JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 112, D12109, doi:10.1029/2006JD008272, 2007

Robustness of proxy-based climate field reconstruction methods

Michael E. Mann,[1] Scott Rutherford,[2] Eugene Wahl,[3] and Caspar Ammann[4]

Received 22 November 2006; revised 30 January 2007; accepted 20 February 2007; published 23 June 2007.

[1] We present results from continued investigations into the fidelity of covariance-based climate field reconstruction (CFR) approaches used in proxy-based climate reconstruction. Our experiments employ synthetic "pseudoproxy" data derived from simulations of forced climate changes over the past millennium. Using networks of these pseudoproxy data, we investigate the sensitivity of CFR performance to signal-to-noise ratios, the noise spectrum, the spatial sampling of pseudoproxy locations, the statistical representation of predictors used, and the diagnostic used to quantify reconstruction skill. Our results reinforce previous conclusions that CFR methods, correctly implemented and applied to suitable networks of proxy data, should yield reliable reconstructions of past climate histories within estimated uncertainties. Our results also demonstrate the deleterious impact of a linear detrending procedure performed recently in certain CFR studies and illustrate flaws in some previously proposed metrics of reconstruction skill.

Citation: Mann, M. E., S. Rutherford, E. Wahl, and C. Ammann (2007), Robustness of proxy-based climate field reconstruction methods, *J. Geophys. Res.*, *112*, D12109, doi:10.1029/2006JD008272.

1. Introduction

[2] There is a substantial recent history in the application of covariance-based climate field reconstruction (CFR) methods to the problem of climatic and paleoclimatic reconstruction. Applications include the infilling of the instrumental surface temperature field [*Reynolds and Smith*, 1994; *Smith et al.*, 1996, 1998; *Rayner et al.*, 1996, 2000, 2003; *Kaplan et al.*, 1997, 1998; *Folland et al.*, 1999, 2000, 2001; *Schneider*, 2001; *Rutherford et al.*, 2003; *Smith and Reynolds*, 2005] and instrumental sea level pressure (SLP) field [*Kaplan et al.*, 2000; *Zhang and Mann*, 2005; *Allan and Ansell*, 2006; *Ansell et al.*, 2006], and reconstruction of paleoclimatic fields of surface temperature, SLP, and continental drought from "proxy" data such as tree rings, corals, ice cores, and historical documentary evidence [e.g., *Fritts et al.*, 1971; *Cook et al.*, 1994; *Mann et al.*, 1998, hereinafter referred to as MBH98; *Mann et al.*, 1999, hereinafter referred to as MBH99; *Luterbacher et al.*, 1999, 2002a, 2002b, 2004, 2006; *Evans et al.*, 2002; *Pauling et al.*, 2003; *Mann and Rutherford*, 2002, hereinafter referred to as MR02; *Xoplaki et al.*, 2005; *Zhang et al.*, 2004; *Rutherford et al.*, 2005; *Mann et al.*, 2005; *Casty et al.*, 2005; *Pauling et al.*, 2006]. Recently developed modifications of CFR include the separate reconstruction of low- and high-frequency components of climate fields [*Rutherford et al.*, 2005, hereinafter referred to as R05; *Smith and Reynolds*, 2005; *Mann et al.*, 2005, hereinafter referred to as M05]. None of the CFR studies mentioned above employed the controversial procedure [see *Wahl et al.*, 2006] recently introduced by *Von Storch* and associates [*Von Storch et al.*, 2004, hereinafter referred to as VS04; *Burger and Cubasch*, 2005, hereinafter referred to as BC05; *Burger et al.*, 2006, hereinafter referred to as BFC06] in which data are linearly detrended prior to calibration. We return to this important point later.

[3] Other statistical methods, such as the simple compositing of multiple proxy series, centered and scaled by the target instrumental series over the modern interval (the so-called composite-plus-scale (CPS) approach), can be used to reconstruct a single time series, such as the Northern Hemisphere (NH) mean temperature series, from proxy climate data [e.g., *Bradley and Jones*, 1993; *Overpeck et al.*, 1997; *Jones et al.*, 1998; *Crowley and Lowery*, 2000; *Esper et al.*, 2002; *Mann and Jones*, 2003; *Cook et al.*, 2004; *Jones and Mann*, 2004; *Moberg et al.*, 2005; *Hegerl et al.*, 2006, 2007] or individual regional temperature series [*Briffa et al.*, 2001]. The fidelity of the CPS approach as a function of proxy signal-to-noise ratio (SNR) was investigated previously by M05.

[4] The CPS approach makes the potentially quite restrictive assumption that all proxy data used are local indicators of the particular climate field (e.g., surface temperature) for which a reconstruction is sought. The CFR approach avoids this potentially restrictive assumption, providing a reconstruction of the entire climate field of interest (e.g., surface temperature field) from a spatially distributed network of climate proxy indicators containing a diverse range of climate signals. The spatial reconstructions can be averaged to yield, e.g., a hemispheric or regional mean temperature series. Additionally and more importantly, the spatial infor-

[1]Department of Meteorology and Earth and Environmental Systems Institute, Pennsylvania State University, University Park, Pennsylvania, USA.
[2]Department of Environmental Science, Roger Williams University, Bristol, Rhode Island, USA.
[3]Department of Environmental Studies, Alfred University, Alfred, New York, USA.
[4]Climate Global Dynamics Division, National Center for Atmospheric Research, Boulder, Colorado, USA.

0148-0227/07/2006JD008272$09.00

mation provided by CFR-based reconstructions can provide better insights into the underlying climate dynamics [e.g., MBH98; *Mann et al.*, 2000; *Delworth and Mann*, 2000; *Shindell et al.*, 2001, 2003, 2004; *Waple et al.*, 2002; *Braganza et al.*, 2003; *Adams et al.*, 2003; *Luterbacher et al.*, 2004; *Xoplaki et al.*, 2005; *Casty et al.*, 2005].

[5] In this study, we follow up on previous investigations by M05 of the fidelity of CFR-based surface temperature reconstructions using simulations of forced climate variability over the past millennium. As in M05, synthetic "pseudoproxy" data derived from the model surface temperature field are used to test the performance of the method in reconstructing the actual model surface temperature history.

[6] Like M05, we make use of a simulation of the National Center for Atmospheric Research (NCAR) Climate System Model (CSM) 1.4 coupled model forced by estimated natural and anthropogenic forcing changes over the 1150 year period A.D. 850–1999 [*Ammann et al.*, 2007]. The CSM1.4 model simulation was spun-up with preindustrial initial conditions, and all important natural and anthropogenic forcings were included. Any potential long-term drift was removed. As discussed in M05, the simulation therefore likely provides a more realistic opportunity for testing CFR approaches than does the uncorrected European Centre Hamburg Ocean Primitive Equation–G (ECHO-G) "Erik" simulation of the past millennium used in several other similar recent studies [VS04; *Zorita and Von Storch*, 2005, hereinafter referred to as ZVS05; BFC06]. The "Erik" simulation was spun-up with modern forcing for a preindustrial initial state, leading to a large, long-term drift [*Osborn et al.*, 2006] which is unlikely to have any counterpart in the true climate history. Furthermore, a key anthropogenic forcing (tropospheric aerosols) was not included, leading to an unrealistically large trend over the 19th–20th centuries [see also *Osborn et al.*, 2006] and exaggerating the change in mean surface temperatures between the calibration period and preceding centuries. A more recent long-term simulation of the ECHO-G model [*Gonzalez-Rouco et al.*, 2006; *Von Storch et al.*, 2006] still suffers from the latter problem. For the above reasons, the CSM simulation likely provides a more realistic opportunity for testing CFR approaches than the GKSS "Erik" simulation. Nonetheless, it is useful to test CFR approaches using both (CSM and GKSS "Erik") simulations to better assess the robustness of methodological performance with respect to differing possible scenarios for the climate of the past millennium.

[7] In this study, we employ tests with pseudoproxy data to examine the sensitivity of reconstruction skill to signal-to-noise ratios, the spatial extent of the pseudoproxy networks, the proxy noise spectrum, and the metrics used to diagnose skill. We also examine the robustness of the results with respect to the particular simulation (CSM 1.4 versus GKSS "Erik") used. While we have focused in this study on annual mean reconstructions of the large-scale surface temperature field using temperature proxies, similar conclusions are likely to hold for seasonally specific or regional reconstructions, for the reconstruction of fields other than surface temperature (e.g., SLP), or for reconstructions that make use of precipitation or mixed temperature/precipitation proxies. Investigating such alternative situations is the subject of additional, ongoing investigations.

[8] In section 2, we describe the revised version of the Regularized Expectation-Maximization ("RegEM") that is employed for CFR in this study. In section 3, we describe the pseudoproxy experiments used to test the performance of this method, and in section 4 we describe the results of these tests. We provide a discussion of the results in section 5, and summarize with our primary conclusions in section 6. The actual and reconstructed surface temperatures, pseudoproxy data, Matlab codes and associated documentation for performing all procedures described are provided at http://www.meteo.psu.edu/~mann/PseudoproxyJGR06. Additional information is available as Auxiliary Material.[1]

2. RegEM CFR Method

2.1. Mathematical Description

[9] The "RegEM" algorithm of *Schneider* [2001] has been used in several CFR applications by Mann and collaborators in recent years [MR02; *Rutherford et al.*, 2003; *Zhang et al.*, 2004; *Zhang and Mann*, 2005; R05; M05]. This algorithm is preferable, in terms of its fundamental statistical properties, to simple truncated principal component analysis (PCA)-based approaches used for CFR in earlier work by MBH98, MBH99 and many other studies. Relationships with these previous approaches are expanded upon in section 2.2.

[10] In RegEM, as in the conventional expectation maximization (EM) algorithm for normal data [e.g., *Little and Rubin*, 1987], a linear regression model relates missing "*m*" and available "*a*" values. Each record $\mathbf{x}$ (consisting of missing and available values) is represented as a row vector within a "data matrix" $\mathbf{X}$ that describes the full multivariate data set. Missing values are related to available values either within that record or contained in other records, through

$$\mathbf{x}_m = \mu_m + (\mathbf{x}_a - \mu_a)\mathbf{B} + \mathbf{e} \quad (1)$$

where $\mathbf{B}$ is a matrix of regression coefficients relating available and missing values within the multivariate data set, and the residual vector $\mathbf{e}$ is a random "error" vector with mean zero and covariance matrix $\mathbf{C}$ to be determined. The rows $\mathbf{x}$ of the data matrix $\mathbf{X}$ can be weighted [e.g., R05] to account for differing area representation of grid box data, or differing error variances.

[11] In each iteration of equation (1), estimates of the mean μ and of the covariance matrix $\sum$ of the data $\mathbf{x}$ are taken as given, and from these, estimates of the matrix of regression coefficients $\mathbf{B}$ and of the residual covariance matrix $\mathbf{C}$ are computed for each record with missing values. In the conventional EM algorithm, the estimate of $\mathbf{B}$ is the conditional maximum likelihood estimate given the estimates of μ and $\sum$. In RegEM, the conditional maximum likelihood estimate of $\mathbf{B}$ is replaced by a regularized estimate, which is necessary in typical CFR applications in which the covariance matrix $\sum$ may be rank-deficient or ill-conditioned. The regression model (1) with estimates of the regression coefficients $\mathbf{B}$ is then used to estimate missing values given the available values, and using the estimated missing values and an estimate of the residual

[1] Auxiliary material data sets are available at ftp://ftp.agu.org/apend/jd/2006jd008272. Other auxiliary material files are in the HTML.

covariance matrix **C**, the estimates of μ and $\sum$ are updated. It should be noted that $\sum$ in this context contains not just the sample covariance matrix of the completed data set, but also, consistent with the model (equation (1)), a contribution due to the residual covariance matrix **C**. The above steps are iterated until convergence. Because the algorithm is iterative in nature, it is nonlinear, and cannot be described in terms of a single linear operator acting upon the data matrix.

[12] In applications of RegEM to proxy-based CFR [MR02; *Rutherford et al.*, 2003, 2005; M05] the rows **x** of the data matrix **X** represent either the standardized proxy (''predictor'') or instrumental (''predictand'') data. The covariances both within and between the proxy and instrumental data series are simultaneously estimated through application of equation (1) to the augmented (proxy + instrumental) data matrix **X**. Statistical reconstruction of the climatic field of interest (e.g., surface temperature) is defined as the estimation of the missing values of the rows of **X** corresponding to the instrumental series, prior to the modern period during which instrumental data are available. By analogy with standard regression approaches to paleoclimate reconstruction [see, e.g., *Jones and Mann*, 2004], one can define a ''calibration'' interval as either a full or partial interval of overlap between the proxy and instrumental data. If a partial interval is used, the remaining interval of overlap can be used to independently compare the reconstruction against the actual withheld instrumental data. Such an interval can thus be defined as a ''verification'' or ''validation'' interval.

[13] An important feature of RegEM in the context of proxy-based CFR is that variance estimates are derived in addition to expected values. Statistical proxy-based climate reconstructions are estimates of expected values of the missing climate data (e.g., surface temperature series) prior to the calibration period, conditioned on the information available in the proxy data. As such, the sample variance of the reconstructed series will necessarily underestimate the true variance since variations about the expected values that are not reflected in the reconstructed time series are a component of the true variance of the (unknown) time series. In RegEM, unlike many other estimates, this contribution is taken into account in estimating the variances in estimated quantities.

[14] Our recent applications favor a hybrid variant of the RegEM approach [see R05; M05] wherein low-frequency (>20 year period) and high-frequency ($\leq$20 year period) variations are processed separately, and then subsequently combined. In practice, whether or not the hybrid procedure is used appears to lead to only very modest differences in skill (see R05 and M05, and also experiments discussed later in section 4.6), but it is necessary in real-world applications where, for example, one wishes to make use of decadally resolved as well as annually resolved records.

2.2. Regularization

[15] As explained by *Schneider* [2001], under normality assumptions, the conventional EM algorithm without regularization converges to the maximum likelihood estimates of the mean values, covariance matrices and missing values, which thus enjoy the optimality properties common to maximum likelihood estimates [*Little and Rubin*, 1987]. In the limit of no regularization, as *Schneider* [2001] further explains, the RegEM algorithm reduces to the conventional EM algorithm and thus enjoys the same optimality properties. While the regularization process introduces a bias in the estimated missing values as the price for a reduced variance (the bias/variance trade-off common to all regularized regression approaches), it is advisable in the potentially ill-posed problems common to CFR. Unlike other current CFR methods, RegEM offers the theoretical advantage that its properties are demonstrably optimal in the limit of no regularization.

[16] There are a number of possible ways to regularize the EM algorithm, including principal component (PC) regression, truncated total least squares regression (TTLS [*Fierro et al.*, 1997]), and ridge regression [*Tikhonov and Arsenin*, 1977; *Hoerl and Kennard*, 1970a, 1970b]. Both ridge regression and TTLS account for observational error in available data (i.e., represent ''errors-in-variables'' approaches), and regularize a total least squares regression under the assumption that relative observational errors are homogeneous [*Golub et al.*, 2000]. In our previous applications to CFR [MR02; *Rutherford et al.*, 2003, 2005; M05], we used the ridge regression procedure as described by *Schneider* [2001]. In this case, regularization is accomplished through use of a ridge parameter h which specifies the degree of inflation $(1 + h^2)$ of the main diagonal of the covariance matrix $\sum$, and therefore determines the degree of smoothing of the estimated missing values. In TTLS, by contrast, regression coefficients are computed in a truncated basis of principal components of the overall covariance matrix $\sum$ and regularization is accomplished through a choice of the truncation parameter K.

[17] The continuous regularization parameter of ridge regression, *Schneider* [2001] speculates, might offer advantages over TTLS, particularly when there is only a small choice of possible truncation parameters. However, we have found that the estimation of optimal ridge parameters is poorly constrained at decadal and longer timescales in our tests with pseudoproxy data, and we have learned that earlier results using ridge regression (e.g., M05) are consequently sensitive to, e.g., the manner in which data are standardized over the calibration period (see discussion below in section 2.3). We have found TTLS to provide more robust results with respect to these considerations. TTLS moreover is considerably more parsimonious in terms of the number of estimated parameters (TTLS requires one truncation parameter per iteration, while ridge regression requires one ridge parameter iteration per record), and thus is far less computationally intensive (typically requiring a factor of ten or more less time for convergence than using ridge regression). For these reasons, we employ TTLS rather than ridge regression in the analyses described in this study.

[18] There are a number of alternative possible objective criteria for choosing the TTLS truncation parameter K. We have found that a conservative choice that works well in practice is to estimate K as corresponding to the number of leading eigenvalues of the calibration period data matrix that lie above the estimated noise continuum. The noise continuum is estimated by a linear fit to the log eigenvalue spectrum. In the low-frequency band, for which there are typically roughly a dozen or less nonzero eigenvalues of the calibration interval data matrix, such a linear fit is not well

constrained. In this case, we found that an even simpler criterion (retaining the first K eigenvalues which resolve 50% of the data variance) works well in practice. While these criteria proved effective and robust throughout our tests, they appeared slightly too conservative at very high signal-to-noise ratios. The investigation of alternative objective criteria for selecting K (e.g., cross validation) represents a worthwhile topic for further investigation.

[19] To further insure regularization of the procedure, the predictand (in the current study, the annual mean ''instrumental'' surface temperature grid box data which are temporally complete over the calibration interval but entirely missing prior that interval) is represented in the data matrix $\mathbf{X}$ by its leading M PC time series, where M is small compared to the total number of nonzero eigenvalues of the calibration period covariance matrix. This step is performed only once, at initiation of the RegEM procedure. M represents the number of distinct modes of variance resolvable from noise, and is objectively determined by the eigenvalue-based noise/signal separation procedure described earlier, applied to the predictor data set over the calibration interval at the initiation of the RegEM procedure. The predictand in the end is then reconstructed through the appropriate eigenvector expansion, using the M reconstructed PC series. In the context of surface temperature reconstructions, the most conservative possible choice $M = 1$ closely approximates the ''index only'' approach discussed in section 2.3, wherein a single quantity (e.g., the hemispheric mean temperatures), rather than a spatial field (e.g., the surface temperature field) is targeted for reconstruction. $M = 1$ in general yields a near optimal hemispheric or global mean reconstruction, but optimal reconstructions of the underlying spatial patterns are generally achieved for a choice $M > 1$ as dictated by the objective criterion discussed above.

[20] There is some resemblance between the procedure described above, and the truncated PCA approach originally used in MBH98. The original MBH98 procedure can be seen as an approximation to the present procedure wherein (1) the iterative Expectation-Maximization procedure is replaced by a single estimation of data covariances between available proxy and instrumental data and (2) simple inverse regression, rather than TTLS, is used to relate the information in the proxy and instrumental data.

2.3. Assumptions

[21] An important assumption implicit in RegEM and other statistical imputation techniques that do not explicitly model the mechanism responsible for missing values is that data are missing at random, which means that the fact that a value is missing does not depend on the missing data values; it does not mean that data must be missing in a spatially or temporally random fashion. The validity of this assumption may face a challenge in CFR where certain series are often selectively missing during earlier periods that are characterized by different mean values from the later periods. The analyses described in the present study (where the ''instrumental data'' are selectively missing during the precalibration interval), as previous experiments with climate model simulation data [*Rutherford et al.*, 2003; M05], show no evidence that this assumption leads to any significant bias in practice.

[22] As in other errors-in-variables methods such as ''total least squares'' (TLS) which has been used in proxy reconstruction studies [*Hegerl et al.*, 2006, 2007], the RegEM method accommodates the existence of errors in both the ''predictors'' (proxy data) and ''predictand'' (instrumental surface temperatures). Indeed, if an ''index only'' approach is taken wherein the predictand is represented by a single index (e.g., the NH mean temperature series) rather then a multivariate (e.g., surface temperature) field, then the RegEM method reduces to a regularized multivariate regression for that index. In this case, the regularized TLS reconstruction approach used by *Hegerl et al.* [2006, 2007] to reconstruct NH mean surface temperature from proxy data can be thought of as a single iteration of a RegEM-like algorithm.

[23] In applications involving the infilling of missing values in instrumental climate fields [e.g., *Schneider*, 2001; *Rutherford et al.*, 2003; *Zhang and Mann*, 2005], it is appropriate to use the original unscaled (e.g., actual surface temperature) series as the records $\mathbf{x}$ in equation (1), as they have common dimensions (e.g., °C) and absolute errors can reasonably be assumed approximately uniform. In paleoclimate applications where the different records $\mathbf{x}$ may represent series with different dimensions (e.g., either an instrumental temperature series in units of °C or a proxy time series with arbitrary units), it is necessary to first standardize (i.e., normalize and center) all records over some common period. When all records have been standardized prior to the analysis without any further weighting, it is implicitly assumed that relative errors are homogenous, i.e., approximately uniform among all records. In applications to proxy-based climate reconstruction, this assumption is unlikely to be strictly valid, as the signal-to-noise ratios for the instrumental and proxy series are different. Explicitly accommodating differing levels of relative error variance in the two constituent data sets, however, makes little difference in practice (Auxiliary Material, section 1).

[24] In M05, all data (pseudoproxy and surface temperature grid box series) were first standardized using their true long-term mean and standard deviations, and all data were centered at each iteration of the RegEM procedure relative to the mean over all data prior to the calibration interval. However, in real world proxy-based reconstructions, the statistics of the surface temperature data themselves are available only over the calibration interval. A fair criticism of the convention adopted by M05 (T. Lee, personal communication, 2006) is that these long-term statistics should thus not be used in standardizing the surface temperature data when testing paleoclimate reconstruction methods. Instead, the standardization of all data (proxy and surface temperature) should, as in the current study, be performed over the calibration interval. When, as in most studies, data are standardized over the calibration period, however, the fidelity of the reconstructions is diminished when employing ridge regression in the RegEM procedure as in M05 (in particular, amplitudes are potentially underestimated; see Auxiliary Material, section 2). However, the revised approach used in the present study, where TTLS is used in place of ridge regression in the RegEM procedure, proves robust with respect to the standardization interval used, and excellent results are achieved standardizing data over the calibration period (Auxiliary Material, section 2).

Implications of these considerations for previous RegEM proxy reconstructions using ridge regression are discussed further in section 2.5.

2.4. RegEM in the Context of Previous Criticisms of CFR

[25] It is worth considering the attributes of the RegEM methodology in the context of previous criticisms that have been published of CFR approaches to paleoclimate reconstruction. BC05 have criticized the truncated PCA-based CFR approach used by MBH98 (and many other studies) for a putative absence of either a "sound mathematical derivation of the model error" or "sophisticated regularization schemes that can keep this [model] error small." These criticisms do not apply to the RegEM method used in more recent work [MR02; *Rutherford et al.*, 2003, 2005; *Zhang et al.*, 2004; M05]. As is clear from the discussion in sections 2.1–2.3, RegEM both employs an objective regularization scheme, and an explicit statistical modeling of errors.

[26] BC05 argue that truncated PCA-based CFR involves a number of potentially subjective procedural choices. One of the "choices" they argue for, detrending the predictand prior to calibration, is simply inappropriate, as discussed in more detail later. The other "choices" argued are irrelevant, however, in the context of RegEM-based CFR. Any motivation for using a subset of PC summaries in representing proxy data networks (the "PCR" choice in BC05) is eliminated in RegEM, because any potential colinearity among predictors is accounted for in the regularization process. The issue of how proxy data networks might first be standardized in the estimation of PCs (the "CNT" choice in BC05; the two conventions considered involve whether one standardizes with respect to the long-term or calibration period statistics of the proxy series prior to PCA) is therefore rendered irrelevant as well. The use of PCs to reduce the predictor set in RegEM in such a case merely acts to eliminate potentially useful information, and is likely to degrade reconstruction skill in general. Nonetheless, PC summaries of proxy networks can be used in RegEM, and whether or not this is done has a minimal influence on the resulting reconstruction, as shown in R05 for the case of the MBH98 proxy data, and as demonstrated in the present study in analyses described in more detail below. The rescaling step (the "RSC" choice in BC05) is not appropriate since regression coefficients are objectively determined through equation (1). Finally, the distinction between inverse and direct regression approaches (the "INV" choice in BC05) is inapplicable since in our applications to paleoclimate reconstruction RegEM represents an errors-in-variables method (albeit one with certain specific variance assumptions; see section 2.2), and can be described neither as pure direct nor pure inverse regression.

[27] As demonstrated below, the RegEM algorithm for proxy-based CFR as implemented in the manner described above, performs remarkably well for a wide range of signal-to-noise ratios, a variety of proxy noise assumptions, a range of proxy network sizes and spatial distributions, differing choices of the modern calibration interval, and using two entirely different climate model simulations. As discussed above, we nonetheless believe that aspects of the algorithm (in particular the relatively simple objective selection rules developed) could be further optimized. We welcome future efforts in this direction by other researchers.

2.5. RegEM Reconstructions With Actual Proxy Data

[28] The considerations discussed in section 2.2 suggest that previous RegEM applications to proxy-based climate reconstruction employing ridge regression for regularization, such as R05, are susceptible to a potential underestimation of reconstruction amplitudes. To investigate this further, we have performed surface temperature reconstructions with RegEM using the same two proxy data sets used in that R05 (the "multiproxy" data set of MBH98, and the gridded "MXD" tree ring latewood density data set of Osborn and coworkers), but incorporating the revised RegEM methodology of this study, which employs TTLS in place of ridge regression. Figure 1 shows the NH annual mean reconstruction using the MBH98 data set, including the "PC/proxy" version of the proxy network in which dense networks of tree ring data are represented in the data matrix $\mathbf{X}$ by their leading PCs (Figure 1a), and the "all proxy" version in which all proxy series are represented individually in the data matrix $\mathbf{X}$ (Figure 1b). In both cases, the amplitude of the resulting NH mean reconstruction is slightly enhanced relative to what is shown in R05, but remains well within the uncertainties of the results shown in R05 and in MBH989. A similar conclusion holds for reconstructions based on the "MXD" tree ring proxy network also used in R05 (see Auxiliary Material, section 3). While the results shown previously in MBH98 and R05 therefore appear relatively robust with respect to methodological considerations, our current analyses (e.g., Auxiliary Material, section 2) show that this need not be true more generally. There is therefore good reason to favor the RegEM implementation described in this study over the previously used implementation in proxy-based CFR.

[29] The similarity between reconstructions based on the "PC/proxy" and "all proxy" versions of the MBH98 network, as an aside, reinforces other recent findings [*Wahl and Ammann*, 2007; *Huybers*, 2005; *Von Storch and Zorita*, 2005] rejecting the claim made elsewhere [*McIntyre and McKitrick*, 2005, hereinafter referred to as MM05] that use of PC summaries to represent proxy data has any significant influence on the MBH98/MBH99 reconstructions.

3. Pseudoproxy Experiments

3.1. Surface Temperature Field

[30] Surface temperature grid box data were available over the interval A.D. 850–1999 interval from the NCAR CSM 1.4 simulation and over the interval A.D. 1000–1989 for the GKSS "Erik" simulation. We regridded the model surface temperature field from both simulations to 5° latitude by longitude grid cells, commensurate with the resolution of gridded observational surface temperature data. We furthermore confined the surface temperature fields to the region over which nearly continuous annual observations are available in the real world from the mid 19th–20th century [see M05]. Application of this criterion to the real world surface temperature data leads to a set of 1312 global temperature land air/sea surface temperature grid boxes which have <30% missing data over the interval 1856–1998, and no single gaps >6 months (Figure 2; see M05 for

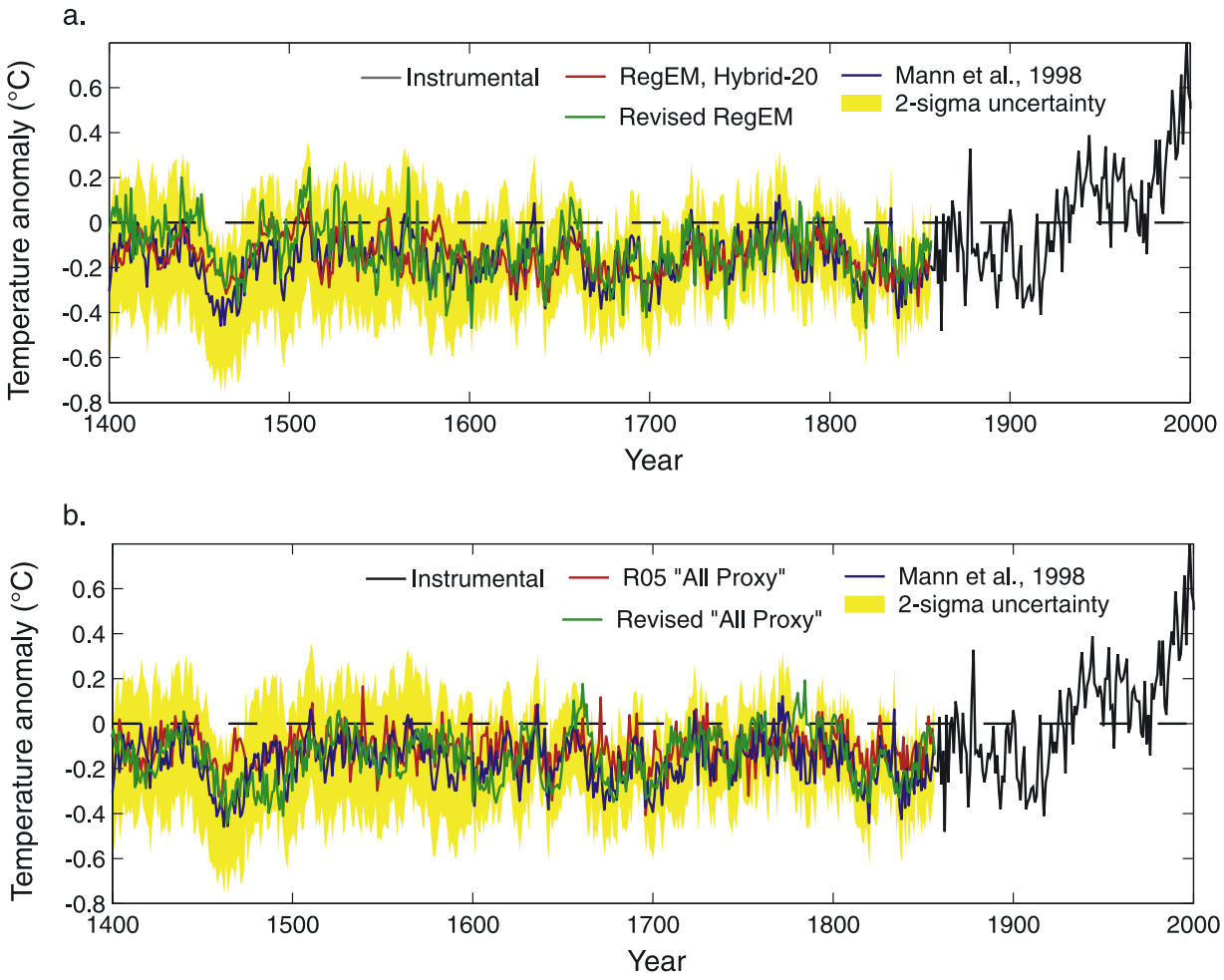

Figure 1. Comparison between the NH annual mean reconstructions of *Rutherford et al.* [2005] with reconstructions that result from using same (MBH) proxy data sets, but incorporating the revised version of the RegEM method discussed in the text. Shading indicated 95% confidence intervals calculated from verification residuals. (a) Comparison of reconstructions using the "PC/proxy" representation of the MBH98 multiproxy network. (b) Comparison of reconstructions using the "all proxy" representation of the MBH98 multiproxy network.

further details). In our experiments here, the model surface temperature field is complete (i.e., there are no temporal gaps) for all 1312 grid boxes used. Temperatures are expressed as anomalies relative to the mean over the 1900–1980 period. The NH mean was defined as the areally weighted mean of all available grid boxes north of the equator, while the Niño3 index was defined as the areally weighted mean over all available grid boxes in the Niño3 region of the tropical Pacific (5°N–5°S, 90–150°W). The difference between the model NH mean calculated over the restricted subdomain spanned by the 1312 grid boxes versus the full model NH mean (i.e., based on averaging over all model grid boxes north of the equator) is relatively small (see Auxiliary Material, section 4). However, use of the restricted grid box region provides a more faithful representation of real-world reconstructions which use the actual available surface temperature records.

3.2. Pseudoproxy Networks

[31] We employed four different spatial networks of pseudoproxies ("A," "B," "C," and "D," see Figure 2). Network A, which we adopt as our standard case, corresponds to the 104 model grid boxes associated with the 104 unique sites of the full MBH98 network of proxy indicators. This network was used in the M05 pseudoproxy experiments. Networks B and C correspond to reduced networks of unique sites used by MBH98 back to A.D. 1400 (18 locations) and by MBH99 back to A.D. 1000 (11 locations), respectively. Network D consists of twice as many (208) grid boxes as network A, including the 104 grid boxes of network A and a set of 104 additional randomly selected grid boxes restricted to the surface temperature domain shown in Figure 2. To simplify the intercomparison of results, we restricted the network B–D experiments to a smaller number of sensitivity analyses than for the standard network A.

[32] As in M05, pseudoproxy time series were formed through summing the annual model grid box temperature series at a given location with an independent realization of noise. The use of independent noise realizations for each location reflects the assumption, in the absence of evidence

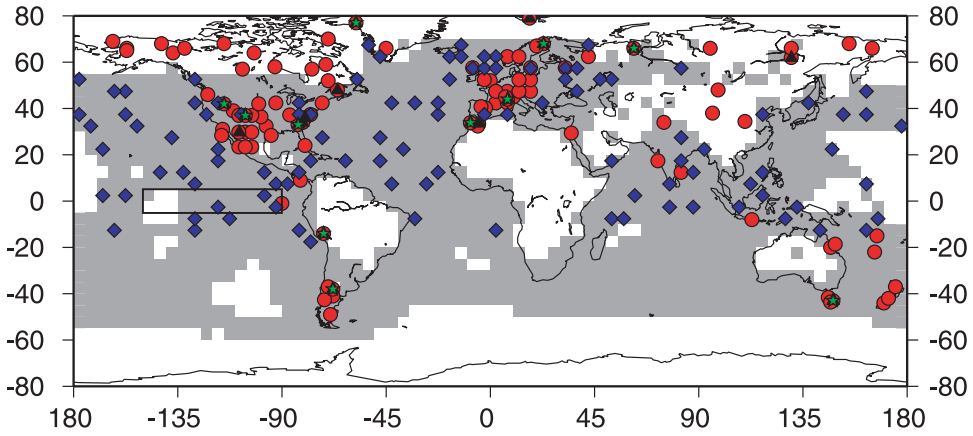

Figure 2. Distribution of data used in study, including model surface temperature field domain used (indicated by gray shading; Niño3 region of eastern tropical Pacific indicated by rectangle). Pseudoproxy locations that correspond with MBH98/99 proxy locations extending back to at 1820 (network A) are shown by circles, those that extend back to A.D. 1400 (network B, 18 locations) are indicated by triangles, and proxies that are available back to A.D. 1000 (network C, 11 locations) are shown by stars. The 208 sites used in pseudoproxy network D correspond to the 104 MBH98/99 sites (circles) and the additional 104 locations indicated by the diamonds.

to the contrary, that the noise processes degrading proxy climate signals (e.g., vital effects in the case of corals, microclimate influences on snow accumulation in the case of ice cores, or forest-level competition dynamics in the case of tree ring measurements) can be assumed to be specific to the proxy record, and not correlated over large spatial scales. An independent set of noise realizations were used in each experiment, unless explicitly noted otherwise.

3.3. Signal Versus Noise Characteristics

[33] Our experiments allowed for various relative noise amplitudes. As in previous work [MR02 and M05] we defined SNR values by the ratio of the amplitudes (in °C) of the grid box temperature "signal" and the added "noise." Experiments were performed for five different values of SNR: 0.25, 0.4 0.5, 1.0 and ∞ (i.e., no added noise). (Note that these SNR values represent broadband (i.e., spectrally averaged) properties. When the spectrum of the underlying climate field is "red" (as with surface temperatures), however, SNR will in general increase with decreasing frequency. This feature is indeed observed for the MBH98 proxy network (see Auxiliary Material, section 5)). Relating SNR to the associated root-mean-square correlation between proxies and their associated local climate signal through $r = \mathrm{SNR}/(1 + \mathrm{SNR}^2)^{1/2}$ gives r = 0.24, 0.37, 0.45, 0.71, and 1.0, respectively, for the five SNR values considered. In the terminology of VS04 and BFC06 who express signal versus noise attributes in terms of the "% noise" (defined as the fraction of the variance in the pseudoproxy series accounted for by the noise component alone) these SNR values correspond to 94%, 86%, 80%, 50%, and 0% noise, respectively. We adopted as our "standard" case SNR = 0.4 (86% noise, r = 0.37) which represents a signal-to-noise ratio than is either roughly equal to or lower than that estimated for actual proxy networks (e.g., the MXD or MBH98 proxy networks; see Auxiliary Material, section 5), making it an appropriately conservative standard for evaluating real-world proxy reconstructions.

[34] Previous studies such as VS04 and M05 have assumed "white" proxy noise, consistent with the assumption that the level of random degradation of climate information recorded by proxies is equal across timescales. However, it is plausible that some proxies, such as certain tree ring data, do possess selective losses of low-frequency variance. Under such conditions, the noise must instead be modeled as a first-order autoregressive "red noise" process with autocorrelation coefficient ρ (see MR02). The ratio of the lowest (i.e., in this case, centennial-scale) and broadband (i.e., frequency-averaged) noise variance is given by the factor $(1 + \rho)/(1 - \rho)$ [e.g., *Gilman et al.*, 1963]. The amplitude ratio is correspondingly given by $\beta = [(1 + \rho)/(1 - \rho)]^{1/2}$. Note that $\beta = 1$ for white noise pseudoproxies ($\rho = 0$) as in VS04 and M05.

[35] Under the assumption of moderate or low signal-to-noise ratios (e.g., lower than about SNR $\approx$ 0.5 or "80% noise"), which holds for the MBH98 proxy network as noted earlier, the value of ρ for the "noise" closely approximates that for the "proxy" (which represents a combination of signal and noise components). We verified this with our pseudoproxy networks. For example, for SNR = 0.4 and imposed noise autocorrelations of ρ = 0.32 and 0.71 we estimated ρ = 0.31 and ρ = 0.65, respectively, from our pseudoproxy networks over the interval A.D. 850–1980. Calculating the average value of ρ for the full network of 112 proxy multiproxy indicators used by MBH98, we determined ρ = 0.29 with standard error ±0.03, indicating ρ = 0.32 to be a conservative (i.e., likely "redder" than in reality) estimate for the MBH98 network.

[36] We investigated the influence of red proxy noise using noise autocorrelations ranging from the approximate estimate for the actual MBH98 network (ρ = 0.32), to the unrealistically high value (ρ = 0.71) recently employed by *Von Storch et al.* [2006]. The corresponding low-frequency noise amplitude inflation factors are $\beta \approx 1.4$ and $\beta \approx 2.4$,

respectively, while the variance inflation factors are $\beta^2 \approx 2.0$ and $\beta^2 \approx 5.8$, respectively. In other words, for the $\rho = 0.32$ value that approximates the actual MBH98 proxy network, the centennial-timescale noise has about twice as much variance as "white noise" proxies with the same overall SNR value, while for the much higher value $\rho = 0.71$ used by *Von Storch et al.* [2006], it has almost six times as much variance.

3.4. Validation and Skill Estimation

[37] For both short (1900–1980) and long (1856–1980) calibration interval choices (corresponding roughly to the calibration intervals used by MBH98 and R05, respectively), the range of temperature variation over the preceding centuries is considerably greater than that observed over the calibration interval in both the NCAR and GKSS simulations. This insures that our analyses provide a rigorous test of the performance of the RegEM CFR method. For the long calibration experiments, we used the entire available precalibration interval (A.D. 850–1855) for statistical validation or "verification." For the short calibration experiments, we alternatively used the long A.D. 850–1855 verification interval, as well as a shorter (A.D. 1856–1899) validation interval.

[38] The "true" long-term skill of the reconstructions can only be diagnosed from the long verification intervals, in which case the skill diagnostics measure the actual long-term goodness-of-fit of the reconstruction. Nonetheless, the more uncertain estimates of skill provided by the short verification intervals, with their greater sampling fluctuations, more realistically reflect verification estimates available in actual proxy reconstruction studies such as MBH98, MBH99 and R05, where the true (instrumental) climate history is only known for the relatively recent past, e.g., since the mid 19th century. For this reason, uncertainties were evaluated as in M05, as the square root of the unresolved verification period temperature variance using the short (1856–1899) validation interval of the short (1900–1980) calibration. We tested the reliability of these uncertainty estimates by performing experiments that were identical in all respects except the particular noise realization used to generate the pseudoproxy network. The resulting reconstructions were found to lie within the originally estimated uncertainties in these cases (Auxiliary Material, section 6).

[39] We evaluated statistical reconstruction skill for all reconstructions using the same three verification skill metrics *RE*, *CE*, and r^2 (the "reduction of error," "coefficient of efficiency," and squared Pearson product-moment correlation coefficient, respectively), as M05. Uncertainties were diagnosed from the variance of the verification residuals as in M05. For reasons discussed in R05 and M05 [see also *Wahl and Ammann*, 2007], *RE* is the preferred measure of resolved variance for diagnosing skill in statistical reconstructions of fields, such as surface temperature which exhibit nonstationary behavior marked by potential changes in mean and variance outside the calibration interval. The alternative *CE* statistic rewards successful prediction of changes in variance but not mean, thus emphasizing high-frequency variability when a short verification interval is used. r^2 rewards neither changes in mean nor variance and is in these respects a flawed metric of reconstruction skill. We nonetheless provide results from r^2 for comparison with the other skill metrics. Expanding on M05, we employed three different alternative diagnostics of statistical skill, using measures of resolved variance in the NH mean, the underlying multivariate spatial field ("mult"), and the Niño3 index, representative of variability associated with the model's approximation to the ENSO phenomenon. For comparison with M05, we diagnosed reconstruction skill and statistical uncertainties for NH (and Niño3) at decadal resolution. For "mult" we diagnosed reconstruction skill at annual resolution.

3.5. Statistical Significance Estimation

[40] Statistical significance of verification skill for single series (NH and Niño3) was determined through Monte Carlo simulations as in M05. This procedure involves the generation of 1000 surrogate random reconstructions with the same lag-one autocorrelation structure, variance and mean as the actual grid box temperature series over the calibration interval. For each noise realization, we project an AR(1) noise process back in time over the validation interval from the first year of the calibration interval. The skill scores resulting for these random reconstructions are tabulated to form a null distribution consistent with AR(1) red noise reconstructions (see Auxiliary Material, section 7, for an example of 10 random AR(1) NH reconstructions compared against the actual NH series using a 1900–1980 calibration and 1856–1899 validation interval). To evaluate full-field verification scores, we employed a spatiotemporal AR(1) red noise model that preserves the actual spatial correlation structure of the surface temperature field. This is accomplished by fitting an AR(1) process to each grid box series over the calibration interval, diagnosing the sequence of innovation terms (that is, the white noise forcing terms), and tabulating the spatial patterns of the innovation term for each year. We then temporally permute the innovations in a spatially coherent manner (by applying the same permutation for all grid boxes in parallel) to generate ensembles of a spatiotemporal AR(1) red noise process that preserves the actual spatial correlation structure within the temperature field.

4. Results

[41] Experiments were performed using the NCAR CSM 1.4 model simulation and the standard pseudoproxy network A for all SNR values (0.25, 0.4, 0.5, 1.0, and ∞) and both short (1900–1980) and long (1856–1980) calibration intervals. Additional experiments were performed for pseudoproxy networks B, C, and D using the standard value SNR = 0.4 (and additionally SNR = ∞ for network D) and "short" calibration interval. A small number of parallel experiments were performed using the GKSS ECHO-G "Erik" simulation employed by VS04 in tests of (1) the influence of red proxy noise (section 4.5) and (2) the detrending data during calibration (section 4.7).

[42] The results of our experiments are summarized below. Validation statistics are provided (Table 1) for all experiments based on each of the skill metrics (*RE*, *CE*, and r^2) and diagnostics (NH, "mult," and Niño3). The parameter choices resulting from application of the selection criteria outlined in section 2.2 are provided in Auxiliary

Table 1. Verification Skill Diagnostics for RegEM CFR Experiments Discussed in Text[a]

						NH Mean			Multivariate			Niño3		
Experiment	Network	SNR	%	ρ	Calibration Period	RE	CE	r^2	RE	CE	r^2	RE	CE	r^2
a	A	∞	0	0	1856–1980	0.96	0.74	0.87	0.51	0.28	0.30	0.90	0.44	0.61
b	A	∞	0	0	1900–1980	0.97	0.81	0.82	0.39	0.09	0.22	0.78	0.17	0.53
	A	∞	0	0	1900–1980[b]	0.94	0.62	0.71	0.27	0.03	0.19	0.69	0.20	0.28
c	D	∞	0	0	1856–1980	0.97	0.80	0.89	0.56	0.30	0.27	0.94	0.60	0.79
d	A	∞	0	0	1856–1980	0.96	0.74	0.87	0.51	0.28	0.30	0.90	0.44	0.61
e	A	1.0	50	0	1856–1980	0.96	0.71	0.86	0.46	0.21	0.23	0.89	0.30	0.53
f	A	0.5	80	0	1856–1980	0.93	0.65	0.83	0.45	0.18	0.19	0.84	0.22	0.47
g	A	0.4	86	0	1856–1980	0.95	0.67	0.74	0.37	0.07	0.14	0.90	0.26	0.55
h	A	0.25	94	0	1856–1980	0.88	0.17	0.34	0.32	−0.02	0.06	0.81	−0.08	0.17
i	A	0.4	86	0	1900–1980	0.95	0.67	0.71	0.36	0.04	0.14	0.80	0.11	0.53
	A	0.4	86	0	1900–1980[b]	0.95	0.66	0.82	0.22	−0.03	0.13	0.74	−0.12	**0.19**
j	B	0.4	86	0	1900–1980	0.86	0.01	0.31	0.26	−0.11	0.04	0.71	−0.43	0.20
	B	0.4	86	0	1900–1980[b]	0.75	−0.79	**0.04**	0.08	−0.21	0.03	0.51	−0.81	**0.00**
k	C	0.4	86	0	1900–1980	0.85	−0.05	0.33	0.25	−0.12	0.04	0.78	−0.04	0.28
	C	0.4	86	0	1900–1980[b]	0.77	−0.66	**0.39**	0.11	−0.18	0.03	0.63	−0.16	**0.09**
l	A	0.4	86	0	1856–1980	0.95	0.67	0.74	0.37	0.07	0.14	0.90	0.26	0.55
m	A	0.4	86	0	1900–1980	0.95	0.67	0.71	0.36	0.04	0.14	0.80	0.11	0.53
	A	0.4	86	0	1900–1980[b]	0.95	0.66	0.82	0.22	−0.03	0.13	0.74	−0.12	**0.19**
n	D	0.4	86	0	1856–1980	0.95	0.66	0.69	0.40	0.15	0.22	0.92	0.45	0.71
o	D	0.4	86	0	1900–1980	0.93	0.58	0.67	0.34	0.07	0.18	0.82	0.28	0.67
	D	0.4	86	0	1900–1980[b]	0.93	0.46	0.64	0.19	−0.07	0.16	0.84	0.37	0.85
p	A	1.0	50	0.32	1856–1980	0.94	0.60	0.85	0.45	0.18	0.22	0.91	0.46	0.63
q	A	1.0	50	0.32	1900–1980	0.92	0.44	0.77	0.28	0.08	0.19	0.83	0.14	0.40
	A	1.0	50	0.32	1900–1980[b]	0.90	0.31	0.61	0.16	−0.11	0.18	0.79	0.15	**0.24**
r	A	0.4	86	0.32	1856–1980	0.94	0.63	0.66	0.31	−0.03	0.15	0.86	0.01	0.45
s	A	0.4	86	0.32	1900–1980	0.93	0.56	0.70	0.35	0.03	0.14	0.82	0.13	0.41
	A	0.4	86	0.32	1900–1980[b]	0.96	0.69	0.81	0.23	−0.02	0.12	0.64	0.16	**0.00**
t (GK)	A	1.0	50	0.32	1856–1980	0.99	0.97	0.98	0.75	0.57	0.64	0.88	0.83	0.91
u (GK)	A	1.0	50	0.32	1900–1980	0.97	0.92	0.96	0.65	0.41	0.54	0.87	0.74	0.90
	A	1.0	50	0.32	1900–1980[b]	0.97	0.64	0.85	0.58	0.13	0.34	0.92	0.62	0.62
v (GK)	A	0.4	86	0.32	1856–1980	0.97	0.91	0.93	0.69	0.48	0.57	0.86	0.75	0.85
w (GK)	A	0.4	86	0.32	1900–1980	0.96	0.90	0.91	0.62	0.36	0.46	0.80	0.55	0.78
	A	0.4	86	0.32	1900–1980[b]	0.96	0.47	0.67	0.55	0.08	0.24	0.82	0.22	**0.30**
x	A	0.4	86	0	1900–1980D	**0.37**	−3.60	0.17	0.13	−0.30	0.05	0.32	−1.03	0.04
	A	0.4	86	0	1900–1980D[b]	0.55	−2.29	**0.09**	−0.02	−0.35	0.06	0.29	−1.01	**0.02**
y (GK)	A	0.4	86	0	1900–1980	0.94	0.93	0.95	0.68	0.46	0.57	0.91	0.80	0.91
	A	0.4	86	0	1900–1980[b]	0.97	0.67	0.79	0.61	0.19	0.41	0.95	0.70	0.91
z (GK)	A	0.4	86	0	1900–1980D	**−0.08**	−1.79	0.37	−0.03	−0.75	0.27	**−0.03**	−0.11	0.44
	A	0.4	86	0	1900–1980D[b]	**0.00**	**−10.8**	0.49	0.03	−0.97	0.34	**0.02**	−0.78	0.64

[a]Unless otherwise indicated, the NCAR CSM 1.4 simulation was used. Unless otherwise indicated, an 850–1855 validation interval has been used. Indicated for each experiment are network, signal-to-noise (SNR) amplitude ratio (associated "% noise" variance also indicated), and noise characteristic (white with $\rho = 0.0$ or red with $\rho = 0.32$). Results using three different skill metrics (NH mean, full multivariate field, and Niño3 region) are provided. Statistical significance from Monte Carlo simulations exceeds $p = 0.05$ level unless otherwise indicated (italics indicate significant at $0.10 > p > 0.05$; boldface indicates not significant at $p = 0.10$ level). Note that certain lines are repeated for organizational clarity (e.g., a and d; i and m, and g and l). The "multivariate" scores represent sums over all Northern Hemisphere grid box series shown in Figure 2. Scores based on summing over both Southern and Northern hemisphere grid boxes and separate low- and high-frequency SNR characteristics of the pseudoproxy networks are provided in the Auxiliary Material (sections 15 and 16, respectively). "D" indicates that predictand (surface temperature field) was linearly detrended prior to calibration. "GK" indicates that GKSS "Erik" simulation was used.

[b]An 1856–1899 verification interval was used in indicated experiment.

Material (section 8). The NH mean reconstructions are shown for each experiment. The Niño3 reconstructions are shown as Auxiliary Material (section 9).

4.1. Perfect Proxies

[43] We examine first the results of pseudoproxy experiments using SNR = ∞ (i.e., "perfect" pseudoproxies, with no added noise; Figure 3a). In this case, the imperfect nature of the reconstruction results purely from the incomplete spatial sampling of the field. In each of the scenarios explored, the reconstructions of NH mean temperature over the reconstructed (A.D. 850–1855) interval closely follow the true histories, and the actual series lie within the estimated uncertainties of the reconstructions. Note that even with "perfect proxies" there is tendency for greater underestimation of the short-term cooling response to explosive volcanic forcing events. As discussed by M05, this appears to arise from the lack of analogs, e.g., the very large explosive eruptions of 1258, 1453, 1601, and 1815, over the relatively volcanically quiescent calibration interval.

[44] With a network of 104 "MBH98" location pseudoproxies (network A), the CFR reconstruction resolves a substantial fraction of variance in surface temperature (Table 1, experiments a–c). *RE*, our primary metric of resolved variance, indicates that 96% (*RE* = 0.96) and 97% (*RE* = 0.97) of the variance in the true NH mean series is resolved depending on whether a short (1900–1980) or long (1856–1980) calibration period was used,

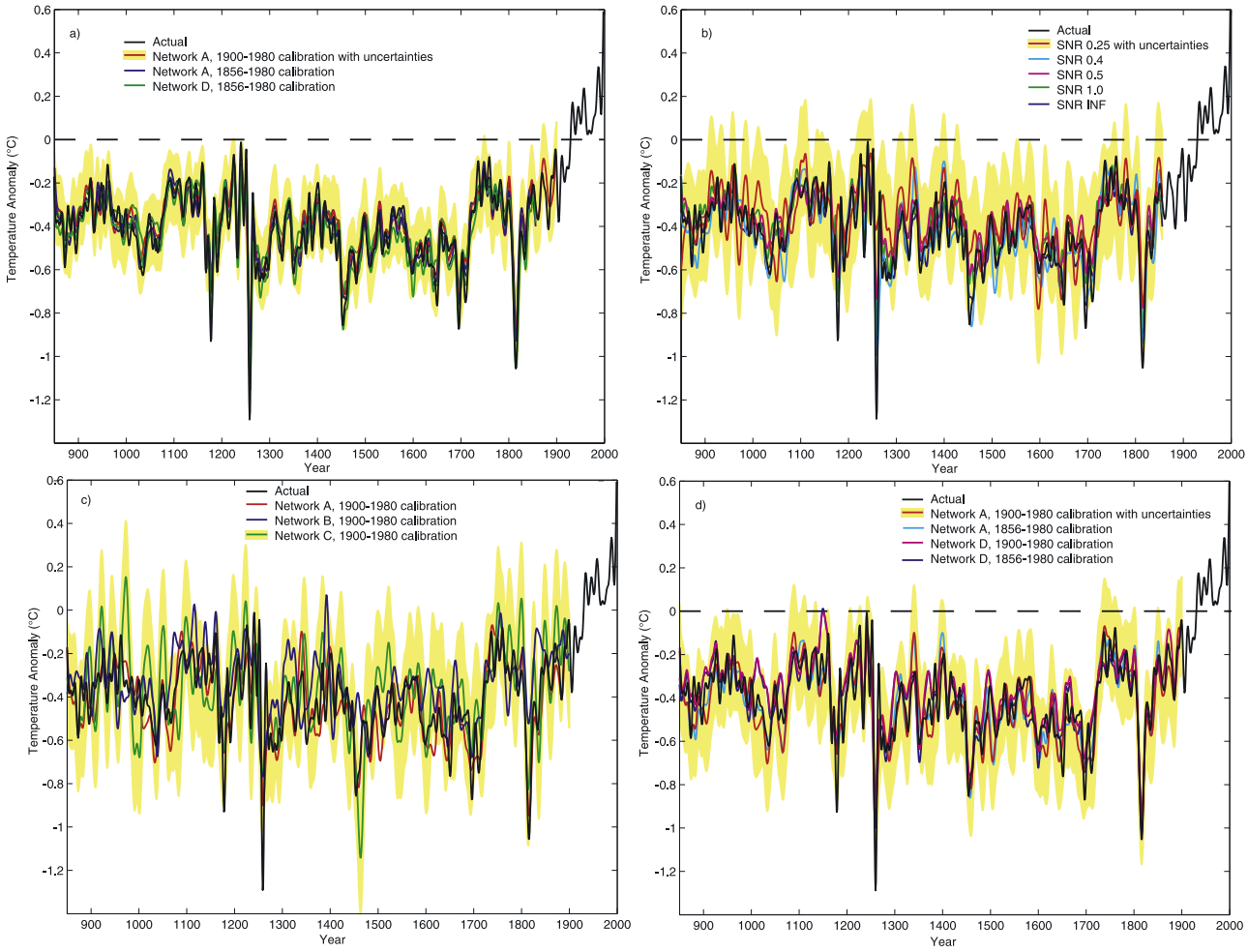

Figure 3. Comparisons of RegEM NH reconstructions (NCAR CSM 1.4 simulation). Here as in all similar plots below, anomalies are expressed relative to 1900–1980 mean (denoted by horizontal line) series have been decadally smoothed using a low-pass filter with cutoff frequency f = 0.1 cycle/year (see M05 for details), shading is used to indicate 95% confidence interval for decadally smoothed series based on short validation period residuals, and the actual model NH mean series (black) is shown for comparison. (a) Comparison of reconstructions using an infinite signal-to-noise ratio (no noise; Table 1, experiments a–c) with short and long calibration periods, and both the "MBH98" network of 104 sites (network A) and a network of double that size (208 random sites, network D). Uncertainties diagnosed from experiment b. (b) Comparison of reconstructions using long (1856–1980) calibration period and varying SNR values (Table 1, experiments d–h). Uncertainties diagnosed from experiment b. (c) Comparison of reconstructions using SNR = 0.4, short calibration period, with the full 1820 network (network A) and the sparser A.D. 1400 (network B) and A.D. 1000 (network C) networks of MBH98 (Table 1, experiments i–k). Uncertainties diagnosed from experiment k. (d) Comparison of reconstructions using SNR = 0.4 and both short (1900–1980) and long (1856–1980) calibration periods, using both the MBH98 network of 104 sites (network A), and a network of double that size (208 random sites, network D) (Table 1, experiments l–o). Uncertainties diagnosed from experiment m.

respectively. Slightly lower scores are indicated using the alternative metrics *CE* (0.74 and 0.81, respectively) and r^2 (0.87 and 0.82, respectively). A more modest fraction of variance is resolved in the Niño3 series and the full annual surface temperature field, and the level of resolved variance is significantly more dependent on the length of the calibration interval in both cases. For Niño3, we have *RE* = 0.90, *CE* = 0.44, and r^2 = 0.61 for the long calibration interval, but *RE* = 0.78, *CE* = 0.17, and r^2 = 0.53 for the short calibration. For "mult," we have *RE* = 0.51, *CE* = 0.28, and r^2 = 0.30 for the long calibration interval, but *RE* = 0.39, *CE* = 0.09, and r^2 = 0.22 for the short calibration. These results suggest, as further demonstrated in other experiments described below that, while hemispheric mean estimates seem relatively insensitive to the calibration interval length in RegEM, use of a longer calibration interval (e.g., 1856–1980) appears to provide significant improvement in reconstruction skill at regional scales. This finding is consistent with the observation that a larger number of degrees of freedom are generally retained in the regulariza-

tion process in the long calibration experiments than in parallel short calibration experiments (see Auxiliary Material, section 8).

[45] It is noteworthy that while doubling the number of pseudoproxies to 208 (i.e., network D experiment) leads to only a very minor (1%) increase in the NH *RE* score, it leads to more pronounced increases in reconstruction skill at regional scales, with increases of 5% and 4% in the "mult" and Niño3 *RE* scores, respectively. Note finally from experiment b, that the skill estimates derived from the short (A.D. 1856–1899) validation interval tend to underestimate the true long-term skill as measured by the long validation interval (A.D. 850–1855). This proves to be a more general pattern in the additional experiments discussed below, indicating that skill estimates (and uncertainties) diagnosed from short validation intervals are very likely to be conservative.

4.2. Impact of Varying SNR

[46] We next consider the results of experiments using the standard network (A) and long (1856–1980) calibration period, but varying the SNR level. For all SNR values (including the SNR = 0.25 or "94% noise," which is almost certainly lower than in actual proxy networks such as those used by MBH98 and R05), we find that the RegEM CFR method faithfully reconstructs the true long-term model NH history (Figure 3b), with the true NH series lying well within the estimated uncertainties of the reconstructed NH series, and with no evidence of any systematic bias in the reconstructed long-term histories. Even for the lowest value SNR = 0.25, a large fraction (88%) of the true long-term NH mean variance is resolved (Table 1, experiments d–h), and a substantial 32% of the total annual variance is resolved in the full surface temperature field. Long-term *CE* scores are, as expected, substantially lower than the corresponding *RE* scores in all experiments, but they remain statistically significant in all cases, and positive in all cases except for the "mult" and Niño3 scores for the SNR = 0.25 experiment.

4.3. Impact of Increasing Proxy Sparseness Back in Time

[47] Using the standard SNR = 0.4 and a short (1900–1980) calibration period (as in MBH98), we next explore the impact of decreases in the size of the pseudoproxy network back in time (as in the MBH98 reconstruction). We compare results for the "full" MBH98 network (A) with 104 unique locations, to the sparser "A.D. 1400" network B with 18 unique locations and the even sparser "A.D. 1000" network C with 11 unique locations. In each case, the main long-term features of the NH mean series (Figure 3c) are captured by the reconstructions. Even for the sparsest network (C), a large (85%) fraction of the true long-term variance is resolved by the NH mean series. Statistically significant skill scores are evident in all experiments, for all metrics and diagnostics, with the exception of r^2 in the short validation experiments (Table 1, experiments i–k). The shortcomings of the r^2 statistic in this context are discussed further in section 5 below.

4.4. Impacts of Calibration Interval Length

[48] Using again the standard SNR = 0.4 value we next explore the impact of varying the length (short 1900–1980 versus long 1856–1980) of the calibration interval. We use the same noise realizations for both the long and short calibration experiments, to insure that differences in skill are associated purely with varying the length of the calibration interval. Experiments are performed for both proxy network A with its 104 predictors and network D with its 208 predictors.

[49] Relatively little sensitivity to the calibration interval is observed for the NH series, and the main features of the long-term series are resolved well for either calibration period for both networks tested (Figure 3d). Long-term validation statistics for all three skill metrics are nearly identical for both short and long calibration intervals for both networks, for the NH series. Significantly more skillful results are obtained for "mult" and Niño3, however using a longer calibration interval (Table 1, experiments l–o). This is especially true using the expanded proxy network D for "mult," where scores improve from $RE = 0.34$ to $RE = 0.40$ (and from $CE = 0.07$ to $CE = 0.15$). Accordingly, as in the "perfect proxy case" SNR = ∞ case, use of the long calibration period leads to a visually clear improvement in the fidelity of the Niño3 reconstructions (see Auxiliary Material, section 9).

4.5. Impact of "Redness" of Proxy Noise

[50] We examined the impact of "red" pseudoproxy noise, modeling the proxy noise series as an AR(1) process using a noise autocorrelation level ($\rho = 0.32$) consistent with estimates for the actual MBH98 proxy network (see section 3.3). Experiments were performed with both the NCAR CSM 1.4 and GKSS "Erik" simulations, using the standard pseudoproxy network A, two different (SNR = 1.0 and SNR = 0.4) noise levels, and both long and short calibration periods. In each case, the long-term NH history was found to be skillfully resolved, with the reconstructed NH series well within the estimated uncertainties of the true NH series for both simulations, both SNR values, and both calibration interval choices (Figure 4). Validation results (Table 1, experiments p–s, NCAR; experiments t–w, GKSS) indicate modest decreases in skill relative to the parallel white proxy noise experiments (compare, e.g., experiments s and i in Table 1). An additional test of the "sparse" pseudoproxy network C yielded similar conclusions (Auxiliary Material, section 10). We found no evidence to support the claim by ZVS05 that realistic "red" pseudoproxy noise leads to any substantial underestimate of low-frequency variability even for the short 1900–1980 calibration interval, using the RegEM approach employed here. Similar results were obtained using higher noise autocorrelation values. Only for the unrealistically high value $\rho = 0.71$ used by *Von Storch et al.* [2006], do we find any notable degradation of reconstruction skill, and here only for the lower SNR = 0.4 value, the shorter (1900–1980) calibration interval, and only one of the two simulations (the NCAR simulation; see Auxiliary Material, section 11). The fact that CFR performance appears largely independent of the proxy noise spectrum in our experiments is consistent with the observation that the signal and noise in question are readily distinguishable by their distinct spatial characteristics, regardless of spectral attributes. The underlying climate signals exhibit large-scale coherence, while the proxy noise realization, as discussed earlier, can,

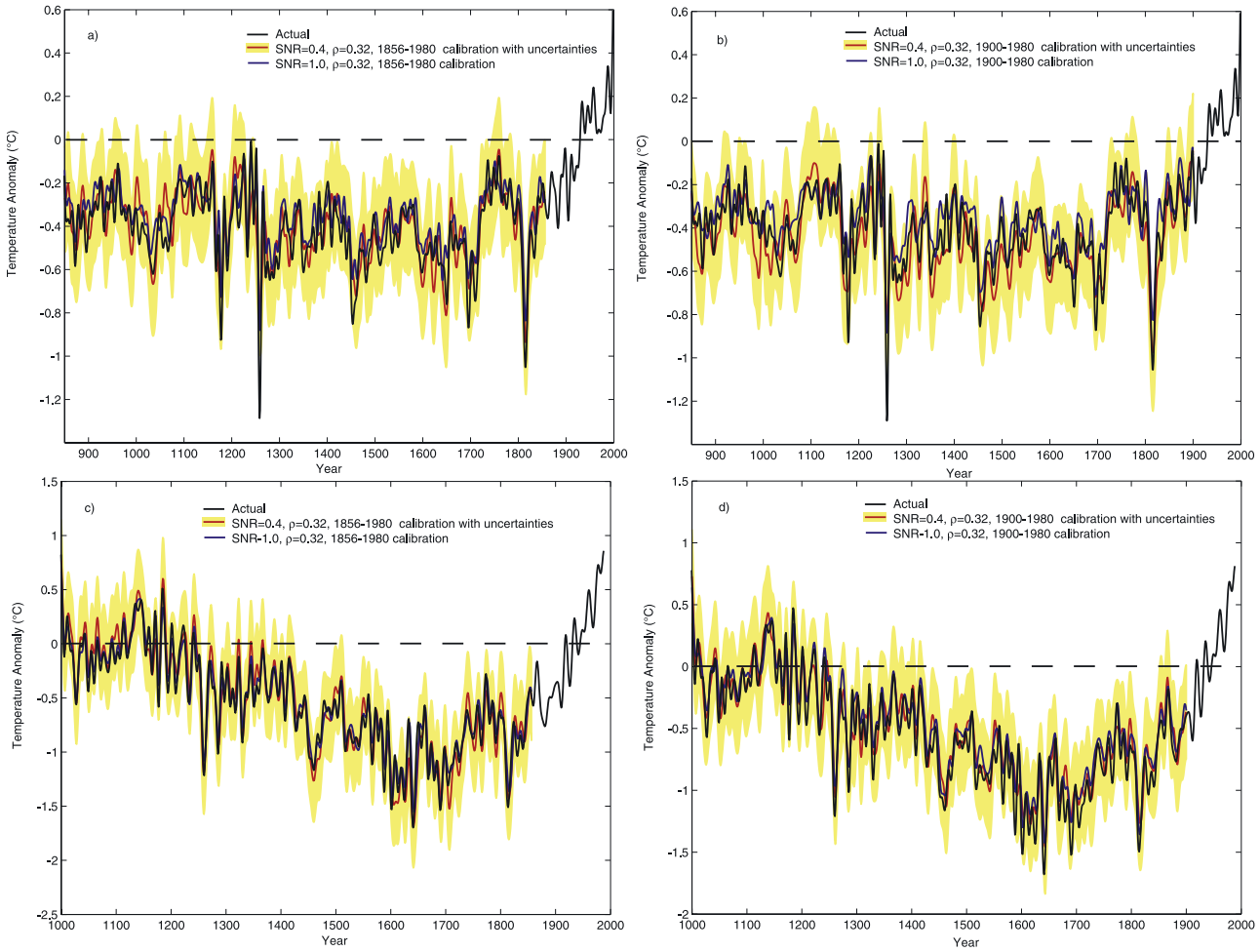

Figure 4. Comparison of NH mean reconstructions using both NCAR CSM 1.4 and GKSS "Erik" simulations, based on network A, "red" proxy noise with $\rho = 0.32$, and two different SNR levels (0.4 and 1.0) (Table 1, experiments p–s for NCAR; experiments t–w for GKSS). (a) NCAR long (1856–1980) calibration, (b) NCAR short (1900–1980) calibration, (c) GKSS long (1856–1980) calibration, and (d) GKSS short (1900–1980) calibration. Uncertainties diagnosed from experiment s for NCAR and experiment w for GKSS.

as is implicit in our experiments, be assumed to be specific to that proxy.

4.6. Additional Methodological Considerations

[51] A number of additional sensitivity analyses were performed using the NCAR CSM simulation (see Auxiliary Material, section 12, for results and analysis details), based on the standard case SNR = 0.4. These analyses included (1) use of the nonhybrid version of the RegEM procedure. The nonhybrid reconstruction is found to yield a skillful long-term reconstruction that lies within estimated uncertainties, but the skill estimates are slightly lower than for the hybrid version of the procedure. This is more pronounced using "red" proxy noise.

[52] We then tested the possible real-world complication wherein proxy quality is variable within a given set of proxy data by (2) allowing the signal-to-noise ratio to vary over the pseudoproxy data set (from SNR = 0.1 to 0.7) for a given (SNR = 0.4) average noise level. In this case, it is worth noting that many pseudoproxies are nearly pure (e.g., 98–99% by variance) noise. We also investigated (3) the sensitivity to area weighting of temperature grid box data, testing both variance and amplitude-based area weighting conventions, alternatively weighting instrumental records in equation (1) by $\cos\varphi^{1/2}$ and $\cos\varphi$, respectively, where φ is the central latitude of the grid box). We further investigated (4) the impact of using an "index only" approach (see section 2.3) where only the NH mean temperature series, rather than the entire spatial surface temperature field, is targeted for reconstruction.

[53] Complementary to the "red" proxy noise analyses of section 4.5, we additionally investigated the impact of (5) "blue" proxy noise (using an AR(1) noise model with $\rho = -0.32$) wherein the pseudoproxies have selectively greater noise amplitude at increasingly short, rather than increasingly long, timescales (as might be the case due, e.g., to biological persistence that in some cases suppresses the sensitivity of tree ring indicators to high-frequency climate forcing). In each of the above tests (2–5), results qualitatively similar to those shown in the main manuscript were

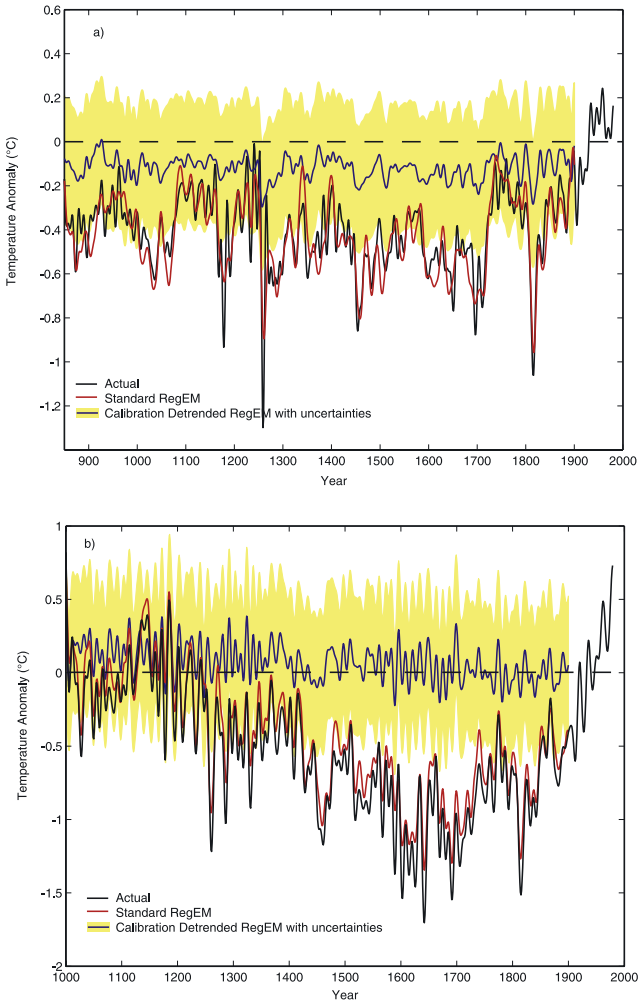

Figure 5. Comparison of NH mean reconstructions using both NCAR CSM 1.4 and GKSS "Erik" simulations, SNR = 0.4 and short (1900–1980) calibration interval, contrasting standard RegEM procedure (Table 1, experiments i and y for NCAR and GKSS, respectively) and "detrended" calibration procedure discussed in text (Table 1, experiments x and z for NCAR and GKSS, respectively). (a) NCAR (uncertainties for detrended calibration diagnosed from experiment x) and (b) GKSS (uncertainties for detrended calibration diagnosed from experiment z).

obtained, suggesting that the RegEM methodology used in the study is robust with respect to a number of potential methodological considerations.

[54] We investigated (6) the impact of using subsets of PC summaries in representing proxy networks rather than the individual proxy records themselves. We considered both (calibration period and long-term) possible standardization conventions discussed in section 2.3. These experiments employed network D with its larger (208) number of pseudoproxy series, retaining as predictors the leading PCs of the pseudoproxy network identified as statistically significant on the basis of the procedure described in section 2.4. These analyses show that use of PCA to reduce the dimensionality of the predictor set yields a similar reconstruction to using all predictors individually, regardless of which of the two standardization conventions are used. Moreover, while use of PC summaries is observed to lead to similar levels of the skill in the NH mean reconstructions, it leads to a significant decrease in the level of regional reconstruction skill, as measured by the "mult" and Niño3 skill diagnostics. This finding is expected for reasons discussed previously (section 2.6). The primary motivation for using a PC subset representation of the predictor networks is dealing with potential colinearity of predictors. However, since the regularization process in RegEM method already accounts for colinearity, the use of a subset of PC summaries simply amounts to throwing away potential useful information that is available in the full data.

[55] Finally, we investigated two additional real-world considerations. We examined (7) the impact of significantly shortening the reconstruction interval. We found that if the reconstruction is performed back to a starting point that is only slightly earlier than the beginning of the calibration interval (e.g., back to A.D. 1800 using a 1900–1980 calibration), the fidelity of the low-frequency component of the reconstruction is inferior in comparison with reconstructions that extend back at least to A.D. 1600. We attribute this finding to the fact that RegEM can make use of considerably greater low-frequency information regarding the covariance structure between predictors if the calibration interval extends back several centuries. While this factor must be weighted against the decreased availability of proxy data further back in time one encounters with real world proxy reconstructions, it suggests that the low-frequency component of reconstructions using less than a few centuries of data are likely to be less reliable than those which make use of a longer reconstruction interval. We also found that (8) "late" short validation experiments (e.g., experiments in which the calibration interval extends from 1856–1936, and the reconstruction over the subsequent interval 1937–1999 is used for validation) yielded inferior results relative to our standard "early" short validation experiments (in which a calibration interval of 1900–1980 is used, and the reconstruction over the prior interval A.D. 1856–1899 is used for validation).

4.7. Impact of Detrending Data Prior to Calibration

[56] In our final series of experiments, we tested the impact of a procedure similar to that used by VS04, BC05, and BFC06 wherein the predictand (i.e., the surface temperature field) is linearly detrended over the calibration interval prior to application of the CFR method (similar results are obtained if both predictand and predictors, i.e., the surface temperature field and pseudoproxies, are both detrended over the calibration interval; see Auxiliary Material, section 13). The experiments were performed for pseudoproxy network A using both the NCAR CSM1.4 and GKSS "Erik" simulations, the standard case SNR = 0.4, and short (1900–1980) calibration interval. While application of the standard RegEM procedure (experiment i, NCAR; experiment y, GKSS) yielded faithful reconstructions for both simulations, use of the detrended calibration procedure fails to capture essentially any of the true low-frequency variability for the GKSS simulation and very little for the NCAR simulation (Figure 5) leading to failure of statistical validation (Table 1, experiment x, NCAR;

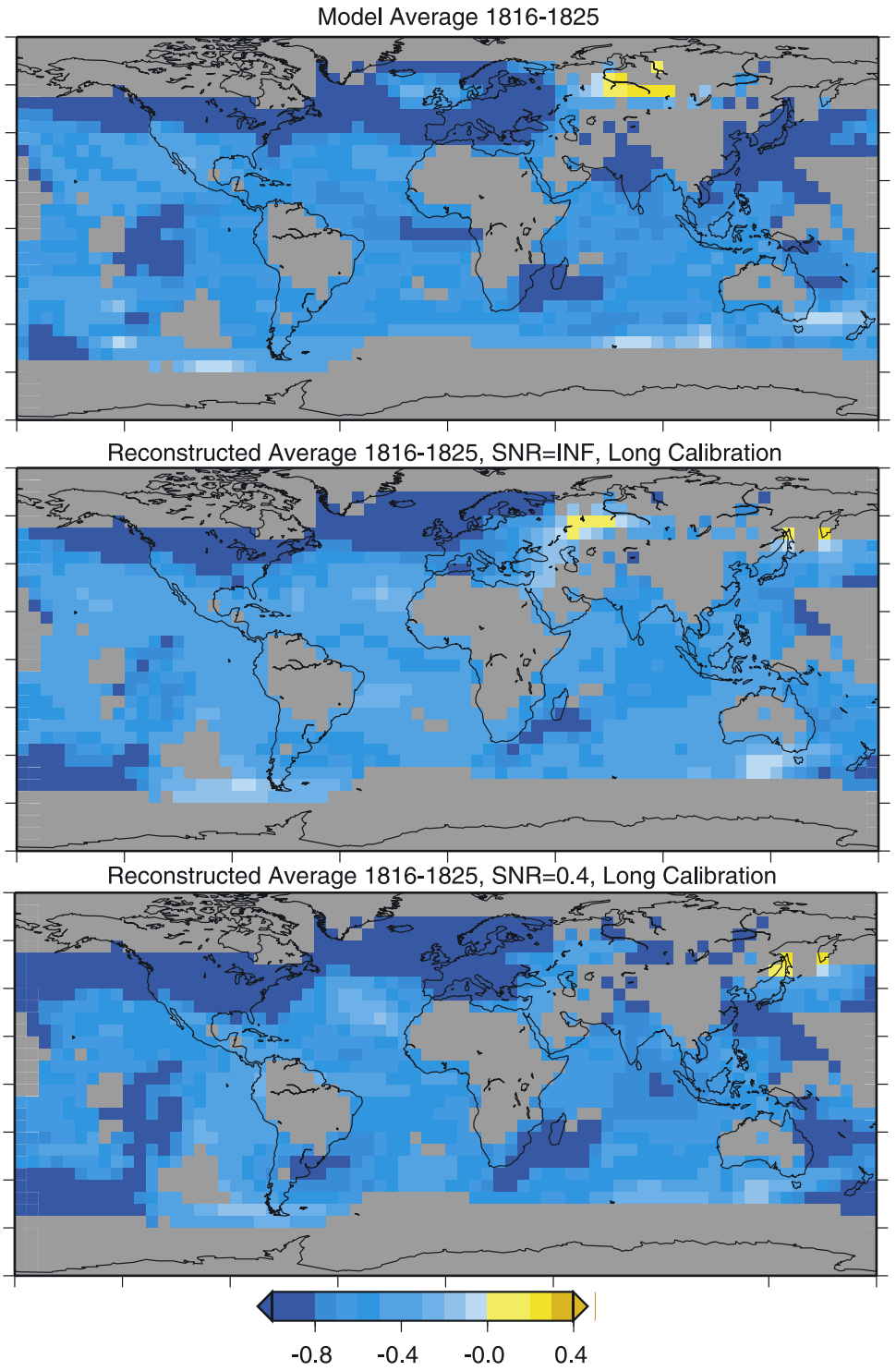

Figure 6. Comparison of actual and reconstructed spatial surface temperature anomaly pattern for decade (1816–1825) following 1815 Tambora eruption (NCAR CSM 1.4 simulation). Shown are results based on long (1856–1980) calibration with pseudoproxy network A and both SNR = ∞ and SNR = 0.4 (i.e., experiments d and g).

experiments y and z, GKSS), with certain exceptions discussed further below in section 5.

4.8. Spatial Patterns

[57] Finally, we examine the spatial patterns of temperature over selected time intervals. The most distinctive decadal event in the NH mean series is the prolonged cooling following the Tambora eruption of 1815 and a sequence of subsequent smaller eruptions. We thus focus on the spatial pattern of this event in both the actual model surface temperature field and in the pseudoproxy reconstructions thereof (Figure 6), based on the standard pseudoproxy

network A and long calibration interval. The 1816–1825 decadal mean temperature pattern (expressed as anomalies relative to the 1900–1980 mean) shows widespread cooling as expected, with a roughly 0.5°C cooling over the tropics but closer to a full degree C cooling over large parts of the extratropical Northern Hemisphere and western Pacific. By contrast, a moderate (<0.5°C) warm anomaly is observed over north central Eurasia. The reconstructions for both perfect proxies (SNR = ∞; experiment d) and our standard proxy noise case (SNR = 0.4; experiment g) reconstruct the main spatial features, including the overall global-scale cooling, and the enhanced cooling in large parts of the extratropical Northern Hemisphere and the western Pacific. Certain smaller-scale details (e.g., the north central Eurasian warm anomaly) are not resolved for the noisy (SNR = 0.4) pseudoproxy case, but are resolved for the noise-free (SNR = ∞) case. The level of unresolved spatial detail is qualitatively consistent with the verification estimates of resolved multivariate variance (Table 1): 51% and 37% for SNR = ∞ and 0.4, respectively, based on the long-term *RE* validation scores. The corresponding spatial pattern of the *RE* statistic (Auxiliary Material, section 14) indicates broad skill in both cases across most of the domain, with the exception of certain regions in the southern ocean.

5. Discussion

[58] The RegEM results presented in this study directly address a number of recently published criticisms of CFR methods and results. Our present results, for example, refute the claims made in certain previous studies that CFR methods intrinsically underestimate low-frequency variability [VS04; ZVS05] or yield nonrobust results [BC05, and BFC06]. As the RegEM approach described in this study is governed by the objective methodology detailed in section 2, it moreover cannot be considered subject to ad hoc and subjective procedural choices such as those used in BC05 and BFC06. It is additionally noteworthy that nearly identical results to those of the MBH98 proxy-based CFR study are reproduced here by applying the RegEM method to the same proxy data set, whether proxy data are used individually, or represented by PC summaries. This result (as well as the separate study by *Wahl and Ammann* [2007]) thus refutes the previously made claim by MM05 that the features of the MBH98 reconstruction are somehow an artifact arising from the use of PC summaries to represent proxy data.

[59] Our findings furthermore support criticisms [e.g., *Wahl et al.*, 2006] of certain recent temperature reconstruction studies such as VS04, BC05, and BFC06 that employed a controversial procedure in which the predictand and/or predictors were linearly detrended prior to calibration. *Wahl et al.* [2006] argue that such a procedure inappropriately removes the primary pattern of coherent large-scale temperature variation from the data, hindering the reconstruction of long-term trends. In the present study, we have corroborated this assertion, demonstrating that application of this procedure produces reconstructions which underestimate long-term trends and fail standard verification metrics where correct implementation of the RegEM CFR procedure readily yields skillful reconstructions.

[60] The experiments performed in this study also provide insights into the relative merit of alternative reconstruction skill diagnostics. The standard verification skill diagnostic *RE* evaluated over the extended (A.D. 850–1855) validation interval affords an accurate measure of the long-term fidelity of the synthetic reconstructions produced. Use of these "long validation" *RE* scores indicates skill (i.e., a level of agreement with the actual series that is statistically significant relative to the null hypothesis of red noise) for all experiments which correctly implement the RegEM method (Table 1, experiments a–w and y). These metrics, moreover, demonstrate as discussed above a lack of skill in those experiments which incorrectly implement the RegEM method, through use of the detrended calibration procedure of VS04, BC05, and BFC06 (Table 1, experiments x and z). These statistical inferences simply confirm the conclusions from visual inspection, that all experiments using the correct RegEM procedure (Figures 3 and 4) produce a skillful reconstruction (i.e., one that agrees with the true series within estimated uncertainties), while those using the detrended calibration procedure (Figure 5) do not. In the latter case (that is, experiments x and z), however, the r^2 metric evaluated over the extended interval fails to reject the reconstructions because of its focus on the relative tracking of the two series at the highest frequency. This is just one example of the errors in statistical inference, discussed further below, that result from using r^2 as a diagnostic of statistical skill in cases where means and variances change over time.

[61] Such extended validation intervals are unfortunately absent in the real world, as widespread instrumental climate data do not extend further back than the mid 19th century. The results from short validation period (e.g., 1856–1899 using a 1900–1980 calibration) are thus more appropriate for gauging the statistical sampling properties of verification scores calculated in actual proxy-reconstruction studies where short validation intervals are used to validate long-term reconstructions. Comparisons of the approximate (short validation) and true (long-term validation) skill metrics for the same reconstruction (e.g., experiments b, i, j, k, o, q, s, u, w, x, y, and z in Table 1) therefore provide insights into the reliability of the various skill metrics using the short available real-world validation intervals. These comparisons demonstrate a general tendency for the short validation tests to yield lower skill scores, even when the long-term fidelity of the reconstruction is excellent both visually and as inferred from the long-term validation scores. These lowered scores result simply from the sampling variations associated with a short (i.e., less than 50 year) validation interval, and nonetheless are statistically significant for all skillful reconstructions (i.e., experiments a–w and y in Table 1) with the notable exception of r^2 scores.

[62] Focusing on the short validation skill scores for the NH series, both *RE* and *CE* are seen to perform well from the point of view of both type I and type II errors of statistical inference. In all experiments using the correctly implemented RegEM method and yielding skillful reconstructions (i.e., experiments a–w and y in Table 1), the short validation *RE* and *CE* scores correctly identify the reconstruction as skillful (i.e., passing at the $\alpha = 0.05$ significance level). This outcome contrasts with the results

for experiments that use the inappropriate procedure of detrending prior to calibration (experiments x and z in Table 1). The short validation *RE* and *CE* scores correctly reject the reconstruction for GKSS simulation experiment z, which captures essentially none of the low-frequency variability and consequently, none of the precalibration interval cooling. The reconstruction from the corresponding NCAR simulation experiment x, does capture some of the low-frequency variability, including some of the 19th century cooling prior to the calibration interval (this feature results from the greater residual low-frequency variability relative to the linear trend during the 20th century calibration period for the NCAR simulation, which provides some limited low-frequency information in the calibration process, see Figure 5). The short validation experiments consequently yield statistically significant *RE* and *CE* scores in this case, though the unusually large negative value $CE = -2.29$ nonetheless implies a poor reconstruction.

[63] The generally reliable performance of *RE* and *CE* as skill diagnostics using short validation periods when inappropriate detrending is avoided contrasts sharply with the clear failure of r^2 in this context. Use of the short validation r^2 produces examples indicating unacceptable type II errors, incorrectly discarding (i.e., not rejecting the null hypothesis of no skill at even the $\alpha = 0.1$ level) the skillful NH mean reconstructions produced in experiments j and k. In the former case, the short validation r^2 score is nearly zero ($r^2 = 0.04$) despite the fact that the reconstruction clearly captures much of the true low-frequency variability (see Figure 3c). The performance of r^2 is similarly unacceptable from the perspective of type I errors, incorrectly accepting (i.e., passing at the $\alpha = 0.1$ level) the poorest NH reconstruction in all of our experiments, the detrended GKSS calibration experiment z. As discussed above, this reconstruction dramatically fails validation based on the other two metrics (*RE* and *CE*). The inappropriate acceptance by r^2 derives from the fact that the high-frequency variability in the reconstructed and actual series is highly correlated. Yet the overall quality of the reconstruction is clearly poor from a visual inspection (compare Figure 5b), highlighting the inappropriateness of using r^2 as a measure of the overall fidelity of a reconstruction.

[64] Our analyses thus expose a fundamental weakness in the use of r^2 as a metric of reconstruction skill [cf. MM05]. *Wahl and Ammann* [2007] note that because r^2 does not account for possible changes in either mean or variance relative to the calibration interval, its use as skill metric can lead to an unacceptably high probability of a type II error (i.e., the false rejection of a skillful reconstruction). Our results confirm and amplify this observation, demonstrating, as discussed above, a pattern wherein skillful long-term reconstructions are erroneously rejected on the basis of an insignificant r^2 statistic diagnosed over a short validation interval. Equally problematic, clearly unskillful reconstructions (i.e., those which reconstruct essentially none of the true low-frequency variability) are erroneously accepted on the basis of an apparently significant r^2 statistic. No such pattern is evident for either *RE* or *CE*. It is consequently apparent that MM05 employed flawed statistical reasoning when they argued for the rejection of reconstructions established as skillful in short validation experiments using a conventional metric (*RE*), based instead on the use of a metric (r^2) that is overly prone to both type I and type II errors.

6. Summary and Conclusions

[65] Using the RegEM CFR procedure favored by Mann and collaborators [e.g., MR02; R03; *Zhang et al.*, 2004; R05; M05] with the regularization scheme described in this study, we have demonstrated that CFR methods used in long-term large-scale paleoclimate reconstruction can produce skillful reconstructions with no evident systematic bias. These findings are based on our use of synthetic "pseudoproxy" data with realistic signal-versus-noise properties that are derived from two entirely independent simulations of the climate of the past millennium, both of which exhibits sizable long-term variations prior to the modern (19th/20th century) calibration intervals used. The actual long-term model histories are skillfully reconstructed, and lie within estimated uncertainties, in all of our experiments employing the correct CFR procedure. Moreover, meaningful reconstructions are not achieved when adopting the detrending procedure used in certain recent studies [VS04, BC05 and BFC06].

[66] The above conclusions are shown to hold for experiments using pseudoproxy networks with substantially lower SNR values (e.g., SNR = 0.25) and significantly redder noise spectra than is evident for actual proxy data networks used in paleoclimate reconstructions such as MBH98 and R05. These conclusions are insensitive to whether all proxies are individually used as predictors, or are represented by PC summaries using either of two possible standardization conventions explored in past work, though we find that the use of PC summaries to represent proxy networks is neither necessary nor beneficial in the context of the RegEM method, as potential colinearity of predictors is implicitly dealt with in the regularization process used in RegEM.

[67] Consistent with other recent studies [*Wahl and Ammann*, 2007; *Wahl et al.*, 2006] we conclude that the original MBH98 and MBH99 reconstructions are robust with respect to methodological considerations. Recently published criticisms of CFR methods are demonstrated not to hold up to independent scrutiny in the context of our experiments with the RegEM method. These previous criticisms instead appear to result from flawed implementation of statistical procedures or errors of statistical inference as detailed above. Comparison of skill diagnostics for the same experiments based on long (A.D. 850–1855) and short (A.D. 1856–1899) validation intervals indicate that the short validation intervals used in actual proxy climate reconstruction studies such as MBH98 and R05 are likely to provide conservative estimates of reconstruction skill and statistical uncertainties.

[68] Nearly a decade later, more than a dozen studies using alternative proxy data and reconstruction methods have, moreover, independently reaffirmed earlier studies such as MBH98, producing millennial or longer hemispheric temperature reconstructions which agree with the those reconstructions within estimated uncertainties. These additional studies support the key conclusion that late 20th century/early 21st century warmth is anomalous not only in the context of the past millennium, but apparently at least the

past 1.5 or 2 millennia [see *Jones and Mann*, 2004; *Moberg et al.*, 2005; *Hegerl et al.*, 2007].

[69] As highlighted by MBH99 and other related studies, current reconstructions rely on an increasingly sparse database of high-quality available proxy data back in time. Analyses of model climate simulation results can help guide strategies for paleoclimate proxy network design by identifying key potential regions (e.g., in the tropics or Southern Hemisphere) from where additional long-term proxy records might best contribute toward decreasing current uncertainties.

[70] We note that the continued emphasis only on hemispheric mean temperature series in many recent studies [e.g., VS04, MM05, BC05, BFC06], is likely to provide only limited physical or dynamical insight into the workings of the climate system. We thus encourage greater future focus on the reconstruction of spatial patterns of climate variability, and on key climate phenomena such as ENSO. We encourage expanded investigations of the issues explored in this study. We invite other researchers to download the source codes (and data) we have provided at http://www.meteo.psu.edu/~mann/PseudoproxyJGR06, and to further explore CFR performance using either the CSM1.4 simulation results, or other appropriately chosen climate model simulation results, using either surface temperatures or other fields (such as precipitation or sea level pressure), and using pseudoproxies such as those used in this study, or generalized to represent possible nonlocal teleconnected and/or nonlinear relationships between proxies and climate.

[71] **Acknowledgments.** This work was supported (M.E.M.) by the National Science Foundation under grant 0542356. The NCAR model simulation was conducted under the NCAR Weather and Climate Impacts Assessment Science Initiative by C.M.A., who acknowledges contributions from C. Tebaldi, L. Mearns, F. Joos, D. Schimel, and B. Otto-Bliesner. The National Center for Atmospheric Research is sponsored by the National Science Foundation. We thank two anonymous reviewers for their helpful comments. We thank E. Zorita for providing the GKSS simulation results and T. Lee for bringing to our attention the potential sensitivity to the standardization conventions used when employing ridge regression. We also thank J. Esper and D. Frank for useful discussions concerning the reliability of various skill metrics.

References

Adams, J. B., M. E. Mann, and C. M. Ammann (2003), Proxy evidence for an El Niño-like response to volcanic forcing, *Nature*, *426*, 274–278.

Allan, R., and T. Ansell (2006), A new globally complete monthly historical gridded mean sea level pressure data set (HadSLP2): 1850–2004, *J. Clim.*, *19*, 5816–5842.

Ammann, C. M., F. Joos, D. Schimel, B. L. Otto-Bliesner, and R. Tomas (2007), Solar influence on climate during the past millennium: Results from transient simulations with the NCAR Climate System Model, *Proc. Natl. Acad. Sci. U. S. A.*, *104*, 3713–3718.

Ansell, T. J., et al. (2006), Daily mean sea level pressure reconstruction for the European-North Atlantic region for the period 1850–2003, *J. Clim.*, *19*, 2717–2742.

Bradley, R. S., and P. D. Jones (1993), 'Little Ice Age' summer temperature variations: Their nature and relevance to recent global warming trends, *Holocene*, *3*, 367–376.

Braganza, K., D. J. Karoly, A. C. Hirst, M. E. Mann, P. Stott, R. J. Stouffer, and S. F. B. Tett (2003), Simple indices of global climate variability and change: Part I - variability and correlation structure, *Clim. Dyn.*, *20*, 491–502.

Briffa, K. R., T. J. Osborn, F. H. Schweingruber, I. C. Harris, P. D. Jones, S. G. Shiyatov, and E. A. Vaganov (2001), Low-frequency temperature variations from a northern tree-ring density network, *J. Geophys. Res.*, *106*, 2929–2941.

Burger, G., and U. Cubasch (2005), Are multiproxy climate reconstructions robust?, *Geophys. Res. Lett.*, *32*, L23711, doi:10.1029/2005GL024155.

Burger, G., I. Fast, and U. Cubasch (2006), Climate reconstruction by regression, *Tellus, Ser. A*, *58*, 227–235.

Casty, C., D. Handorf, and M. Sempf (2005), Combined winter climate regimes over the North Atlantic/European sector 1766–2000, *Geophys. Res. Lett.*, *32*, L13801, doi:10.1029/2005GL022431.

Cook, E., J. Esper, and R. D'Arrigo (2004), Extra-tropical Northern Hemisphere land temperature variability over the past 1000 years, *Quat. Sci. Rev.*, *23*, 2063–2074.

Cook, E. R., K. R. Briffa, and P. D. Jones (1994), Spatial regression methods in dendroclimatology: A review and comparison of two techniques, *Int. J. Climatol.*, *14*, 379–402.

Crowley, T. J., and T. S. Lowery (2000), How warm was the Medieval Warm Period? A comment on 'Man-made versus natural climate change,' *Ambio*, *39*, 51–54.

Delworth, T. L., and M. E. Mann (2000), Observed and simulated multidecadal variability in the Northern Hemisphere, *Clim. Dyn.*, *16*, 661–676.

Esper, J., E. R. Cook, and F. H. Schweingruber (2002), Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability, *Science*, *295*, 2250–2253.

Evans, M. N., A. Kaplan, and M. A. Cane (2002), Pacific sea surface temperature field reconstruction from coral δ^{18}O data using reduced space objective analysis, *Paleoceanography*, *17*(1), 1007, doi:10.1029/2000PA000590.

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.

Folland, C. K., D. E. Parker, A. W. Colman, and R. Washington (1999), Large scale modes of ocean surface temperature since the late nineteenth century, in *Beyond El Niño: Decadal and Interdecadal Climate Variability*, edited by A. Navarra, pp. 73–102, Springer, Berlin.

Folland, C. K., N. Rayner, P. Frich, T. Basnett, D. Parker, and B. Horton (2000), Uncertainties in climate data sets—A challenge for WMO, *WMO Bull.*, *49*, 59–68.

Folland, C. K., et al. (2001), Global temperature change and it uncertainties since 1861, *Geophys. Res. Lett.*, *28*, 2621–2624.

Fritts, H. C., T. J. Blasing, B. P. Hayden, and J. E. Kutzbach (1971), Multivariate techniques for specifying tree-growth and climate relationships and for reconstructing anomalies in paleoclimate, *J. Appl. Meteorol.*, *10*, 845–864.

Gilman, D. J., F. J. Fuglister, and J. M. Mitchell Jr. (1963), On the power spectrum of red noise, *J. Atmos. Sci.*, *20*, 182–184.

Golub, G. H., P. C. Hansen, and D. P. O'Leary (2000), Tikhonov regularization and total least squares, *SIAM J. Matrix Anal. Appl.*, *21*, 185–194.

Gonzalez-Rouco, F., H. Beltrami, E. Zorita, and H. Von Storch (2006), Simulation and inversion of borehole temperature profiles in surrogate climates: Spatial distribution and surface coupling, *Geophys. Res. Lett.*, *33*, L01703, doi:10.1029/2005GL024693.

Hegerl, G. C., T. J. Crowley, W. T. Hyde, and D. J. Frame (2006), Climate sensitivity constrained by temperature reconstructions over the past seven centuries, *Nature*, *440*, 1029–1032.

Hegerl, G. C., T. Crowley, M. Allen, W. T. Hyde, H. Pollack, J. Smerdon, and E. Zorita (2007), Detection of human influence on a new 1500 yr climate reconstruction, *J. Clim.*, *20*, 650–666.

Hoerl, A. E., and R. W. Kennard (1970a), Ridge regression: Applications to non-orthogonal problems, *Technometrics*, *12*, 69–82.

Hoerl, A. E., and R. W. Kennard (1970b), Ridge regression: Biased estimation for non-orthogonal problems, *Technometrics*, *12*, 55–67.

Huybers, P. (2005), Comment on "Hockey sticks, principal components, and spurious significance" by S. McIntyre and R. McKitrick,, *Geophys. Res. Lett.*, *32*, L20705, doi:10.1029/2005GL023395.

Jones, P. D., and M. E. Mann (2004), Climate over past millennia, *Rev. Geophys.*, *42*, RG2002, doi:10.1029/2003RG000143.

Jones, P. D., K. R. Briffa, T. P. Barnett, and S. F. B. Tett (1998), High-resolution palaeoclimatic records for the last millennium: Integration, interpretation and comparison with General Circulation Model control run temperatures, *Holocene*, *8*, 455–471.

Kaplan, A., Y. Kushnir, M. A. Cane, and M. B. Blumenthal (1997), Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperatures, *J. Geophys. Res.*, *102*, 27,835–27,860.

Kaplan, A., M. A. Cane, Y. Kushnir, and A. C. Clement (1998), Analyses of global sea surface temperature 1856–1991, *J. Geophys. Res.*, *103*, 18,567–18,589.

Kaplan, A., Y. Kushnir, and M. A. Cane (2000), Reduced space optimal interpolation of historical marine sea level pressure, *J. Clim.*, *13*, 2987–3002.

Little, R. J. A., and D. B. Rubin (1987), *Statistical Analysis With Missing Data: Series in Probability and Mathematical Statistics*, 278 pp., John Wiley, New York.

Luterbacher, J., C. Schmutz, D. Gyalistras, E. Xoplaki, and H. Wanner (1999), Reconstruction of monthly NAO and EU indices back to AD 1675, *Geophys. Res. Lett.*, *26*, 2745–2748.

Luterbacher, J., et al. (2002a), Extending North Atlantic Oscillation reconstructions back to 1500, *Atmos. Sci. Lett.*, *2*, 114–124, doi:10.1006/asle.2001.0044.

Luterbacher, J., E. Xoplaki, D. Dietrich, R. Rickli, J. Jacobeit, C. Beck, D. Gyalistras, C. Schmutz, and H. Wanner (2002b), Reconstruction of sea level pressure fields over the eastern North Atlantic and Europe back to 1500, *Clim. Dyn.*, *18*, 545–561.

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*, 1499–1503.

Luterbacher, J., et al. (2006), Mediterranean climate variability over the last centuries: A review, in *The Mediterranean Climate: An Overview of the Main Characteristics and Issues*, edited by P. Lionello, P. Malanotte-Rizzoli, and R. Boscolo, pp. 27–148, Elsevier, Amsterdam, Netherlands.

Mann, M. E., and P. D. Jones (2003), Global surface temperatures over the past two millennia, *Geophys. Res. Lett.*, *30*(15), 1820, doi:10.1029/2003GL017814.

Mann, M. E., and S. Rutherford (2002), Climate reconstruction using ''pseudoproxies'', *Geophys. Res. Lett.*, *29*(10), 1501, doi:10.1029/2001GL014554.

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.

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.

Mann, M. E., E. Gille, R. S. Bradley, M. K. Hughes, J. T. Overpeck, F. T. Keimig, and W. Gross (2000), Global temperature patterns in past centuries: An interactive presentation, Earth Interact., *4*. (Available at http://EarthInteractions.org)

Mann, M. E., S. Rutherford, E. Wahl, and C. Ammann (2005), Testing the fidelity of methods used in proxy-based reconstructions of past climate, *J. Clim.*, *18*, 4097–4107.

McIntyre, S., and R. McKitrick (2005), Hockey sticks, principal components, and spurious significance, *Geophys. Res. Lett.*, *32*, L03710, doi:10.1029/2004GL021750.

Moberg, A., D. M. Sonechkin, K. Holmgren, N. M. Datsenko, and W. Karlen (2005), Highly variable Northern Hemisphere temperatures reconstructed from low and high-resolution proxy data, *Nature*, *433*, 613–617.

Osborn, T. J., S. C. B. Raper, and K. R. Briffa (2006), Simulated climate change during the last 1000 years: Comparing the ECHO-G general circulation model with the MAGICC simple climate model, *Clim. Dyn.*, *27*, 185–197.

Overpeck, J., et al. (1997), Arctic environmental change of the last four centuries, *Science*, *278*, 1251–1256.

Pauling, A., J. Luterbacher, and H. Wanner (2003), Evaluation of proxies for European and North Atlantic temperature field reconstructions, *Geophys. Res. Lett.*, *30*(15), 1787, doi:10.1029/2003GL017589.

Pauling, A., J. Luterbacher, C. Casty, and H. Wanner (2006), 500 years of gridded high-resolution precipitation reconstructions over Europe and the connection to large-scale circulation, *Clim. Dyn.*, *26*, 387–405.

Rayner, N. A., E. B. Horton, D. E. Parker, C. K. Folland, and R. B. Hackett (1996), Version 2.2 of the global sea-ice and sea surface temperature data set, 1903–1994, *Clim. Res. Tech. Note*, *74*, 43 pp., Natl. Meteorol. Libr., Bracknell, U. K.

Rayner, N. A., D. E. Parker, P. Frich, E. B. Horton, C. K. Folland, and L. V. Alexander (2000), SST and sea-ice fields for ERA40, in *Proceedings of the Second International WCRP Conference on Reanalyses, Wokefield Park, Reading, UK, 23–27 August 1999*, *WCRP-109, WMO/TD 985*, pp. 18–21, World Meteorol. Organ., Geneva, Switzerland.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan (2003), Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, *J. Geophys. Res.*, *108*(D14), 4407, doi:10.1029/2002JD002670.

Reynolds, R. W., and T. M. Smith (1994), Improved global sea surface temperature analyses using optimum interpolation, *J. Clim.*, *7*, 929–948.

Rutherford, S., M. E. Mann, T. L. Delworth, and R. Stouffer (2003), Climate field reconstruction under stationary and nonstationary forcing, *J. Clim.*, *16*, 462–479.

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 methodology, 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*, 853–871.

Shindell, D. T., G. A. Schmidt, M. E. Mann, D. Rind, and A. Waple (2001), Solar forcing of regional climate change during the Maunder Minimum, *Science*, *7*, 2149–2152.

Shindell, D. T., G. A. Schmidt, R. L. Miller, and M. E. Mann (2003), Volcanic and solar forcing of climate change during the preindustrial era, *J. Clim.*, *16*, 4094–4107.

Shindell, D. T., G. A. Schmidt, M. E. Mann, and G. Faluvegi (2004), Dynamic winter climate response to large tropical volcanic eruptions since 1600, *J. Geophys. Res.*, *109*, D05104, doi:10.1029/2003JD004151.

Smith, T. M., and R. W. Reynolds (2005), A global merged land-air-sea surface temperature reconstruction based on historical observations (1880–1997), *J. Clim.*, *18*, 2021–2036.

Smith, T. M., R. W. Reynolds, R. E. Livezey, and D. C. Stokes (1996), Reconstruction of historical sea surface temperatures using empirical orthogonal functions, *J. Clim.*, *9*, 1403–1420.

Smith, T. M., R. E. Livezey, and S. S. Shen (1998), An improved method for analyzing sparse and irregularly distributed SST data on a regular grid: The tropical Pacific Ocean, *J. Clim.*, *11*, 1717–1729.

Tikhonov, A. N., and V. Y. Arsenin (1977), *Solution of Ill-Posed Problems*, *Scripta Series in Mathematics*, 258 pp., V. H. Winston and Sons, Inc., Palm Beach, Fla.

Von Storch, H., and E. Zorita (2005), Comment on ''Hockey sticks, principal components, and spurious significance'' by S. McIntyre and R. McKitrick, *Geophys. Res. Lett.*, *32*, L20701, doi:10.1029/2005GL022753.

Von Storch, H., E. Zorita, J. M. Jones, Y. Dimitriev, F. Gonzalez-Rouco, and S. F. B. Tett (2004), Reconstructing past climate from noisy data, *Science*, *306*, 679–682.

Von Storch, H., E. Zorita, J. M. Jones, F. Gonzalez-Rouco, and S. F. B. Tett (2006), Response to comment on ''Reconstructing past climate from noisy data,'' *Science*, *312*, 529c.

Wahl, E. R., and C. M. Ammann (2007), Robustness of the Mann, Bradley, Hughes reconstruction of surface temperatures: Examination of criticisms based on the nature and processing of proxy climate evidence, *Clim. Change*, in press.

Wahl, E. R., D. M. Ritson, and C. M. Ammann (2006), Comment on ''Reconstructing past climate from noisy data,'' *Science*, *312*, 529b.

Waple, A., M. E. Mann, and R. S. Bradley (2002), Long-term patterns of solar irradiance forcing in model experiments and proxy-based surface temperature reconstructions, *Clim. Dyn.*, *18*, 563–578.

Xoplaki, E., J. Luterbacher, H. Paeth, D. Dietrich, N. Steiner, M. Grosjean, and H. Wanner (2005), European spring and autumn temperature variability and change of extremes over the last half millennium, *Geophys. Res. Lett.*, *32*, L15713, doi:10.1029/2005GL023424.

Zhang, Z., and M. E. Mann (2005), Coupled patterns of spatiotemporal variability in Northern Hemisphere sea level pressure and conterminous U.S. drought, *J. Geophys. Res.*, *110*, D03108, doi:10.1029/2004JD004896.

Zhang, Z., M. E. Mann, and E. R. Cook (2004), Alternative methods of proxy-based climate field reconstruction: Application to the reconstruction of summer drought over the conterminous United States back to 1700 from drought-sensitive tree ring data, *Holocene*, *14*, 502–516.

Zorita, E., and H. Von Storch (2005), Methodical aspects of reconstructing non-local historical temperatures, *Mem. Soc. Astron. It.*, *76*, 794–801.

C. Ammann, Climate Global Dynamics Division, National Center for Atmospheric Research, 1850 Table Mesa Drive, Boulder, CO 80305, USA.

M. E. Mann, Department of Meteorology, 523 Walker Building, Pennsylvania State University, University Park, PA 16802, USA. (mann@psu.edu)

S. Rutherford, Department of Environmental Science, Roger Williams University, Bristol, RI 02809, USA.

E. Wahl, Department of Environmental Studies, Alfred University, Alfred, NY 14802, USA.