Towards a semantic web of paleoclimatology

Julien Emile-Geay

Department of Earth Sciences, University of Southern California, Los Angeles, CA

Jason A. Eshleman

IO Informatics, Berkeley, CA

Abstract. The paleoclimate record is information-rich, yet signifiant technical barriers currently exist before this record can be used to answer scientific questions. Here we make the case for a universal format to structure paleoclimate data. A simple example demonstrates the scientific utility of such a self-contained way of organizing coral data and meta-data. This example is generalized to a universal ontology that may form the backbone of an open-source, open-access and crowd-sourced paleoclimate database. The format is designed to enable semantic searches, and is expected to accelerate discovery on topical scientific problems like climate extremes, the characteristics of natural climate variability, and climate sensitivity to various forcings.

1. Introduction

Science is based on building on, reusing and openly criticising the published body of scientific knowledge. For science to effectively function, and for society to reap the full benefits from scientific endeavours, it is crucial that science data be made open.
Panton Principles [*Murray-Rust et al.*, 2010]

In the past decade or so, the paleoclimate community has embraced these principles to a remarkable extent, so much so that a vast number of proxy records from all parts of the world are now publicly available through online databases like the International Tree Ring Data Bank [*ITRDB*, 2009], the National Climatic Data Center[1] or PANGAEA [2]. The benefits of making this information freely available are immense. In principle, it may:

- facilitate reproduction of published results, a core principle of modern science
- enable the exploration of all available data, and ease comparisons between all relevant proxy records through meta-searches.
- serve as the basis for multiproxy reconstructions of past climates
- widen the of citizen participation to individuals outside the relatively small community of climate science specialists

Yet, as currently archived, this information does not allow to answer relatively simple questions like: " How anomalous is twentieth century climate in the context of the past 2,000 years?" or "What is the relationship between climate forcing and observed climate variations in the paleoclimate record?". To be sure, no simple answer is expected to either question; nonetheless, a persistent and seldom-recognized barrier to the optimal use of this information is that the diversity of proxies, measurements, and chronological information requires a flexible format that maintains sufficient metadata. So far, no universal format has been established to do so optimally; the synthesis of climate proxy information is therefore labor-intensive, unnecessarily tedious and increases the probability of introducing human errors.

Such obstacles to the access to, and analysis of, paleoclimate data would largely vanish if data and metadata were archived in a machine-readable format, as has been demonstrated in other scientific domains (e.g. BioDash [*Neumann and Quan*, 2006], QuakeML [*Schorlemmer et al.*, 2004], GenBank [*Benson et al.*, 2011], MAGE-ML [*Spellman et al.*, 2002]). Microarray experimental metadata and parameters can be searched and complete datasets rapidly retrieved from two centralized best-practice repositories, Gene Expression Omnibus (GEO) [*Edgar et al.*, 2002] and ArrayExpress [*Parkinson et al.*, 2007]. This ability to rapidly construct novel datasets with all appropriate data and metadata has enabled data-mining and machine-learning algorithms to be applied to a much wider dataset than ever possible before. Such work, to give but a few examples, has allowed for the discovery and validation of robust biomarkers for organ transplant rejection [*Chen et al.*, 2010], characterize gene function [*Pierre et al.*, 2010], and detect patterns of gene coexpression [*Adler et al.*, 2009] . Similar breakthroughs would be possible in climate science if online archives conformed to relatively simple principles.

In this article we discuss the benefits of designing a universal, self-contained and flexible format for the archival of paleoclimate data. To further motivate this technical operation and illustrate its scientific merit, we start by describing a currently existing format used in recent publications, demonstrating how it can enable rapid information extraction. We then propose desirable extensions of this format, motivated by broad scientific goals. Finally, we discuss the pros and cons of our approach and outline an achievable path towards the stated objectives.

2. The power of well-organized data

2.1. Database format

The USC climate dynamics group recently compiled all publicly available coral records and archived them into a format suitable for the Matlab programming language. The format organizes data as a Matlab structure with fields describing the proxy class (in this case `class: 'CORAL'`), the type of measurement (e.g. $\delta^{18}O$, Sr/Ca , extension rate, fluorescence), includes the raw chronology (`chron`) and raw data (`data`) as vectors, as well as metadata such as latitude (`lat`, a scalar), longitude (`lon`, a scalar), the site name (`site`, a character string), a quick bibliographic reference (`reference`, a string), and a cite key to a bibliographical software like EndNote or BibTeX. For a concrete example, consider the Sr/Ca record from Palmyra Island from *Nurhati et al.* [2011].

Copyright 2012 by the American Geophysical Union.

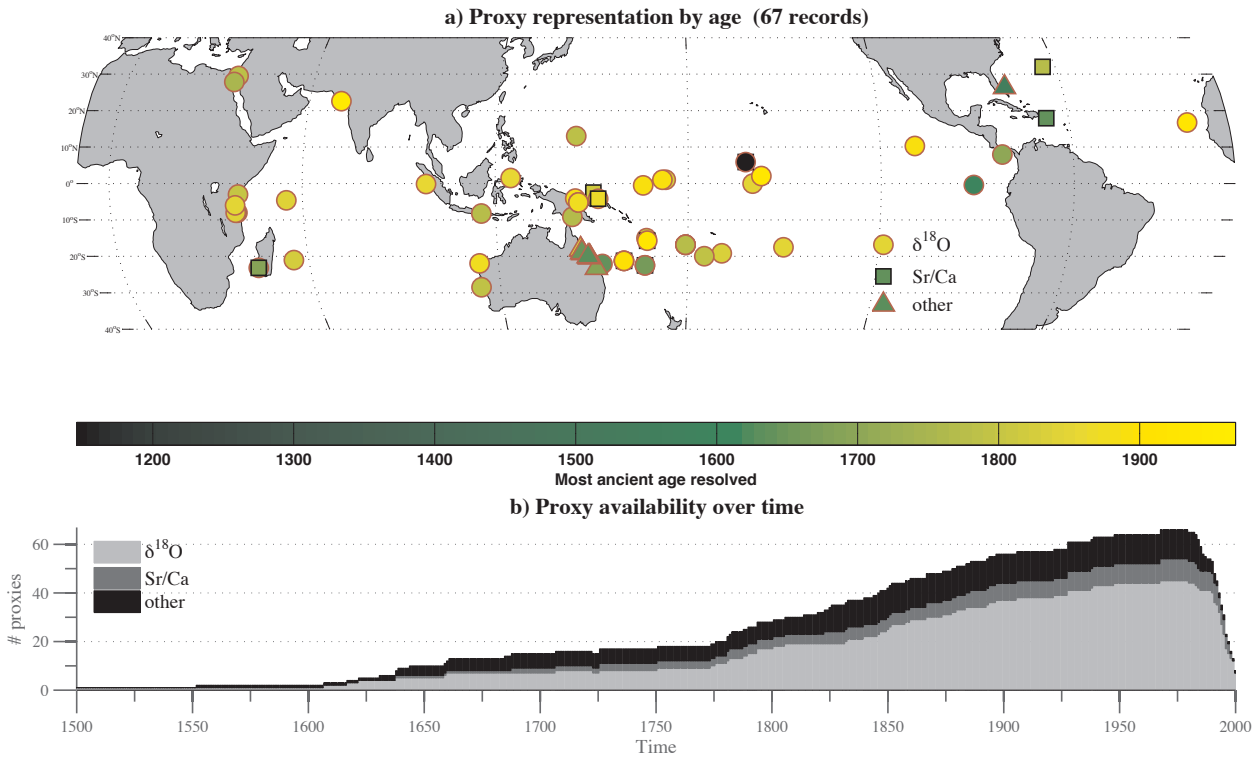

Figure 1. An information-rich representation of the coral database introduced herein. a) Location of each record, stratified by proxy types δ^{18}O (circles), Sr/Ca (squares), "other", (triangles, covering fluorescence and extension rate)) coded by shape and most ancient age resolved. b) Proxy availability over time, against stratified by proxy type.

The corresponding database entry reads:

```
class: 'CORAL'
type: 'Sr/Ca'
chron: [1347x1 double]
data: [1347x1 double]
year_i: 1886
year_f: 1998
raw_res: 1.0000
lat: 5.8833
lon: 197.9167
site: 'Palmyra'
reference: 'Nurhati et al [2011]'
cite_key: 'Nurhati2011'
```

Accompanying the aforementioned fields are others that are calculated from them and enable a rapid parsing of the database. These include the initial year present in the chronology (`year_i`), the final year (`year_f`), the average resolution in months (`raw_res`). Such quantities will soon prove useful to select records on the basis of their time coverage or resolution. In addition to these fields, one may want to process the raw proxy information to obtain more meaningful quantities. For instance, the data may have been collected at uneven time intervals, so it would need to be interpolated onto a regular time grid for certain types of analysis (e.g. spectral analysis, EOF analysis) to be performed. In addition, a monthly-resolved record such as this one exhibits a prominent seasonal cycle that one may want to remove prior to analysis. This can be done very simply by adding the corresponding data vectors `chron_reg` and `data_reg`, the new resolution `res` (in this case, `res = ... raw_res = 1`) and a vector of monthly anomalies `anom`).

```
data_reg: [1347x1 double]
chron_reg: [1347x1 double]
res: 1.0000
anom: [1347x1 double]
```

This strategy had the advantage of enabling a "journaled" approach to data processing; it makes clear which transformations (e.g. standardization, detrending, deseasonalization) were applied to each proxy record and enables users to go back to the raw data and apply their own processing steps if so desired. Let us now see what scientific information may be derived from this form of organization.

2.2. Analysis

The first thing one might want to do is visualize the spatiotemporal coverage of the database. This may be done in a few lines of code, and is illustrated in Fig 1.

Next, one may want to inspect the data themselves. It is easy to annualize all records (using DJF averages for those subannually-resolved records) and assemble them all into a single matrix, which is plotted in Fig 2 to obtain a synoptic view of the dataset. The well-documented tendency for negative δ^{18}O and Sr/Ca values in the late twentieth century (probably due to large-scale warming and/or reorganizations of the hydrological cycle, see *Gagan et al.* [2000]) is then immediately apparent.

Finally, one may want to perform a more revealing analysis of climate variability as portrayed by these records. The extensive meta-data and the structured array syntax allow the user to easily subset records according to several criteria, e.g. finding records containing δ^{18}O measurements only, with a quarterly resolution or finer, at least 100 years in length and overlapping. This leaves 25 records out of the original 67 (Table 1), covering the period 1893 to 1984. These records can be readily annualized (taking December-January-February averages for monthly records, cold season averages otherwise), assembled into a matrix and an empirical orthogonal function (EOF) analysis [*Lorenz*, 1956] may be readily performed.

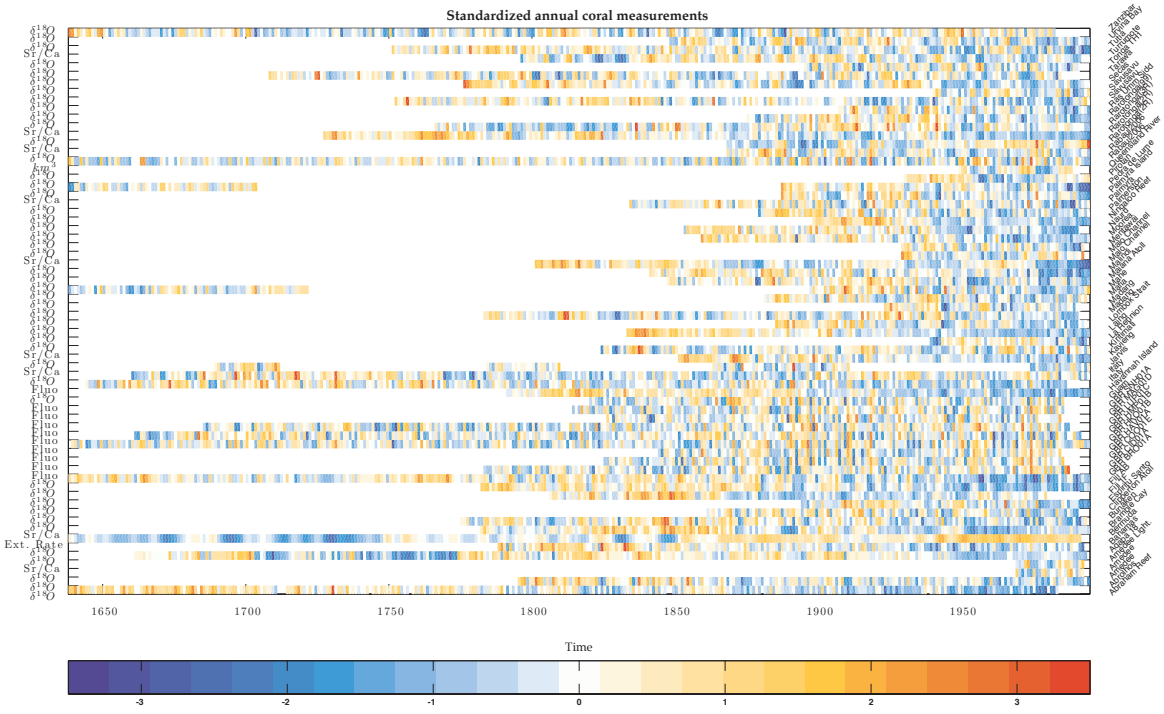

Figure 2. Synopsis of the coral matrix. Colors represent Z-scores with records sorted in alphabetical order.

Table 1. List of coral records used for the EOF analysis of Fig 3. The records were obtained from the full database (illustrated in Fig 1) after applying the following conditions: δ^{18}O only, quarterly resolution or finer, at least 100 years long, and overlapping. Time resolution is expressed in months. The quantity ρ represents the linear correlation to local temperature derived from the HadSST2 dataset [*Rayner et al.*, 2006]. Correlations are assessed using a non-parametric Monte Carlo test against 1000 isopersistent red noise timeseries ; significant correlations are denoted in bold.

	Site	Type	Resolution	Latitude	Longitude	Range	ρ	Reference
1	Abrolhos	$\delta^{18}O$	1	28°S	114°E	1794–1994	**-0.27**	*Kuhnert et al.* [1999]
2	Amedee Light.	$\delta^{18}O$	3	22°S	166°E	1660–1993	**-0.24**	*Quinn et al.* [1998]
3	Bunaken	$\delta^{18}O$	1	2°N	125°E	1860–1990	**-0.21**	*Charles et al.* [2003]
4	Clipperton Atoll	$\delta^{18}O$	1	10°N	109°W	1893–1994	**-0.27**	*Linsley et al.* [2000]
5	Guam	$\delta^{18}O$	1	13°N	145°E	1790–2000	**-0.32**	*Asami et al.* [2005]
6	Ifaty	$\delta^{18}O$	2	23°S	44°E	1659–1995	**-0.11**	*Zinke et al.* [2004]
7	La Reunion	$\delta^{18}O$	2	21°S	55°E	1832–1995	-0.07	*Pfeiffer et al.* [2004]
8	Laing	$\delta^{18}O$	3	4°S	145°E	1884–1993	**-0.37**	*Tudhope et al.* [2001]
9	Lombok Strait	$\delta^{18}O$	1	8°S	116°E	1782–1990	**-0.13**	*Charles et al.* [2003]
10	Madang	$\delta^{18}O$	3	5°S	146°E	1880–1993	**-0.22**	*Tudhope et al.* [2001]
11	Mafia	$\delta^{18}O$	2	8°S	40°E	1622–1722	**-0.20**	*Damassa et al.* [2006]
12	Mahe	$\delta^{18}O$	1	5°S	55°E	1846–1995	**-0.32**	*Charles et al.* [1997]
13	Maiana Atoll	$\delta^{18}O$	2	1°N	173°E	1840–1994	**-0.46**	*Urban et al.* [2000]
14	Mentawai	$\delta^{18}O$	1	0°N	99°E	1858–1997	**-0.37**	*Abram et al.* [2008]
15	Moorea	$\delta^{18}O$	2	18°S	150°W	1852–1990	-0.06	*Boiseau et al.* [1998]
16	Ningaloo Reef	$\delta^{18}O$	2	22°S	114°E	1878–1995	**-0.24**	*Kuhnert et al.* [2000]
17	Palmyra	$\delta^{18}O$	1	6°N	162°W	1886–1998	**-0.62**	*Cobb et al.* [2001]
18	Rabaul2006	$\delta^{18}O$	1	4°S	152°E	1867–1997	**-0.23**	*Quinn et al.* [2006]
19	Rarotonga(2R)	$\delta^{18}O$	1	21°S	160°E	1726–1996	**-0.18**	*Linsley et al.* [2008]
20	Rarotonga(3R)	$\delta^{18}O$	1	21°S	160°E	1874–2000	**-0.20**	*Linsley et al.* [2006]
21	Ras Umm Sidd	$\delta^{18}O$	2	28°N	34°E	1751–1995	**-0.24**	*Felis et al.* [2000]
22	Secas	$\delta^{18}O$	1	8°N	82°W	1707–1984	**-0.11**	*Linsley et al.* [1994]

The first mode of this EOF analysis is presented in Fig 3. As expected, the spatial pattern of sea-surface temperature (SST) associated with this mode bears a strong resemblance to El Niño-Southern Oscillation (ENSO), as confirmed by its temporal expression (PC1), which displays maxima and minima coincident with known ENSO events. The MTM spectrum of the mode reveals a relatively strong annual component and dominant interannual variability, consistent with what is independently known about ENSO [e.g. *Sarachik and Cane*, 2010]. The relative area covered by each circle illustrates the magnitude of the EOF coefficients ("loadings"), while the color refers to their sign (after multiplying δ^{18}O values by -1 so that negative excursions in oxygen isotopes correspond to positive temperature anomalies). Their signs are consistent with the ENSO thermal signal expected in each isotopic record, though the disparate amplitudes reflect the varying influence of other factors (climatic or not).

That a pan-tropical network of subannually-resolved coral records may capture ENSO variability has been amply demonstrated before [*Evans et al.*, 2000, 2002]. It is merely confirmed by the present analysis, with the difference that this information was extracted in less than 200 lines of code (including visualiza-

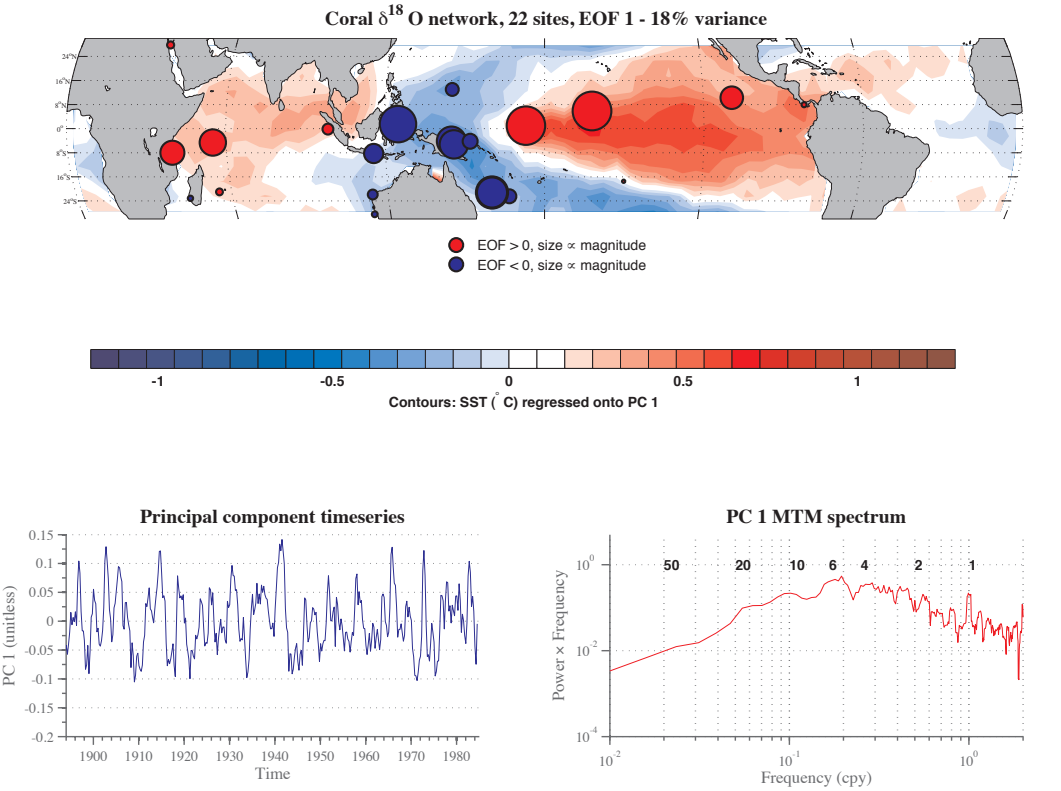

Figure 3. First EOF mode of the network of 25 δ^{18}O records presented in Table 1. (top) EOF coefficients (blue < 0, red > 0) overlain on a map of HadSST2 DJF temperature regressed onto the first principal component. δ^{18}O values were multiplied by -1 so that positive excursions correspond to warming temperature; (bottom left) timeseries of the first principal component; (bottom right) Multi-taper spectrum [*Thomson*, 1982] of the first principal component; note that the y-axis is the product of the power by the frequency so that the relative area under the spectrum is preserved in this logarithmic scale. Numbers refers to the period of oscillation in years

tion) and can be easily modified for different data selection criteria, which amount to a mere logical indexing of the Matlab structured array. The main reason for the conciseness of this code is the consistent format associated with each record, which enable very easy manipulations of the dataset. In the next section, we describe how simple extensions of this principle of data organization may enable much richer scientific questions to be asked of paleoclimate proxy networks.

3. The paleoclimate database of the future

Given the imperfections of each proxy type, the value of inferring climate information from multiple proxy classes is no longer in doubt [e.g. *Mann*, 2002; *Li et al.*, 2010; *Wahl et al.*, 2010], and indeed forms the basis of surface temperature reconstructions of the past millennium and beyond [*Jones et al.*, 2009; *NRC*, 2006], which have wide implications in the public sphere. The recent community focus embodied by the PAGES 2K project[3] illustrates the widespread recognition of the importance of pooling together all the proxy information available over the Common Era to enable new scientific discoveries to be made.

Three leading questions motivated the PAGES 2K data collation initiative:

1. What did the main patterns and modes of climate variability on sub-decadal to orbital timescales look and operate like?

2. How does climate variability and extreme events relate to the important primary forcing factors, namely orbital, solar and volcanic?

3. What feedbacks operated to modulate the climate response?

In an ideal world, the first two questions could be answered relatively easily by applying machine-learning and signal-processing algorithms to an internally-consistent database gathering all paleoclimate records contributed to date. Our world, as it stands, is far from ideal, but relatively simple steps, which we describe next, should improve upon this situation.

3.1. Attributes of an ideal database

An ideal data format would conform to the following attributes:

1. **Parsability**: the format should be self-contained (hence machine-readable), and would therefore enable a *semantic web* [*Berners-Lee et al.*, 2001] of paleoclimate information.

2. **Universality**: the format should be platform-independent (readable on all computer and operating systems), and language-independent (readable in major programming languages like Fortran, C, MATLAB, R, Python, IDL and NCL/PyNGL). It should also be accessible via an OPeNDAP protocol [4] to ensure that users rely on the most up-to-date information prior to conducting analysis. A natural way to achieve this is to code the database in a flexible, universal markup language like XML, which satisfies all three requirements.

3. **Extensibility**: the format should require a minimum set of fields to appropriately define a paleoclimate record, but allow for the database to grow organically as more records are added, or – equally important – as more metadata are added to existing records.

4. **Citability**: a legitimate concern amongst scientists who generate paleoclimate observations is that their work be given due recognition in subsequent analyses and syntheses. Currently, the NCDC relies on an honor system, but one could imagine easier and more stringent tools. This could include, as in the database described in Section 2, electronic references to publications (such a digital object identifiers, EndNote or BibTeX cite keys) embedded in the data format. This would allow the automatic citation of peer-reviewed articles as well as data citations [*Killeen, T.*, 2012] whenever a data record is being used for analysis. For instance, Table 1

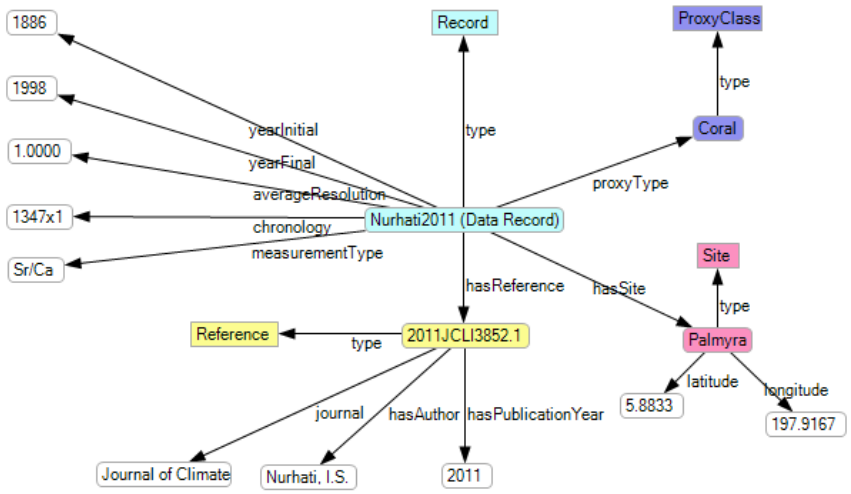

Figure 4. Preliminary ontology describing relationships between the data and meta-data fields of the *Nurhati et al.* [2011] climate record. Several fields are viewed as instances of larger classes (ProxyClass,Site,Reference), which would allow computers to perform operations on all records within a specific class (e.g. if the measurement type is δ^{18}O , or if the proxy class is 'Tree Ring Width', or if the resolution is less than 3 months, etc). All records in such a database would be bound to each other by similar links, allowing machines to automatically process any form of query involving existing information. Such a design would also allow growth, by adding records and/or additional information about each record.

was created automatically in Matlab (using LaTeX conventions) to make due recognition of scientific work an automatic part of paleoclimate data analysis. This principle should be foundational to a universal paleoclimate data format.

5. **Ergonomy**: The format should be easy to use, update and manage.

In practice, two elements are necessary to a semantic database. The first is an *ontology*, which describes the objects and the relationships between them. The ontology provides a standardized vocabulary within the domain. A minimal ontology can specify those elements of the data necessary to describe the experimental landscape; it further prevents semantic confusion on terminology (e.g. defining the characteristics a "chronology"). When extended, an ontology also allows for formal reasoning and inference about data based on a hierarchy of classes, known properties of these classes, and relationships between particular instances of these classes. As an illustration, a basic ontology describing relationships between the data and meta-data fields of the climate record described in Section 2.1 is given in Fig 4.

The second element needed is a way to store this information. A natural way to achieve this is to code metadata and data as a Resource Description Framework (RDF), a standard for modeling data interchange on the web, which itself can be expressed as an implementation of an eXtensible Markup Language (XML). A graph-data format for representing information, RDF is in fact the backbone of a semantic web (Fig 4). There are already libraries for creating, writing, storing, searching, and analyzing RDF databases available in open-source languages like R and Python [5]. It is readily serializable as XML text files. Open-source database backend "triple stores" also exist that make it possible to store RDF in a manner that facilitates distributed querying. When made available as SPARQL endpoints, the data within can be queried across the internet. While much of this is also true of storing data in a relational database, storing data in a triplestore does not require *a priori* schema, unlike relational databases. Both the ontology and data in RDF are expressed as a series of three component statements or "triples": a subject of the statement, a predicate expressing a property of the subject, and an object related to the subject via this predicate. For instance, in the expression "trees only grow on land", the subject of the statement is "trees", the object "land", and the predicate relating the two is "only grow on"). Objects of one statement can also be subjects of another statement, thus creating a limitless graph (or network) of information. To extend a dataset requires only adding additional triples. Querying a triplestore can be done without prior knowledge of the structure of the data. Perhaps most importantly, dramatically fewer limits, if any at all, are placed on the extensibility of a dataset when compared to a relational database where the initial choice of schema can limit the types of information that can be recorded and the types of queries that can be performed. Currently, strong initiatives are underway to share graph-model data in the domain of Health Care and Life Sciences (HCLS) data, including the Bio2RDF and Linked Open Drug Data (LODD) initiatives, but it is easy to envision a near-future where paleoclimate records will be parsable via knowledge-search engines like Wolfram Alpha[6].

3.2. Building a universal database

We now consider practical steps to achieving these objectives. The standard should be widely-adopted, which requires collective discussion and community consensus. This paper serves as a blueprint for the elaboration of a format of optimal use to the scientific community.

An important problem to address is to ensure that scientist who take the time to upload their data to online repositories a) conform to this new data standard ; and b) do not face significant additional time burden for doing so. This can be done most effectively by providing them with spreadsheet or web-form templates, which does not require any programming expertise and will automatically organize the data into a form that can be converted to the desired one. The complexity of the data network in the backend should be largely or wholly invisible to the scientist sharing data. Once uploaded, however, the full extent of this data network complexity, including any raw and processed data as well as metadata, should be easily ported directly into analysis tools in a manner similar to that employed by the BioConductor package [*Gentleman et al.*, 2004], where entire microarray datasets can be imported directly from ArrayExpress or GEO into R for analysis.

Another important consideration is scalability: it should be easy to nucleate the database using existing efforts (like that above) or a similar initiative in R (vpIR, T. Ault., *pers. comm.* 2012). It should also be easy to grow given a backbone. This leads to the definition of a *minimal information standard* (part of the ontology) necessary to define a paleoclimate record. At a minimum, it would seem that the record should include: a chronology (timeseries of dates, one date per sample), one measurement timeseries (one proxy observation per sample), latitude, longitude, site name, and bibliographic reference. Of course, considerably more information could be added, and indeed, should be. For instance, age model uncertainties in non-annual proxies are best assessed if the actual radiometric dates are available, not just the one age model selected for analysis and publication. One may therefore need to define a standard format for geoscientific chronologies as part of the ontology, incorporating raw measurements so that the chronology can be updated in the light of different calibrations (e.g. updates to the "radiocarbon curve") as was done (manually) in *Anchukaitis and Tierney* [2012]. This process is currently labor-intensive, and need not be. Additionally, one should be able to add as much metadata as is thought relevant. One could think of adding information about experimental protocol, proxy interpretation, uncertainties, instrument precision, sample number, reproducibility, as well as tracking various processing steps applied to the raw data (e.g. isotopic corrections, standardization, detrending) as illustrated in Section 2. This method of defining a minimal information standard has proved successful with microarray data, where MIAME (Minimal Information About Microarray Experiments, [*Brazma et al.*, 2001]) specifies those core elements for recording data and its provenance, experimental protocols necessary, and versioning.

4. Discussion

This article has proposed a new, self-contained way to describe and store paleoclimate data. It is intended as a blueprint to spark discussion among the community, and to hopefully lead to the implementation and deployment of one such technology in the near future.

The large-scale adoption of such a system would profoundly affect the way we do science. As demonstrated in Section 2.2, a rich data format means that more information can be extracted with less effort from the user. The XML-based format we propose here would enable proxy records to be viewed as instances of certain classes of objects, thus enabling object-oriented programming to be applied to paleoclimatology. A universal, open-source data format would quickly enable software libraries to be built around those objects, and would apply, by design, to all objects within a class. Making this code widely available (perhaps on the same Web servers as the data themselves) would enable paleoclimate scientists to easily apply sophisticated data analysis tools to their own data before sharing it, even without much programming knowledge.

In a nutshell, a semantic web for paleoclimatology would allow the power of automation to enter the study of past climates. This would first change the way paleoclimate records are compared to each other, a task done routinely in paleoclimate studies. Currently, each new record is typically stacked against existing ones and a comparison is made. The choice of records used in this comparison is partially dictated by scientific considerations (e.g. "proxies sensitive to the Pacific Decadal Oscillation"), and partially dictated by ease of access. A universal data format would ensure that all the information relevant to a comparison would be gathered using objective criteria, and using the most appropriate methods (e.g. taking time uncertainties directly into account). Further, a semantic web would radically change the way paleoclimate information is used to evaluate general circulation models (GCMs), an endeavor at the heart of initiatives like the Paleoclimate Modeling Intercomparison Project (PMIP). It would also transform the integration of paleoclimate records with instrumental variables (i.e. climate reconstruction), again allowing records to be included in a reconstruction only if they fit certain objective criteria (e.g. "give me all annually-resolved records located within 30 degrees of the equator"). Coupled to process-based models of proxy behavior, it would eventually enable the assimilation of paleoclimate data into climate model trajectories. All these tasks presently require sophisticated tools and a daunting amount of human resources, chiefly provided by volunteers. These tasks would be considerably eased by agreeing upon a common data standard, whose overhead cost would be negligible once the submission of proxy records to online databases is designed to structure the data to conform to these standards (via web forms or spreadsheet templates). Indeed, since a majority of paleoclimate scientist already send their data to the NCDC or PANGAEA, there would be little additional work involved for them once the submission process is standardized.

On one hand, this semantic web is expected to bring many benefits: increased relevance of paleoclimate data to future climate projections, increased transparency, increased accessibility, increased reproducibility, and increased educational opportunities. On the other hand, one may easily imagine a scenario whereby inadequate algorithms will be thoughtlessly applied to crunch through the database and come up with absurd answers to irrelevant questions. Obviously, automation will never be a substitute for knowledge, and expert judgment will always be critical to the interpretation of any result derived from paleoclimate records. The solution to such problems is more collaboration between all members of the paleoclimate community ; a common language to describe such data can only help this objective. Furthermore, malicious or ignorant use of climate data is already performed routinely by climate change deniers, and a paleoclimate semantic web would do little to change this.

The evolution we propose parallels that of other data-driven fields like bioinformatics, whose recent history has shown both the power and limits of automation. A semantic web of paleoclimatology would create much of the same opportunities, and likely much of the same pitfalls. Yet, in this field as in others, the advantages seem to far outweigh the disadvantages: a self-descriptive database of paleoclimate proxies means that paleoclimatologists will be able to put their data in a larger, more useful context, spend less time accessing data, and more time thinking about the scientific implications of their results. Who wouldn't want that?

Acknowledgments. JEG acknowledges Tiffany Tsai and Nasim Mirnateghi for help with assembling the coral database. This work was partially supported by grant NA10OAR4310115 from the National Oceanographic and Atmospheric Administration. Data and code necessary to generate the figures of this article is available on (URL TBD).

Notes

1. `http://www.ncdc.noaa.gov/paleo/data.html`
2. `http://www.pangaea.de/`
3. `http://www.pages-igbp.org/workinggroups/2k-network`
4. `http://opendap.org/`
5. `http://code.google.com/p/rdflib/`
6. `http://wolframalpha.com`

References

Abram, N. J., M. K. Gagan, J. E. Cole, W. S. Hantoro, and M. Mudelsee (2008), Recent intensification of tropical climate variability in the Indian Ocean, *Nature Geoscience*, *1*, 849–853, doi:10.1038/ngeo357.

Adler, P., R. Kolde, M. Kull, A. Tkachenko, H. Peterson, J. Reimand, and J. Vilo (2009), Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods, *Genome Biol.*, *10*(12), R139.

Anchukaitis, K., and J. Tierney (2012), Identifying coherent spatiotemporal modes in time-uncertain proxy paleoclimate records, *Climate Dynamics*, *in revision*.

Asami, R., T. Yamada, Y. Iryu, T. M. Quinn, C. P. Meyer, and G. Pauley (2005), Interannual and decadal variability of the western Pacific sea surface condition for the years 1787-2000: Reconstruction based on stable isotope record from a Guam coral, *Journal of Geophysical Research*, *110*, doi:10.1029/2004JC002555.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers (2011), GenBank, *Nucleic Acids Res.*, *39*(Database issue), D32–37.

Berners-Lee, T., J. Hendler, and O. Lassila (2001), The semantic web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, *Scientific American*, pp. 29–37.

Boiseau, M., A. Juillet-Leclerc, P. Yiou, B. Salvat, P. Isdale, and M. Guillaume (1998), Atmospheric and oceanic evidences of El Niño-Southern Oscillation events in the south central Pacific Ocean from coral stable isotopic records over the last 137 years, *Paleoceanography*, *13*, 671–685, doi:10.1029/98PA02502.

Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron (2001), Minimum information about a microarray experiment (MIAME)-toward standards for microarray data, *Nat. Genet.*, *29*(4), 365–371.

Charles, C. D., D. E. Hunter, and R. G. Fairbanks (1997), Interaction Between the ENSO and the Asian Monsoon in a Coral Record of Tropical Climate, *Science*, *277*(5328), 925–928, doi:10.1126/science.277.5328.925.

Charles, C. D., K. Cobb, M. D. Moore, and R. G. Fairbanks (2003), Monsoon-tropical ocean interaction in a network of coral records spanning the 20th century, *Marine Geology*, *201*, 207–222.

Chen, R., T. K. Sigdel, L. Li, N. Kambham, J. T. Dudley, S. C. Hsieh, R. B. Klassen, A. Chen, T. Caohuu, A. A. Morgan, H. A. Valantine, K. K. Khush, M. M. Sarwal, and A. J. Butte (2010), Differentially expressed RNA from public microarray data identifies serum protein biomarkers for cross-organ transplant rejection and other conditions, *PLoS Comput. Biol.*, *6*(9).

Cobb, K. M., C. D. Charles, and D. E. Hunter (2001), A central tropical Pacific coral demonstrates Pacific, Indian, and Atlantic decadal climate connections, *Geophys. Res. Lett.*, *28*, 2209–2212, doi:10.1029/2001GL012919.

Damassa, T. D., J. E. Cole, H. R. Barnett, T. R. Ault, and T. R. McClanahan (2006), Enhanced multidecadal climate variability in the seventeenth century from coral isotope records in the western Indian Ocean, *Paleoceanography*, *21*, PA2016, doi:10.1029/2005PA001217.

Edgar, R., M. Domrachev, and A. E. Lash (2002), Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Res.*, *30*(1), 207–210.

Evans, M. N., A. Kaplan, and M. A. Cane (2000), Intercomparison of coral oxygen isotope data and historical sea surface temperature (SST): Potential for coral-based SST field reconstructions, *Paleoceanogr.*, *15*, 551–562.

Evans, M. N., A. Kaplan, and M. A. Cane (2002), Pacific sea surface temperature field reconstruction from coral $\delta^{18}O$ data using reduced space objective analysis, *Paleoceanogr.*, *17*, 7–1.

Felis, T., J. Pätzold, Y. Loya, M. Fine, A. H. Nawar, and G. Wefer (2000), A coral oxygen isotope record from the northern Red Sea documenting NAO, ENSO, and North Pacific teleconnections on Middle East climate variability since the year 1750, *Paleoceanography*, *15*, 679–694, doi:10.1029/1999PA000477.

Gagan, M. K., L. K. Ayliffe, J. W. Beck, J. E. Cole, E. R. M. Druffel, R. B. Dunbar, and D. P. Schrag (2000), New views of tropical paleoclimates from corals, *Quaternary Science Reviews*, *19*(1-5), 45 – 64, doi:DOI:10.1016/S0277-3791(99)00054-2.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang (2004), Bioconductor: open software development for computational biology and bioinformatics, *Genome Biol.*, *5*(10), R80.

ITRDB (2009), National Climatic Data Center, World Data Center for Paleoclimatology, Tree Ring Database.

Jones, P., K. Briffa, T. Osborn, J. Lough, T. van Ommen, B. Vinther, J. Luterbacher, E. Wahl, F. Zwiers, M. Mann, G. Schmidt, C. Ammann, B. Buckley, K. Cobb, J. Esper, H. Goosse, N. Graham, E. Jansen, T. Kiefer, C. Kull, M. Kuttel, E. Mosley-Thompson, J. Overpeck, N. Riedwyl, M. Schulz, A. Tudhope, R. Villalba, H. Wanner, E. Wolff, and E. Xoplaki (2009), High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects, *The Holocene*, *19*(1), 3–49, doi:10.1177/0959683608098952.

Killeen, T. (2012), Data citation in the geosciences.

Kuhnert, H., J. Pätzold, B. Hatcher, K. . H. Wyrwoll, A. Eisenhauer, L. B. Collins, Z. . R. Zhu, and G. Wefer (1999), A 200-year coral stable oxygen isotope record from a high-latitude reef off Western Australia, *Coral Reefs*, *18*, 1–12, doi:10.1007/s003380050147.

Kuhnert, H., J. Pätzold, K.-H. Wyrwoll, and G. Wefer (2000), Monitoring climate variability over the past 116years in coral oxygen isotopes from Ningaloo Reef, Western Australia, *International Journal of Earth Sciences*, *88*, 725–732, doi:10.1007/s005310050300.

Li, B., D. W. Nychka, and C. M. Ammann (2010), The value of multi-proxy reconstruction of past climate, *J. Amer. Statist. Assoc.*, *105*, 883–911.

Linsley, B. K., R. B. Dunbar, G. M. Wellington, and D. A. Mucciarone (1994), A coral-based reconstruction of intertropical convergence zone variability over Central America since 1707, *J. Geophys. Res.*, *99*, 9977–9994, doi:10.1029/94JC00360.

Linsley, B. K., L. Ren, R. B. Dunbar, and S. S. Howe (2000), El Niño Southern Oscillation (ENSO) and decadal-scale climate variability at 10°Nin the eastern Pacific from 1893 to 1994: A coral-based reconstruction from Clipperton Atoll, *Paleoceanography*, *15*, 322–335, doi:10.1029/1999PA000428.

Linsley, B. K., A. Kaplan, Y. Gouriou, J. Salinger, P. B. deMenocal, G. M. Wellington, and S. S. Howe (2006), Tracking the extent of the South Pacific Convergence Zone since the early 1600s, *Geochemistry, Geophysics, Geosystems*, *7*, 5003–+, doi:10.1029/2005GC001115.

Linsley, B. K., P. Zhang, A. Kaplan, S. S. Howe, and G. M. Wellington (2008), Interdecadal-decadal climate variability from multicoral oxygen isotope records in the South Pacific Convergence Zone region since 1650 A.D., *Paleoceanography*, *23*(26), A262,219+, doi:10.1029/2007PA001539.

Lorenz, E. N. (1956), Empirical orthogonal functions and statistical weather prediction, *Scientific Report 1, Statistical Forecasting Project. 110268*, Massachusetts Institute of Technology Defense Doc. Center.

Mann, M. E. (2002), CLIMATE RECONSTRUCTION: The Value of Multiple Proxies, *Science*, *297*(5586), 1481–1482, doi:10.1126/science.1074318.

Murray-Rust, P., C. Neylon, R. Pollock, and J. Wilbanks (2010), Panton principles, principles for open data in science.

Neumann, E. K., and D. Quan (2006), BioDash: a semantic web dashboard for drug development, *Pacific Symposium on Biocomputing*, pp. 176–187, PMID: 17094238.

NRC (2006), *Surface Temperature Reconstructions for the Last 2000 Years*, The National Academies Press, Washington, D.C.

Nurhati, I. S., K. M. Cobb, and E. Di Lorenzo (2011), Decadal-Scale SST and Salinity Variations in the Central Tropical Pacific: Signatures of Natural and Anthropogenic Climate Change, *Journal of Climate*, *24*(13), 3294–3308, doi:10.1175/2011JCLI3852.1.

Parkinson, H., M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans, and A. Brazma (2007), ArrayExpress–a public database of microarray experiments and gene expression profiles, *Nucleic Acids Res.*, *35*(Database issue), D747–750.

Pfeiffer, M., O. Timm, W.-C. Dullo, and S. Podlech (2004), Oceanic forcing of interannual and multidecadal climate variability in the southwestern Indian Ocean: Evidence from a 160 year coral isotopic record (La Réunion, 55°E, 21°S), *Paleoceanography*, *19*, 4006–+, doi:10.1029/2003PA000964.

Pierre, M., B. DeHertogh, A. Gaigneaux, B. DeMeulder, F. Berger, E. Bareke, C. Michiels, and E. Depiereux (2010), Meta-analysis of archived DNA microarrays identifies genes regulated by hypoxia and involved in a metastatic phenotype in cancer cells, *BMC Cancer*, *10*, 176.

Quinn, T. M., T. J. Crowley, F. W. Taylor, C. Henin, P. Joannot, and Y. Join (1998), A multicentury stable isotope record from a New Caledonia coral: Interannual and decadal sea surface temperature variability in the southwest Pacific since 1657 A.D., *Paleoceanography*, *13*, 412–426, doi:10.1029/98PA00401.

Quinn, T. M., F. W. Taylor, and T. J. Crowley (2006), Coral-based climate variability in the Western Pacific Warm Pool since 1867, *Journal of Geophysical Research (Oceans)*, *111*, 11,006–+, doi:10.1029/2005JC003243.

Rayner, N., P. Brohan, D. Parker, C. Folland, J. Kennedy, M. Vanicek, T. Ansell, and S. Tett (2006), Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: the HadSST2 data set, *J. Clim.*, *19*(3), 446–469.

Sarachik, E. S., and M. A. Cane (2010), *The El Niño-Southern Oscillation Phenomenon*, Cambridge University Press.

Schorlemmer, D., A. Wyss, S. Maraini, S. Wiemer, and M. Baer (2004), QuakeML - An XML schema for seismology, *Tech. rep.*, Swiss Seismological Service.

Spellman, P. T., M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W. L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B. J. Aronow, A. Robinson, D. Bassett, C. J. Stoeckert, and A. Brazma (2002), Design and implementation of microarray gene expression markup language (MAGE-ML), *Genome Biol.*, *3*(9), RESEARCH0046.

Thomson, D. J. (1982), Spectrum estimation and harmonic analysis, *Proc. IEEE*, *70(9)*, 1055–1096.

Tudhope, A. W., C. P. Chilcott, M. T. McCulloch, E. R. Cook, J. Chappell, R. M. Ellam, D. W. Lea, J. M. Lough, and G. B. Shimmield (2001), Variability in the El Niño- Southern Oscillation Through a Glacial-Interglacial Cycle, *Science*, *291*, 1511–1517, doi:10.1126/science.1057969.

Urban, F. E., J. E. Cole, and J. T. Overpeck (2000), Influence of mean climate change on climate variability from a 155-year tropical Pacific coral record, *Nature*, *407*, 989–993.

Wahl, E. R., D. M. Anderson, B. A. Bauer, R. Buckner, E. P. Gille, W. S. Gross, M. Hartman, and A. Shah (2010), An archive of high-resolution temperature reconstructions over the past 2+ millennia, *Geochem. Geophys. Geosyst.*, *11*(1), doi:10.1029/2009GC002817.

Zinke, J., W.-C. Dullo, G. A. Heiss, and A. Eisenhauer (2004), ENSO and Indian Ocean subtropical dipole variability is recorded in a coral record off southwest Madagascar for the period 1659 to 1995, *Earth and Planetary Science Letters*, *228*, 177–194, doi:10.1016/j.epsl.2004.09.028.

Julien Emile-Geay, Department of Earth Sciences, University of Southern California, 3651 Trousdale Parkway, ZHS 275. Los Angeles, CA 90089 - 0740 e-mail: julieneg@usc.edu