LETTERS

Testing the Fidelity of Methods Used in Proxy-Based Reconstructions of Past Climate

MICHAEL E. MANN

Department of Environmental Sciences, University of Virginia, Charlottesville, Virginia

SCOTT RUTHERFORD

Department of Environmental Science, Roger Williams University, Bristol, Rhode Island

EUGENE WAHL

Department of Environmental Studies, Alfred University, Alfred, New York

CASPAR AMMANN

Climate Global Dynamics Division, National Center for Atmospheric Research, Boulder, Colorado*

(Manuscript received 4 January 2005, in final form 8 June 2005)

ABSTRACT

Two widely used statistical approaches to reconstructing past climate histories from climate "proxy" data such as tree rings, corals, and ice cores are investigated using synthetic "pseudoproxy" data derived from a simulation of forced climate changes over the past 1200 yr. These experiments suggest that both statistical approaches should yield reliable reconstructions of the true climate history within estimated uncertainties, given estimates of the signal and noise attributes of actual proxy data networks.

1. Introduction

Two distinct types of methods have primarily been used to reconstruct past large-scale climate histories from proxy data. One group, so-called climate field reconstruction (CFR) methods, assimilates proxy records into a reconstruction of the underlying patterns of past climate change (e.g., Fritts et al. 1971; Cook et al. 1994; Mann et al. 1998, henceforth MBH98; Evans et al. 2002; Luterbacher et al. 2002; Rutherford et al. 2005; Zhang et al. 2004). The other group, simple so-called composite-plus-scale (CPS) methods (Bradley and Jones 1993; Jones et al. 1998; Crowley and Lowery 2000; Briffa et al. 2001; Esper et al. 2002; Mann and Jones 2003, henceforth MJ03; Crowley et al. 2003), composites a number of proxy series and scales the resulting composite against a target (e.g., Northern Hemisphere temperature) instrumental series. CFR methods offer the advantage of estimating spatial patterns, while CPS methods involve a simpler statistical procedure.

It is difficult to compare the performance of alternative approaches to proxy-based climate reconstruction because no adequate "ground truth" is available for evaluating the fidelity of long-term reconstructions. Climate model simulations can, however, be used to provide a simulated ground truth, resting on physically based first principles, which can be exploited for the testing of competing methods. With a climate model simulation, we can sample the field of interest (e.g., the surface temperature field) at selected locations (i.e., model gridboxes), to produce a set of regional climate time series. Proxy data, however, contain considerable noise due both to the imperfect nature of the proxy data themselves and the inherent climate noise that ex-

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author address: Prof. Michael E. Mann, Department of Meteorology, The Pennsylvania State University, 503 Walker Building, University Park, PA 16802-5013.
E-mail: mann@psu.edu

ists at the local spatial scales represented by actual proxy sites. We thus add a noise component, of appropriate amplitude, to represent the real-world uncertainties that exist in the relationship between proxies and regional climate. The resulting synthetic proxies or "pseudoproxies" then represent imperfect local climate indicators that possess similar properties to actual proxy data. We can then apply different climate reconstruction methods to pseudoproxy networks that are constructed to have properties similar to those that have been used in past studies to produce hypothetical reconstructions. Observing how well the actual climate model results can be recovered from the reconstruction over an independent (verification) period provides us then with an objective assessment of the reliability of the method, given the assumed properties of the pseudoproxy data. Surprisingly, these types of exercises have not previously been applied to assess the performance of CPS methods. They have been used, however, to investigate the fidelity of CFR methods (Mann and Rutherford 2002; Rutherford et al. 2003; Zorita et al. 2003; von Storch et al. 2004).

The conclusions of such exercises are only meaningful if the climate model results are reasonably realistic. Most previous studies employing these tests suggest that that CFR methods are likely to produce faithful estimates of long-term trends given the statistical attributes inferred for actual proxy data and the history of climate forcing estimated over past centuries (e.g., Crowley 2000). One study by von Storch et al. (2004, henceforth VS04), however, concludes that a substantial bias may arise in proxy-based estimates of long-term temperature changes using CFR methods. VS04 based this conclusion on experiments using a simulation of the GKSS coupled model [similar experiments described by VS04 using an alternative simulation of the third Hadley Centre Coupled Ocean–Atmosphere General Circulation Model (HadCM3) coupled model showed little such bias]. The GKSS simulation was forced with unusually large changes in natural radiative forcing in past centuries [the peak-to-peak solar forcing changes on centennial time scales ($\sim$1 W m^{-2}) were about twice that used in other studies (e.g., Crowley 2000) and much larger than the most recent estimates ($\sim$0.15 W m^{-2}; see Lean et al. 2002; Foukal et al. 2004)]. A substantial component of the low-frequency variability in the GKSS simulation, furthermore, appears to have been a "spinup" artifact: the simulation was initialized from a very warm twentieth-century state at A.D. 1000, prior to the application of preanthropogenic radiative forcing, leading to a long-term drift in mean temperature (Goosse et al. 2005). CFR methods are known to perform poorly in capturing patterns of variability that are entirely or largely missing during the calibration period (Rutherford et al. 2003). The long-term model drift in the GKSS simulation contributes an unphysical pattern of variance in early centuries that is likely almost entirely absent from the later twentieth-century calibration period used by VS04. The large changes in solar forcing assumed by VS04 also occur largely before the twentieth century. These arguably unrealistic features in the GKSS simulation make the simulation potentially inappropriate for use in testing climate reconstruction methods.

We investigate here both the CFR and CPS approaches, using networks of synthetic pseudoproxy data (see Mann and Rutherford 2002) constructed to have attributes similar to actual proxy networks used in past CFR and CPS studies, respectively. The pseudoproxy data are derived from a simulation of the climate of the past millennium (A.D. 850–1999) using the National Center for Atmospheric Research (NCAR) Climate System Model (CSM) version 1.4 coupled ocean–atmosphere model (Ammann et al. 2005, manuscript submitted to *Proc. Natl. Acad. Sci.,* henceforth A05), which we believe to be appropriately realistic for use in testing climate reconstruction methods. The model has a climate sensitivity of 2°C/(2 × CO_2), somewhat lower than that for the GKSS model [3.2°C/(2 × CO_2)]. Unlike VS04, however, the model has been forced with estimated natural (volcanic + solar) and anthropogenic radiative forcing estimates that are within conventionally accepted ranges, though perhaps slightly higher than used in most other simulations (see Jones and Mann 2004). The surface temperature field has, moreover, been corrected for the small long-term drift present (A05). This simulation is characterized by forced Northern Hemisphere (NH) annual mean temperature variations in past centuries that are modestly greater in amplitude than most other simulations (Jones and Mann 2004), ensuring that the simulation represents an appropriately challenging, but fair, test bed for investigating the fidelity of statistical climate reconstruction approaches. (The actual and reconstructed NH series, temperature pseudoproxy and gridbox series for both CPS and CFR experiments are available at http://fox.rwu.edu/~rutherfo/supplements/Pseudoproxy05/.)

2. Methods

a. CPS method

Our implementation of the CPS method is similar to that used in past studies (e.g., Bradley and Jones 1993; Jones et al. 1998; Crowley and Lowery 2000; Briffa et al. 2001; Esper et al. 2002; MJ03) wherein roughly a

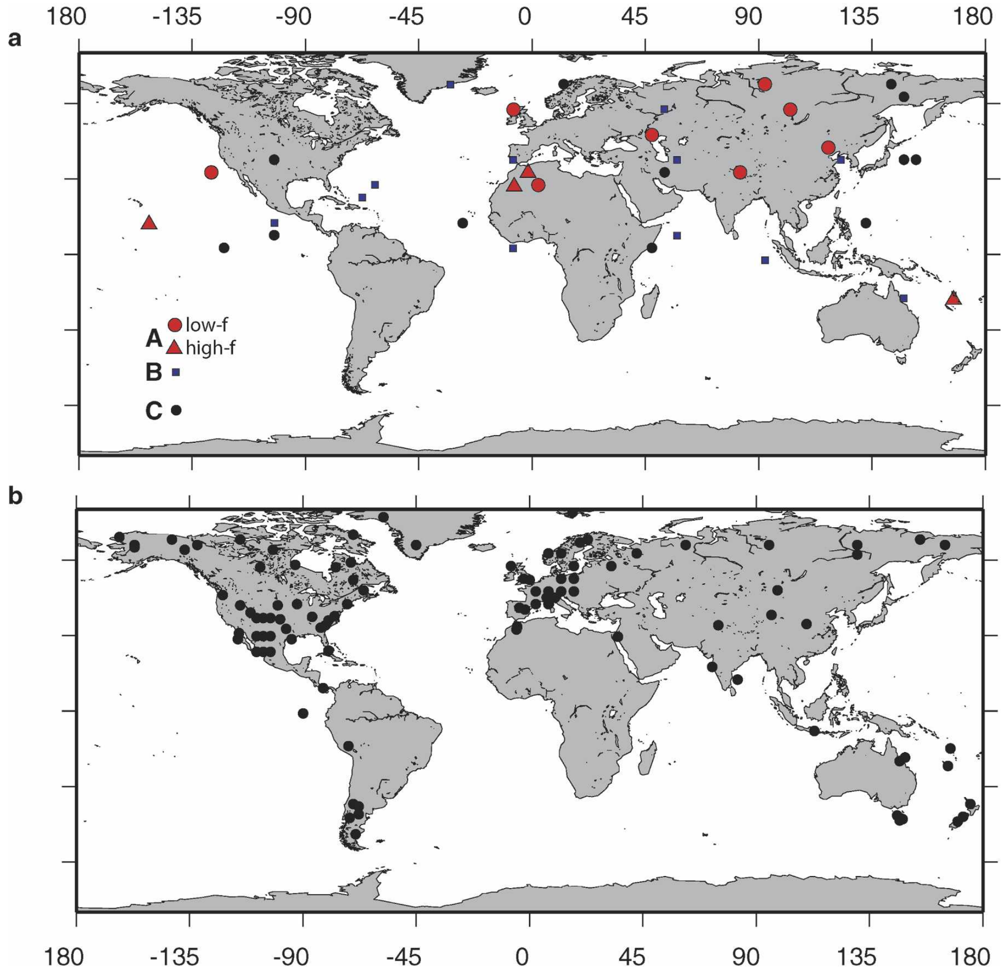

Fig. 1. Spatial distributions of (a) each of the three random networks of 12 pseudoproxies used in the various CPS experiments (A, B, and C) and (b) the 104 pseudoproxies used in the CFR experiments. For CPS experiment A, subsets of pseudoproxies used in the separate low- and high-frequency composites for the M05 emulation experiment are indicated.

dozen proxy series, each of which is assumed to represent a linear combination of local temperature variations and an additive "noise" component, are composited (typically at decadal resolution; see, e.g., Bradley and Jones 1993; Crowley and Lowery 2000; Esper et al. 2002; MJ03) and scaled against an instrumental hemispheric mean temperature series during an overlapping "calibration" interval to form a hemispheric reconstruction. The rationale for such an approach is that the surface temperature field over, for example, the Northern Hemisphere only possesses a dozen or less spatial degrees of freedom on interannual and longer time scales (e.g., Briffa and Jones 1993). In the CPS method, the composite is typically scaled (e.g., Bradley and Jones 1993; Crowley and Lowery 2000; Esper et al. 2002; MJ03) to have the same variance as the target instrumental series. Such scaling is based on the assumption that the contributing proxy series can be treated as actual local temperature indicators, rather than simply a set of statistical predictors. The CPS approach thus differs from a multivariate regression approach wherein the reconstruction is guaranteed to have *less* variance than the target series over the calibration interval.

We formed three different networks (A, B, and C) of a dozen randomly selected model grid boxes (restricted to the region of the Northern Hemisphere and Tropics that is sampled by the twentieth-century instrumental record to simulate the coverage of real-world proxy networks; see Fig. 1a). Pseudoproxy time series (see supplementary Fig. 1 at http://dx.doi.org/10.1175/jcli3564.s1) were formed through summing each grid box annual mean temperature series with a realization of white noise (a reasonable assumption for representing observational error—the effects of the noise "color" were investigated by Mann and Rutherford 2002), allowing for various relative amplitudes of noise variance (expressed as a signal-to-noise ratio (SNR) of amplitudes; see Mann and Rutherford 2002). Experiments with SNR = 1.0 were performed for all three networks. For the first network (A), experiments were performed for four different values of SNR: 0.25, 0.5, 1.0, and ∞ (i.e., no added noise).

Following the typical CPS procedure (e.g., Bradley and Jones 1993; Crowley and Lowery 2000; Esper et al. 2002; MJ03), each pseudoproxy series was decadally smoothed (using a 10-point Butterworth filter with optimal boundary constraints as described by Mann 2004) and standardized. A weighted composite was formed based alternatively on (a) a simple areal weighting (i.e., weighting by cosine of the latitude of the gridbox), and (b) areal weighting coupled with an additional weighting by the calibration period correlation with the corresponding observed gridbox temperature series (as in MJ03). The composite was then scaled to have the same mean and decadal standard deviation as the actual NH series over a modern calibration interval (we used both 1856–1980 and 1900–80 calibration intervals, representative of the varying calibration intervals that have been used in actual reconstructions).

We calculated the reduction of error (RE) reconstruction skill diagnostics during both the calibration period and an independent, precalibration "verification" period (see, e.g., Cook et al. 1994; Rutherford et al. 2005, and references therein). We also calculated alternative coefficient of efficiency (CE) and squared Pearson correlation (r^2) verification skill metrics. While *RE* is considered a superior metric for evaluating statistical reconstruction skill (see Rutherford et al. 2005), verification RE values tend to be enhanced in these particular experiments as a result of the unusually large pre-twentieth-century mean temperature changes that occur in the model output and are captured by the reconstructions. These mean changes are larger than those observed in the actual instrumental data. We therefore also used a highly conservative estimate of unresolved variance provided by $1 - r^2$ (along with the more conventional $1 - \mathrm{RE}$) to estimate statistical uncertainties as conservatively as possible. For experiments using a 1900–80 calibration interval, an 1855–99 verification interval was used, to mimic the verification intervals that have been used in previous studies (Mann et al. 1998; Mann and Rutherford 2002; Rutherford et al. 2003, 2005). For experiments using an 1856–1980 calibration interval, the full available 850–1855 period was instead used for verification. Statistical significance of resolved variance metrics were calculated through Monte Carlo simulations based on the null hypothesis of (order one autoregressive) AR(1) red noise with the serial correlation and variance of the actual model NH series over the calibration period.

b. Climate field reconstruction method

Our implementation of the CFR approach makes use of the regularized expectation maximization (RegEM) method of Schneider (2001), which has been applied to CFR in several recent studies (Mann and Rutherford 2002; Rutherford et al. 2003; Zhang et al. 2004; Rutherford et al. 2005). The method is similar to principal component analysis (PCA)-based approaches (e.g., MBH98; Evans et al. 2002; Luterbacher et al. 2002) but employs an iterative estimate of data covariances to make more complete use of the available information (see Schneider 2001). As in Rutherford et al. (2005), we tested (i) straight application of RegEM, (ii) a "hybrid frequency-domain calibration" approach that employs separate calibrations of high (shorter than 20-yr period) and low frequency (longer than 20-yr period) components of the annual mean data that are subsequently composited to form a single reconstruction, and (iii) a "stepwise" version of RegEM in which the reconstruction itself is increasingly used in calibrating successively older segments ("MATLAB" source codes available at Web site given above). As in Rutherford et al. (2005), the reconstruction was initialized with the mean of all of the predictors (i.e., the pseudoproxies) over the precalibration interval. This choice speeds convergence, but the final reconstruction is insensitive, as it must be to be reliable, to the initial condition (see supplementary table and supplementary Fig. 2 at http://dx.doi.org/10.1175/jcli3564.s2).

The RegEM method has been shown to yield a very similar hemispheric mean temperature reconstruction to that of MBH98 when applied to the same network of multiple proxy indicators (Rutherford et al. 2005; Fig. 2), suggesting that conclusions regarding real-world proxy reconstructions are not likely to be dependent upon the precise method of CFR used. Similar experiments to those performed here, but using the MBH98 PCA-based approach, are described by Ammann et al. (2005, unpublished manuscript, henceforth AWT05). Following Zorita et al. (2003), VS04, and Rutherford et al. (2005), our CFR analyses make use of annual mean data and the precise MBH98 multiproxy network locations. The actual MBH98 proxy network becomes sparser back in time. Previous studies (Mann and Rutherford 2002; Rutherford et al. 2003; Zorita et al. 2003) have used pseudoproxy experiments to investigate the influence of increasing sample sparseness on CFR re-

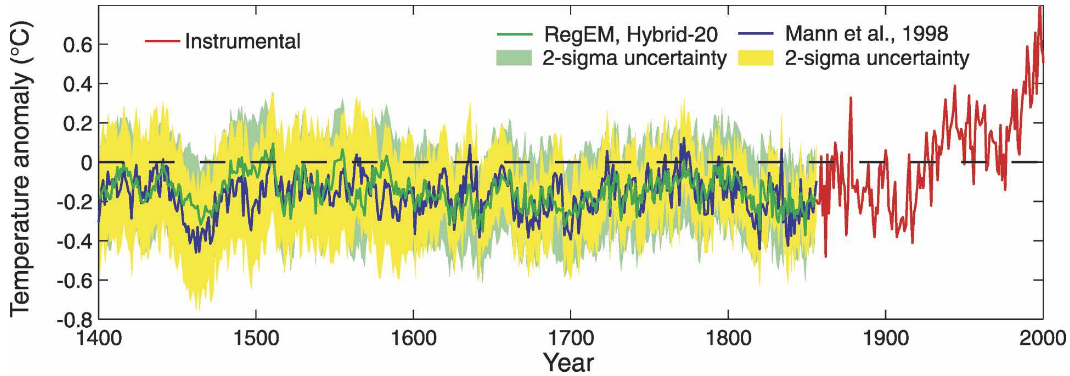

FIG. 2. Comparison between annual mean NH reconstructions using the MBH98 multiproxy dataset and (i) PCA-based method (blue) of MBH98 and (ii) hybrid frequency-domain RegEM method (green) of Rutherford et al. (2005). The two reconstructions share 74% of their variance during the 1400–1855 overlap period (RE = 0.74 with the MBH98 reconstruction as the reference series, and using the 1900–71 overlap in calibration periods as the reference interval). Instrumental annual mean NH series of Jones et al. (1998) is shown for comparison (red), along with the two standard error uncertainty intervals for both reconstructions.

sults. Here, as in VS04, we have for simplicity used a distribution of sites corresponding to the full network (all proxies available back through the early nineteenth century) used by MBH98.

Using the method described above, we formed pseudoproxy indicators at the 104 unique locations (Fig. 1b) used by MBH98. While the pseudoproxy indicators were, as above, computed from the surface temperature field, it should be noted that with CFR methods, the proxy data need not reflect local surface temperatures; any indicator that is statistically related to one or more large-scale patterns expressed in the surface temperature field (e.g., a coral record reflecting salinity changes in the western tropical Pacific related to the El Nino–Southern Oscillation) can be used in a CFR reconstruction of the surface temperature field (see, e.g., Mann et al. 1998; Rutherford et al. 2005). We used the same four values, SNR = 0.25, 0.5, 1.0, and ∞,

TABLE 1. Comparisons of calibration and verification reconstruction skill for both CPS and CFR experiments using pseudoproxy networks of varying SNR, spatial distribution, calibration interval, and (in CPS case) weighting scheme. Statistical significance of verification skill metrics from Monte Carlo simulations exceeds the $p = 0.05$ level unless otherwise indicated (italicized: significant at $0.10 > p > 0.05$; boldface: not significant at $p = 0.10$ level).

Type	Network	SNR	Weighting	Calibration period	Calibration statistics RE	Verification statistics RE	CE	r^2
CPS	12 proxies (A)	∞	Uniform	1856–1980	0.74	0.93	0.63	0.77
CPS	12 proxies (A)	∞	Uniform	1900–80	0.77	0.84*	−0.09*	0.64*
CPS	12 proxies (A)	1.0	Uniform	1856–1980	0.73	0.89	0.45	0.68
CPS	12 proxies (A)	1.0	Uniform	1900–80	0.72	0.79*	−0.43*	*0.43**
CPS	12 proxies (A)	1.0	Correlation	1900–80	0.71	0.78*	−0.48*	*0.44**
CPS	12 proxies (A)	1.0	Correlation	1856–1980	0.72	0.88	0.42	0.67
CPS	12 proxies (B)	1.0	Correlation	1856–1980	0.74	0.91	0.58	0.67
CPS	12 proxies (C)	1.0	Correlation	1856–1980	0.69	0.87	0.38	0.58
CPS	12 proxies (A)	0.5	Correlation	1900–80	0.34	*0.46**	*−2.62*	*0.38**
CPS	12 proxies (A)	0.5	Correlation	1856–1980	0.42	0.83	0.17	0.58
CPS	12 proxies (A)	0.25	Correlation	1900–80	0.52	**−0.11***	**−6.54***	**0.16***
CPS	12 proxies (A)	0.25	Correlation	1856–1980	0.30	0.58	−1.04	0.20
CFR	MBH98	∞	NA	1856–1980	NA	0.97	0.83	0.85
CFR	MBH98	1.0	NA	1900–80	NA	0.93*	0.55*	0.62*
CFR	MBH98	1.0	NA	1856–1980	NA	0.96	0.80	0.83
CFR	MBH98	0.5	NA	1856–1980	NA	0.94	0.72	0.74
CFR	MBH98	0.25	NA	1856–1980	NA	0.87	0.39	0.38

* An 1856–99 verification interval was used in this experiment (an A.D. 850–1855 verification interval was used for experiments employing an 1856–1980 calibration interval).

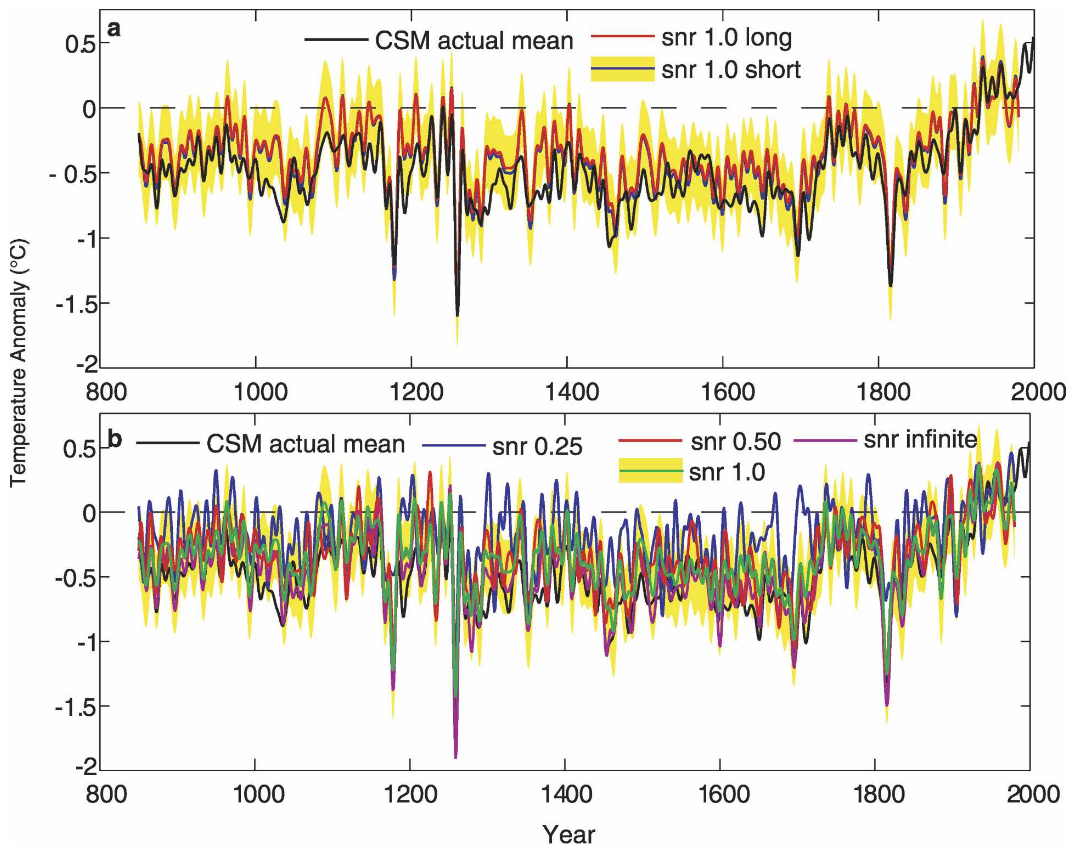

FIG. 3. Pseudoproxy reconstruction of decadal mean NH mean temperature based on CPS approach. We show results using (a) the SNR value (1.0) most consistent with actual proxy networks (results for network A shown) and two alternative calibration intervals (1856–1980 and 1900–80), and (b) all four SNR values and an 1856–1980 calibration interval. Statistical uncertainties (shading of two standard error regions) for SNR = 1.0 and actual model decadal mean NH series are also shown.

and the same two alternative calibration intervals described above. The statistical verification procedure and uncertainty estimation procedures described above were applied to the (decadally smoothed) CFR results. It should be noted that, in contrast with the CPS method, traditional calibration metrics of resolved variance are not available in RegEM. Statistical significance of verification skill metrics was estimated using the same Monte Carlo significance estimation procedure described above.

3. Results

a. CPS method

In the CPS experiments, results for SNR = 1.0 most closely resembled those obtained for actual proxy reconstructions (Mann et al. 1998; Rutherford et al. 2005; Mann and Jones 2003), with respect to the NH mean variance resolved (see Table 1) during the calibration (roughly 70%) and verification periods (roughly 45%–70%). By contrast, lower SNR values (i.e., 0.25) yielded significantly lower estimates of reconstruction skill in calibration (<50%) and verification (<20%) than indicated for actual reconstructions. For SNR = 1.0, the reconstructions (Fig. 3a) are observed to be relatively insensitive to whether the short (1900–80) or long (1856–1980) calibration interval is used, the specific locations of the 12 pseudoproxy sites (comparisons of the results for the three different networks are provided in Table 1 and Fig. 4), or which of the two weighting schemes was used in the composite (see Table 1). It is noteworthy that the reconstructions faithfully capture the large, short-term coolings associated with large explosive volcanic forcing events that take place (see A05) during for example, the late twelfth century, mid-thirteenth century, mid-fifteenth century, and early nineteenth century. At lower SNR (0.25 or 0.5), use of a combined nineteenth-/twentieth-century calibration interval (as in Mann and Jones 2003) appears to yield a more reliable reconstruction than a twentieth-century-

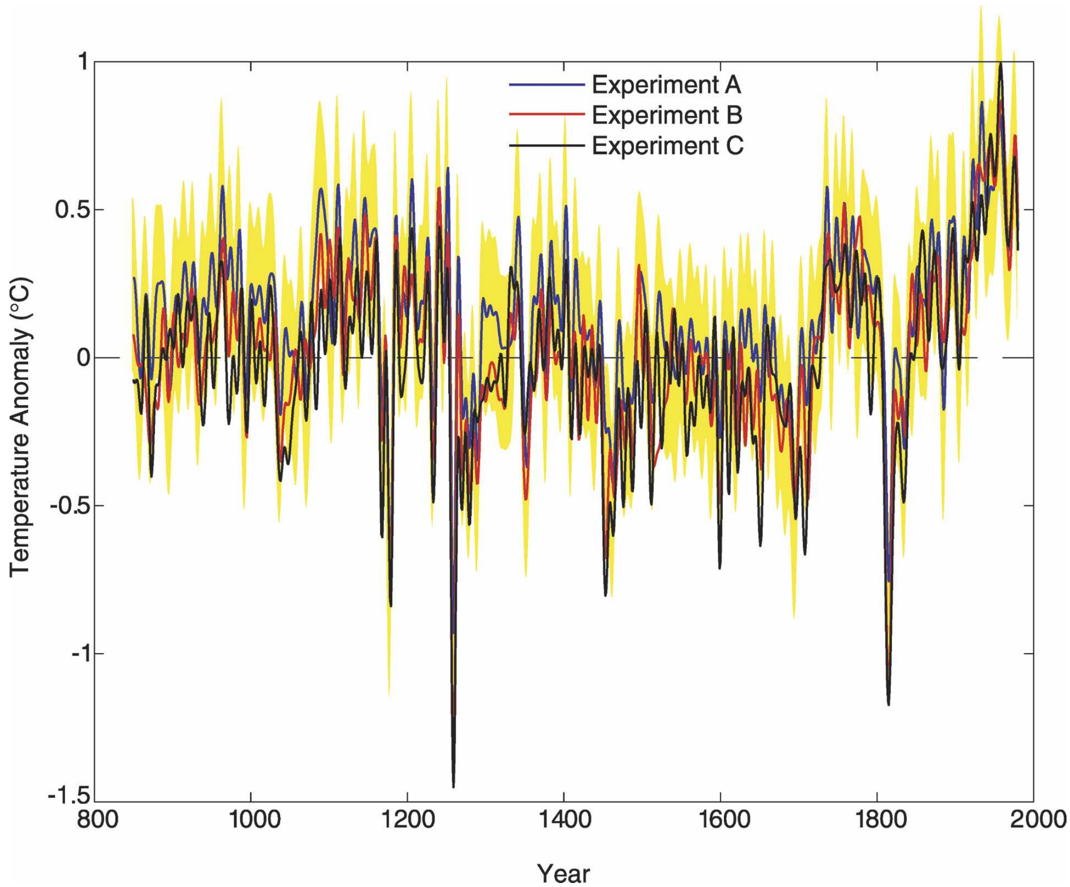

FIG. 4. Comparison of CPS reconstructions for SNR = 1.0 (1856–1980 calibration period) for using the three different random networks—A, B, and C—shown in Fig. 1.

only interval (as in Bradley and Jones 1993; Jones et al. 1998; Crowley and Lowery 2000). This finding appears consistent with other studies of the sensitivity of the CPS method to the length of the calibration interval (Esper et al. 2005). For SNR = 1.0, a modest underestimate of long-term changes in mean is evident, though this bias is well within the estimated uncertainties. This result is consistent with the suggestion that simple CPS, or regression-based methods employing a short calibration period, are likely to underestimate long-term variability (Osborn and Briffa 2004; it should be noted that this argument does not apply to CFR methods). Only for the lowest SNR ratio (0.25) does this underestimate become significant (Fig. 3b).

A variant on the CPS approach was recently described by Moberg et al. (2005, henceforth M05). In this alternative approach, proxy series are first standardized by their nominal standard deviations, and separate composites are formed in both low-frequency (>80 yr) and high-frequency (<80 yr) bands, making use in each case of only multidecadal/centennial resolution and only annual resolution (tree ring), proxy indicators, respectively. The two composites are then simply added, and the combined result is scaled against the modern instrumental record. M05 demonstrated that application of such an approach to a set of long-term proxy data appears to indicate greater low-frequency variability than is evident in other (see, e.g., Fig, 5 of Jones and Mann 2004) existing reconstructions. We have used the same pseudoproxy networks employed in our tests of the conventional CPS method described above to investigate whether this greater low-frequency variability is likely to be real, or spurious, in nature.

Since the low-resolution records represent effectively low-passed versions of whatever underlying annual climate process they represent, they have less broadband variance than would be present in a corresponding annually resolved record. Thus, when both the high-resolution and low-resolution indicators are each separately standardized by their nominal standard deviations, the variance in the low-resolution records is effectively inflated relative to that of the high-

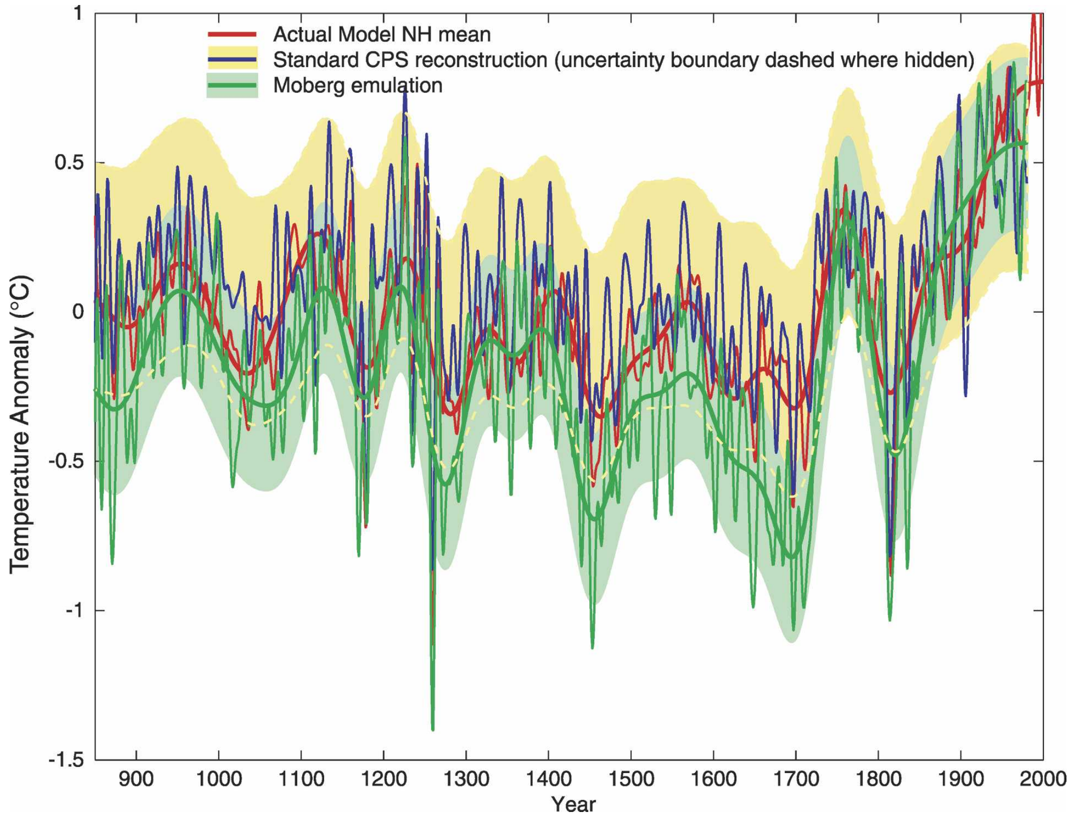

FIG. 5. Comparison of reconstructions using conventional CPS approach (blue) and our emulation of the M04 procedure (green), based on network used in CPS experiment A with SNR = 0.5. Shown for comparison is the true model NH history (red). Thick curves highlight the low-frequency (>80-yr time scale) variations, based on low-pass filtering at $f = 0.025$ cycles yr^{-1} using the method described by Mann (2004). Shading indicates two standard error uncertainties for the >80-yr time-scale variations in the two reconstructions.

resolution records. Since the sum of the separate high-frequency and low-frequency composites is scaled to have the same variance as the instrumental record, this means that the low-frequency variance is inflated, and the high-frequency variance is deflated, relative to the instrumental temperature record. In other words, the low-frequency variability in the reconstruction is likely to be artificially inflated relative to the true low-frequency temperature variations.

To test this, we simulated the M05 procedure for the case SNR = 0.5 (network A) described above. We first split the set (see Fig. 1) into one-third (four proxies) for the high-frequency (<80-yr period) composite, and two-thirds (eight proxies) for the low-frequency (>80-yr period) composite and additionally smoothed the eight low-frequency proxies using an 80-yr low-pass filter (Mann 2004) to simulate the low-resolution character of the various proxies (e.g., sediment cores) used by M04 in their low-frequency composite. We then standardized all series by their nominal standard deviation and composited the high- and low-frequency components separately. Finally, we added the high- and low-frequency composites and scaled the resulting series to have the same mean and decadal variance as the model's actual NH series over the 1856–1980 interval. Figure 5 shows the resulting reconstruction along with the reconstruction resulting from application of the conventional CPS procedure (i.e., as shown in Fig. 3b). While the conventional CPS reconstruction modestly underestimates the true variance as discussed earlier, the resulting reconstruction nonetheless remains well within estimated uncertainties of the true series. In contrast, the emulated M05 reconstruction lies near or outside of the 95% confidence interval over too wide a region to be consistent with the actual model climate and generally overestimates the true low-frequency variability. The existence of such an overestimation bias when using the M05 scheme appears to depend on the SNR value, the particular random network used, and which subsets of indicators are reserved for the two different frequency bands. However, the fact that the scheme significantly overestimates the true low-

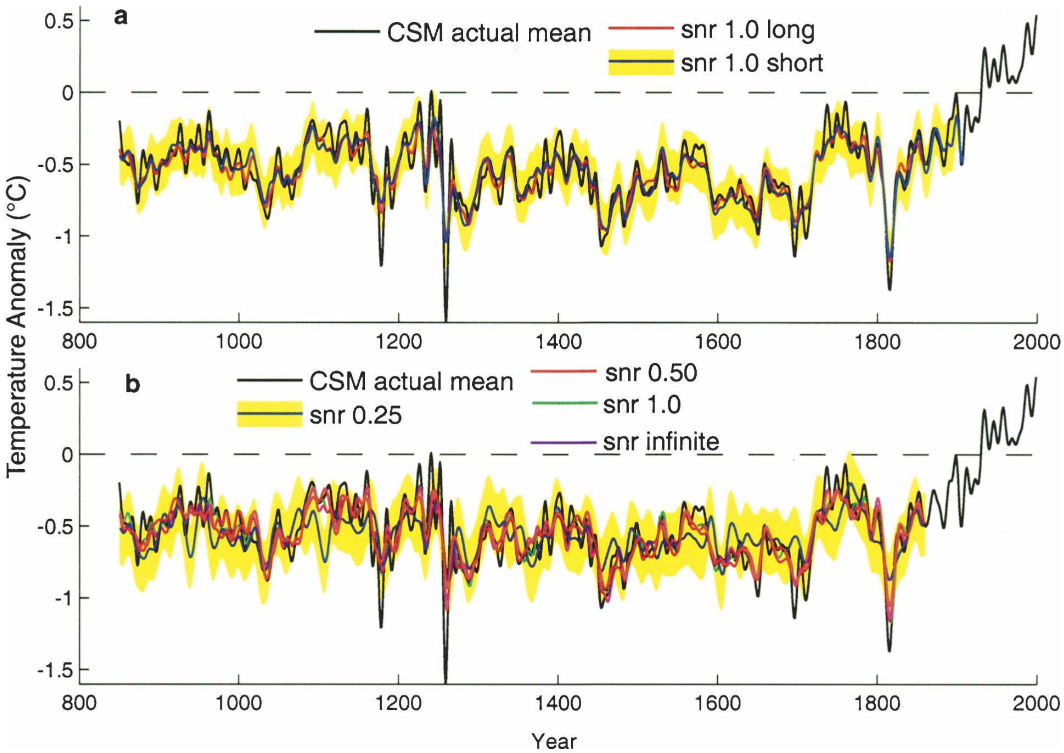

FIG. 6. Pseudoproxy reconstruction of NH mean temperature based on CFR approach compared with model NH series (decadally smoothed data are shown) for (a) SNR = 1.0 using two alternative calibration intervals (1856–1980 and 1900–80). Uncertainties (yellow shading indicates two standard error regions) are shown for the latter of the two calibration intervals, based on 1856–99 verification period residuals. (b) All four SNR values and an 1856–1980 calibration interval. Statistical uncertainties (shading of two standard error regions) are shown for lowest SNR value (0.25) based on pre-1856 verification residuals.

frequency variability in some situations indicates that it cannot be trusted to yield a reliable reconstruction when applied to real-world proxies. Additional experiments (see supplementary Fig. 3 at http://dx.doi.org/10.1175/jcli3564.s3) indicate that adjustment of the M05 procedure to additionally weight proxies by the calibration period correlation with the corresponding actual gridbox temperature series (as used in earlier experiments described above, and as used in MJ03) can potentially ameliorate the overestimation bias.

b. CFR method

We make use of the hybrid frequency-domain version of the RegEM algorithm in the CFR experiments described below. Other versions of the RegEM algorithm (straight RegEM and stepwise RegEM) yielded similar results (see supplementary table at http://dx.doi.org/10.1175/jcli3564.s4). As with the CPS experiments, SNR = 1.0 (Fig. 6a) yielded a similar verification resolved variance (60%–80% depending on calibration period used) to that observed for actual proxy reconstructions. As in the CPS experiments, there is a slight sensitivity to which (short or long) calibration period is used, with a long calibration period (as used by Rutherford et al. 2005) yielding a moderately more skillful reconstruction (Table 1).

There is a surprising insensitivity, however, of the low-frequency features of the reconstruction to the precise value of SNR in the CFR experiments; for no SNR value was there any evidence of a systematic underestimate of low-frequency variability (Fig. 6b), a result that conflicts dramatically with the claims of von Storch et al. (2004). As discussed above, we suspect that the reason for the difference lies, at least partly, in certain unrealistic features of the GKSS simulation used by VS04. The possible additional impacts of the use of differing CFR methods are also currently being explored (AWT05).

The CFR approach does appear to systematically underestimate the amplitude of the larger volcanic cooling events, a finding that is unsurprising. The climatic response to volcanic forcing represents a particular challenge for the CFR method, since only a small number of moderate volcanic forcing events are contained within the calibration interval, and these events are quite small in comparison with the much larger events present in the precalibration interval. The spatial patterns of response to volcanic forcing are consequently not well captured in the calibration period. CFR approaches have been shown to yield a systematic underestimate of variance under precisely such circumstances (Rutherford et al. 2003). This bias, as expected, becomes more prominent at lower values of SNR.

4. Conclusions

We find no evidence for the suggestion (e.g., VS04) that real-world proxy-based temperature reconstructions are likely to suffer from any systematic underestimate of low-frequency variability. Our findings suggest that both standard methods that have been used in proxy-based reconstruction (CPS and CFR) are likely to provide a faithful estimate of actual long-term hemispheric temperature histories, within estimated uncertainties. On the other hand, apparent complementary strengths and weaknesses emerge in comparing the two approaches. The CFR approach systematically underestimates high-frequency changes associated with the cooling response to explosive volcanism but shows no evidence of any systematic bias in the reconstructed low-frequency variability. The CPS approach, by contrast, shows no tendency to underestimate volcanic cooling but does exhibit a tendency to moderately underestimate long-term variations in the mean, although this becomes significant only at signal-to-noise ratios lower than is inferred for real-world proxy data. We encourage similar future experiments using coupled ocean–atmosphere models and appropriate experimental designs for simulating proxy-based climate reconstruction approaches.

Acknowledgments. This research was partly supported (MEM and SR) by the NOAA- and NSF-supported Earth Systems History program. The model simulations were conducted under the NCAR Weather and Climate Impacts Assessment Science Initiative by CMA and ERW, who acknowledge contributions from C. Tebaldi, L. Mearns, F. Joos, D. Schimel, and B. Otto-Bliesner.

REFERENCES

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.

Briffa, K. R., and P. D. Jones, 1993: Global surface air temperature variations during the twentieth century: Part 2, implications for large-scale high-frequency palaeoclimatic studies. *Holocene,* **3,** 77–88.

——, 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.

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., 2000: Causes of climate change over the past 1000 years. *Science,* **289,** 270–277.

——, 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.

——, S. K. Baum, K.-Y. Kim, G. C. Hegerl, and W. T. Hyde, 2003: Modeling ocean heat content changes during the last millennium. *Geophys. Res. Lett.,* **30,** 1932, doi:10.1029/2003GL017801.

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.

——, D. C. Frank, R. J. Wilson, and K. R. Briffa, 2005: Effect of scaling and regression on reconstructed temperature amplitude for the past millennium. *Geophys. Res. Lett.,* **32,** L07711, doi:10.1029/2004GL021236.

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

Foukal, P., G. North, and T. Wigley, 2004: A stellar view on solar variations and climate. *Science,* **306,** 68–69.

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. Meteor.,* **10,** 845–864.

Goosse, H., T. J. Crowley, E. Zorita, C. M. Ammann, H. Renssen, and E. Driessschaert, 2005: Modeling the climate of the last millennium: What causes the differences between simulations? *Geophys. Res. Lett.,* **32,** L06710, doi:10.1029/2005GL022368.

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

——, 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.

Lean, J. L., Y.-M. Wang, and N. R. Sheeley Jr., 2002: The effect of increasing solar activity on the Sun's total and open magnetic flux during multiple cycles: Implications for solar forcing of climate. *Geophys. Res. Lett.,* **29,** 2224, doi:10.1029/2002GL015880.

Luterbacher, J., and Coauthors, 2002: Reconstruction of sea level pressure fields over the Eastern North Atlantic and Europe back to 1500. *Climate Dyn.,* **18,** 545–561.

Mann, M. E., 2004: On smoothing potentially non-stationary climate time series. *Geophys. Res. Lett.,* **31,** L07214, doi:10.1029/2004GL019569.

——, and S. Rutherford, 2002: Climate reconstruction using 'pseudoproxies.' *Geophys. Res. Lett.,* **29,** 1501, doi:10.1029/2001GL014554.

——, and P. D. Jones, 2003: Global surface temperature over the past two millennia. *Geophys. Res. Lett.,* **30,** 1820, doi:10.1029/2003GL017814.

——, R. S. Bradley, and M. K. Hughes, 1998: Global-scale temperature patterns and climate forcing over the past six centuries. *Nature,* **392,** 779–787.

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., and K. R. Briffa, 2004: The real color of climate change? *Science,* **306,** 621–622.

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

——, ——, 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. Climate,* **18,** 2308–2329.

Schneider, T., 2001: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Climate,* **14,** 853–871.

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.

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

Zorita, E., F. Gonzalez-Rouco, and S. Legutke, 2003: Testing the Mann et al. (1998) approach to paleoclimate reconstructions in the context of a 1000-yr control simulation with the ECHO-G coupled climate model. *J. Climate,* **16,** 1378–1390.