doi.org/10.22498/pages.26.2.73

# Accelerating progress in proxy-model synthesis using open standards

Gregory Hakim[1], S. Dee[2], J. Emile-Geay[3], N. McKay[4] and K. Rehfeld[5]

Weather prediction has undergone a "quiet revolution" in recent decades (Bauer et al. 2015), fueled by increasing observations and the capability to assimilate them into increasingly sophisticated numerical models. Paleoclimatology today is at the cusp of such a revolution, moving away from a focus on single-site studies to the assimilation of global, multiproxy data streams into climate models (e.g. Brönnimann et al. 2013; Goosse 2016; Hakim et al. 2016; Franke et al. 2017) thanks to (1) advances in data assimilation (DA) methodology; (2) open, standardized paleoclimate datasets; and (3) proxy system models (PSMs). A critical element of DA that allows this synthesis involves mapping information from climate models to proxy measurements through PSMs (e.g. Dee et al. 2015). DA weighs the information from proxies against a climate-model simulation of the proxy value, and spreads that information in space and to other climate variables (Fig. 1). Future progress depends strongly on openness and standardization of paleoclimate proxy data, so we describe here the dependence of DA on open paleoclimate data, emerging standards, and ideas for accelerating progress.

## Openness in data sharing and standardization

The currently highly heterogeneous nature of the proxy records is the main limitation to DA progress. Improvements involve three components common in data science: (1) data distribution, (2) data standardization, and (3) data-revision tracking.

Over the past two decades, distribution of paleoclimate proxy data has migrated from individual scientists sharing their data to centralized data centers, such as the World Data Service for Paleoclimatology, the International Tree Ring Databank, Neotoma, and Pangaea; however, large amounts of data have not yet been transferred to public repositories. Curated versions of paleoclimate data from these centers and from the literature, through quality control and screening, have proven critical to recent synthesis efforts (e.g. PAGES 2k Network projects). However, because these curated versions do not track uniquely from the original proxy data, future efforts either have to work with these "forks" from the source, or substantially duplicate effort by returning to the original data. Having the ability to track data from the source through the forks would allow for robust branching without returning to sources to begin anew.

Climate model output is available in standard format (NetCDF), with conventions for units and variable naming (cfconventions.org). Ongoing efforts combine PSMs in a standardized and open-source framework (e.g. PRYSM; Dee et al. 2015), but such standardization is just beginning for paleoclimate data. For example, the Linked Paleo Data (LiPD) format (McKay and Emile-Geay 2016 and this issue) provides a universal, flexible container for a wide range of paleoclimate data. Because LiPD's structure and terminology are inspired by the PSM framework, it is a natural format for DA codes, since LiPD metadata can direct PSM selection for a particular dataset. Although the emergence of LiPD offers the potential for a large increase in efficiently using proxy data in DA applications, most proxy data remain to be converted to LiPD format.

## Future directions

Data standardization is the area where the greatest immediate impact can be experienced. Widespread adoption of LiPD across proxy archives would greatly facilitate the reuse of proxy data and synthesis efforts, as would standardized revision histories. As much as revision tracking has transformed productivity in software development with distributed version control software such as Git, similar practices for proxy data are compelling.

One speculative future direction involves decentralized ledgers for proxy data. Cryptographically secure ledgers, such as Bitcoin's blockchain, contain unalterable revision history that do not depend upon a central authority. For paleoclimate proxy data, this technology could be used to allow anyone to correct errors and, through a consensus algorithm, add revisions to the public ledger. One can imagine motivating public participation with micropayments of Bitcoin. A small amount of funding distributed in this way could offer rapid progress to cleaning the "bugs" from proxy data archives, with the added benefit of citizen scientist participation in paleoclimate research.

AFFILIATIONS

[1]Department of Atmospheric Sciences, University of Washington, Seattle, USA
[2]University of Texas, Institute for Geophysics, Austin, USA
[3]University of Southern California, Los Angeles, USA
[4]Northern Arizona University, Flagstaff, USA
[5]Institute of Environmental Physics, Ruprecht-Karls-Universitaet Heidelberg, Germany

CONTACT

Gregory Hakim: ghakim@uw.edu

REFERENCES

Bauer P et al. (2015) Nature 525: 47-55

Brönniman S et al. (2013) PAGES news 21: 74-75

Dee S et al. (2015) J Adv Model Earth Sys 7: 1220-1247

Franke J et al. (2017) Sci data 4: 170076

Goosse H (2016) J Adv Model Earth Sys 8: 1501-1503

Hakim GJ et al. (2016) J Geophys Res Atmos 121: 6745-6764

McKay NP, Emile-Geay J (2016) Clim Past 12: 1093-1100

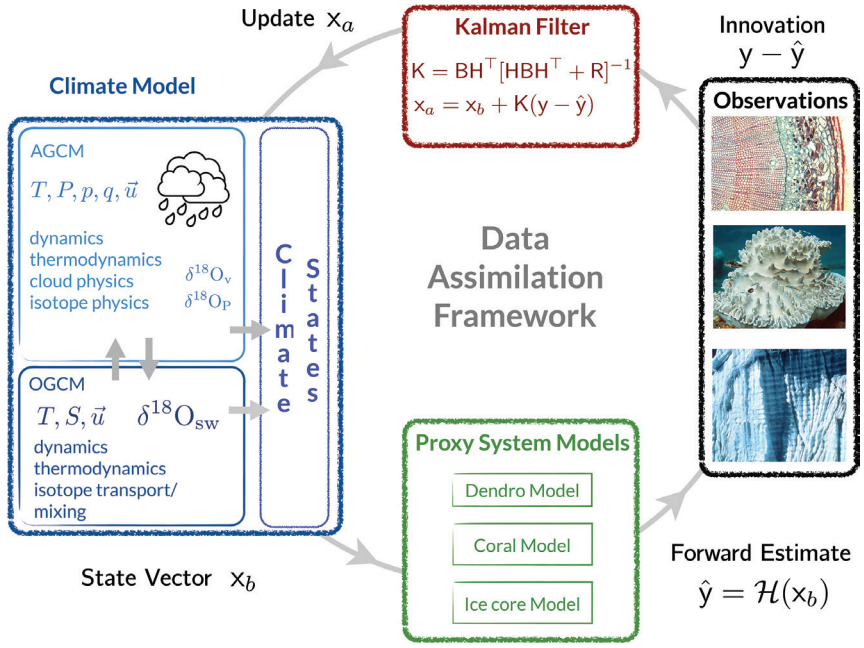


**Figure 1:** DA uses climate variables to estimate proxy values using PSMs, which can then be compared with the actual proxy values. The difference between these values (new information about the climate state) is weighted ("K") by the error in the proxies relative to the estimate from the climate model; most important, DA also spreads this information in space and to other climate variables. From Hakim et al. (2016).