AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1002/2016JD024751

The last millennium climate reanalysis project: Framework and first results

Gregory J. Hakim[1], Julien Emile-Geay[2], Eric J. Steig[1,3], David Noone[4], David M. Anderson[5], Robert Tardif[1], Nathan Steiger[1], and Walter A. Perkins[1]

[1]Department of Atmospheric Sciences, University of Washington, Seattle, Washington, USA, [2]Department of Earth Sciences, University of Southern California, Los Angeles, California, USA, [3]Department of Earth and Space Sciences, University of Washington, Seattle, Washington, USA, [4]College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, Oregon, USA, [5]Monterey Bay Aquarium Research Institute, Monterey, California, (USA)

Key Points:
- Data assimilation climate field reconstruction skillful against out-of-sample instrumental data and proxies
- Reconstruction skill is highest in the tropics and lowest over Northern Hemisphere land areas
- Multivariate reconstruction of 1808/1809 volcanic cooling associated with PNA pattern in 500 hPa geopotential height

Supporting Information:
- Supporting Information S1

Correspondence to:
G. J. Hakim,
ghakim@uw.edu

Citation:
Hakim, G. J., J. Emile-Geay, E. J. Steig, D. Noone, D. M. Anderson, R. Tardif, N. Steiger, and W. A. Perkins (2016), The last millennium climate reanalysis project: Framework and first results, *J. Geophys. Res. Atmos.*, *121*, 6745–6764, doi:10.1002/2016JD024751.

Received 15 JAN 2016
Accepted 16 MAY 2016
Accepted article online 24 MAY 2016
Published online 22 JUN 2016

©2016. The Authors.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Abstract An "offline" approach to DA is used, where static ensemble samples are drawn from existing CMIP climate-model simulations to serve as the prior estimate of climate variables. We use linear, univariate forward models ("proxy system models (PSMs)") that map climate variables to proxy measurements by fitting proxy data to 2 m air temperature from gridded instrumental temperature data; the linear PSMs are then used to predict proxy values from the prior estimate. Results for the LMR are compared against six gridded instrumental temperature data sets and 25% of the proxy records are withheld from assimilation for independent verification. Results show broad agreement with previous reconstructions of Northern Hemisphere mean 2 m air temperature, with millennial-scale cooling, a multicentennial warm period around 1000 C.E., and a cold period coincident with the Little Ice Age (circa 1450–1800 C.E.). Verification against gridded instrumental data sets during 1880–2000 C.E. reveals greatest skill in the tropics and lowest skill over Northern Hemisphere land areas. Verification against independent proxy records indicates substantial improvement relative to the model (prior) data without proxy assimilation. As an illustrative example, we present multivariate reconstructed fields for a singular event, the 1808/1809 "mystery" volcanic eruption, which reveal global cooling that is strongly enhanced locally due to the presence of the Pacific-North America wave pattern in the 500 hPa geopotential height field.

1. Introduction

The paleoclimate record provides abundant evidence for significant decadal to centennial variability in a host of climate variables [e.g., *Mann et al.*, 1998; *Mayewski et al.*, 2004; *Jones et al.*, 2009; *Masson-Delmotte et al.*, 2013], but the instrumental record is too short to adequately characterize this variability. Yet most applications—notably, the evaluation of decadal-to-century scale variability simulated by general circulation models (GCMs)—are facilitated by spatially gridded estimates of climate fields such as surface air temperature, sea level pressure, geopotential height, and precipitation. Consequently, there is a pressing need for methods that can accurately estimate such fields from the sparse, noisy, and indirect observations of past climates obtained from climate proxy records.

Since the high-profile work of *Mann et al.* [1998, 1999], many investigators have applied a range of methods to infer climate fields from proxy observations, a process known as climate field reconstruction (CFR) [*Tingley et al.*, 2012, and references therein]. These methods fall into three main categories:

Regression Models. The most traditional approach employs multivariate regression of climate variables onto the proxy records, in which the relationship between proxies and fields of interest are obtained from a well-observed "training" period and are assumed to be stationary [e.g., *Mann et al.*, 1998, 1999; *Evans et al.*, 2002; *Luterbacher et al.*, 2004; *Cook et al.*, 2004; *Rutherford et al.*, 2005; *Mann et al.*, 2009; *Cook et al.*, 2010; *Barriopedro et al.*, 2011]. A principal challenge is that the grid size often greatly exceeds the number of samples available for calibration. This underdetermined setting introduces a major challenge: the sample covariance matrix used to quantify the relationship between proxies and climate fields is poorly constrained and needs to be regularized to yield meaningful estimates of past climates. A popular regularization approach consists of truncating the canonical expansion of the covariance matrix using singular value decomposition. Other approaches include the expectation-maximization algorithm [*Dempster et al.*, 1977; *Schneider*, 2001;

Little and Rubin, 2002] regularized by *ridge regression* [*Hoerl and Kennard*, 1970a, 1970b; *Tikhonov and Arsenin*, 1977], and *truncated total least squares* (TTLS) [*Van Huffel and Vandewalle*, 1991; *Fierro et al.*, 1997]. Other approaches employ canonical correlation analysis [*Smerdon et al.*, 2010] and Markov random fields [*Guillot et al.*, 2015]. Since regularization is not unique, these various approaches yield significantly different estimates of past climate fields [*Smerdon et al.*, 2011; *Wang et al.*, 2014, 2015; *Smerdon et al.*, 2015]. Another difficulty with these regression models derives from the fact that they express climate fields as a function of proxy values, whereas reality works in the opposite direction (e.g., tree growth responds to climate variability; climate does not respond to tree growth). *von Storch et al.* [2004] point out that this functional assignment leads to reconstructions that overly damp the amplitude of past variability. Inverse regression [e.g., *Christiansen*, 2010] proposes to alleviate this issue but suffers from the opposite problem: potentially unbounded estimates of past variability [*Tingley and Li*, 2012].

Bayesian Hierarchical Models. Another major challenge to climate field reconstruction concerns the relationship between proxies and climate variables, which can be nonlinear, thresholded, and multivariate [*Evans et al.*, 2013, 2014]. As a consequence, it can be useful to include knowledge on these aspects into an inference process at the data level of a Bayesian hierarchical model. Bayes's theorem offers a natural way to invert the proxy-climate relation, yielding a fully probabilistic analysis of the predicted climate field that incorporates all known sources of error [*Tingley and Huybers*, 2010a, 2010b]. The resulting algorithm, Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time (BARCAST), has been used to reconstruct hemispheric-scale surface temperature over the past 600 years [*Tingley and Huybers*, 2013]. Regularization of the covariance estimation problem comes from imposing analytical covariance functions as priors, whose parameters are estimated from the data. There are, however, limitations to this approach:

1. It includes inference about only one climate field (temperature, in this case); generalizations to multiple dependent fields are difficult.
2. The inference is computationally demanding, even for the simplest hierarchical models.
3. While the BARCAST framework allows many possible extensions, it has so far been applied only to covariance functions that vary with distance, which cannot easily represent the highly anisotropic character of many atmospheric and oceanic features (e.g., mountain ranges, coastlines, jets, and currents).

Data Assimilation. Paleoclimate data assimilation (PDA) [see, e.g., *Dirren and Hakim*, 2005; *Annan et al.*, 2005; *Goosse et al.*, 2006; *Ridgwell et al.*, 2007; *Widmann et al.*, 2010; *Bhend et al.*, 2012; *Steiger et al.*, 2014], provides a dynamically motivated alternative to the CFR problem. PDA uses model-simulated climate states to measure the novel information in proxy data and to spatially spread that information through all climate variables subject to the dynamical constraints of the climate model. This has several advantages:

1. PDA can infer multiple climate fields simultaneously. While specifying a statistical model that links several fields together is possible in principle, the physical basis of GCMs provides a more natural framework to represent such relationships.
2. In general, PDA does not assume stationary teleconnections, unlike most purely statistical approaches.
3. PDA uses dynamical models to infer spatial relationships within and between climate fields. The fields so reconstructed are, ipso facto, dynamically consistent.
4. PDA allows flexibility in accounting for the dependence of proxies on multiple state parameters (e.g., temperature and moisture controls on tree-ring width, a common climate proxy) and therefore can better constrain multivariate states.
5. PDA admits proxies having different timescales (e.g., annual and decadal resolution) to participate in a reconstruction, without interpolation or smoothing [*Steiger and Hakim*, 2015].

PDA is conveniently posed within a Bayesian framework [*Wikle and Berliner*, 2007] and offers the opportunity to inject more information at the process level (via GCMs) and at the data level (via proxy system models PSMs, [*Evans et al.*, 2013; *Dee et al.*, 2015]) than other approaches. PDA may be understood as a limiting case of a Bayesian hierarchical model like BARCAST, using climate models as a covariance estimation device. A caveat is that the resulting prior inherits some potentially large biases from climate models [*Flato et al.*, 2013].

This paper describes the PDA-based framework for the Last Millennium climate Reanalysis (LMR) project. In the lineage of DA-based reconstructions of meteorological and oceanographic fields [see, e.g., *Kalnay et al.*, 1996; *Compo et al.*, 2011; *Dee et al.*, 2014; *Carton and Giese*, 2008; *Köhl*, 2015], the LMR aims to extend the

realm of regularly gridded climate fields beyond the instrumental record by making use of the information contained in paleoclimate proxies of the past 1000 years or more.

The two main objectives of this paper are the following: first, to document the essential elements of the PDA methodology for paleoclimate reconstruction, which will be used (and improved) in future applications to the archives of existing paleoclimate data, and second, to present results that illustrate and validate the method. With regard to the second objective, the paper focuses on a detailed comparison to existing reconstructions, reanalyses, and instrumental temperature data sets. In this initial effort, we consider a relatively limited subset of proxies and climate models and a simplified representation of proxy—climate–variable relationships. Despite these simplifications, we show that this initial PDA CFR performs on par with reanalysis products based on instrumental observations. Furthermore, the results identify a climate anomaly associated with a previously unconfirmed volcanic event in 1808/1809, which is examined to succinctly illustrate the advantages of a multivariate dynamically constrained reconstruction.

The remainder of the paper is organized as follows. In section 2 we present the PDA methodology as implemented by the LMR. Section 3 presents preliminary results on two select climate fields, 2 m air temperature, and 500 hPa geopotential height. The sensitivity of the results to various aspects of the method is investigated in section 4, followed by a case study documenting the dynamical context of the "mystery" volcanic eruption of 1808/1809 [*Guevara-Murua et al.*, 2014], which illustrates the advantages of the multivariate aspect of the PDA approach. We conclude with a discussion of the merits and limitations of the assimilation system and provide perspective on opportunities to improve estimates of historical climate using a synthesis of advanced models and comprehensive collections of high-quality proxy data.

2. Method

Our experimental design consists of paleoclimate data assimilation applied to the PAGES2K data set described in *PAGES2K Consortium* [2013]. There are five aspects discussed subsequently: the data assimilation method, the proxy data, the prior data, modeling the proxy data from the prior data, and error analysis. The PDA framework is graphically summarized in Figure 1.

2.1. Paleoclimate Data Assimilation

In general, data assimilation starts with some prior estimate of the field(s) to be analyzed. This may be an "uninformed" prior, such as a random sample from a model climatology, or a highly informed prior, such as a short-term forecast from a very accurate weather analysis. In either case, the prior provides an independent estimate of the true value, and the difference between these two estimates of the true value (from the proxy and the prior) is used to update the prior. In order to estimate the observation, a forward model is used to map from the climate variables to the measurement (e.g., linearly interpolating a regularly gridded temperature field to a measurement taken at a specific location).

The method we employ here is similar to that used in pseudoproxy experiments by *Steiger et al.* [2014] but for real proxies and a number of different calibration and prior data sets. Specifically, we use an "offline" (no cycling) [*Oke et al.*, 2002; *Evensen*, 2003; *Bhend et al.*, 2012; *Steiger et al.*, 2014; *Matsikaris et al.*, 2015] data assimilation approach and a fixed prior to solve the update equation of the Kalman filter [see, e.g., *Kalnay*, 2003],

$$\mathbf{x}^a = \mathbf{x}^p + \mathbf{K}[\mathbf{y} - \mathcal{H}(\mathbf{x}^p)]. \tag{1}$$

Here $\mathbf{x}^p$ and $\mathbf{x}^a$ are the prior and analysis state vectors, respectively, which contain the climate variables of interest averaged to a particular time period [*Huntley and Hakim*, 2010], including scalars and grid point data for spatial fields. Vector $\mathbf{y}$ contains the proxy data, and $\mathcal{H}(\mathbf{x}^p)$ is a vector estimate of the proxies based on the proxy system model $\mathcal{H}$. The innovation, $\mathbf{y} - \mathcal{H}(\mathbf{x}^p)$, represents the new information from the proxies not already contained in the prior. This new information is weighted against the prior by the Kalman gain matrix

$$\mathbf{K} = \mathbf{B}\mathbf{H}^{\mathrm{T}}\left[\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right]^{-1} \tag{2}$$

where $\mathbf{B}$ is the covariance matrix for the prior data, $\mathbf{R}$ is the error covariance matrix for the proxy data, and $\mathbf{H}$ is the linearization of $\mathcal{H}$ about the mean value of the prior. We employ an ensemble square root approach [*Whitaker and Hamill*, 2002], which uses a sample estimate for the prior covariance and solves for the ensemble mean and perturbations about the ensemble mean separately. This approach treats $\mathbf{R}$ as a

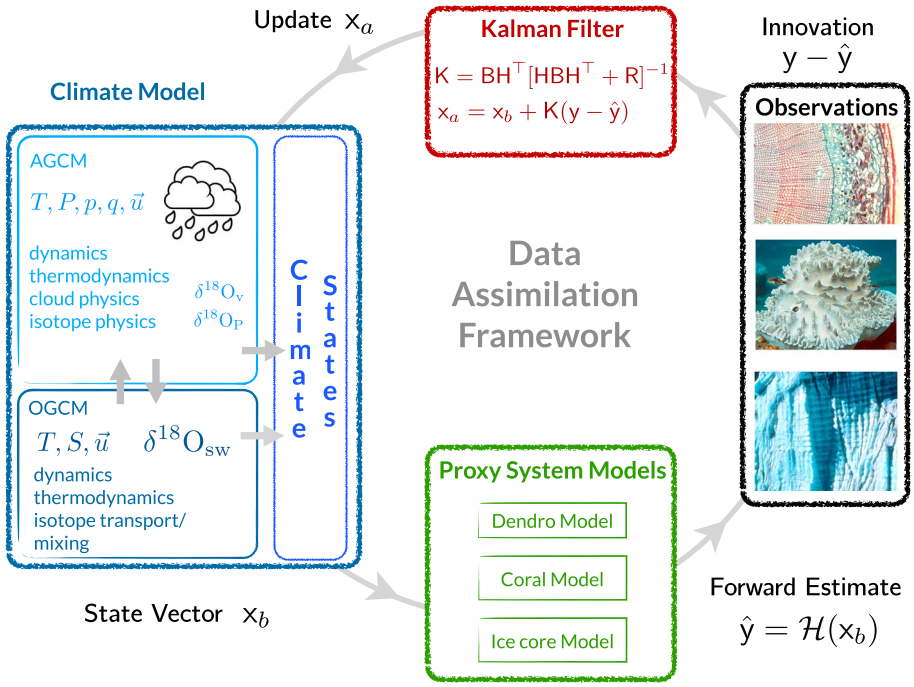

Figure 1. Conceptual framework for the Last Millennium Reanalysis, outlining our paleoassimilation approach. Starting from the prior (a collection of simulated climate states) from which random draws are pulled, the states are mapped to proxy space via a proxy system model (PSM). These predictions $\hat{y}$ are compared to the actual proxy measurements y to compute the innovation, $\hat{y} - y$. These innovations are then used to update the prior via the Kalman filter equations, which also update the error covariance. The cycle is repeated many (10^4) times to sample the distribution of the prior ensemble.

diagonal matrix, and the diagonal values represent the error variance for each proxy (defined in section 2.4). A 100-member ensemble is used, and results are insensitive to this choice provided that the ensemble has at least 50 members.

As an explicit illustration of how multivariate fields are recovered, consider a single tree ring width proxy measurement for 1 year and assume a PSM that depends only on 2 m air temperature. We take the prior to consist of 2 m air temperature at the location of the proxy and a globally gridded 500 hPa geopotential height field. The prior fields consist of an ensemble of annual-mean values, randomly drawn from a long climate simulation. From this, we should expect that the prior estimate of the proxy, $\mathcal{H}(\mathbf{x}^p)$, will differ considerably from the proxy value; i.e., the innovation will be large. How much weight the innovation gets in equation (1) depends on **R** (see equation (2)), which in this example is simply a scalar variance, r. Given the innovation, which in this case is a scalar value, **K** determines the weight and transfers information from the innovation (in units of tree ring width) to the 500 hPa geopotential height field. Consider the 500 hPa geopotential height at a single point in the global grid and call it x. From (2),

$$\mathbf{BH}^{\mathsf{T}} \sim \frac{1}{n-1}\mathbf{x}(\mathbf{Hx}^p)^{\mathsf{T}} \tag{3}$$

where n is the ensemble size and **x** is a row vector containing the ensemble of n values of 500 hPa geopotential heights (x) at the point, with the ensemble mean removed. The right side of equation (3) represents the covariance between the 500 hPa height at the point and the prior estimate of the proxy. In the denominator of **K**, $\mathbf{HBH}^{\mathsf{T}}$ is a scalar, representing the variance of the ensemble estimate of the proxy, $\mathrm{var}(y^e)$ (directly comparable to r). Therefore, we update the ensemble mean 500 hPa geopotential height at the point by

$$x^a = x^p + \frac{\mathrm{cov}(x^p, y^e)}{\mathrm{var}(y^e) + r}(y - y^e) \tag{4}$$

where $y^e = \mathbf{Hx}^p$ is the prior estimated proxy; i.e., this equation says that the analysis 500 hPa geopotential height at the point is determined by linearly regressing the prior estimate against the innovation.

Table 1. Gridded Data Sets Used to Form a Prior in the Data Assimilation

Data Set	Acronym	References
Community Climate System Model version 4, CMIP5 last millennium simulation	CCSM4	*Taylor et al.* [2012] and *Landrum et al.* [2013]
Max-Planck-Institute Earth System Model paleomode (MPI-ESM-P), CMIP5 last millennium simulation	MPI	*Taylor et al.* [2012]
NOAA-CIRES twentieth century reanalysis, version 2c	20CR-V2	*Compo et al.* [2011]
ECMWF reanalysis of the twentieth century	ERA-20C	*Dee et al.* [2014]

2.2. Paleoclimate Proxy Data

For proxy data we use the multiproxy database developed by the *PAGES2K Consortium* [2013]. Only time series labeled annually resolved have been considered, representing a maximum of 465 time series out of the original 508 in the database. Advantages of this network over previous compilations are its emphasis on global coverage, long time series, and multiple proxies (in the subset used here: tree ring width, mixed latewood density, ice core oxygen isotope ratio, ice core hydrogen isotope ratio, coral oxygen isotopes, coral luminescence, lake sediment varve thickness, marine sediment magnesium to calcium ratio, and speleothem laminae thickness). The PAGES network selection benefits from expert knowledge, where regional groups "identified the proxy climate records that they found were best suited for reconstructing annual or warm-season temperature variability within their region, using a priori established criteria" [*PAGES2K Consortium* 2013]. Potential shortcomings of the database include the still-incomplete geographic coverage (Africa is not represented, and the oceans are poorly sampled). Moreover, some of the proxies have multiple influences in addition to temperature, and the proxies differ in their ability to capture low-frequency variability. The PAGES database consists of time series in their native units (e.g., ring width and isotopic ratio), and to provide the most direct comparison to the PAGES results, we use exactly the same time series to develop the univariate linear proxy system models (see section 2.4). A revised PAGES2k database, with many improvements including expanded ocean coverage, is under development http://www.pages-igbp.org/ini/wg/2k-network/data and will be used in future iterations of the LMR project.

2.3. Prior Data

The data forming the prior state vector are taken from various sources, which serves as a basis for sensitivity experiments described in section 4. These sources, summarized in Table 1, range from existing long preindustrial simulations from the Coupled Model Intercomparison Phase 5 (CMIP5) project [*Taylor et al.*, 2012], to Twentieth Century reanalysis data sets [*Compo et al.*, 2011; *Dee et al.*, 2014] . The 100-member ensemble used in the data assimilation solver is drawn randomly from a chosen prior source, and the same sample is used for every year. Recall that this sampling strategy reflects an ensemble forecast having no skill over the model climatology and thus represents a random sample from the model climate. For our control reconstruction, we use the CCSM4 last millennium (CCSM4-LM) simulation, which is a coupled atmosphere-ocean-sea ice simulation covering the 850–1850 C.E. time period. It includes variable long-lived greenhouse gases and stratospheric aerosols reflecting known volcanic eruptions. Incoming solar variability is also included as well as prescribed anthropogenic changes to global land cover. For more details, see *Landrum et al.* [2013]. Monthly data are averaged to a calendar year, and the temporal mean over the entire data set is removed. The resulting annual-mean anomalies are then spatially truncated in spherical harmonic space to T42. For the last millennium simulations, this leaves 1000 truncated spatial fields from which we draw the 100-member ensemble random sample.

Using the offline PDA approach provides several advantages over an online approach. First, the computational cost of a 100-member ensemble of millennial-scale simulations is barely feasible at the current time, which means that, at best, only a single reconstruction could be performed. Since current models have little predictive skill on the annual timescale of the proxies considered here, there is little benefit to making such forecasts; the prior ensembles in this case are well approximated by the "climate" prior we use by randomly sampling a long simulation of the same model. An additional computational advantage derives from the fact that since each year is reconstructed independently, the solver can be efficiently parallelized. A second advantage derives from the fact that ensembles from different models can be used as a way to assess the sensitivity of solutions to the source of the prior. Multimodel ensembles, which consist of samples drawn from multiple models, are not considered here. Since we use the same prior ensemble for every year, a third advantage is that the prior anomalies have exactly zero temporal anomaly; all trends and temporal structure in the

Table 2. Gridded Temperature Anomaly Data Sets Used for the Calibration of Proxy System Models and for the Verification of Reconstructions

Data Set	Acronym	Reference
NASA Goddard Institute for Space Studies Surface Temperature Analysis	GISTEMP	*Hansen et al.* [2010]
Met Office Hadley Centre Climatic Research Unit Temperature, version 4.4.0.0	HadCRUT4	*Morice et al.* [2012]
Berkeley Earth Surface Temperatures	BE	*Rohde et al.* [2013]
NOAA Merged Land-Ocean Surface Temperature Analysis, version 3.5.4	MLOST	*Smith et al.* [2008]

reconstructions derive from the proxies. Finally, we note that since a forecast model does not need to be initialized, we may be selective about the climate variables we reconstruct (e.g., 500 hPa geopotential height), which is a subspace of that required for model initialization.

2.4. Proxy System Models

An essential feature of assimilation is the direct comparison between the measurement (proxy value) and an estimate of the measurement based on the prior. For PDA, this comparison requires a mapping from the climate state variables (e.g., temperature and precipitation) to the quantity that forms the proxy record (e.g., tree ring width). This mapping is achieved by forward modeling the proxy, which is the reverse of the commonly used inverse approach where climate variables are estimated from the proxies. This proxy system model (PSM) can range in sophistication from a simple statistical relationship to more complex mechanistically based models. Indeed, PSMs [*Evans et al.*, 2013; *Dee et al.*, 2015] are analogous to the "forward models" used in the data assimilation and estimation literature. In the interest of establishing a baseline measure of performance, in this paper we use a simple linear approach, where each proxy is linearly regressed against a temperature calibration data set during the instrumental period. Such linear, univariate PSMs take the form:

$$y = \beta_0 + \beta_1 T' + \epsilon \tag{5}$$

where T' is the annual-mean 2 m air temperature anomaly from a calibration data set, β_0, β_1 are the intercept and slope, and ϵ is a Gaussian random variable with zero mean and variance σ^2. Calibration temperatures concurrent with available proxy data are taken at the grid point nearest the proxy location and the ordinary linear least squares solution determines parameters β_0, β_1, and σ. Table 2 summarizes the various gridded temperature data sets used for calibration in this study.

A critically important aspect of this approach is that the variance of the regression residuals, σ^2, is used to define the diagonal elements of matrix **R** in equation (2). This ensures that the proper weight is applied to the innovation through the Kalman gain. Proxies for which the fit equation (5) is poor have large residual variance and get relatively less weight. As a result, this approach allows for the use of all available proxy data, without having to prefilter the proxies to select only those having a significant correlation with 2 m air temperature. In fact, we find similar results in both cases (not shown), but in the interest of simplicity and transparency, we include all proxies to the PDA solver.

It would, of course, be better to fit the proxy data to the climate field to which it is sensitive (e.g., soil moisture for many tree ring width chronologies), but calibration data for these other fields are not as readily available as for temperature. We leave this refinement for future work, along with tests involving more-sophisticated PSMs.

We take the opportunity here to emphasize that the role of the PSM in the PDA algorithm is to predict the proxy value from prior data (i.e., the randomly drawn states from the climate simulation). It is essential to appreciate that the data on which the PSM is calibrated are independent of the prior data. Specifically, the PSM is fit on data from the instrumental period for which there is a trend, but the resulting linear model is used to forecast proxy data with a predictor that comes from millennial-scale preindustrial simulations that have little or no trend. We will return to this point in section 4.

2.5. Error Analysis

2.5.1. Proxy Sampling

Ensemble PDA provides a natural source for uncertainty quantification in terms of the perturbations about the ensemble mean. Nevertheless, we take additional steps in both the construction of the experiments and the verification of the results, to account for uncertainty. In the solution method, we randomly sample from the proxies 100 times to yield a sample of 100 reconstructions; we shall refer to these as Monte Carlo realizations. Specifically, we randomly sample 75% of the proxies to assimilate for the entire 2000 years for a

given prior. This has two advantages. First, it provides a measure of uncertainty due to the proxies that participate in the reconstruction. This is rarely, if ever, done in reanalysis projects and is made possible here by the offline PDA approach. Second, the remaining 25% of the proxies provide a source for independent verification, which is particularly useful during the preinstrumental time period. We note that because we have 100 different Monte Carlo realizations, each proxy chronology may participate as both an assimilated and independent proxy. Details of this error analysis are provided in section 2.5.3.

2.5.2. Measures of Skill

We verify results for 2 m air temperature and 500 hPa geopotential height against both the calibration and reanalysis data sets summarized in sections 2.3 and 2.4 using two primary verification metrics: correlation and coefficient of efficiency. Given a time series of n values of a climate variable for reconstruction values x, and verification values, v, the correlation is defined by

$$\mathrm{corr} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})(v_i-\bar{v})}{\sigma_x\sigma_v} \tag{6}$$

and the coefficient of efficiency (CE) by [*Nash and Sutcliffe*, 1970]

$$\mathrm{CE} = 1 - \frac{\sum_{i=1}^{n}(v_i-x_i)^2}{\sum_{i=1}^{n}(v_i-\bar{v})^2} \tag{7}$$

Here an overbar represents a mean value and σ the standard deviation. We note that the correlation measures signal timing and is not affected by errors in signal amplitude or bias. CE is affected by these factors, and as such, it is a useful measure for these aspects of the reconstruction. CE sensitivity to bias depends on the definition of a reference time period to define anomalies, which may be nonoverlapping between proxies, calibration data, and prior data.

A known issue with CE is that if the means over the verification period for the proxies and the verification data are different, then *CE* can be negative even when the reconstruction is skillful [e.g., *Cook et al.*, 1999]. For example, returning to equation (7) in the case $x_i = a$, a constant, we find

$$\mathrm{CE}(x_i = a) = -\frac{(\bar{v}-a)^2}{\sigma_v^2} \tag{8}$$

which is negative semidefinite. For $a = 0$, as in the case for our millennial-scale simulation priors with respect to their time mean, $\mathrm{CE} < 0$ when verified against proxies, which in general have a nonzero time mean relative to the same baseline. As a result, for the reconstructed fields, the basis for evaluation should be the reference CE values for the prior; negative values may still reflect improvement upon the prior. When discussing the results, we therefore include both the CE measure and a ΔCE, which is the difference between CE and its value for the prior.

We note that having ensembles admits a wide range of probabilistic verification measures, but with a large range of topics to consider, we opt to leave this to a future report with one exception. Ensemble calibration is an important consideration, as it provides a measure of the performance of the PDA technique. Ensemble calibration is defined in observation space by the ratio of the mean square error of the ensemble mean to the average ensemble variance [e.g., *Houtekamer et al.*, 2005]. In the limit of a large ensemble, this ratio should be in unity, which indicates that the ensemble is drawn from true distribution. Typically, this ratio is larger than one because the ensemble has too little spread for the given ensemble mean error [*Hamill*, 2001]. We find that the LMR ensembles are well calibrated in both the prior and the reconstructed values. Verified against proxies during 1880–2000, the distribution of ensemble calibration ratios has a median value of 1.03 for both the prior and reconstructed fields; for 0–1879 C.E., the ratio for both is 1.17.

2.5.3. Independent Proxy Data

As mentioned above, 25% of the proxies in each experiment are withheld from data assimilation. After the reconstruction is complete, the linear PSM equation (5) is used on the reconstructed temperature field to predict the value of the independent proxies. The predicted values can then be verified against the proxy chronologies themselves using the correlation and CE performance measures.

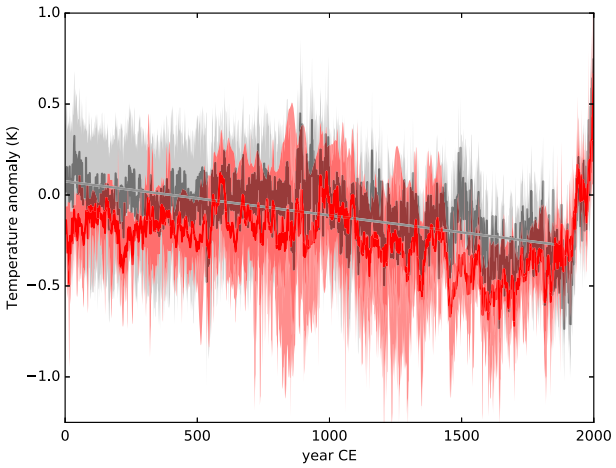

Figure 2. Comparison between LMR Northern Hemisphere 2 m air temperature (black line) and an average of 17 reconstructions (red line) summarized in IPCC AR5 (WG1, Figure 5.7). The LMR mean is taken over 100 realizations of 100-member ensembles for the reconstruction using MLOST for PSM calibration and CCSM4 LM simulation for the prior. Anomalies are defined relative to the 1950–1980 time mean. The gray band shows the 5–95% percentile range for LMR, and the red band shows the range of data in the 17 reconstructions. The straight solid gray line shows LMR trend over 0–1850 C.E. fit by least squares.

3. Results

We begin with results for 2 m air temperature over the 0–2000 C.E. period, before proceeding to verification of this field during 1880–2000 and finally to verification results for 500 hPa geopotential height. The "control" reconstructions are based on the PSM calibration against the MLOST gridded instrumental product, and prior data drawn from the CCSM4-LM simulation.

3.1. Two Meter Air Temperature

In order to place our results in the context of previously published reconstructions, we show in Figure 2 (and in the supporting information) the 2 m air temperature averaged over the Northern Hemisphere for the LMR and a range of studies quoted in Intergovernmental Panel on Climate Change Fifth Assessment Report (IPCC AR5) (Working Group 1 Figure 5.7 and Table 5.A.6). The LMR is in reasonable agreement with several previous reconstructions, with the closest three correlating with the LMR at 0.55 [*Mann et al.*, 2008],

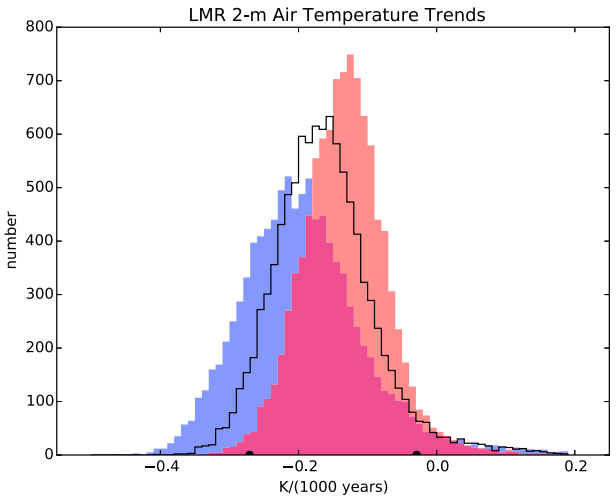

Figure 3. Histograms of Northern Hemisphere (blue), Southern Hemisphere (red), and global-mean 2 m air temperature trend over 0–1850 C.E. (units K/1000 years). Samples apply to 100 realizations of 100-member ensembles (10,000 samples total). Black dots on the abscissa show the 5–95% percentile range for the global mean.

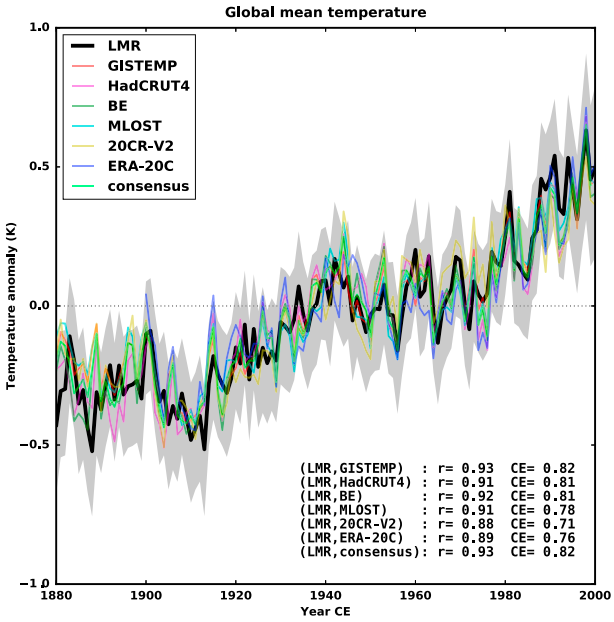

Figure 4. Comparison between LMR 2 m global-mean air temperature and other analyses (GISTEMP: NASA GISS surface temperature analysis; HadCRUT4: Hadley Center/Climate Research Unit at the University of East Anglia temperature data set version 4; 20CR-V2: NOAA twentieth Century Reanalysis version 2; BE: Berkeley Earth surface temperatures; MLOST: NOAA merged land-ocean surface temperature analysis; ERA-20C: ECMWF reanalysis of the twentieth century; Consensus: average of all but LMR. Gray band shows 5–95% percentile range for LMR.)

0.58 [*Shi et al.*, 2013], and 0.64 [*Mann et al.*, 2009]. Comparing against the consensus mean over the 14 reconstructions, half of the reconstructions correlate at 0.7 or higher with the mean, whereas LMR correlates with this consensus mean at 0.66 (see supporting information). Apparently, most other reconstructions are more similar to each other than they are with the LMR. As in many previous reconstructions, LMR has a cooling trend over the 0–2000 C.E. time period. Superimposed upon this trend are multicentennial oscillations, including features consistent with the nominal Medieval Climate anomaly (MCA) and Little Ice Age (LIA).

The probability density of the trend as estimated by 10,000 reconstruction realizations (100 Monte Carlo samples each having a 100-member ensemble) reveals a cooling trend about twice as large in the Northern Hemisphere compared to the Southern Hemisphere (Figure 3). We refrain from speculating on the source of these trends and their difference but note that the proxy data distribution is heavily weighted toward the Northern Hemisphere. Further investigation using equal-sized samples for both hemispheres is left for future work with larger proxy data sets.

For verification during the instrumental period, 1880–2000, we consider first the global-mean 2 m air temperature at annual resolution (Figure 4). ERA-20C data period begins in 1900, its verification results apply to the 1900–2000 time period. The LMR sample mean (over 10,000 realizations) agrees very well with available reanalysis products, with correlations of 0.88–0.93 and C.E. values of 0.71–0.82; the "consensus" verification data set consists of the average over all verification data. We note that even though the PSM for the LMR reconstruction was calibrated against MLOST data, the results verify best against GISTEMP (bootstrap-estimated error in the correlation estimates is <0.005). To measure the uncertainty in the verification data sets themselves, Figure 5 shows correlation and C.E. values using each data set as the reference standard. This reveals broad agreement between the data sets, particularly for the consensus data set, which verifies best in both measures. The worst agreement concerns verification of 20CR-V2 and ERA-20C against each other, which results from poor-quality analyses over land (see supporting information). Given that 20CR-V2 and ERA-20C utilize 6 h assimilation on surface pressure measurements using operational weather forecasting models, and online DA algorithms, it is remarkable that the LMR performs nearly as well as these products with an offline PDA approach and only 447 noisy proxy chronologies.

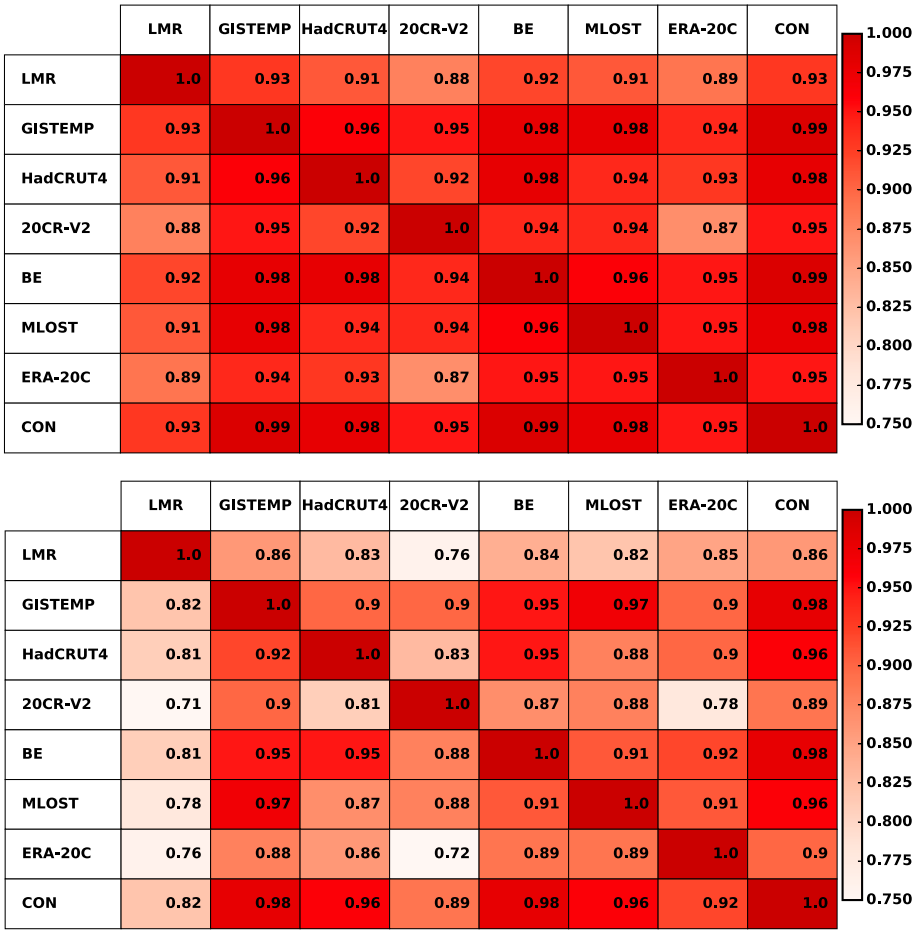

Figure 5. Verification of global-mean 2 m air temperature during 1880–2000 C.E. in terms of (top) correlation and (bottom) coefficient of efficiency for analyses defined in Tables 1 and 2. Verification data sets are given in row space, and data sets being verified are given in column space.

Figure 6 yields insight into how much of the LMR performance is dependent on capturing the 1880–2000 trend in 2 m air temperature. Results show that the LMR 1880–2000 trend is larger than the verification data sets by about 10–20%. Detrending both the LMR and verification data yields correlations of about 0.7 and C.E. values of about 0.4–0.5; the one exception is 20CR-V2, against which LMR compares with lower skill. Although the 1880–2000 trend contributes about 50% of the skill in CE, the remainder evidently comes from the variability on interannual to decadal timescales.

Verification of the full gridded 2 m air temperature field (Figure 7) shows high correlation nearly globally, with lower values over portions of land masses. C.E. in particular shows lower skill over land, and a larger range of values between results for 20CR-V2 and GISTEMP (see also supporting information). We note that near many of these land areas, such as over western North America, the correlation results are large and positive, which indicates good agreement in anomaly timing. This implies that the poor CE score is due to bias and/or to a problem with the amplitude of the signal. Given that many of the proxies in this location are moisture-sensitive trees, which we have modeled with a temperature-based PSM, a bias in the reconstruction is perhaps not surprising. Another possible contribution to this issue stems from the fact that tree growth responds primarily to summertime conditions, whereas the contribution to annual-mean midlatitude temperature variance is dominated by wintertime synoptic variability. A more complete investigation of these issues is left for future research.

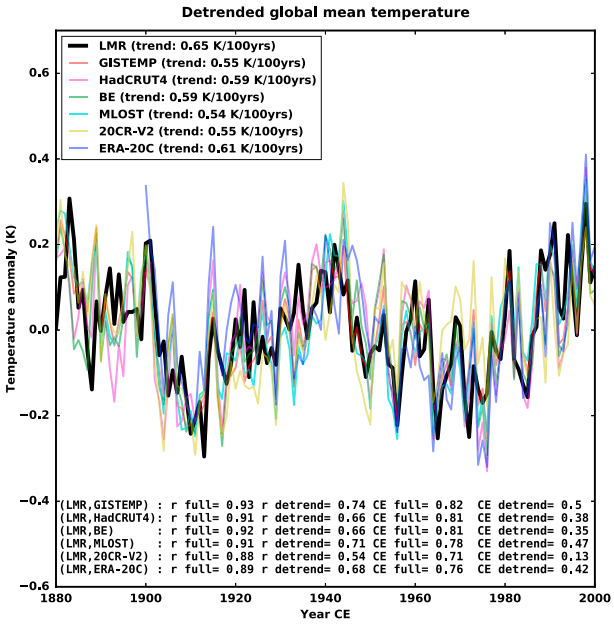

Figure 6. Verification of LMR global-mean 2m air temperature against other analyses shown in Figure 5 for the trend and detrended signal components. The trend is defined by the least squares linear fit over the time interval, and the detrended signal has this trend removed.

Another measure of fidelity in the spatial skill of the LMR 2 m air temperature field derives from spatial anomaly correlation for each year (Figure 8, left column). For most years, the LMR shows skill at reconstructing spatial patterns, and histograms of these time series (Figure 8, right column) reveal mean values mostly around 0.2.

Finally, we verify the LMR directly against proxy data for both assimilated and independent proxy chronologies (Figure 9). These results are based on direct comparisons between proxy values and estimates of the proxy values from the LMR using PSM equation (5) on the annually resolved LMR reconstructed 2 m air temperature field. In addition to comparing results for both the assimilated and independent proxy records, we also compare results during the 1880–2000 PSM calibration period with 0–1879 C.E. (blue and red distributions, respectively). Moreover, we also present the distributions for prior estimate of the proxy values (black lines), which show zero correlation and negative CE scores as anticipated by equation (8).

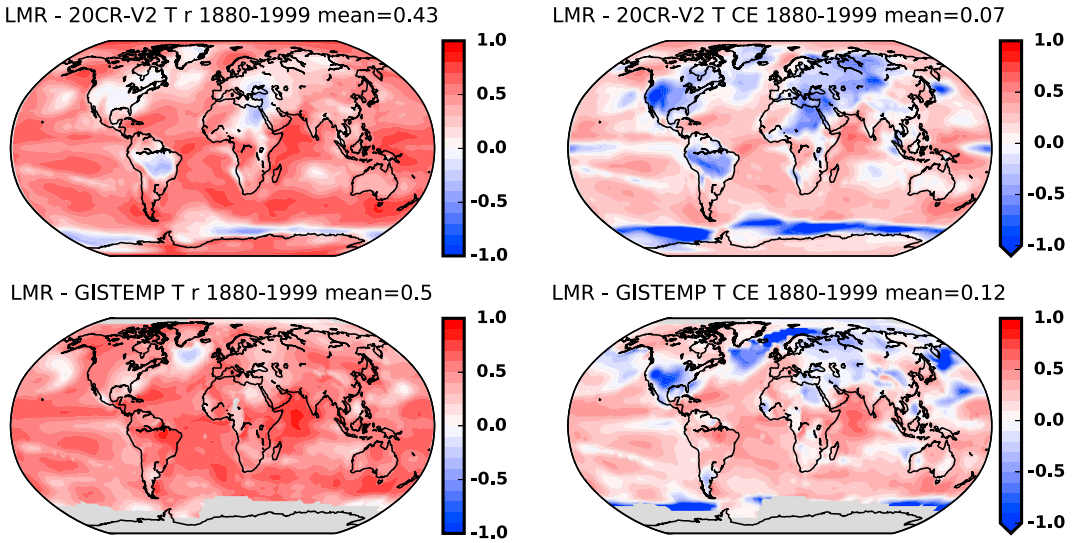

Figure 7. Verification of LMR 2 m air temperature against (top row) 20CR-V2 and (bottom row) GISTEMP. Shown are (left column) time series correlation and (right column) coefficient of efficiency.

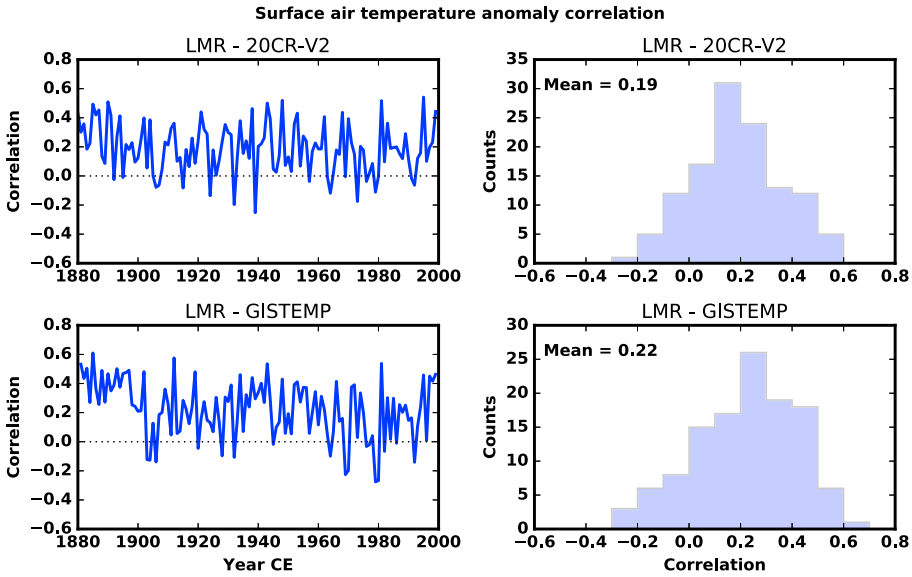

Figure 8. Spatial anomaly correlation between LMR 2 m air temperature and (top row) 20CR-V2 and (bottom row) GISTEMP. (left column) Time series of annual-mean spatial correlation is shown and (right column) histograms of these values.

For the assimilated proxies, results are very similar for the calibration and precalibration time period for both correlation and C.E. The correlation distributions are mostly positive with median values around 0.3. The change in CE from the prior (Figure 9, top right) is peaked near zero, with a long tail toward positive values. Results for nonassimilated proxies (Figure 9, bottom row) are similar to those for the assimilated proxies, with smaller correlations values, and similar values for change in CE.

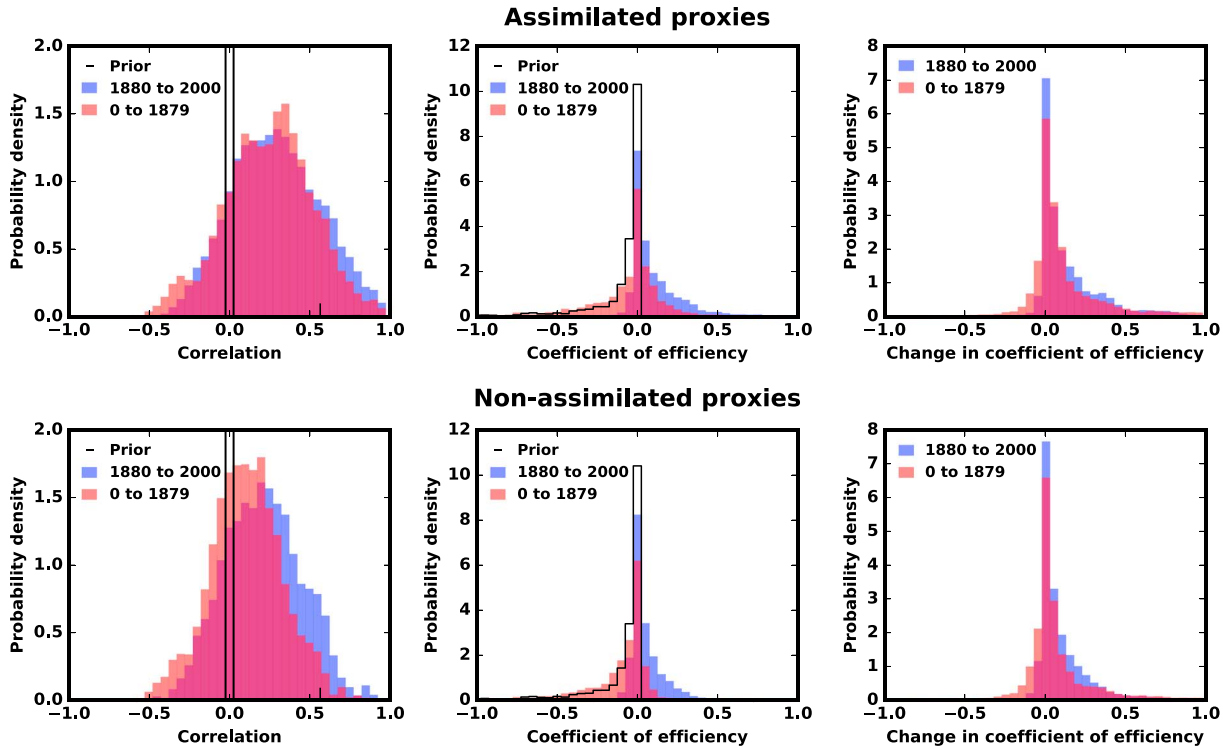

Figure 9. Verification of LMR against (top row) assimilated and (bottom row) nonassimilated proxy data. (left column) Correlation is shown, (middle column) coefficient of efficiency, and (right column) coefficient of efficiency difference (LMR analysis CE minus prior C.E.). Results for 1880–2000 (calibration period) are shown in blue, and those for 0–1879 in red. Proxy values are estimated from the LMR analysis using the proxy system model.

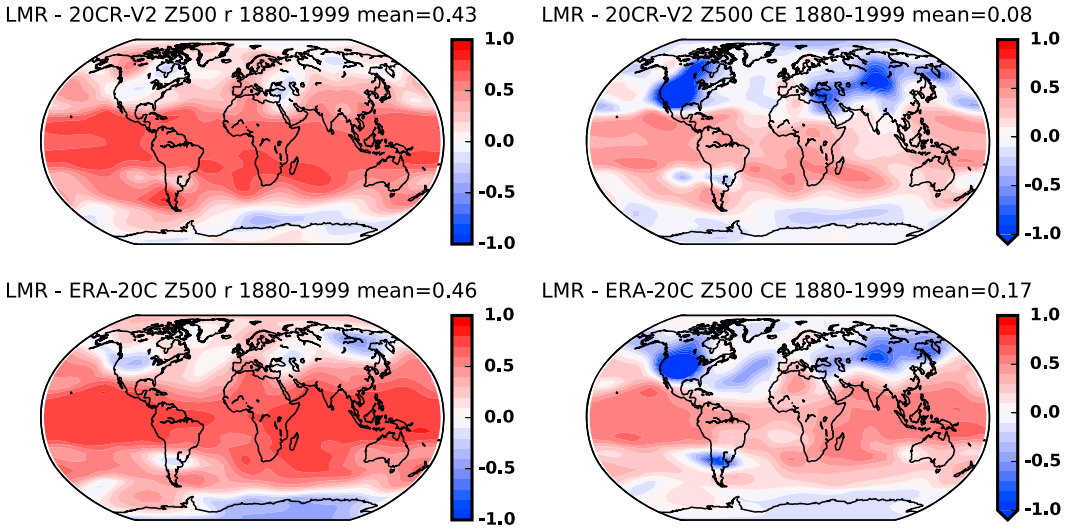

Figure 10. Verification of LMR 500 hPa geopotential height anomalies against (top row) 20CR-V2 and (bottom row) ERA-20C for (left column) time series correlation and (right column) coefficient of efficiency.

3.2. Geopotential Height (500 hPa)

Verification of LMR results for 500 hPa geopotential height is limited to 20CR-V2 and ERA-20C reanalyses. Full spatial verification reveals, as for 2 m air temperature, that the region of low skill in the Northern Hemisphere is over North America and Asia, with moderate skill over the North Pacific ocean (Figure 10). Anomaly correlation is also similar to 2 m air temperature, with perhaps slightly lower skill overall (Figure 11). These results show that, over average, there is significant skill in LMR reconstructions of 500 hPa geopotential height. We hypothesize that the greater skill in the tropics derives from a mostly hydrostatic response over a deep layer due to an environment that is actively mixed by moist convection. In other words, there is a deep response to surface temperature in the tropics compared to high latitudes, where surface temperature anomalies rapidly decorrelate in the vertical, which reduces the effectiveness of surface proxies to constrain the mass field over a deep layer.

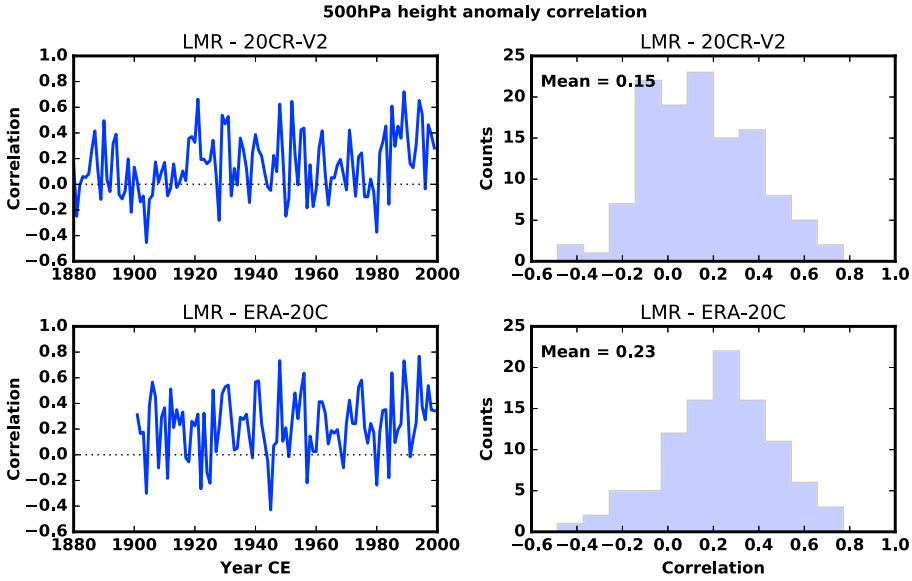

Figure 11. Spatial anomaly correlation between LMR 500 hPa geopotential height anomalies and (top row) 20CR-V2 and (bottom row) ERA-20C (left column) as a function of time and (right column) as a histogram showing the distribution of values over all years.

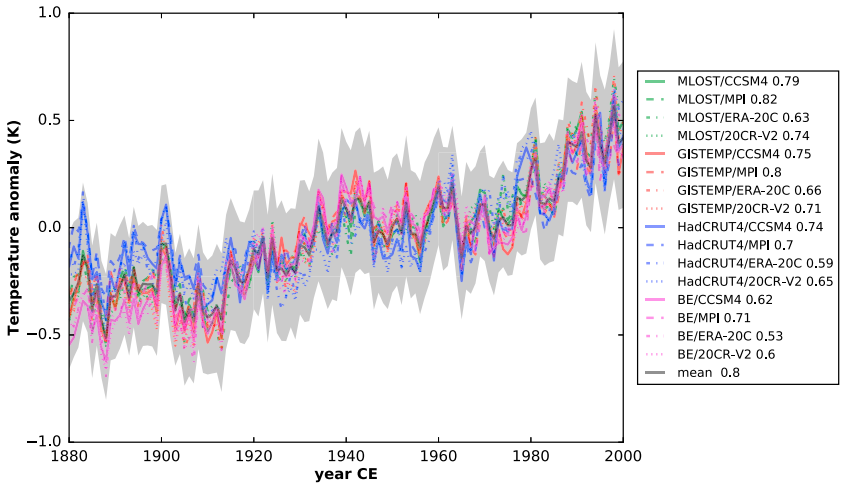

Figure 12. Sensitivity of global-mean temperature to calibration and prior data sources. Legend entries for each experiment are denoted by calibration/prior data source. Each curve denotes the mean over 100 realizations that randomly sample over 75% of the proxy data using a 100-member ensemble for each. CE verification against consensus mean is shown in the legend.

4. Sensitivity Analysis

We turn now to the sensitivity of the results to each component of the PDA framework: PSM, prior, and proxies. In order to effectively summarize the results, we consider only global-mean air temperature and note that qualitatively similar findings emerge when other dynamical fields are considered.

We combine the sensitivity analysis for the PSM calibration and prior data into a set of 16 experiments, each with 100 Monte Carlo realizations for 100-member ensembles (160,000 reconstructions in all), for all combinations of calibration and prior data (Figure 12). For each calibration data set, the linear PSM equation (5) is fit and used to predict the proxy values from the prior data for that experiment. For example, the MLOST/MPI experiment uses PSMs calibrated on MLOST 2 m air temperature data and the MPI last millennium simulation as the prior; i.e., proxy values are predicted from the MPI using the MLOST-calibrated PSMs.

Results show little sensitivity, and close correspondence to the control reconstruction verified in section 3.1. In all cases, verification is against the consensus global-mean temperature shown in Figure 4. There is, however, an apparent correspondence between calibration and prior data and skill in the reconstruction. Skill for 20CR-V2 and ERA-20C priors are clearly lower than for other priors. This is potentially surprising, since the other priors pertain to preindustrial simulations whereas 20CR-V2 and ERA-20C apply to the verification time period. We suspect that the source of this lower skill relates to the lower skill on annual-mean 2 m air temperature over continents (compare Figure 7 (top row) and Figure 7 (bottom row), and see supporting information) because these reanalysis products rely primarily on surface pressure measurements rather the surface temperature data. The MPI prior has the highest skill for three of the four calibration data sets considered. Skill for each calibration data set averaged over all prior data sets is 0.75, 0.73, 0.67, and 0.62 for MLOST, GISTEMP, HadCRUT4, and BE, respectively. The reconstruction with the highest skill, MLOST/MPI, is the only one with higher skill than the grand ensemble average over all 160,000 experiments.

Sensitivity to proxy data has already been considered by randomly sampling the entire network. Here we consider the sensitivity to proxy type. Since proxy data from trees dominate the PAGES2k data set, we perform two experiments: one with all proxies other than trees (27 chronologies total), and one with only tree data (19 tree ring density chronologies and 401 tree ring width chronologies; 420 chronologies total). In both cases, all of the proxy data available were used (i.e., we did not withhold 25% as in all other experiments), and we use MLOST PSM calibration and the CCSM4 for the prior, as for the control. We find that the reconstruction with only trees captures most of the signal in the case with all proxies (Figure 13). Despite having only 27 chronologies, the reconstruction without trees still has a positive skill, but with smaller amplitude in the main signals (trend and multidecadal variability).

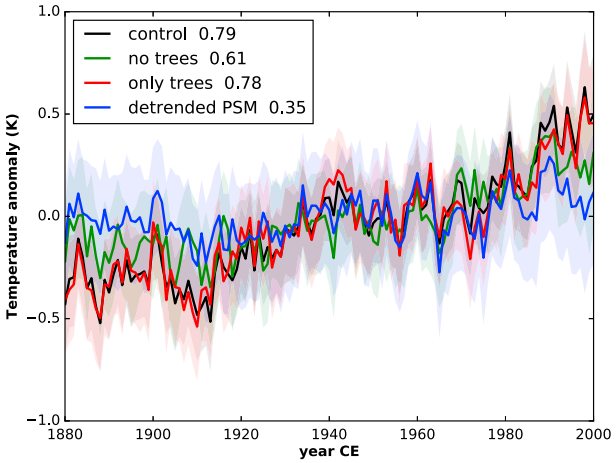

Figure 13. Sensitivity of global-mean temperature to proxy source (trees only, no trees) and the effect of eliminating the trend from the PSM calibration.

Finally, the blue curve in Figure 13 gives the result when the proxy PSMs are fit on detrended data. In this case, both the proxy and the calibration data are detrended over the entire calibration period, and the proxy error variance, diagonal elements of **R** in equation (2), is defined relative to the detrended time series. Since the signal decreases relative to the noise in this PSM construction, the proxies have less influence, resulting in a shallower trend and less variability in the reconstruction. Our interpretation of this result is that the twentieth century trend provides a large signal on which to calibrate the regression-based PSM, increasing signal over noise, which is particularly large on interannual timescales. Moreover, the fact that the control experiment, with PSMs calibrated with the trend, captures variability independent of the trend (Figure 6) indicates that the PSM is not simply tuned to capture the trend. Furthermore, we remind the reader that the PSM predicts the proxies from the prior data, which is not only out of sample but is also entirely external to the calibration data set since it comes from a simulation from a climate model for a different period of time. Therefore, there is no a priori training of the PSMs that can artificially induce "known" trends.

5. Multivariate Reconstruction Example

A fundamental advantage of the PDA approach is to enable dynamically consistent reconstructions of multivariate states from the proxy record. We demonstrate this by examining multivariate fields for a specific historical event: the mystery volcanic eruption of 1808/1809 [*Guevara-Murua et al.*, 2014]. The global 2 m air temperature and 500 hPa geopotential height field are shown in Figure 14 for 1808–1813, with anomalies relative to the 10 year mean preceding 1808. Warm conditions are evident during 1808, particularly in the eastern tropical Pacific ocean and northern Europe, with mainly weak 500 hPa geopotential height anomalies except near Antarctica. Global cooling takes place in 1809, with locally stronger cooling over Canada where 500 hPa geopotential heights are anomalously low. Cooling intensifies in 1810, particularly over the northern portion of Northern Hemisphere land areas. It appears that the regions of eastern North America were largely unaffected on these timescales, with largest anomalies over western North America. The 500 hPa geopotential height field is dominated by an extratropical wave train closely resembling the Pacific-North America (PNA) pattern [*Wallace and Gutzler*, 1981], which is known to respond to thermal forcing in the tropical Pacific Ocean [e.g., *Hoerling and Kumar*, 2002]. This pattern persists and slowly evolves through 1813, with locally intensified cooling over most land areas by this time.

The global-mean temperature evolution during the 30 year period surrounding the 1808/1809 eruption clearly shows the abrupt cooling in 1809, with reinforced cooling around 1815 and 1820 before recovering around 1823 (Figure 15). Presumably, the 1815 cooling enhancement is due to the eruption of Mount Tambora, which from this analysis appears to simply reinforce the cooling event established by the 1808/1809 eruption. Also shown in Figure 15 is the uncertainty from both the ensemble (shading) and a sample over experiments using different combinations of calibration data for the PSMs and prior data in the

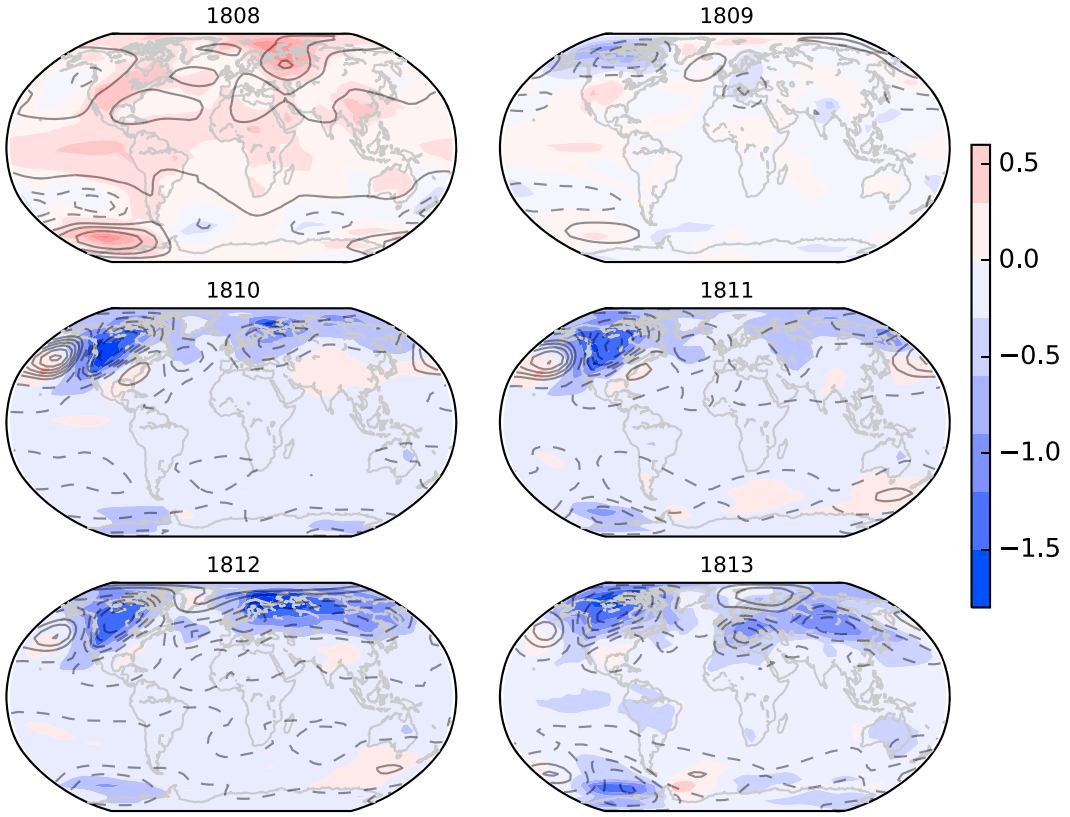

Figure 14. Example of multifield reconstruction for a specific event: the volcanic eruption of late 1808. Colors show 2 m air temperature anomalies (units: K), and contours show 500 hPa geopotential height anomalies (contours every 5 m; negative values dashed). Anomalies apply to the 10 year mean prior to the eruption and to the MLOST/CCSM4 reconstruction.

reconstruction (colors). It is interesting that little uncertainty exists prior to 1808, with all of the reconstructions closely clustered. Uncertainty increases rapidly following the 1808/1809 eruption, with the largest spread of about 0.3 K apparent about 5 years after the eruption.

Uncertainty in the spatial fields from the range of calibration/prior reconstructions shown in Figure 15 indicates that the primary features for 1810 are robust (Figure 16). Specifically, all reconstructions have a PNA

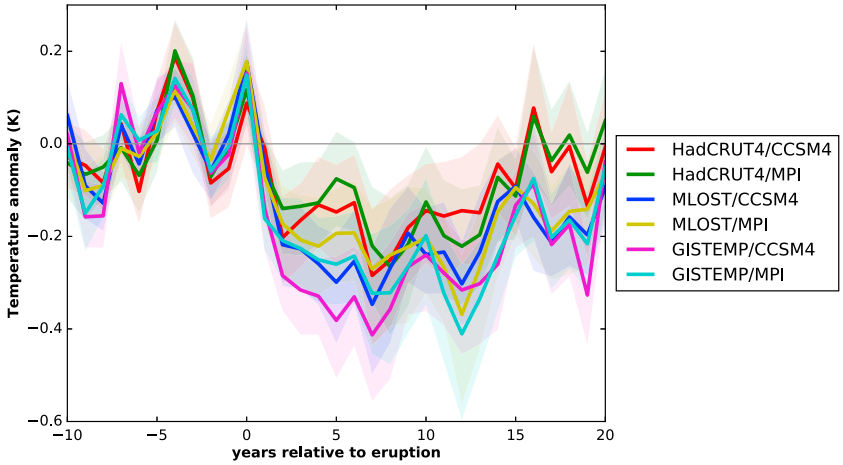

Figure 15. Sensitivity of global-mean 2 m air temperature anomalies to the calibration and prior data (shown in the caption as calibration/prior). Anomalies are relative to 10 year period prior to the 1808 eruption, which is denoted by year 0 on the abscissa.

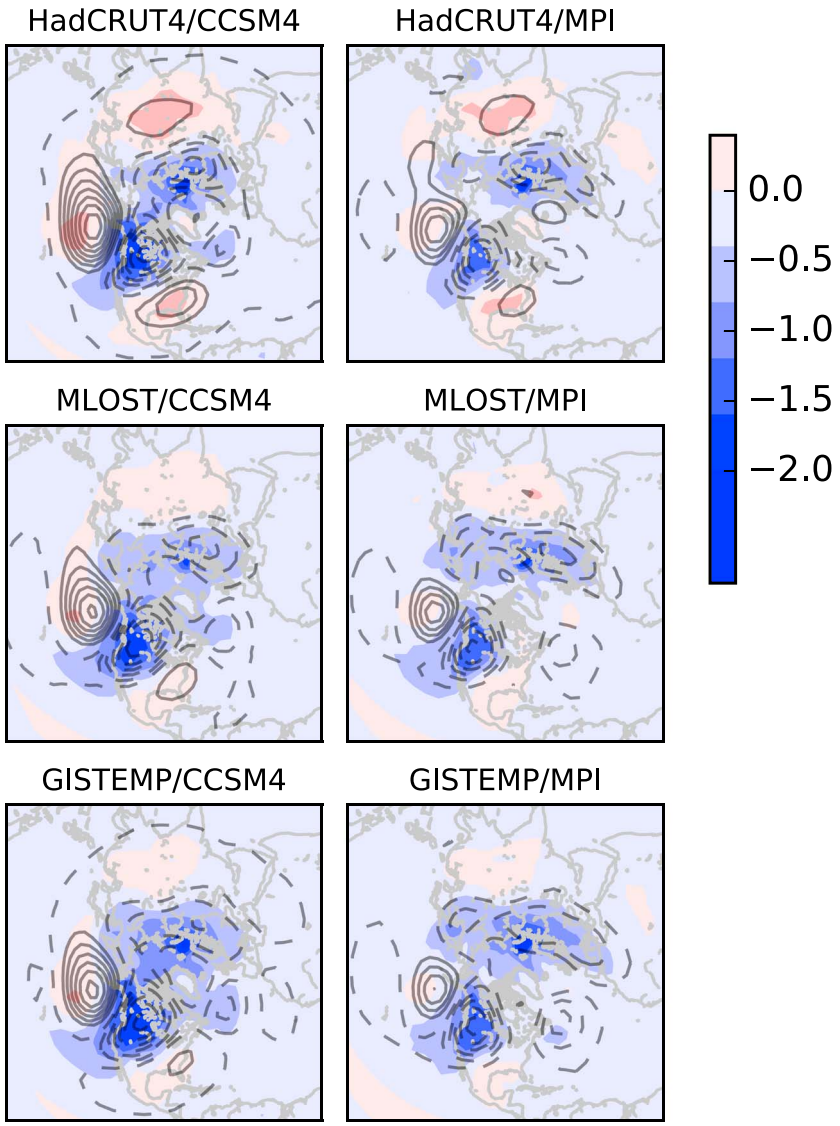

Figure 16. Sensitivity of global-mean 2 m air temperature (units: K) and 500 hPa geopotential height anomaly fields (contours every 5 m; negative values dashed) to the calibration and prior data (shown in the titles as calibration/prior). All panels pertain to the year 1810.

pattern in the 500 hPa geopotential height field, although they differ in the amplitude of the ridge over the North Pacific ocean and the trough over western North America. Although the 2 m air temperature field corresponds to a roughly equivalent barotropic response with the 500 hPa geopotential height field, it is interesting to note that the anomalies in the 2 m air temperature field with the North Pacific ridge are smaller that for the North American trough, as might be expected for the land-ocean contrast associated with these features. Comparing the reconstructions that are warmest (HadCRUT4/MPI) and coldest (GISTEMP/CCSM4-LM) in the global-mean 2 m air temperature suggests that, at least for the Northern Hemisphere, these differences are associated with relatively minor changes in the spatial pattern.

6. Concluding Summary

The purpose of this paper was to document an approach to paleoclimate climate field reconstruction using a data assimilation technique that weights data from proxies against climate model simulations and to verify the assimilated reconstructions from this approach primarily during the instrumental era. The approach involves linear, univariate proxy system models and an "offline" (no cycling) approach to data assimilation. Both of these approximations are chosen to establish a baseline for reconstruction skill as a reference to

measure future improvements by generalizing these approximations. We performed reconstructions using proxy data drawn from the PAGES2K [*PAGES2K Consortium*, 2013] proxy data set and CMIP5 climate model simulations. Reconstructions are compared to previous reconstructions of Northern Hemisphere mean 2 m air temperature, and a sample of reanalysis products in both space and time.

For the control reconstruction, the PSMs are fit against the MLOST temperature field, and a 100-member prior ensemble is drawn randomly from the CCSM4 Last Millennium simulation. When compared to previous reconstructions of Northern Hemisphere mean 2 m air temperature for 0–2000 C.E., we find agreement in gross aspects, such as millennial-scale cooling, but also that many of the previous reconstructions are more similar to each other than they are to the LMR. This is due, perhaps, to the fact that the previous reconstructions share methods that are more similar to each other than they are to the LMR. Verification against six commonly used reanalysis products at annual resolutions reveals greatest skill in the global mean, with less skill in the spatial field, particularly over North America and Eurasia. Verification against independent proxy records withheld from data assimilation indicate improvement over the baseline skill of the prior.

Reconstructions for 16 combinations of calibration and prior data show little sensitivity in the results. This is true for both the time series of global-mean 2 m air temperature and for spatial patterns of geopotential height and temperature during the 1808/1809 volcanic eruption used as an example. The success of the LMR in reproducing persistent cooling, and robust geopotential height anomalies, following a volcanic eruption is an important demonstration of the ability of the PDA approach to capture specific climate fluctuations and therefore offers confidence in the LMR reconstructions over the duration of the 2000 year proxy data set. The largest impact on the results is obtained when the linear PSMs are fit on detrended data, both for proxy and calibration data. In this case, the reconstructed global-mean 2 m air temperature has a much reduced trend and decadal variability, which is a result of less weight on the proxies during data assimilation due to relatively larger errors for the PSM calibration. Given that the control, calibrated including the trend, shows substantial skill when the trend is removed (Figure 3) suggests that the trend simply provides greater signal-to-noise ratio when calibrating the PSMs.

Having established a baseline LMR reconstruction, we will focus future efforts on using a vastly expanded proxy network, adopting more comprehensive PSMs such as PRYSM [*Dee et al.*, 2015], accounting for seasonal dependencies, using isotope-enabled GCMs as priors, and cycling data assimilation.

Acknowledgments

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). Climate simulations were generated as part of the Paleoclimate Model Intercomparison (PMIP3) project. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. CMIP data used in this paper may be obtained from the Earth System Grid Federation at http://esgf.llnl.gov/. Access to data and software related to research in this paper are published at the doi: 10.17911/S9WC7N. Comments 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.

References

Annan, J., J. Hargreaves, N. Edwards, and R. Marsh (2005), Parameter estimation in an intermediate complexity Earth system model using an ensemble Kalman filter, *Ocean Model.*, *8*(1), 135–154.

Barriopedro, D., E. M. Fischer, J. Luterbacher, R. M. Trigo, and R. García-Herrera (2011), The hot summer of 2010: Redrawing the temperature record map of Europe, *Science*, *332*(6026), 220–224, doi:10.1126/science.1201224.

Bhend, J., J. Franke, D. Folini, M. Wild, and S. Brönnimann (2012), An ensemble-based approach to climate reconstructions, *Clim. Past*, *8*(3), 963–976, doi:10.5194/cp-8-963-2012.

Carton, J., and B. Giese (2008), A reanalysis of ocean climate using Simple Ocean Data Assimilation (SODA), *Mon. Weather Rev.*, *136*(8), 2999–3017.

Christiansen, B. (2010), Reconstructing the NH mean temperature: Can underestimation of trends and variability be avoided?, *J. Clim.*, *24*(3), 674–692, doi:10.1175/2010JCLI3646.1.

Compo, G., et al. (2011), The twentieth century reanalysis project, *Q. J. R. Meteorol. Soc.*, *137*, 1–28.

Cook, E., C. Woodhouse, C. Eakin, D. Meko, and D. Stahle (2004), Long-term aridity changes in the western United States, *Science*, *306*(5698), 1015–1018.

Cook, E. R., D. M. Meko, D. W. Stahle, and M. K. Cleaveland (1999), Drought reconstructions for the continental United States*, *J. Clim.*, *12*(4), 1145–1162.

Cook, E. R., K. J. Anchukaitis, B. M. Buckley, R. D. D'Arrigo, G. C. Jacoby, and W. E. Wright (2010), Asian monsoon failure and megadrought during the last millennium, *Science*, *328*(5977), 486–489, doi:10.1126/science.1185188.

Dee, D., M. Balmaseda, G. Balsamo, R. Engelen, A. Simmons, and J. Thépaut (2014), Toward a consistent reanalysis of the climate system, *Bull. Am. Meteorol. Soc.*, *95*(8), 1235–1248.

Dee, S., J. Emile-Geay, M. N. Evans, A. Allam, E. J. Steig, and D. M. Thompson (2015), Prysm: An open-source framework for proxy system modeling, with applications to oxygen-isotope systems, *J. Adv. Model. Earth Syst.*, *7*(3), 1220–1247, doi:10.1002/2015MS000447.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm, *J. R. Stat. Soc. B*, *39*(1), 1–38.

Dirren, S., and G. J. Hakim (2005), Toward the assimilation of time-averaged observations, *Geophys. Res. Lett.*, *32*, L04804, doi:10.1029/2004GL021444.

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, *Paleoceanography*, *17*, 1007, doi:10.1029/2000PA000590.

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

Evans, M. N., J. E. Smerdon, S. E. Tolwinski-Ward, A. Kaplan, and J. F. González-Rouco (2014), Climate field reconstruction uncertainty arising from multivariate and nonlinear properties of predictors, *Geophys. Res. Lett.*, *41*, 9127–9134, doi:10.1002/2014GL062063.

Evensen, G. (2003), The ensemble Kalman filter: Theoretical formulation and practical implementation, *Ocean Dyn.*, *53*(4), 343–367.
Fierro, R. D., G. H. Golub, P. C. Hansen, and D. P. O' Leary (1997), Regularization by truncated total least squares, *SIAM J. Sci. Comput.*, *18*, 1223–1241.
Flato, G., et al. (2013), Evaluation of climate models, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. F. Stocker et al., pp. 741–866, Cambridge Univ. Press, Cambridge, U. K., and New York, doi:10.1017/CBO9781107415324.020.
Goosse, H., H. Renssen, A. Timmermann, R. Bradley, and M. Mann (2006), Using paleoclimate proxy-data to select optimal realisations in an ensemble of simulations of the climate of the past millennium, *Clim. Dyn.*, *27*, 165–184, doi:10.1007/s00382-006-0128-6.
Guevara-Murua, A., C. Williams, E. Hendy, A. Rust, and K. Cashman (2014), Observations of a stratospheric aerosol veil from a tropical volcanic eruption in December 1808: Is this the unknown 1809 eruption?, *Clim. Past*, *10*(5), 1707–1722.
Guillot, D., B. Rajaratnam, and J. Emile-Geay (2015), Statistical paleoclimate reconstructions via Markov random fields, *Ann. Appl. Stat.*, *9*(1), 324–352, doi:10.1214/14-AOAS794.
Hamill, T. M. (2001), Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Weather Rev.*, *129*(3), 550–560.
Hansen, J., R. Ruedy, M. Sato, and K. Lo (2010), Global surface temperature change, *Rev. Geophys.*, *48*, RG4004, doi:10.1029/2010RG000345.
Hoerl, A. E., and R. W. Kennard (1970a), Ridge regression: Biased estimation for non-orthogonal problems, *Technometrics*, *12*, 55–67.
Hoerl, A. E., and R. W. Kennard (1970b), Ridge regression: Applications to non-orthogonal problems, *Technometrics*, *12*, 69–82, correction, 12, 723.
Hoerling, M. P., and A. Kumar (2002), Atmospheric response patterns associated with tropical forcing, *J. Clim.*, *15*(16), 2184–2203.
Houtekamer, P., H. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen (2005), Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations, *Mon. Weather Rev.*, *133*(3), 604–620.
Huntley, H. S., and G. J. Hakim (2010), Assimilation of time-averaged observations in a quasi-geostrophic atmospheric jet model, *Clim. Dyn.*, *35*(6), 995–1009.
Jones, P., et al. (2009), High-resolution palaeoclimatology of the last millennium: A review of current status and future prospects, *Holocene*, *19*(1), 3–49.
Kalnay, E. (2003), *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge Univ. Press, Cambridge, U. K.
Kalnay, E., et al. (1996), The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*(3), 437–470.
Köhl, A. (2015), Evaluation of the GECCO2 ocean synthesis: Transports of volume, heat and freshwater in the Atlantic, *Q. J. R. Meteorol. Soc.*, *141*(686), 166–181.
Landrum, L., B. Otto-Bliesner, E. Wahl, A. Conley, P. Lawrence, N. Rosenbloom, and H. Teng (2013), Last millennium climate and its variability in CCSM4, *J. Clim.*, *26*(4), 1085–1111.
Little, R. J. A., and D. B. Rubin (2002), *Statistical Analysis With Missing Data*, Wiley Series in Probability and Statistics, New York. [Available at http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471183865.html.]
Luterbacher, J., D. Dietrich, E. Xoplaki, M. Grosjean, and H. Wanner (2004), European seasonal and annual temperature variability, trends, and extremes since 1500, *Science*, *303*(5663), 1499–1503, doi:10.1126/science.1093877.
Mann, M. E., R. S. Bradley, and M. K. Hughes (1998), Global-scale temperature patterns and climate forcing over the past six centuries, *Nature*, *392*, 779–787, doi:10.1038/33859.
Mann, M. E., R. S. Bradley, and M. K. Hughes (1999), Northern Hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations, *Geophys. Res. Lett.*, *26*, 759–762, doi:10.1029/1999GL900070.
Mann, M. E., Z. Zhang, M. K. Hughes, R. S. Bradley, S. K. Miller, S. Rutherford, and F. Ni (2008), Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia, *Proc. Natl. Acad. Sci.*, *105*(36), 13,252–13,257.
Mann, M. E., Z. Zhang, S. Rutherford, R. S. Bradley, M. K. Hughes, D. Shindell, C. Ammann, G. Faluvegi, and F. Ni (2009), Global signatures and dynamical origins of the little ice age and medieval climate anomaly, *Science*, *326*(5957), 1256–1260, doi:10.1126/science.1177303.
Masson-Delmotte, V., et al. (2013), Information from paleoclimate archives, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, chap. 5*, edited by T. Stocker et al., pp. 383–464, Cambridge Univ. Press, Cambridge, U. K., and New York.
Matsikaris, A., M. Widmann, and J. H. Jungclaus (2015), On-line and off-line data assimilation in palaeoclimatology: A case study, *Clim. Past*, *11*, 81–93.
Mayewski, P. A., et al. (2004), Holocene climate variability, *Quat. Res.*, *62*(3), 243–255.
Morice, C., J. Kennedy, N. Rayner, and P. Jones (2012), Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 dataset, *J. Geophys. Res.*, *117*, D08101, doi:10.1029/2011JD017187.
Nash, J., and J. Sutcliffe (1970), River flow forecasting through conceptual models: Part I — A discussion of principles, *J. Hydrol.*, *10*(3), 282–290.
Oke, P. R., J. S. Allen, R. N. Miller, G. D. Egbert, and P. M. Kosro (2002), Assimilation of surface velocity data into a primitive equation coastal ocean model, *J. Geophys. Res.*, *107*(C9), 3122, doi:10.1029/2000JC000511.
PAGES2K Consortium (2013), Continental-scale temperature variability during the past two millennia, *Nat. Geosci.*, *6*(5), 339–346, doi:10.1038/ngeo1797.
Ridgwell, A., J. Hargreaves, N. R. Edwards, J. Annan, T. M. Lenton, R. Marsh, A. Yool, and A. Watson (2007), Marine geochemical data assimilation in an efficient Earth system model of global biogeochemical cycling, *Biogeosciences*, *4*(1), 87–104.
Rohde, R., R. Muller, R. Jacobsen, S. Perlmutter, A. Rosenfeld, J. Wurtele, J. Curry, C. Wickman, and S. Mosher (2013), Berkeley Earth temperature averaging process, *Geoinfor. Geostat. An Overview*, *1*(2), 1–13, doi:10.4172/2327-4581.1000103.
Rutherford, S., M. E. Mann, T. J. Osborn, R. S. Bradley, K. R. Briffa, M. K. Hughes, and P. D. Jones (2005), Proxy-based Northern Hemisphere surface temperature reconstructions: Sensitivity to method, predictor network, target season, and target domain., *J. Clim.*, *18*, 2308–2329.
Schneider, T. (2001), Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values, *J. Clim.*, *14*(5), 853–871, doi:10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2.
Shi, F., B. Yang, A. Mairesse, L. von Gunten, J. Li, A. Bräuning, F. Yang, and X. Xiao (2013), Northern Hemisphere temperature reconstruction during the last millennium using multiple annual proxies, *Clim. Res.*, *56*, 231–244.
Smerdon, J. E., A. Kaplan, D. Chang, and M. N. Evans (2010), A pseudoproxy evaluation of the CCA and RegEM methods for reconstructing climate fields of the last millennium, *J. Clim.*, *23*(18), 4856–4880, doi:10.1175/2010JCLI3328.1.
Smerdon, J. E., A. Kaplan, E. Zorita, J. F. González-Rouco, and M. N. Evans (2011), Spatial performance of four climate field reconstruction methods targeting the Common Era, *Geophys. Res. Lett.*, *38*, L11705, doi:10.1029/2011GL047372.
Smerdon, J. E., S. Coats, and T. R. Ault (2015), Model-dependent spatial skill in pseudoproxy experiments testing climate field reconstruction methods for the Common Era, *Clim. Dyn.*, *46*, 1921–1942, doi:10.1007/s00382-015-2684-0.

Smith, T., R. Reynolds, T. Peterson, and J. Lawrimore (2008), Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880–2006), *J. Clim.*, *21*(10), 2283–2296.
Steiger, N., and G. Hakim (2015), Multi-time scale data assimilation for atmosphere-ocean state estimates, *Clim. Past Discuss.*, *11*, 3729–3757.
Steiger, N. J., G. J. Hakim, E. J. Steig, D. S. Battisti, and G. H. Roe (2014), Assimilation of time-averaged pseudoproxies for climate reconstruction, *J. Clim.*, *27*(1), 426–441.
Taylor, K., R. Stouffer, and G. Meehl (2012), An overview of CMIP5 and the experiment design, *Bull. Am. Meteorol. Soc.*, *93*(4), 485–498.
Tikhonov, A. N., and V. Y. Arsenin (1977), *Solution of Ill-Posed Problems, Scripta Series in Mathematics*, V. H. Winston, 258 pp., Washington.
Tingley, M. P., and P. Huybers (2010a), A Bayesian algorithm for reconstructing climate anomalies in space and time. Part 1: Development and applications to paleoclimate reconstruction problems, *J. Clim.*, *23*, 2759–2781, doi:10.1175/2009JCLI3015.1.
Tingley, M. P., and P. Huybers (2010b), A Bayesian algorithm for reconstructing climate anomalies in space and time. Part 2: Comparison with the regularized expectation-maximization algorithm, *J. Clim.*, *23*, 2782–2800, doi:10.1175/2009JCLI3016.1.
Tingley, M. P., and P. Huybers (2013), Recent temperature extremes at high northern latitudes unprecedented in the past 600 years, *Nature*, *496*(7444), 201–205, doi:10.1038/nature11969.
Tingley, M. P., and B. Li (2012), Comments on "reconstructing the nh mean temperature: Can underestimation of trends and variability be avoided?", *J. Clim.*, *25*(9), 3441–3446, doi:10.1175/JCLI-D-11-00005.1.
Tingley, M. P., P. F. Craigmile, M. Haran, B. Li, E. Mannshardt, and B. Rajaratnam (2012), Piecing together the past: Statistical insights into paleoclimatic reconstructions, *Quat. Sci. Rev.*, *35*, 1–22, doi:10.1016/j.quascirev.2012.01.012.
Van Huffel, S., and J. Vandewalle (1991), *The Total Least Squares Problem: Computational Aspects and Analysis*, Frontiers in Applied Mathematics, vol. 9, SIAM, Philadelphia, Pa.
von Storch, H., E. Zorita, J. M. Jones, Y. Dimitriev, F. González-Rouco, and S. F. B. Tett (2004), Reconstructing past climate from noisy data, *Science*, *306*, 679–682, doi:10.1126/science.1096109.
Wallace, J. M., and D. S. Gutzler (1981), Teleconnections in the geopotential height field during the Northern Hemisphere winter, *Mon. Weather Rev.*, *109*, 784–812.
Wang, J., J. Emile-Geay, D. Guillot, J. E. Smerdon, and B. Rajaratnam (2014), Evaluating climate field reconstruction techniques using improved emulations of real-world conditions, *Clim. Past*, *10*(1), 1–19, doi:10.5194/cp-10-1-2014.
Wang, J., J. Emile-Geay, D. Guillot, N. P. McKay, and B. Rajaratnam (2015), Fragility of reconstructed temperature patterns over the Common Era: Implications for model evaluation, *Geophys. Res. Lett.*, *42*, 7162–7170, doi:10.1002/2015GL065265.
Whitaker, J. S., and T. M. Hamill (2002), Ensemble data assimilation without perturbed observations, *Mon. Weather Rev.*, *130*(7), 1913–1924.
Widmann, M., H. Goosse, G. Schrier, R. Schnur, and J. Barkmeijer (2010), Using data assimilation to study extratropical Northern Hemisphere climate over the last millennium, *Clim. Past*, *6*(5), 627–644.
Wikle, C. K., and L. M. Berliner (2007), A Bayesian tutorial for data assimilation, *Phys. D*, *230*(1), 1–16.