core, ice core, coral, speleothem, and tree ring archives, using published data primarily from the NOAA Paleoclimatology archive and a set of tree ring width chronologies standardized from raw International Tree Ring Data Bank ring width series. In contrast to many previous paleo compilations, the data were not selected (screened) on the basis of their environmental correlation, multi-century length, or other attributes. The inclusion of proxies sensitive to moisture and other environmental variables expands their use in data assimilation. A preliminary calibration using linear regression with mean annual temperature reveals characteristics of the proxy series and their relationship to temperature, as well as the noise and error characteristics of the records. The additional records are structured as individual files in the NOAA Paleoclimatology format and archived at NOAA Paleoclimatology (Anderson et al. 2018) and will continue to be improved and expanded as part of the LMR Project. The additions represent a four-fold increase in the number of records available for assimilation, provide expanded geographic coverage, and add additional proxy variables. Applications include data assimilation, proxy system model development, and paleoclimate reconstruction using climate field reconstruction and other methods.

Keywords: paleoclimatology; climate reconstructions; tree rings; data assimilation

Background

Climate and environmental reconstructions from paleo proxy records often emphasize collections of proxy records. Collections expand the spatial and temporal extent of the data, reduce noise, and can enable the reconstruction of multiple climate variables. Following the pioneering multi-proxy temperature reconstruction by Mann et al. (1998), many different reconstructions, based on unique and overlapping collections of proxy records, have been assembled. Most reconstructions have focused on the Northern Hemisphere land surface where observations are concentrated (Mann, et al. 2008; PAGES 2k Consortium 2013; Masson-Delmotte et al. 2013). Recent compilations have expanded networks in the Southern Hemisphere (Neukom et al. 2014), the oceans (Tierney et al. 2015), the Arctic (McKay and Kaufman, 2014), and the globe (PAGES Consortium, 2017). Compilations of moisture-sensitive proxies are being developed using multiple proxies (Smerdon et al. 2017) and tree ring databases (Cook et al. 2009; Cook et al. 2010). Some of these moisture-sensitive chronologies are included here through incorporation of the Breitenmoser et. al. (2014) collection. Previously, the Last Millennium Reanalysis (LMR) (Hakim et al. 2016; Emile-Geay et al. 2018) used the PAGES 2k Consortium (2013) temperature-sensitive multi-proxy database, and subsequently used the

PAGES 2k Consortium (2017) update (consisting of 571 records) archived at NOAA (PAGES Consortium (2017a). Additions were sought that would expand the geographic coverage and the sensitivity to different environmental variables. The additional data should be amenable to the development of quantitative proxy system models, including proxy error variance estimates needed for assimilation. Moreover, the data must be structured (i.e. consistently formatted) so that thousands of records, each consisting of tens to hundreds of years of observations, can be parsed by the assimilation algorithm. For these reasons we developed a method (and available code (**Table 1**)) to quality-control, standardize, and structure a set of records consisting primarily of existing (published) paleo proxy time series.

Our goal is to produce a collection of paleoclimate proxy time series, in their original data units, with minimal processing, add additional metadata about the series characteristics, and provide structured time series in individual files for reanalysis, reconstruction, proxy system model development, and other applications. The additions described here are merged with a collection of 571 records described previously (PAGES 2k Consortium 2017) in a processing step of the LMR code to create a comprehensive, integrated and assimilation-ready proxy database termed LMRdb v1.0.0. The scope of the additions is broad and, in contrast to the PAGES 2k Consortium (2017) data, minimally screened (screening refers to inclusion or rejection of series based on criteria). We added proxies from five archives (tree rings, corals and sclerosponges, ice cores, lake cores, and speleothems (**Table 2**)). The original series are annually, sub-annually, or super-annually sampled. Short records (less than 30 years length) are included and the maximum length spans the Holocene (10,000 years). All of the records are hypothesized to have some relationship with environmental variables that can be quantified as either an empirical linear univariate or bivariate regression or a physically based proxy system model. The LMR framework includes a data pre-processing step (which can include screening or filtering records based on certain criteria), and a proxy system model step. The model step develops a statistical model relating proxy observations to environmental conditions and quantifies uncertainty in observations required during the assimilation step. The code is designed so that this statistical approach can be replaced by more sophisticated proxy system models. In the final assimilation step the proxy observations are compared with estimates derived from the prior state using the proxy system model. The additional records described here are all-encompassing, from many proxy sources and locations, of varying length and resolution, minimally-screened, for the broadest application. The files are structured and archived as NOAA-formatted files (NOAA Paleoclimatology 2018) to ensure that sufficient metadata accompany the records so that transformations to other formats, such as LiPD (McKay and Emile-Geay 2016), are readily achievable. The original references appear in the header of individual proxy files and should be cited when a record is used.

The selection criteria for the additional records are guided by the needs of the LMR and by the needs of proxy system model development (Dee et al. 2015; Evans et al. 2013). The LMR project aims to make code and data publically available to facilitate use and reuse, thus only publicly accessible or already-archived data are considered. In selecting data to add we targeted data for which forward proxy system models have already been developed, including tree rings (VS-lite; Tolwinski-Ward et al. 2011), corals (Thompson et al. 2011), and ice cores (Dee et al. 2015). In addition to their use in assimilation, availability of proxy data can improve proxy system model development (Dee et al. 2015). Two recent paleo data assimilation efforts (Hakim et al. 2016; Franke et al. 2017) have been designed to use proxy data that are annually or better resolved and that overlap with the instrumental period, over which required characteristics such as the error can be assessed and regression models developed. A third assimilation effort (Steiger et al. 2018) used a draft version of the data described here. The additions focus on proxy time series that overlap the instrumental period. All three assimilation efforts are designed to analyze many of the variables within the climate system

Table 1: Code used to process and analyze the collection and available on GitHub (Hakim et al., 2018).

Code	Purpose
LMR_proxy_preprocess.py	Process NOAA text files, perform QC, standardize on LMR conventions, build data frames and save frames as Python Pickle files.
LMR_PSMbuild.py	Build proxy system models using user-specified calibration files and options. Currently linear univariate and bivariate regressions are supported; classes allow for other models to be developed.
summarize_proxy_database.py	Calculates summary statistics for the collection organized by archive types listed in Table 2.
summarize_linearPSM.py	Analyzes linear univariate regression models.
Map_proxies.py	Maps proxy records.

Table 2: Assignment of original data variables to LMR archive and variable categories. (Note that some entries here are associated with processing of PAGES 2k Consortium (2017) data).

Archive	Variable	Original Variable Name
Tree Rings	WidthPages2	trsgi
Tree Rings	WidthBreit	trsgi
Tree Rings	WoodDensity	max_d, min_d, early_d, earl_d, late_d, MXD, density
Tree Rings	Isotopes	d18O
Corals and Sclerosponges	d18O	d18O, delta18O, d18o, d18O_stk, d18O_int, d18O_norm, d18o_avg, d18o_ave, dO18, d18O_4
Corals and Sclerosponges	SrCa	Sr/Ca, Sr_Ca, Sr/Ca_norm, Sr/Ca_anom, Sr/Ca_int
Corals and Sclerosponges	Rates	ext, calc, calcification, calcification rate, composite
Ice Cores	d18O	d18O, delta18O, delta18o, d18o, d18o_int, d18O_int, d18O_norm, d18o_norm, dO18, d18O_anom
Ice Cores	dD	deltaD, delD, dD
Ice Cores	accumulation	accum, accumu
Ice Cores	MeltFeature	MFP, melt
Lake Cores	Varve	varve', 'varve_thickness', 'varve thickness', 'thickness'
Lake Cores	Biomarkers	Uk37, TEX86, tex86
Lake Cores	GeoChem	Sr/Ca, Mg/Ca, Cl_cont
Lake Cores	Misc	RABD660_670, X_radiograph_dark_layer, massacum
Marine Cores	d18O	d18O
Marine Cores	tex86	tex86
Marine Cores	uk37	uk37, UK37
Speleothems	d18O	d18O
Bivalve	d18O	d18O

that covary with proxy records, encouraging incorporation of proxies sensitive to moisture availability, sunlight, surface pressure, water isotope ratios, and other aspects of the climate.

Data Processing

The workflow consists of 1) adding metadata and information to candidate records needed to meet NOAA archive requirements and structuring the data time series and metadata in the NOAA text format and, and 2) preprocessing the data to identify duplicates and errors, standardize variable names, and average the data to annual values. The workflow is described in the following steps followed by archive-specific notes.

Step 1. Obtain proxy time series and metadata and format in concordance with the standard developed for NOAA text files (NOAA Paleoclimatology Program 2018). Additional metadata and notes were added but the data were not altered. The Breitenmoser et al. (2014) chronologies and metadata were obtained from the authors as Matlab files. The sensitivity of a series to the environment during a specific season was provided by PAGES 2k Consortium (2017) authors, assigned by us for the Breitenmoser et al. (2014) trees (described below), or obtained if it existed from the original metadata.

Step 2. Process individual files. Identify and eliminate duplicates (see notes below): translate the original archive types and variable names to LMR types and names, compute annual January-December means (optionally April–March), optionally gaussianize the time series, make coarse checks on the time variable, standardize the time variable to Year C.E., read metadata (site name, latitude, longitude, elevation, investigators, variable attributes (including variable name and units)), read the numerical data (age and values), check the earliest year and most recent year metadata fields with the numerical data, arrange series from oldest to youngest, and identify data as missing (e.g. data reported as NaN) for years where valid data are not available. The guassianize option is provided because some paleo proxies, notably lake records, have a baseline zero

value with excursions from zero which are not successfully modeled by a linear regression. Ranking the data by point value and transforming so the distribution has a mean equal to zero while preserving the rank can improve the linear regression model. Data frames are then built consisting of 1) metadata and 2) numerical data (i.e. the actual proxy time series). The data frames are available from the LMR project documentation web pages (Tardif et al. 2018).

A critical step in compiling heterogeneous paleo data from multiple sources is to determine which variables represent the same quantity, and can be treated identically by proxy system models. The identification of similar variables is listed in **Table 2**, and accomplished in the processing step, leaving the variable names in the archive unchanged. In cases where suspect data or metadata were found the data series was rejected. No data values were changed. A Python program labeled LMR_proxy_preprocess.py performs this step. The processed data are not archived because many different processing options may be selected. Instead, the raw data are archived and the processing code made available (Hakim et al. 2018). The approach facilitates adding additional raw data that may be processed in the same way. A working version of the processed data is available for use with the LMR code (Tardif et al. 2018). The production version of the code, including the preprocessing code, is available from GitHub (Hakim et al. 2018).

The primary emphasis of the processing step and its description here is to document how the data are transformed, and how heterogenous time conventions and variable names are treated so that disparate data sets produced by different authors can be made amenable to data assimilation, while preserving the original characteristics of the data in the archive. These transformations involve several choices regarding processing that may be changed depending on the application.

Archive-specific Processing

Tree Rings

A refined set of International Tree Ring Data Bank (ITRDB) chronologies was produced by Breitenmoser et al. (2014) for use in data assimilation and forward modeling using VS-lite, and these chronologies were incorporated with minor changes and metadata added after receiving the data from the authors as Matlab files. Breitenmoser et al. (2014) developed their chronologies by first editing the raw ring width data and metadata and then applying a uniform and consistent detrending procedure and compositing to produce site chronologies. The editing increased the total number of raw ring width files that could be processed but eliminated a small number of files that contained unresolvable issues. The application of uniform processing to all sites produces a collection that differs from existing ITRDB chronologies which have been developed by different authors using different types of detrending in the standardization. As described in Breitenmoser et al. (2014) the program ARSTAN (Cook 1985) was used for processing. A hierarchical approach was used to select the ARSTAN processing options. The first preference was the negative exponential detrend, and the second preference linear detrending where appropriate. A smoothing spline with a low-frequency cut-off was used in some cases. The chronologies were developed using a biweight robust mean estimate, with the variance stabilized. After applying two quality-control criteria (every point in a chronology must be developed from at least 8 points, and the 1901–1970 period must be fully represented) the processing yielded 2287 chronologies. ARSTAN processing was only performed on the years 1600 and after in Breitenmoser, et al. (2014). Approximately 500 chronologies extend prior to 1600. For these longer series, we processed the raw ring widths obtained from the authors using the 2014 update of the ARSTAN program (44xp), selecting processing options that yielded a close match during the post-1600 interval. In most cases the selected coefficients were the ARSTAN default. For the first detrending, the default negative exponential option was selected (option 1, negative exponential curve ($k > 0$)). If that created any divide-by-zero issues, a smoothing spline was applied (with a value of –75, denoting the use of a percent smoothing rather than an absolute length). Variance stabilization was also applied (ARSTAN menu option [13][1]). No data transformation was used (ARSTAN menu item [3]). The result yields 2287 consistently processed and minimally quality-screened chronologies. We eliminated the chronologies that appear in the PAGES 2k Consortium (2017) database because for all duplicate proxy records (including ice cores and corals) we retained the PAGES 2k Consortium (2017) version owing to its greater scrutiny and quality control (PAGES 2k Consortium, 2017).

Breitenmoser et al. (2014) calculated VS-lite parameters (i.e. M1/M2 and T1/T2 soil moisture and temperature thresholds), as well as VS-Lite-based determinations of each chronologies' sensitivity to climate (precipitation/temperature) and these statistics were transferred from the matlab files to the metadata. Generalized seasonalities are included in the metadata according to the following rule: tropical trees (23.5°S to 23.5°N) were denoted as year-round records, NH trees (23.5°N to 90°N) denoted Boreal summer records (JJA), and SH trees (90°S to 23.5°S) denoted Austral summer records (DJF).

Corals

Twenty-seven coral records were added, including the Palmyra coral oxygen isotope series obtained from Linked Earth (http://wiki.linked.earth/Palmyra.Cobb.2013) (documentation appears in the file metadata). Since the publication of Cobb et al. (2003), many hundreds of measurements have been made at seasonal resolution at the same sites. In parallel, revisions to the U/Th dates have substantially altered the timing of the record. The updated record was published in Emile-Geay et al. (2013), but was not incorporated in the Ocean2k synthesis (Tierney et al. 2015).

Speleothems

Twenty speleothem records identified during discussions at the first LMR workshop were added. In contrast to the PAGES speleothem records the additions were not screened for temperature sensitivity. The speleothem records are from low latitudes including six from the Asian monsoon region.

Ice Cores

We added 77 ice core oxygen isotope records with the ultimate goal of using them with forward models (Dee et al. 2015). In contrast to the 39 PAGES 2k Consortium (2017) records they are not screened for temperature.

Data Description and Linear Regression Summary

The geographical and temporal distribution of the additional data and the complete database are shown in **Figure 1**. The archive types are unevenly distributed in time and space. Coral records (27) dominate the tropical ocean, ice core records (88) dominate at latitudes greater than 60 degrees, and the middle northern latitudes are dominated by tree rings. The Breitenmoser et al. (2014) chronologies expand the previous number of records four-fold. The additions are dominated by trees from the Northern Hemisphere mid-latitude continental areas. Going backwards in time the data density decreases rapidly from ca. 1875 to a much smaller number of sites prior to 1600, however some of the sites with data at 1600 also have observations back to 1 CE. The addition of the Breitenmoser et al. (2014) chronologies particularly increases the number of observations in the 20th century. Tree ring chronologies decrease in number in the latter part of the 20th century because many of the chronologies were collected prior to 1980. There are less than ten series within each of the following archive variable categories (**Table 3**): tree ring isotopes, bivalve $\delta^{18}O$, ice core accumulation, ice core melt feature, lake core varve, lake core misc, and marine core $\delta^{18}O$.

Within the current LMR framework, statistical relationships between proxies and key climate variables (e.g. temperature) are used as forward models (i.e. proxy system models) within the data assimilation procedure. To provide a preliminary assessment of the sensitivity of the additional data for this report, linear univariate regressions were calculated for individual proxy time series using annually-averaged calibration data (we used mean annual temperature from GISTEMP v4 (GISTEMP Team, 2018)) available over the instrumental-era (1850 to 2015). The regression parameters are similar to the parameters used by Hakim et al. (2016) on PAGES 2k Consortium (2013) proxy time series. The models are formulated as the relationship (slope (m) and intercept) between the expected proxy value (y_e) and the independent environmental variables (x, in this case temperature) (for example, $y_e = m^*x$ + intercept). Regression statistics are included on the 'dashboard' graphical data summary pages and the archive summaries. Correlations (R^2) relate the expected to the observed values. The program labeled LMR_PSMbuild.py performs this step. The linear univariate annual regressions are intended to be a first step in data exploration, extending the approach used by Hakim et al. (2016) to the the additional data. Other models, particularly models including moisture related variables, can be explored and may be more appropriate for the moisture-sensitive proxies that we have added. These comparisons will be reported elsewhere.

The development of series-specific regression models reduces the number of records that can be assimilated from 2290 to 2244. The linear regression to annual mean temperature used by Hakim et al. (2016) is only one of many possible relationships with the environment, and indeed the purpose of structuring this collection and including moisture-sensitive proxies such as the Breitenmoser et al. (2014) chronologies and ice core oxygen isotopes and coral oxygen isotopes is to support exploration of alternative models, including bivariate regression models (temperature and moisture) and other proxy system models (Dee et al. 2015). Cases where a regression was not produced are due to lack of observations, either proxy observations or calibration observations near the proxy location during the target calibration interval (1850–2015). This reduces the ice core records from 77 to 61, corals from 19 to 11, and speleothem records from 20 to 6 (**Table 3**). The criteria used here follow Hakim et al. (2016): a linear regression of the annualized proxy values to annual mean temperature of the corresponding grid cell from the GISTEMP v4 temperature product,

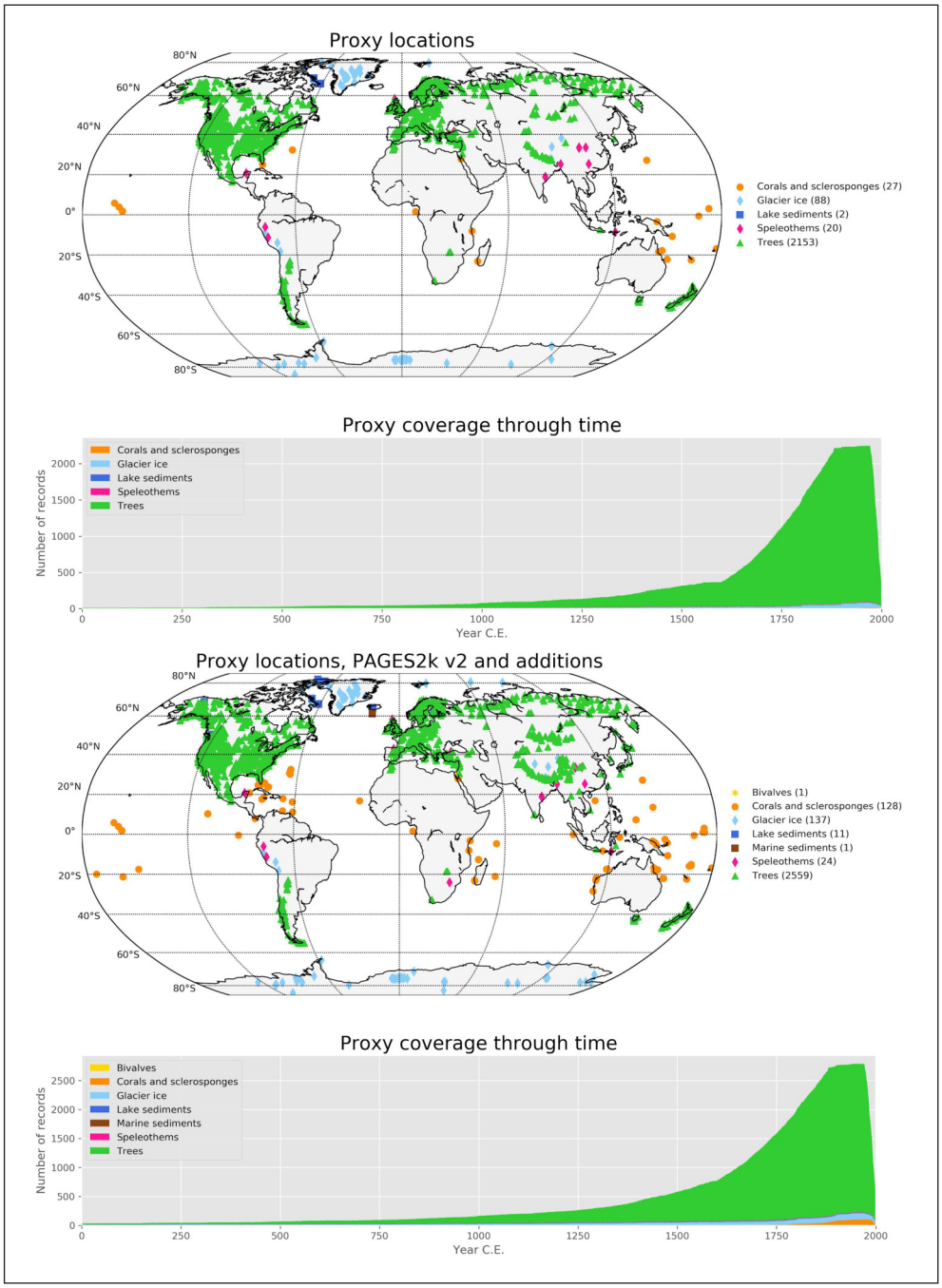

Figure 1: Proxy site locations (coded by archive) and proxy coverage through time for the additional data (top) and the complete database LMRdb 1.0.0 (bottom).

for the period 1850–2015, with a requirement of annual proxy resolution, and 50% data available over the period 1850–2015. No minimum correlation threshold is set for models. Records with poor fits have less weight in the assimilation.

The regression summary statistics (**Table 4**) support consideration of all these archive types in assimilation. Statistics based on fewer than 10 records are struck-through and not described here, however we present the R^2 information on the dashboards for each record. Several measures are useful in assessing how well the regression models fit the observations. The correlation (R^2) directly reflects the fit of the regression model and appears for each record in the online supplement. The signal to noise ratio is used in archive summaries to compare different proxies. The coefficient of efficiency used by us (Hakim. et al. 2016) assesses the skill of the assimilation analysis. Here it is applied to the point-based reconstructions. Coral $\delta^{18}O$ (37.82) has the highest signal-to-noise (SNS) ratio, followed by ice core $\delta^{18}O$. These records also have the highest coefficient of efficiency. The mean SNS decreases from 4.3 to 3.5 and mean CE decreases from 0.033 to 0.030 for chronologies

Table 3: Processing of additional proxy record files.

Proxy Variable	PAGES 2k	Additions	Calibrated Additions
Bivalve_d18O	1	0	0
Corals and Sclerosponges_Rates	8	1	1
Corals and Sclerosponges_SrCa	28	7	3
Corals and Sclerosponges_d18O	65	19	11
Ice Cores_Accumulation	0	3	3
Ice Cores_MeltFeature	2	0	0
Ice Cores_d18O	39	77	61
Ice Cores_dD	8	8	5
Lake Cores_Misc	3	0	0
Lake Cores_Varve	6	2	2
Marine Cores_d18O	1	0	0
Speleothems_d18O	4	20	6
Tree Rings_Isotopes	0	1	1
Tree Rings_WidthBreit	0	2152	2152
Tree Rings_WidthPages2	347	0	0
Tree Rings_WoodDensity	59	0	0
Total	571	2290	2244

Table 4: Regression summary statistics for each proxy type. Signal to noise ratio (SNS) and coefficient of efficiency (CE) minimum (Min), Mean, and maximum (Max), based on linear regression with respect to annual temperature and comparison to observed values. Struck-through values are based on fewer than 10 proxy records.

Proxy	Min	Mean	Max	Min	Mean	Max
~~Corals and Sclerosponges_SrCa~~	~~2.57~~	~~13.70~~	~~35.39~~	~~0.03~~	~~0.17~~	~~0.42~~
Corals and Sclerosponges_d18O	0.00	37.82	175.81	0.00	0.24	0.62
~~Ice Cores_Accumulation~~	~~1.61~~	~~12.32~~	~~32.74~~	~~0.02~~	~~0.14~~	~~0.39~~
Ice Cores_d18O	0.00	5.39	33.96	0.00	0.06	0.25
~~Ice Cores_dD~~	~~2.07~~	~~6.42~~	~~17.21~~	~~0.05~~	~~0.08~~	~~0.14~~
~~Lake Cores_Varve~~	~~2.89~~	~~5.03~~	~~7.16~~	~~0.03~~	~~0.04~~	~~0.06~~
~~Speleothems_d18O~~	~~0.01~~	~~2.06~~	~~9.61~~	~~0.00~~	~~0.02~~	~~0.08~~
~~Tree Rings_Isotopes~~	~~16.16~~	~~16.16~~	~~16.16~~	~~0.13~~	~~0.13~~	~~0.13~~
Tree Rings_WidthBreit	0.00	3.51	44.73	0.00	0.03	0.28

when all trees (irrespective of their sensitivity to temperature or moisture) are considered compared to the chronologies screened for the sensitivity to temperature (not shown). This suggests that alternate regressions, involving moisture in addition to temperature, could provide more accurate proxy modeling. Nevertheless, differences remain small and the similar goodness-of-fit measures compared to the extensively screened PAGES 2k Consortium (2017) tree ring width set (not shown) justifies their consideration in the LMR data assimilation framework, particularly given the enhanced geographic coverage these chronologies provide, for example in the Southern Andes and Tasmania (**Figure 1**). Breitenmoser et al. (2014) note that 7% of the analyzed land grid cells contain only temperature sensitive series, 5% contain only moisture-limited series, and 37% contain both temperature and moisture-limited chronologies. The record-specific regression equations, along with the R^2 values, are reported in the online supplement (data dashboards).

Data Records

The principal product is the set of 2290 structured files consisting of metadata and data in the NOAA text format archived at NOAA Paleoclimatology at the National Centers for Environmental Information (Anderson et al. 2018). The file naming convention for the non-tree records is yyxxxxzz.txt, where yy is the last two digits of the publication year, xxxx is a four letter abbreviation of the site name or code, and zz is a two digit sequence number to account for multiple records from the same publication and site. An exception to this convention is the Palmyra coral record, which is named palmyra2013.txt. The site naming convention for the Breitenmoser tree chronologies is a variable-length concatenation of the region, underscore, the ITRDB code, and the letter B (to indicate the source (Breitenmoser et al. 2014)). The file extension is txt. In the NOAA metadata Collection_Name is this filename without the extension. We assigned or wrote metadata for each file according to the NOAA Paleo metadata v1.0 instructions (NOAA Paleoclimatology, 2018). In the notes section of the Breitenmoser et al. (2014) tree chronologies, we added the VS-lite parameters reported by Breitenmoser et al. (2014) coded as key/value pairs following this example using terminology from Breitenmoser et al. (2014): {"VSLite_parameters": {"T1": "5.55447702235", "T2": "18.6547994501", "M1": "0.022191930217", "M2": "0.313406788974"}. The expected seasonality for the Tree Rings was added to the metadata notes section according to the following rule. The seasonality is all months [1, ..12] if the site is located in the tropics and assigned Boreal summer [6, 7, 8] if located in the Northern Hemisphere and Austral summer [11, 12, 1] if located in the Southern Hemisphere, as mentioned earlier.

We emphasize that the data records (files) contain the original data series obtained from the source listed in the metadata. We anticipate that users wanting the processed time series will use the Python code LMR_proxy_preprocess.py (with the options, such as making the annual mean, that suits their application).

Site Dashboards

Single-page summaries similar to those developed for the PAGES 2k Consortium data are provided in the online supplement (https://www.atmos.uw.edu/~hakim/LMR/proxy_dashboards/), and include a graph of the time series and a summary of the metadata. The complete metadata for each site are included in the data files in the NOAA text format. The dashboard title identifies the archive type, the site name, and a label consisting of core identifier:variable. The dashboards are grouped by archive type and alphabetized according to the core identifier in the title. This core identifier is also the name of the corresponding data file, simplifying the task of browsing the data. The dashboards show the time series of the raw and annualized data, with lines connecting the annualized data when data is present for consecutive years. Note that the time values for annualized data are given as integer years while subannual data has decimals corresponding to the exact date, which sometimes results in a slight visual offset between the raw and annualized data. The dashboards also show the proxy location, some metadata, and scatter plots with linear regressions between the annualized proxy and annual, JJA, and DJF-mean GISTEMP temperature for all valid years between 1850–2015. The regression with the best fit (highest R^2) is indicated by a bold-font title. As discussed earlier, some of the proxies are moisture sensitive, so a stronger relationship between the proxy and climate may be found through alternate regressions. Regardless, univariate temperature regressions offer an initial perspective on the proxies. The metadata shown on the dashboards includes information from the structured files, namely the study name, investigators, site name, latitude, longitude, elevation (m), the archive type (**Table 2**), and the proxy measurement (original variable name). The oldest (C.E.) and youngest (C.E.) metadata fields were calculated by us and refer to the annualized data. The "climate variable" metadata field reported for tree rings specifies whether the proxy was classified as primarily temperature-limited or moisture-limited according to the analysis in Breitenmoser et al. (2014). The data records (files) contain original investigator-reported values.

Usage Notes

1. The original references found in the NOAA text format files should be cited when these data are used.
2. Additional information about the records (metadata) beyond that provided in this archive exists in the files and supplementary information provided by Breitenmoser et al. (2014), and in the original references listed in the metadata.
3. Users wanting the processed data instead of the raw data provided in these files can process the archive using the Python code LMR_proxy_preprocess.py which produces two Pickle files, one containing the annualized time series and one containing the metadata. Code is available on GitHub (Hakim et al. 2018).
4. Users wanting to add or modify a collection for use with LMR code can remove or add NOAA formatted files to their own directory containing a copy of the collection. The code recognizes

the original variables listed in **Table 2**, and the following time units: ['age', 'age_int', 'year',\'y_ad', 'Age_AD', 'age_AD', 'age_AD_ass', 'age_AD_int', 'Midpt_year', 'AD',\'age_yb1950', 'yb_1950', 'yrb_1950',\'kyb_1950',\'yb_1989', 'age_yb1989',\'yb_2000', 'yr_b2k', 'yb_2k', 'ky_b2k', 'kyb_2000', 'kyb_2k', 'kab2k', 'ka_b2k', 'kyr_b2k',\'ky_BP', 'kyr_BP', 'ka_BP', 'age_kaBP', 'yr_BP', 'calyr_BP', 'Age(yrBP)', 'age_calBP', 'cal yr BP']. Files not using these time and variable names will not be processed.

5. In addition to the code used for data assimilation, LMR code can be used to analyze the collection (**Table 2**). Proxy system models can be built, and the characteristics of the data and the models can be analyzed using LMR_PSMbuild.py, summarize_proxy_database.py, and summarize_linearPSM.py codes.

Summary and Application

We describe 2290 additions to 571 previously available proxy records, assembled and structured to facilitate reanalysis of the last millennium. The two data sets are merged by processing software to form the Last Millennium Reanalysis dataset version LMRdb v1.0.0. The additions provide a four-fold increase in the number of sites, improve the data density in the Southern Hemisphere, and massively increase the density of Northern Hemisphere tree ring chronologies spanning the last two centuries. The collection consists of observations in their native units and is all-encompassing, minimally-screened, and minimally-processed. Many of the additional records, notably tree ring widths and ice and cave speleothem oxygen isotopes are sensitive to precipitation and other hydrologic variables, and are expected to enable the reanalysis of variables other than temperature, including precipitation, drought indices, and surface pressure. The intended applications include 1) the development and testing of proxy system models, 2) climate reconstruction using various methods such as climate field reconstructions, and 3) reanalysis combining proxy system models and reconstructions derived from model simulations.

Additional Files

The additional files for this article can be found as follows:

- **Archived Data.** Example data file 14kiri01a. DOI: https://doi.org/10.5334/dsj-2019-002.s1
- **Proxy data and metadata.** Corals and Sclerosponges|Butaritari Atoll, Gilbert Islands|14kiri01a:d180. DOI: https://doi.org/10.5334/dsj-2019-002.s2

Acknowledgements

Comment and suggestions from two anonymous referees were helpful in revision and are gratefully acknowledged. We also thank the LMR advisory panel, Kim Cobb, Kevin Anchukaitis, Gil Compo, Michael N. Evans, and Thorsten Kiefer, for their guidance and suggestions.

Funding Information

This research was supported by grants from the National Science Foundation (grant AGS-1304263 to the University of Washington) and the National Oceanic and Atmospheric Administration (grant NA14OAR4310176).

Competing Interests

The authors have no competing interests to declare.

References

Anderson, DM, Tardif, R, Horlick, K, Erb, MP, Hakim, GJ, Noone, D, Perkins, WA and **Steig, EJ.** 2018. *Last millenium reanalysis multi-proxy database.* Available at: https://www.ncdc.noaa.gov/paleo/study/24611 (Accessed: 11 July 2018).

Breitenmoser, P, Brönnimann, S and **Frank, D.** 2014. 'Forward modelling of tree-ring width and comparison with a global network of tree-ring chronologies'. *Clim. Past*, 10(2): 437–449. Copernicus Publications. DOI: https://doi.org/10.5194/cp-10-437-2014

Cook, ER. 1985. *A time series approach to tree-ring standardization.* University of Arizona.

Cook, ER, Anchukaitis, KJ, Buckley, BM, D'Arrigo, RD, Jacoby, GC and **Wright, WE.** 2010. 'Asian Monsoon Failure and Megadrought During the Last Millennium'. *Science*, 328(5977): 486–489. DOI: https://doi.org/10.1126/science.1185188

Dee, S, Emile-Geay, J, Evans, MN, Allam, A, Steig, EJ and **Thompson, DM.** 2015. 'PRYSM: An open-source framework for PRoxY System Modeling, with applications to oxygen-isotope systems'. *Journal of Advances in Modeling Earth Systems*, 7(3): 1220–1247. DOI: https://doi.org/10.1002/2015MS000447

Emile-Geay, J, Cobb, KM, Mann, ME and **Wittenberg, AT.** 2013. 'Estimating Central Equatorial Pacific SST Variability over the Past Millennium. Part I: Methodology and Validation'. *Journal of Climate*, 26: 2302–2328. DOI: https://doi.org/10.1175/JCLI-D-11-00510.1

Emile-Geay, J, Hakim, GJ, Steig, EJ, Noone, D, Anderson, D, Tardif, R, Perkins, W, Steiger, N, Horlick, K and **Erb, MP.** 2018. *Last Millennium Reanalysis*. Available at: https://www.researchgate.net/project/Last-Millennium-Reanalysis (Accessed: 3 March 2018).

Evans, MN, Tolwinski-Ward, SE, Thompson, DM and **Anchukaitis, KJ.** 2013. 'Applications of proxy system modeling in high resolution paleoclimatology'. *Quaternary Science Reviews*, 76: 16–28. DOI: https://doi.org/10.1016/j.quascirev.2013.05.024

Franke, J, Brönnimann, S, Bhend, J and **Brugnara, Y.** 2017. 'Data Descriptor : A monthly global paleo-reanalysis of the atmosphere from 1600 to 2005 for studying past climatic variations'. *Scientific Data*, 4: 1–19. Available at: https://doi.org/10.1038/sdata.2017.76

GISTEMP TEAM. 2018. *GISS Surface Temperature Analysis (GISTEMP)*. Available at: https://data.giss.nasa.gov/gistemp/ (Accessed: 22 May 2017).

Hakim, GJ. 2018. *Last millennium reanalysis (LMR) framework code base*. Available at: https://github.com/modons/LMR (Accessed: 20 August 2018).

Hakim, GJ, Emile-Geay, J, Steig, EJ, Noone, D, Anderson, DM, Tardif, R, Steiger, N and **Perkins, WA.** 2016. 'The last millennium climate reanalysis project: Framework and first results'. *Journal of Geophysical Research: Atmospheres*, 121(12): 6745–6764. Wiley-Blackwell. DOI: https://doi.org/10.1002/2016JD024751

Mann, ME, Bradley, RS and **Hughes, MK.** 1998. 'Global-scale temperature patterns and climate forcing over the past six centuries'. *Nature*, 392: 779–787. DOI: https://doi.org/10.1038/33859

Mann, ME, Zhang, Z, Hughes, MK, Bradley, RS, Miller, SK, Rutherford, S and **Ni, F.** 2008. 'Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia'. *Proceedings of the National Academy of Sciences of the United States of America*, 105(36): 13252–7. DOI: https://doi.org/10.1073/pnas.0805721105

Masson-Delmotte, V, Schulz, M, Abe-Ouchi, A, Beer, J, Ganopolski, A, González Rouco, JF, Jansen, E, Lambeck, K, Luterbacher, J, Naish, T, Osborn, T, Otto-Bliesner, B, Quinn, T, Ramesh, R, Rojas, M, Shao, X and **Timmermann, A.** 2013. 'Information from Paleoclimate Archives'. In: Stocker, TF, Qin, D, Plattner, G-K, Tignor, M, Allen, SK, Boschung, J, Nauels, A, Xia, Y, Bex, V and Midgley, PM (eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Mckay, NP and **Emile-Geay, J.** 2016. 'Technical note: The Linked Paleo Data framework – a common tongue for paleoclimatology'. *Clim. Past*, 12: 1093–1100. DOI: https://doi.org/10.5194/cp-12-1093-2016

McKay, NP and **Kaufman, DS.** 2014. 'An extended Arctic proxy temperature database for the past 2,000 years'. *Scientific Data*, 1: 1–10. DOI: https://doi.org/10.1038/sdata.2014.26

Neukom, R, Gergis, J, Karoly, DJ, Wanner, H, Curran, M, Elbert, J, González-rouco, F, Linsley, BK, Moy, AD, Mundo, I, Raible, CC, Steig, EJ, Van Ommen, T, Vance, T, Villalba, R, Zinke, J and **Frank, D.** 2014. 'Inter-hemispheric temperature variability over the past millennium'. *Nature Climate Change*, 4(5): 362–367. May. DOI: https://doi.org/10.1038/NCLIMATE2174

NOAA Paleoclimatology. 2018. *WDC Paleo Data Submission template, and sample data description*. Available at: https://www1.ncdc.noaa.gov/pub/data/paleo/templates/noaa-wdc-paleo-template-instructions.txt (Accessed: 3 March 2018).

PAGES 2k Consortium. 2013. 'Continental-scale temperature variability during the past two millennia'. *Nature Geoscience*, 6(5). DOI: https://doi.org/10.1038/ngeo1797

PAGES 2k Consortium. 2017. 'A global multiproxy database for temperature reconstructions of the Common Era'. *Scientific Data*, 4(170088): 1–33. DOI: https://doi.org/10.1038/sdata.2017.88

PAGES 2k Consortium. 2017a. *PAGES 2k Global 2,000 Year Multiproxy Database*. Available at: https://www.ncdc.noaa.gov/paleo/study/21171 (Accessed: 4 December 2018).

Smerdon, JE, Luterbacher, J, Phipps, SJ, Anchukaitis, KJ, Ault, T, Coats, S, Cobb, KM, Cook, BI, Colose, C, Felis, T, Gallant, A, Jungclaus, JH, Konecky, B, LeGrande, A, Lewis, S, Lopatka, AS, Man, W, Mankin, JS, Maxwell, JT, Otto-Bliesner, BL, Partin, JW, Singh, D, Steiger, NJ, Stevenson, S, Tierney, JE, Zanchettin, D, Zhang, H, Atwood, AR, Andreu-Hayles, L, Baek, SH, Buckley, B, Cook, ER, Arrigo,

R, Dee, SG, Griffiths, M, Kulkarni, C, Kushnir, Y, Lehner, F, Leland, C, Linderholm, HW, Okazaki, A, Palmer, J, Piovano, E, Raible, CC, Rao, MP, Scheff, J, Schmidt, GA, Seager, R, Widmann, M, Williams, AP and **Xoplaki, E.** 2017. 'Comparing proxy and model estimates of hydroclimate variability and change over the Common Era'. *Climate of the Past Discussions*, 13(12): 1–70. DOI: https://doi.org/10.5194/cp-2017–37

Steiger, NJ, Smerdon, JE, Cook, ER and **Cook, BI.** 2018. 'A reconstruction of global hydroclimate and dynamical variables over the Common Era'. *Scientific Data*, 5: 1–15. DOI: https://doi.org/10.1038/sdata.2018.86

Tardif, R, Perkins, WA, Hakim, GJ, Brennan, MK, Badgeley, JA, Parsons, LA and **Erb, MB.** 2018. *Last millennium reanalysis docs v3.0*. Available at: https://atmos.washington.edu/~hakim/LMR/docs/index.html.

Thompson, DM, Ault, TR, Evans, MN, Cole, JE and **Emile-Geay, J.** 2011. 'Comparison of observed and simulated tropical climate trends using a forward model of coral d18O'. *Geophysical Research Letters*. DOI: https://doi.org/10.1029/2011GL048224

Tierney, JE, Abram, NJ, Anchukaitis, KJ, Evans, MN, Giry, C, Kilbourne, KH, Saenger, CP, Wu, HC and **Zinke, J.** 2015. 'Tropical sea surface temperatures for the past four centuries reconstructed from coral archives'. *Paleoceanography*, 30: 226–252. October. DOI: https://doi.org/10.1002/2014PA002717

Tolwinski-Ward, SE, Evans, MN, Hughes, MK and **Anchukaitis, KJ.** 2011. 'An efficient forward model of the climate controls on interannual variation in tree-ring width'. *Climate Dynamics*, 36(11–12): 2419–2439. DOI: https://doi.org/10.1007/s00382-010-0945-5

How to cite this article: Anderson, DM, Tardif, R, Horlick, K, Erb, MP, Hakim, GJ, Noone, D, Perkins, WA and Steig, E. 2019. Additions to the Last Millennium Reanalysis Multi-Proxy Database. *Data Science Journal*, 18: 2, pp. 1–11. DOI: https://doi.org/10.5334/dsj-2019-002

Submitted: 08 July 2018 **Accepted:** 12 December 2018 **Published:** 07 January 2019

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS