Proxy-based Northern Hemisphere Surface Temperature Reconstructions: Sensitivity to Method, Predictor Network, Target Season, and Target Domain

Rutherford, S.[1*], Mann, M.E.[2], Osborn, T.J.[3], Bradley, R.S.[4], Briffa, K.R.[3], Hughes, M.K.,[5] Jones, P.D.[3]

1. Graduate School of Oceanography, University of Rhode Island Narragansett, R.I.
2. Dept. of Environmental Sciences, University of Virginia, Charlottesville, VA.
3. Climatic Research Unit, University of East Anglia, Norwich, U.K.
4. Dept. of Geosciences, University of Massachusetts, Amherst, MA.
5. Laboratory of Tree-Ring Research, University of Arizona, Tucson, AZ.

*Now At: Dept. of Environmental Sciences, Roger Williams University, Bristol, R.I.

Submitted to *Journal of Climate* July 23, 2003
Revision submitted to *Journal of Climate* June 29, 2004

ABSTRACT:

Results are presented from a set of experiments designed to investigate factors that may influence proxy-based reconstructions of large-scale temperature patterns in past centuries. The factors investigated include (1) the method used to assimilate proxy data into a climate reconstruction, (2) the proxy data network used, (3) the target season, and (4) the spatial domain of the reconstruction. Estimates of hemispheric-mean temperature are formed through spatial averaging of reconstructed temperature patterns that are based on either the local calibration of proxy and instrumental data or a more elaborate multivariate climate field reconstruction approach. The experiments compare results based on the global multi-proxy data set used by Mann and co-workers, with results obtained using the extratropical Northern Hemisphere (NH) maximum latewood tree-ring density set used by Briffa and co-workers. Mean temperature reconstructions are compared for the full NH (tropics and extratropics, land and ocean), and extratropical continents only, and varying target season (cold-season half year, warm-season half year, and annual mean). The comparisons demonstrate dependence of reconstructions on seasonal, spatial, and methodological considerations, emphasizing the primary importance of the target region and seasonal window of the reconstruction. The comparisons support the generally robust nature of several, previously published, estimates of NH mean temperature changes in past centuries, and suggest that further improvements in reconstructive skill are most likely to arise from an emphasis on the quality, rather than quantity, of available proxy data.

1. Introduction:

Evidence of how climate has changed in past centuries can inform our assessment of the anthropogenic role in observed 20th century warming (e.g. Folland et al., 2001). The lack of widespread instrumental surface temperature estimates prior to the mid 19th century (e.g. Jones et al., 1999) places particular emphasis on the need to reconstruct the history of climate changes accurately, which can only be achieved via the careful use of long-term empirical evidence. Such empirical evidence comes from "proxies" of climate variability derived from the environment itself and from documentary evidence (Le Roy Ladurie, 1971; Wigley et al., 1981; Crowley and North, 1991; Bradley and Jones, 1995; Bradley, 1999, Jones et al., 2001a).

Particularly useful in this context are high-resolution (annually or seasonally-resolved) proxies such as tree rings (e.g. Fritts et al., 1971; Fritts, 1991; Briffa et al., 1994;1998ab;2001), corals (e.g. Evans et al., 2002; Hendy et al., 2002), ice cores (O'Brien et al., 1995; Appenzeller et al., 1998; Meeker and Mayewski, 2002), lake sediments (Hughen et al., 2000), and long documentary and instrumental series (Pfister et al., 1998; Luterbacher et al., 1999), all of which may be combined into "multi-proxy" assemblages (Bradley and Jones, 1993; Overpeck et al., 1997; Mann 2002ab; Mann et al., 1998; Crowley and Lowery, 2000; Folland et al., 2001; Jones et al., 1998, 2001a; Cook et al., 2002; Luterbacher et al., 2002). A critical advantage of using such high-resolution proxy data is the possibility of comparing the proxies against long temporally-overlapping instrumental records both to estimate the climate signal in the data (calibration) and independently test the reliability of the signal (verification or cross-validation).

Annually-resolved proxy indicators have been used to reconstruct spatial climate

fields such as sea level pressure (SLP) (Fritts, 1991; Luterbacher et al., 2002; Meeker and Mayewski, 2002), terrestrial surface air temperature (SAT) (Briffa et al., 1994;1998a; 2002b), continental drought (Cook et al., 1999), sea surface temperature (SST) (Evans et al, 2002) and the combined global SAT/SST temperature field (Mann et al., 1998;1999). These reconstructed fields have been spatially-averaged to yield estimates of hemispheric mean temperature (e.g. Osborn et al., 2004; Mann et al., 1998;1999) or circulation/SST indices such as the Niño3 index of El Niño/Southern Oscillation (ENSO) (Mann et al., 2000a,b) and the North Atlantic Oscillation (NAO, Luterbacher et al., 2001, 2002; Cook, 2002). Unlike hemispheric mean reconstructions, spatial field reconstructions retain vital information which can provide insight into the mechanisms or forcing underlying observed variability (e.g. Briffa et al., 1994, 2002a,b; Cook et al., 1997; Delworth and Mann, 2000; Shindell et al., 2001; Waple et al., 2002; Braganza et al., 2002).

Annually-resolved proxy networks have also been used to directly reconstruct indices of climate variability such as the North Atlantic Oscillation (NAO) (D'Arrigo et al., 1993; Appenzeller et al., 1998; Cullen et al., 2001; Mann, 2002b; Cook et al., 2002), the Pacific Decadal Oscillation (PDO) (Biondi et al., 2001; Gedalof et al., 2002), ENSO [including the Niño3 (Mann et al., 2000a,b) and Southern Oscillation (Stahle et al., 1998) indices], and hemispheric mean temperature series (Jacoby and D'Arrigo, 1989; Bradley and Jones, 1993; Overpeck et al., 1997; Jones et al., 1998; Briffa et al., 1998a; 2001; 2002a; Crowley and Lowery, 2000; Mann and Jones, 2003). Such approaches are potentially limited by the assumed relationship between local variables recorded by the proxies (temperature and precipitation) and larger-scale climate patterns, since the relationship between local and large-scale influences may change over time (e.g. Jones et

al., 2003b).

Of particular interest in this study are various recent reconstructions of NH temperature from proxy data networks (Bradley and Jones, 1993; Overpeck et al., 1997; Briffa et al., 1998a,b;2001; Jones et al., 1998; Mann et al., 1998;1999; Mann, 2002a; Crowley and Lowery, 2000). Most reconstructions show notable overall similarity (Mann, 2000; 2001; 2002a; Briffa and Osborn, 2002; Jones et al., 1998, 2001a; Folland et al., 2001; Mann and Jones, 2003; Mann et al., 2003a,b). For example, the late 20th century warmth is unprecedented in the context of the past 1000 years in all reconstructions given the published estimates of uncertainty in the reconstructions (e.g. Folland et al., 2001; Jones et al., 2001a; Mann et al., 2003b; Cook et al., 2004). In addition, the empirical reconstructions generally show considerable similarity to independent climate model simulations (Free and Robock, 1999; Crowley, 2000; Shindell et al., 2001; Gerber et al., 2003; Bertrand et al., 2002; Bauer et al., 2003), with isolated exceptions (Gonzalez-Rouco et al., 2003).

Some differences do exist, however, among hemispheric temperature reconstructions, with certain reconstructions (e.g. Esper et al., 2002) indicating greater peak cooling in past centuries than others (see also Briffa and Osborn, 2002; Mann and Hughes, 2002; Mann, 2002a; Mann et al., 2003b). It is important to try to understand the sources of the differences between the various NH temperature reconstructions. This undertaking is complicated by the fact that several distinct factors in varying combinations could be responsible for the differences between reconstructions. One factor (1) is the method employed to assimilate the information from proxy data networks into a reconstruction of past climate. The simplest method is to construct an unweighted

average of a set of "standardized" proxy series believed to represent a particular quantity (e.g. temperature, or an index of ENSO). The single composite series can then be scaled against an appropriate target index. For example, a composite of proxy indicators known (or assumed, or shown by correlation) to reflect local surface temperatures can be scaled against the instrumental northern hemispheric mean temperature record during the period that proxy and instrumental data overlap. The scaled series is then interpreted as an NH mean temperature reconstruction based on the proxy data (e.g. Bradley and Jones, 1993; Jones et al., 1998; Crowley and Lowery, 2000; Mann and Jones, 2003). Similarly, one can composite indicators believed to be sensitive to ENSO and scale the composite to the instrumental SOI index to yield an SOI reconstruction (Stahle et al., 1998). Alternatively, a large number of local or regional regressions between proxy indicators and instrumental data can be used to build up a reconstruction of an entire field. Such "local calibration" approaches assume a local relationship between predictor (e.g. maximum tree-ring latewood density) and climate variable (e.g. summer surface air temperature) (e.g. Briffa et al., 1998a; 2001;2002a,b).

A more elaborate approach is to use a Climate Field Reconstruction (CFR) technique (see Smith et al., 1996; Kaplan et al., 1997; Schneider, 2001; Mann and Rutherford, 2002; Rutherford et al., 2003) to reconstruct a large-scale field from a proxy data network through multivariate calibration of the large-scale information in the proxy data network against instrumental data (see also Fritts et al., 1971; Guiot, 1985;1988; Fritts, 1991; Cook et al., 1994; Mann et al., 1998; 1999; 2000a,b; Mann and Rutherford, 2002; Luterbacher et al., 2002a,b; Evans et al., 2002; Pauling et al., 2003). The CFR approach does not assume any *a priori* local relationship between proxy indicator and the

climatic field being reconstructed. For example, a proxy sensitive to convection/rainfall in the central tropical Pacific (indicative of ENSO variability) can be used to calibrate the surface temperature patterns associated with ENSO even though the proxy itself is not related to local temperature. In this manner, a large-scale climate field can often be efficiently reconstructed through CFR techniques from a relatively modest network of indicators (e.g. Bradley, 1996; Evans et al., 1998, Mann and Rutherford, 2002; Zorita et al., 2003). Such methods arguably depend more heavily on assumptions about the stationarity of relationships between proxy indicators and large-scale patterns of climate variability than the "local calibration" approach. Model experiments suggest this probably is not problematic for the range of variability inferred for recent past centuries (Rutherford et al., 2003). Reconstructions of the more distant past (e.g. the mid-Holocene—Bush, 1999; Clement et al., 2000) would require, however, a more careful consideration of stationarity issues.

A second complicating factor (2) in comparing different reconstructions involves the potentially different character of the proxy network used to produce the reconstruction. Some proxy networks consist of only one specific type of proxy information (e.g. tree-ring maximum latewood density measurements—Briffa et al., 1998a,b;2001), while other "multiproxy" networks combine several types of proxy information (tree ring width and density measurements, indicators derived from corals, ice cores, lake sediments, and historical documentary records—Mann et al., 1998;1999;2000a,b). Proxy networks can differ in their sensitivity to specific meteorological variables. Extratropical high-latitude tree-ring networks typically provide warm-season temperature information, while tree-ring information from lower-latitude

semi-arid/Mediterranean or tropical environments, corals, ice cores, and documentary records provide variable seasonal information regarding distinct climate variables. In addition, proxy networks often differ dramatically in the number of proxies used, ranging from a handful of very long proxies (Jones et al., 1998; Crowley and Lowery, 2000; Esper et al., 2002) to a potentially much larger (hundreds—e.g., Mann et al., 1998;1999; Briffa et al., 1998a,b;1999;2001) but temporally-variable set of proxies. The sampling error in hemispheric estimates based on the latter is likely to be smaller than that in the former, but errors will increase back in time leading to expanding uncertainties in earlier periods (e.g. Mann et al., 1999; Jones et al., 2001a).

An additional factor (3) is the target season of the reconstruction (annual mean, boreal warm season or boreal cold season), as discussed by Briffa and Osborn (2002) and Jones et al. (2003a). The target season is, to some extent, constrained by the particular proxies used. However, it is not always possible to know the precise mix of seasonal information in the proxy network *a priori*. This is particularly true in large-scale CFR where a precipitation-sensitive proxy may, for example, be an important predictor of a large-scale temperature pattern, as discussed above. In this case, the optimal target seasonal window can nonetheless be evaluated through calibration and cross-validation exercises (e.g. Mann et al., 2000b).

A final related factor (4) is the target region of the reconstruction. Tropical SSTs are, for example, typically less variable than extratropical, continental surface air temperatures, so a reconstruction targeting the entire NH (land and ocean, tropics and extratropics—e.g. Jones et al., 1998; Mann et al., 1999; Crowley and Lowery, 2000; Mann and Jones, 2003) is likely to yield smaller amplitude variability than one targeting

extratropical continental regions only (e.g. Briffa et al., 1998a,b; 1999;2001; Esper et al., 2002).

One approach used to compare reconstructions based on proxy data with different seasonal or regional emphases is to rescale the reconstruction against an appropriate target index. For example, a reconstruction based on extratropical land-only proxies might still be rescaled to the full NH instrumental mean series, or a reconstruction based on annual proxies might still be scaled to a warm-season instrumental hemispheric mean series (Briffa and Osborn, 2002). However, there are some pitfalls to this approach. Any similarity between patterns of temperature change in different seasons and regions over the instrumental record may be relatively unique to the late 19th and 20th century. Seasonal temperature trends show greater differences in prior centuries (see e.g. Jones et al., 2003a; Luterbacher et al., 2004), and pre-anthropogenic, natural forcing appears to have a different spatial and seasonal temperature signature from anthropogenic forcing (Shindell et al., 2003). Moreover, although tropical and extratropical temperature trends are similar during the instrumental period, there is some evidence that they may have been quite different in past centuries (Hendy et al., 2002; Cobb et al., 2003). More sophisticated approaches to dealing with differing seasonal and spatial emphases are thus preferable.

The intent of this study is to provide a systematic assessment of the relative impacts of these four factors on published large-scale surface temperature reconstructions. We do this based on the use of two different reconstruction techniques: (a) local calibration, and (b) large-scale CFR. We analyze two nearly independent networks of predictors, one that is globally extensive (land and ocean, tropical and

extratropical), represents various seasons, and consists of multiple proxy types ("multiproxy"—Mann et al., 1998), and another which is extratropical and terrestrial, is reflective primarily of warm-season conditions, and based entirely on Maximum Latewood Tree-Ring density ("MXD"-Briffa et al., 1998a,b, 2001, 2002a,b). For each of the two proxy networks, reconstructions are performed for three different target seasons (annual mean, boreal cold-season and boreal warm-season), and the resulting NH mean temperature reconstructions are compared based on averages over distinct spatial domains (full NH land and ocean, extratropical land regions only). Additional insights are obtained from comparisons with other published NH temperature reconstructions (Mann et al., 1998;1999; Briffa et al., 1998a,b;2001; Osborn et al., 2004; Esper et al., 2002).

2. Data

A. Instrumental Surface Temperature Data

We use the 5° latitude by 5° longitude Climatic Research Unit (CRU) grid-box surface temperature dataset available from 1856-present to calibrate and reconstruct the surface temperature field from proxy data networks (Note: The surface temperature data are available at http://fox.rwu.edu/~rutherfo/supplements/jclim2003a). The data consist of surface air temperature (SAT) over land and sea surface temperature (SST) over the oceans (Jones et al., 1999, 2001b; Jones and Moberg, 2003). We use the HadCRUT dataset (Jones et al., 2001b) rather than the more recently published HadCRUT2 (Jones and Moberg, 2003) version. We use the restricted period 1856-1971 for calibration as discussed below, but consider the resulting reconstructions in the context of the entire record (1856-1998). This instrumental surface temperature dataset exhibits some

differences from the older instrumental surface temperature dataset (Jones et al., 1994) used by Mann et al. (1998).

We averaged the monthly mean data into annual (calendar), boreal warm (April to September) and cold (October to March) seasonal averages, the target seasons for subsequent reconstructions. In the case of cold-season averages, our convention is to designate the year as corresponding to the early (October-December) rather than late (Jan-Mar) half of the 6 month interval**.** For example, the 1815 cold season is October 1815 through March 1816. The averaged (and raw) data are both temporally and spatially incomplete due to a lack of available data at a given location and time. In particular, the data coverage during the 19th Century is relatively sparse compared to that of the latter half of the 20th Century. To produce a complete instrumental field we infilled the missing instrumental values for seasonal and annual mean values using the RegEM method as described by Schneider (2001) [see also Mann and Rutherford, 2002; Rutherford et al., 2003; Zhang et al., 2004]. Only the Northern Hemisphere data (grid centers at 2.5°N through 67.5°N: the instrumental data are extremely sparse poleward of 67.5°N) were used in this study. Spatial means, including the NH mean, are constructed from areally-weighted averages of the grid box data. The correlation between the time series of the NH mean based on available data only and the RegEM infilled field is r=0.98 (over the 1856-1971 period).

B. Proxy Data

We used two largely independent predictor networks to assess the sensitivity of the temperature reconstructions to the network used. The first of these is a "multiproxy" dataset used by Mann and co-workers (Mann et al. 1998, 1999, 2000a,b; Mann 2002b) to

reconstruct global patterns of annual mean surface temperature (SAT over land and SST over ocean) in past centuries. The second data set consists entirely of tree-ring maximum latewood density ("MXD") data used by Briffa and coworkers (Briffa et al., 1998a,b, 2001, 2002a,b, Osborn et al., 2004) to reconstruct extratropical terrestrial warm-season SAT. Strictly speaking, the two networks are not entirely independent, because they share a small number of tree-ring density series (19, or, 4.6% of the 415 series used by Mann et al. (1998) and 6% of the 387 series contributing to the MXD network, are common to both networks). In addition, many of the ring-width series from Russia used by Mann et al. (1998) were from sites for which density data were used by Briffa and coworkers. We also prepared a third "combined" network by combining both networks.

i. Multiproxy/PC dataset

The multiproxy/PC network (Mann et al., 1998) is a combination of annually-resolved proxy indicators including tree ring chronologies (ring width and density), ice cores (stable isotope, ice melt and ice accumulation data), coral records (stable isotope and fluorescence data) and long historical and instrumental records (temperature and precipitation) from the tropics and extratropics of both hemispheres. (Note: The data in the multiproxy/PC network are available at http://fox.rwu.edu/~rutherfo/supplements/jclim2003a.) The individual proxies in the network were chosen not for their reliability as local indicators of temperature, but for their potential relationship with some seasonal meteorological or climatic variable tied to larger-scale patterns of climate, and surface temperature, change. In areas with spatially dense tree-ring networks, principal components analysis was used to extract the leading principal components (PCs) from the network. Although 415 individual proxy series

were used, data reduction by using leading PCs of tree-ring networks results in a smaller set of 112 indicators in the Multiproxy/PC network available back to 1820 (Figure 1a), with a decreasing number of indicators available progressively farther back in time. Twenty two of the indicators (representing 95 individual proxy series) extend back to at least AD 1400. Many of the indicators in the network end at or near 1980, motivating a termination of the calibration interval at 1980 by Mann et al. (1998), with a modest subset of series terminating between 1971 and 1980 infilled by persistence of the final available value through to 1980. We terminate the calibration period at 1971 in this study to avoid any possible influence of the infilling process used by Mann et al. (1998).

It should be noted that some falsely reported putative "errors" in the Mann et al. (1998) proxy data claimed by McIntyre and McKitrick (2003) are an artifact of (a) the use by these latter authors of an incorrect version of the Mann et al. (1998) proxy indicator dataset, and (b) their misunderstanding of the methodology used by Mann et al. (1998) to calculate PC series of proxy networks over progressively longer time intervals. In the Mann et al. (1998) implementation, the PCs are computed over different time steps so that the maximum amount of data can be used in the reconstruction. For example, if a tree-ring network comprises 50 individual chronologies that extend back to AD 1600 and only 10 of those 50 extend to AD 1400 then calculating one set of PCs from 1400 to 1980 (the end of the Mann et al. (1998) calibration period) would require the elimination of 40 of the 50 chronologies available back to AD 1600. By calculating PCs for two different intervals in this example (1400-1980 and 1600-1980) and performing the reconstruction in a stepwise fashion, PCs of all 50 series that extend back to AD 1600 can be used in the reconstruction back to AD 1600 with PCs of the remaining 10 chronologies used to

reconstruct the period from 1400-1600. The latter misunderstanding led McIntyre and McKitrick (2003) to eliminate roughly 70% of the proxy data used by Mann et al. (1998) prior to AD 1600, including 77 of the 95 proxy series used by Mann et al. (1998) prior to AD 1500. This elimination of data gave rise to spurious, anomalous warmth during the 15th century in their reconstruction, sharply at odds with virtually all other empirical and model-based estimates of hemispheric temperature trends in past centuries (see e.g. Jones and Mann, 2004).

ii. Maximum Latewood Density (MXD) Dataset

The MXD network (Briffa et al., 2001; 2002a,b) is primarily a reflection of growing (warm-season) conditions, though some limited cold-season information is also apparent in the data (Briffa et al., 2002a). The version of the MXD dataset used here was compiled using a combination of grid-box estimates based on traditionally-standardized MXD records (with limited low-frequency information) and regional estimates developed to retain low-frequency information (Osborn et al., 2004) (Note: The data in the MXD network are available at http://fox.rwu.edu/~rutherfo/supplements/jclim2003a.). The latter were developed using the Age-Band Decomposition (ABD) method of standardization, wherein density data from trees of similar ages are averaged to create long chronologies with minimal effect of tree age and size (Briffa et al., 2001). The ABD method is designed to preserve low frequency information in tree ring data that may be reduced when more traditional methods to remove long-term growth trends are used (see Cook et al., 1995). Because the age-banding method requires large numbers of samples throughout the time period being studied, it has been applied only at a regional scale for the MXD network used here, rather than at the level of the 387 original site chronologies.

Osborn et al. (2004) therefore worked first with the traditionally-standardized data at the individual chronology scale, gridded them to provide values in 115 5° by 5° grid boxes (26 available back to AD 1400) in the extratropical NH (Figure 1b). They then developed temperature reconstructions by the local calibration of the MXD grid box data against the corresponding instrumental grid box temperatures. The "missing" low-frequency temperature variability was then identified as the difference between the 30-year smoothed regional reconstructions of Briffa et al. (2001) and the corresponding 30-year smoothed regional averages of the gridded reconstructions. Osborn et al. (2004) add this "missing" low frequency variability to each grid box in a region. After roughly 1960, the trends in the MXD data deviate from those of the co-located instrumental grid-box SAT data for reasons that are not yet understood (Briffa et al., 1998b; 2003; Vaganov et al., 1999). To circumvent this complication, we use only the pre-1960 instrumental record for calibration/cross-validation of this dataset in the CFR experiments.

3. CFR Reconstruction Method

A. RegEM Approach

Various mathematical techniques have been applied to the problem of CFR from sparse data (Smith et al., 1996; Kaplan et al., 1997; Schneider 2001), including applications to paleoclimate field reconstruction (Cook et al., 1994; Mann et al., 1998; Luterbacher et al., 2002a,b; Evans et al., 2002). Here we use the Regularized Expectation Maximization (RegEM) method described by Schneider (2001), which offers several theoretical advantages over other methods of CFR. (Note: Matlab scripts are available at http://fox.rwu.edu/~rutherfo/supplements/jclim2003a.) The RegEM method is an

iterative method for estimating missing data through the estimation of means and covariances from an incomplete data field to impute missing values in a manner that makes optimal use of the spatial and temporal information in the dataset. When a reconstruction is sought from proxy data based on calibration against modern instrumental measurements, the combined (proxy-plus-instrumental dataset) can be viewed as an incomplete data matrix, which contains both instrumental data (surface temperature grid-box values arranged with rows representing the years and columns representing grid boxes) and proxy data (proxy time series with rows representing the years and columns representing the proxy used). The columns of the matrix (i.e., the instrumental grid point data and proxy indicators) are standardized to have zero mean and unit standard deviation over the calibration interval.

Missing values in this matrix represent the unknown pre-instrumental surface temperature values, and are considered as values to be imputed through an iterative infilling of the data matrix making use of the covariance information between all available (instrumental and proxy) data. By analogy with conventional paleoclimate reconstruction approaches (see e.g. Rutherford and Mann, 2003), a calibration interval can be defined as the time interval over which the proxy and instrumental data overlap, while a verification interval is defined by additional cross-validation experiments in which an appropriate subset of the available instrumental data are withheld from the calibration process (e.g., through their specification as missing values in the initial matrix). Schneider (2001) provides a detailed description of the RegEM algorithm, including a comparison with conventional methods such as principal components regression, and application to the infilling of missing values in climate field data, while

Rutherford et al. (2003), Mann and Rutherford (2002), and Zhang et al. (2004) discuss specific applications to paleoclimate reconstruction. The RegEM method has been shown to perform well even in the presence of non-stationary climate forcing, as long as the leading patterns underlying low-frequency variability are captured in calibration (Rutherford et al., 2003).

We have modified the application of the method in two ways to improve its performance for long-term CFR. This includes first, implementing a stepwise approach, where we reconstruct the field back in time in discrete steps to accommodate changing availability of data and second, incorporating a hybrid frequency domain approach where both the proxies and the instrumental calibration data are decomposed into two frequency bands prior to reconstruction. These modifications are discussed below.

B. Stepwise Modification of RegEM

The RegEM approach was in all cases applied in a stepwise fashion back in time to make increasingly better use of low-frequency information in the calibration process back in time. The reconstruction is performed one step at a time, using all available climate field information (both instrumental field and proxy-reconstructed extension thereof) in the calibration process for the reconstruction of each subsequent step back in time. For example, in the first reconstruction step the (infilled) instrumental data are available from 1856 to 1971 and the proxies extend back to AD 1400. This leaves 455 years in which the entire 1008 NH grid boxes (2.5°N to 67.5°N at 5° centers) are "missing". Rather than reconstruct all 455 years at one step, we first reconstruct 1800-1855, producing a complete NH field (1008) grid boxes from 1800 to 1971. We then use the completed 1800-1971 data as input into the next step with the proxies extending back

to 1700. In this step the interval 1700-1799 is reconstructed. The process continues until the reconstruction is complete back to the targeted beginning date (in this case, AD 1400). In the case of the Multiproxy/PC and combined networks, the step lengths are constrained by the network because the PCs of the dense tree-ring networks are recalculated over discrete time intervals. Thus the Multiproxy/PC and combined networks requires some type of stepwise approach with step lengths dictated by the calculation of the PCs. For consistency with the Mann et al. (1998) approach we use the same step lengths here for both the Multiproxy/PC and Combined networks. The MXD network has no such constraints and we chose a step length of 100 years, but the results are insensitive to the exact step length chosen.

Due to the shortness of the instrumental record, one can not gauge the relative performance of the stepwise vs. non-stepwise approaches through cross-validation experiments using the actual instrumental record. Instead, we used a network of synthetic proxy data ("pseudoproxy") data derived from long control and forced integrations of the Geophysical Fluid Dynamics Laboratory's R30 coupled ocean-atmosphere model (Knutson et al., 2000) to test the relative performance of the two methods. We used the approach described by Mann and Rutherford (2002) to derive networks of synthetic proxy data from the model surface temperature field. In these tests 450 years of the control run were combined with 150 years of the forced run to create a continuous and complete temperature field qualitatively similar in character to reconstructed temperature histories over the past six centuries. We constructed 112 pseudo-proxies (the same number as is in the multiproxy/PC indicator network back to 1820), from the modeled temperature field, and selected an increasingly sparse subset of

the 112 indicators back in time to emulate the decrease in the size of the actual proxy networks back in time. The final 150 years were used for calibration to reconstruct the preceding 450 years using only the information available in the pseudo-proxy network. The stepwise approach performed as well as or better than the non-stepwise approach in cross-validation in each of these experiments. The results of these pseudoproxy experiments give us some confidence that the primary conclusions presented in this study are insensitive to whether the stepwise or non-stepwise approach is used.

C. Hybrid Frequency-Domain Modification of RegEM

We modified the RegEM method (Schneider, 2001; Mann and Rutherford, 2002; Rutherford et al., 2003) to employ a hybrid frequency-domain calibration approach, in which the combined proxy/instrumental data set is split into two distinct data sets, through application of a lowpass filter to the data. The lowpass component of the data defines the low-frequency component, while the residual defines the high-frequency component. The frequency-split boundary can be varied arbitrarily, but reasonable constraints on the appropriate choice are, at the high-frequency end, the Nyquist frequency (f=0.5 cycle/yr for annual or seasonal mean data) divided by two or so (i.e., f=0.25 cycle/yr) and, at the low-frequency end, the Rayleigh frequency (f=0.01 cycle/yr for e.g. 100 years of data) multiplied by two or so (i.e., f=0.02 cycle/yr). This corresponds to a high-frequency/low-frequency band split at periods between 4 years, and 50 years for a 100 year interval. As described below, cross-validation experiments motivate the choice f=0.05 cycle/yr (20 year period) for the split frequency in almost all cases.

There are two primary motivations for the hybrid frequency-domain approach. Different types of proxy data exhibit fundamentally different frequency-domain fidelity characteristics (Jones et al., 1998). Conventionally standardized tree-ring data, if based on short constituent segments are unlikely to resolve century or longer timescale variability (e.g. Cook et al., 1995), while very conservatively standardized tree-ring data based on long constituent segments may resolve century-scale and longer variability (Briffa et al., 1996; Cook et al., 1995). Other proxy indicators, such as annually-laminated lake sediments or ice core variables subject to diffusion (Fisher et al., 1996), may preferentially resolve decadal and lower-frequency variability (e.g. Bradley, 1999). Furthermore, some proxies may themselves exhibit different climate responses at low and high frequencies (e.g. LaMarche, 1974; Osborn and Briffa, 2000; Hughes and Funkhouser, 2003). The underlying patterns of climate variability may also exhibit timescale dependence. Interannual timescale variability may be dominated by processes such as ENSO and the NAO, while lower-frequency variability may be dominated by modes involving the overturning ocean circulation (e.g. Delworth and Mann, 2000) or the response to global radiative forcing (e.g. Crowley, 2000). Distinguishing between patterns of high- and low-frequency variability may thus provide a more efficient means of calibration of the large-scale patterns of climate variability, and permit the use of a wider range of natural archives.

Our hybrid frequency-domain calibration approach involves the use of two distinct frequency bands in the calibration process. In the limit of an increasingly large number of distinct frequency bands, this approach would become analogous to the spectral canonical regression approach described by Guiot (1985), in which the

calibration process is performed explicitly in the frequency domain rather than the time domain. In such a case, however, the small number of statistical degrees of freedom in calibrating the lowest-frequency bands of variance leads to a poorly constrained characterization of variability in the lowest-frequencies. Employing a two-band hybrid calibration approach represents a tradeoff between the ability to adequately distinguish distinct patterns of variability with respect to timescale, and yet retain adequate statistical degrees of freedom to characterize and calibrate both bands of variability.

The RegEM method is applied separately to the calibration of proxy and instrumental data in the high- and low-frequency bands. The results of the two independent reconstructions are then recombined to yield a complete reconstruction. Each proxy record is weighted by a bandwidth retention factor defined as the percent of its total variance within the particular frequency band under consideration. For example, a proxy record dominated by interannual variability, with very little low-frequency variability (e.g. a very data-adaptive standardized tree-ring record) would be assigned a high weight (near one) in the high-frequency band calibration and a low weight (near zero) in the low-frequency band calibration. Generally, all proxy series have weights between zero and one in each frequency band with greater weight in the frequency band with the greatest concentration of variance in the unfiltered series. This approach ensures that, for example, a proxy with a small amount of variability in the low-frequency band (which might be residual noise) does not have the same impact as a proxy with much greater low-frequency variability. However, it has the disadvantage that a high-frequency-dominated record containing a nonetheless faithful record of low-frequency

fluctuations (e.g. an indicator of ENSO wherein the interannual variability is intrinsically dominant) might be unduly discounted.

D. Experimental Approach

We performed an array of RegEM CFR experiments based on different target seasons and proxy networks, and tested variants of the approach including (a) both conventional and hybrid frequency-domain approach, the latter with varying split frequency, (b) allowing and not allowing for lags between predictor (proxy indicator) and predictand (instrumental surface temperature data) and (c) use of both pre-whitened and raw predictor/predictand data (see e.g. Cook et al., 1999; Zhang et al., 2004). We compared our results against previous reconstructions based on common predictor datasets (multiproxy/PC and MXD), and alternative reconstruction methodologies [the eigenvector-based CFR approach of Mann et al. (1998) and the local-calibration approach used by (Osborn et al., 2004), respectively] to assess the impact of using different reconstruction methodologies with common data. We areally-averaged spatial reconstructions over both the full NH domain and sub-regions of the domain (e.g. extratropics and/or continents only) to examine the sensitivity of NH "hemispheric mean" estimates to the actual region sampled.

CFR experiments were performed using each of the three proxy networks, Multiproxy/PC, MXD and Combined, three seasonal target windows for the surface temperature predictand (boreal warm season, boreal cold season, and annual calendar mean). In the hybrid frequency domain approach, alternative frequency boundaries were

tested (5, 10, 20 and 25 year period) within the practical constraints discussed in section 3c.

We also performed experiments in which the proxy indicators were lagged (both forward and backward) relative to the instrumental data, under the assumption that some proxies may reflect, at least in part, a lagged or running average response to climate. We lagged the proxy data at -1, 0 and +1 years both independently and in various combinations (e.g. -1 and 0 only, +1 and 0 only, and -1,0, and +1). Use of lagged versions of the proxy indicator network in addition to the nominal network itself increases the effective size of the predictor network. Including the proxy network at both lag 0 and at lag -1, for example, produces a maximum predictor network of 224 indicators (twice the nominal maximum of 112 indicators) for the multiproxy/PC network.

We also performed reconstructions in which predictors were pre-whitened prior to calibration, followed by the re-introduction of the estimated level of serial correlation into the predictand. In drought reconstructions based on tree-ring networks, this procedure has been show to lead to modest improvements in reconstructive skill (Cook et al., 1999; Zhang et al., 2004).

The relative skill of the reconstructions with respect to the different variants of the CFR approach are addressed by cross-validation experiments described in section 4a.

E. Cross-Validation Procedure

A series of verification diagnostics were calculated to evaluate the skill of the reconstructions. First, we conducted full field verifications by removing the instrumental

surface temperature data from the CFR analysis between 1856 through 1900 and reconstructing the surface temperature field over that interval using only the information in the predictor networks calibrated during the 20th century (1901-1971 for the PC/multiproxy network for reasons discussed below, and 1901-1960 for the MXD and network, for reasons discussed earlier). We calculated verification scores using only available instrumental data from grid boxes that were 95 percent complete (a total of 210 brid boxes) prior to initial infilling with RegEM (see section 2A). We refer to these verification scores as the "full field" verification scores. We assessed verification scores both for the full predictor network (available back to at least 1820), and using the increasingly sparse predictor networks available on a century-by-century basis, to assess the fidelity of the reconstruction back in time as the predictor network becomes increasingly sparse. We refer to these verification scores as the "available predictor" scores. As a cross-check, experiments were also performed for the PC/multiproxy network for both the full network and the sparse network available back to 1400 in which an earlier period 1856-1928 was used for calibration and the more recent 1929-1971 period was used for cross-validation. In these cases, the cross validation scores are equal to or better than those for the standard verification period of 1856-1900.

We also used 10 long, annual-mean instrumental grid-box temperature series, nine of which are from western Europe and England with one from North America, to extend cross-validation exercises back into to 1755 (3 of the 10 records are available back to that date, and all 10 are available back to at least 1820) on a more spatially-restricted basis. These instrumental records are part of the multiproxy/PC predictor network used by Mann et al. (1998) and this study. However, these records can also serve as verification

(for the annual mean reconstructions) by removing them from the predictor network and reconstructing them using the information available in the other (non-instrumental) proxy predictors.

The Reduction of Error (Lorenz, 1956; Fritts, 1976) statistic (RE) and Coefficient of Efficiency (CE) (Cook et al., 1994) have been favored as diagnostics of reconstructive skill in most previous climate/paleoclimate reconstruction work (e.g. Fritz, 1976; Cook et al., 1994; Mann et al., 1998; Rutherford et al., 2002; Luterbacher et al., 1999l; 2002a,b; Pauling et al., 2003). RE and CE are defined by:

$$\mathrm{RE} = 1.0 - \left[\frac{\sum (x_i - \hat{x}_i)^2}{\sum (x_i - \overline{x}_c)^2}\right],$$

$$\mathrm{CE} = 1.0 - \left[\frac{\sum (x_i - \hat{x}_i)^2}{\sum (x_i - \overline{x}_v)^2}\right],$$

The sums are over the reconstructed values, $\overline{x}_c$ is the mean of the calibration period, $\overline{x}_v$ is the mean of the verification period, $\hat{x}$ is the reconstructed value, and x is the actual value (see Cook et al., 1994, for further details). Verification scores were computed for both spatial means (e.g. NH mean) of the temperature field (in which case the sums extend over years) and for the full multivariate field (in which case the sums extend over both years and gridboxes, of which there are 210 in the "full field" cross-validation, and 10 in the extended cross-validation).

If the reconstruction is simply the mean of the calibration period, RE=0, which is the threshold for no skill in the reconstruction. Similarly, if the reconstruction is simply the mean of the verification period, CE=0. Thus, depending on the standard, the zero values of these statistics define the threshold for "skill" in the reconstruction. CE≥0 is a

more challenging threshold since, unlike RE, CE does not reward the reconstruction of an observed change in mean relative to the calibration period.

For each experiment, we calculated RE (Tables 1, 2 and 3)and CE (available at http://fox.rwu.edu/~rutherfo/supplements/jclim2003a) verification skill diagnostics. While one could seek to estimate verification skill with the square of the Pearson correlation measure (r^2), this metric can be misleading when, as is the case in paleoclimate reconstructions of past centuries, changes are likely in mean or variance outside the calibration period. To aid the reader in interpreting the verification diagnostics, and to illustrate the shortcomings of r^2 as a diagnostic of reconstructive skill, we provide some synthetic examples which show three possible reconstructions of a series and the *RE*, *CE* and r^2 scores for each (Supplementary material available at http://fox.rwu.edu/~rutherfo/supplements/jclim2003a).

To test whether or not the RegEM reconstruction approach might systematically either overestimate or underestimate the variance in the reconstruction, we performed an additional set of verification experiments in which the reconstructions were systematically rescaled by an inflation factor between 0.5 and two (where a factor of one leaves the reconstruction unchanged) after calibration. If there were any systematic overestimate or underestimate of variance in the calibration process, improved verification statistics should be achieved for scale factors significantly different from one. Instead, we found that the optimal scale factor was close to unity for reconstructions using each of the three networks. The optimal RE statistic (Supplementary material available at http://fox.rwu.edu/~rutherfo/supplements/jclim2003a) for the Northern Hemisphere mean is centered approximately at unity for the annual (Combined network)

reconstruction, slightly below unity (approximately 0.75) for the warm-season (MXD network) reconstruction, and slightly greater than unity (approximately 1.25) for the cold-season (multiproxy/PC network) reconstruction. These results indicate that any substantial (i.e., factor of two or greater) underestimate of variance is unlikely for all three reconstructions.

We estimated self-consistent uncertainties using the "available predictor" verification residuals for each grid box back in time after establishing that the residuals were consistent with Gaussian white noise (Supplementary material available at http://fox.rwu.edu/~rutherfo/supplements/jclim2003a). Gridbox uncertainties were propagated to estimate the uncertainty in spatial means, taking into account spatial correlation.

4. Results

A. Comparisons between Variants of RegEM Approach

We first considered the sensitivity of the results to the effect of pre-whitening predictors and predictand prior to calibration. In two test cases (multiproxy/PC annual and MXD boreal warm season) such a procedure did not result in any consistent improvement of the verification scores. We thus concluded that this step was both unnecessary and, given the importance of faithfully retaining low-frequency variance, probably undesirable in this context, since the approach admits only a limited representation of the temporal dependence structure of the data.

We then considered the impact of allowing for lagged relationships between predictor and predictand (including combinations of lags, e.g. 0 and +1). The cross-validation exercises indicated that a lag of zero (i.e., no lag) produced the optimal skill

diagnostics in all cases, with the following provisos for the cold-season reconstructions. Because the cold-season mean encompasses parts of two calendar years, it is important to define the cold-season convention. For the MXD network, optimal results were achieved for cold-season reconstructions when predictors were temporally aligned with the predictand during the year in which the cold season *ends*. This finding is not surprising since a tree growing during the warm season cannot respond to the climate of the following cold season, but can potentially respond to the climate of the preceding cold season through antecedent soil moisture or soil temperature conditions. For the multiproxy/PC cold-season reconstructions, two lag choices give similar verification scores, "case 1" in which predictors are aligned with the predictand during the year in which the cold-season *ends* and "case 2" in which predictors are aligned with the predictand during the year in which the cold-season *begins*. We adopt "case 2" because, though it performs slightly worse for the multivariate statistics (*RE* and *CE* lower by about 0.03), it performs considerably better for the hemispheric mean statistics (*RE* higher by 0.04 and *CE* higher by 0.08). Case 2 nonetheless seems inappropriate from a biological response point of view, and suggests the importance of a more general approach, beyond the scope of the present study, which allows for variable lags among the different indicators that make up the multiproxy network. Apart from the interannual variability, the hemispheric mean reconstruction is not sensitive to the choice of Case 1 or Case 2. Henceforth, only the optimal results with respect to choice of lag, as described above, are presented for the various seasonal reconstructions based on the various predictor networks.

We then examined the dependence of skill on the frequency band split boundary (5, 10, 20 and 25 year period) used in the hybrid frequency-band calibration approach, finding the 20 year period boundary to give superior results in almost all cases (cross-validation skill was either equal or greater than that for any other choice in all cases). We thus consider henceforth in this study both the standard non-hybrid method (referred to as "non-hybrid"), and the hybrid method with a f=0.05 cycle/year (20 year period) frequency boundary (referred to as "hybrid-20"). A comparison of the NH mean temperature reconstruction for the two approaches (non-hybrid and hybrid-20) is shown in Figure 2 for the multiproxy/PC network, annual mean reconstruction. While the two reconstructions are seen to be broadly similar, the hybrid-20 reconstruction exhibits greater low-frequency variability, particularly prior to AD 1600 when the multiproxy network becomes relatively sparse. The hybrid-20 reconstruction is observed in this case (see discussion below) to demonstrate greater skill in cross-validation for the earlier centuries, suggesting that the greater variability is likely meaningful. As discussed below, whether the non-hybrid or hybrid-20 approach gives optimal results generally depends on the particular predictor network and target season used in the reconstruction.

Another point that must be made is that, although we settle on "optimal" reconstructions, it is not always clear from the verification scores which network, lag, and method implementation (hybrid or non-hybrid) is the "optimal" for a given situation. One set of possibilities (network and lag, for example) may produce a better NH mean verification than another, but at the cost of a degraded multivariate verification, or the hybrid method may outperform the non-hybrid with a sparse network, but the opposite might be true with a more extensive network. In short, it can be difficult to determine

which is the "best" reconstruction when verification skill differences are small. In light of this consideration, we present reconstructions below for each network and season, but recognize that there is a larger suite of reconstructions that might be acceptable based on verification scores. In addition, it is not possible to perform verification experiments on long time scales due to the limitations of the instrumental data. Although we use the few long instrumental records that are available for verification, spatially extensive, long-timescale verification can only be done using output from long GCM simulations.

B. Comparisons of RegEM Results For Different Networks and Seasonal Windows

The results of the cross validation exercises for the various experiments are summarized in Table 1 for the full network available back to 1820, and Table 2 for the increasingly sparse "available predictor" networks back in time (*CE* statistics provided in supplementary material available at http://fox.rwu.edu/~rutherfo/supplements/jclim2003a). The hybrid-20 exhibits the greatest skill (*RE*=0.72 and *CE*=0.46) for annual hemispheric mean reconstructions using the multiproxy/PC network back to 1820. The non-hybrid method, however, exhibits better multivariate skill (*RE*=0.22 and *CE*=0.04), but similar hemispheric mean and multivariate skill if instrumental predictors are withheld (Table 1). However, for proxy networks available further back in time (Table 2), the hybrid-20 approach produces cross-validation skill as good as or better than the non-hybrid approach. We thus favor the use of hybrid-20 for long-term annual reconstructions using the multiproxy/PC network. Similarly, hybrid-20 is favored for the annual reconstructions using the combined network prior to AD 1600. By contrast, the non-hybrid approach is favored by the cross-

validation results for the MXD-based summer temperature reconstructions. The separate selection of optimal weightings in distinct frequency bands afforded by the hybrid-20 approach likely offers a greater advantage for a network of multiple proxy types (multiproxy/PC network) than for a more homogenous (tree-ring MXD) proxy network.

While the different networks often differ by less than 0.03 in terms of cross-validation *RE* or *CE* scores, and comparisons of multivariate and hemispheric-mean skill scores sometimes lead to differing inferences, a few general conclusions can be drawn from the cross-validation results summarized in Tables 1 and 2: (1) The hybrid-20 approach produces the best verification skill scores in general, but there are important exceptions (i.e., the MXD summer temperature reconstructions); (2) the multiproxy/PC network appears best suited for annual and cold-season reconstructions ; (3) as expected on the basis of previous work (e.g. Briffa et al., 2002b), the MXD network appears best suited for annual and warm season reconstructions and appears to provide the best warm-season reconstructions of all 3 networks ; (4) the Combined network exhibits the best skill of all networks in annual reconstruction, and skill in cold-season reconstruction that is comparable to the multiproxy/PC network. The fact that the MXD network outperforms the Combined network for the warm season indicates that the addition of more proxy series does not always produce better results, even if some of those additional proxies are of high quality (e.g. the long instrumental series in the multiproxy/PC network).

The fact that the combined network performs, at best, only marginally better than the two independent networks alone suggests that most of the degrees of freedom in the Northern Hemisphere surface temperature field are already sampled by either network

alone. We conclude that the primary limiting factor governing the skillfulness of current proxy-based large-scale temperature reconstructions may be the quality of the network (e.g. the effective signal-to-noise ratios of the proxy data, and the availability of records in key regions such as the tropical Pacific), rather than the size of the network. This conclusion is reinforced by a comparison of the verification skill for the multiproxy/PC network with and without the long instrumental records (Table 1), which emphasizes that a few high quality indicators can significantly improve reconstructive skill. Furthermore, the addition of poor or inappropriate seasonal indicators to a network can degrade the skill of the reconstruction.

C. Comparisons Between RegEM Results and Results with Other Methods

We compare results using the RegEM method with previously published results that used the same proxy networks but different methods (and a potentially different target region). In these comparisons, we control for the proxy network used and the target season.

i. Comparison With Mann et al. (1998) Annual NH Reconstruction

Although both the Mann et al. (1998) and RegEM methods make use of covariance information in the calibration/reconstruction process, they do so in a quite different manner (Schneider et al., 2001; Rutherford et al., 2003). We compared the RegEM reconstruction with the Mann et al. (1998) surface temperature reconstruction employing the same predictor network, the same calendar annual target season, and same global target region as Mann et al. (1998). We eliminated the infilled values from AD 1400-1403 used by MBH98 to complete one of the Jacoby and D'Arrigo (1989) 'Northern

Treeline' series back to AD 1400. This is easily done in the RegEM method by treating those values as missing, something that could not have been done in MBH98. We terminated the calibration period in 1971 to address the criticism by McIntyre and McKitrick (2003) of the use by MBH98 of a modest number of infilled missing proxy values in the PC/Multiproxy network between 1971 and 1980. However, we also show the verification results for the case where the calibration interval ends in 1980 for direct comparison with the Mann et al. (1998) results. Cross-validation results are compared (Table 3) with those of Mann et al. (1998) for the PC/multiproxy network available back to 1820, using the same (219) grid boxes used for verification by Mann et al. (1998) over the period 1856 to 1900 (top section of Table 3) and for the 10 temperature grid boxes available back to 1820 (bottom section of Table 3—in this case long instrumental indicators have been withheld from the predictor network). These comparisons indicate similar levels of skill in the RegEM (both non-hybrid and hybrid-20) and Mann et al. (1998) reconstructions, with any preference dependent on the precise metric of reconstructive skill. We deduce from the "available predictor" skill diagnostics (Table 2) that the hybrid-20 RegEM reconstruction is increasingly preferable over the non-hybrid reconstruction as the predictor network becomes sparser back in time (for annual NH temperature, at least).

A remarkably close similarity is observed (Figure 3) between the RegEM and Mann et al. (1998) NH annual mean surface temperature reconstructions. The two reconstructions are indistinguishable well within their 2-sigma uncertainties. The RegEM NH reconstruction using all available individual proxy records (rather than replacing spatially

dense tree-ring networks with their leading principal components as in the MBH98 Multiproxy/PC network) again yields nearly indistinguishable estimates (Figure 2). The close reproducibility of the MBH98 reconstruction based on both (a) the use of an independent CFR method and (b) the use of the individual proxies used by MBH98 rather than the Multiproxy/PC representation used by MBH98, discredits the arguments put forth by McIntyre and McKitrick (2003) in support of their putative "correction" to the MBH98 reconstruction.

ii. Comparison Involving Previous MXD-based Warm-Season Extratropical NH Reconstruction

Here we compare the RegEM warm season MXD-based NH mean reconstruction with that of Osborn et al. (2004), the latter based on an areally-weighted mean of 115 locally-calibrated MXD 5° by 5° grid boxes (Figure 1b). This reconstruction (Figure 4) is similar, though not identical, to that presented by Briffa et al. (2001) using the same MXD data; the minor differences arise because Briffa et al. (2001) used a principal component regression of regionally-averaged MXD data, rather than the average of locally-calibrated reconstructions generated by Osborn et al. (2004). In this comparison we control for the proxy network (both use the MXD network) and the target season (both target the boreal warm season mean) and investigate the effects of both the target region and reconstruction method.

Figure 4a compares the Osborn et al. (2004) MXD reconstruction and the RegEM hybrid-20 NH reconstruction of the full NH mean. The Osborn et al. reconstruction exhibits greater interannual variability and is on average slightly cooler in past centuries

than the RegEM reconstruction. Since the proxy network and the target season are identical, the observed differences must be due to a combination of differing method and target region. To progressively control for target region we first mask the RegEM spatial reconstruction for only the terrestrial extratropical (i.e., north of 20°N) grid boxes (Figure 4a) and finally the precise 115 grid boxes averaged by Osborn et al. to obtain a hemispheric mean reconstruction (Figure 4b). The latter masking of the RegEM reconstruction yields a hemispheric mean estimate that is nearly indistinguishable from the Osborn et al. reconstruction, suggesting that the initial differences evident in Figure 4a result largely from differing initial target regions. The remaining modest differences (Figure 4b), which are mostly evident during the relatively data sparse initial centuries, are presumably due to the differences between methods (RegEM CFR method versus spatial average of the locally-calibrated grid box data).

Finally, we include a comparison with an alternative warm-season continental surface temperature reconstruction based on an even more restricted spatial distribution (a maximum of 14 sites) of tree-ring width data (Esper et al., 2002). This reconstruction exhibits greater variability than most other published reconstructions (see Briffa and Osborn, 2002; Mann and Hughes, 2002; Mann, 2002a; Mann et al., 2003a,b). However, when restricted to the grid box locations corresponding to the modest number of sites used in this reconstruction (excepting one grid box that is unavailable from the instrumental record and one that is outside our reconstruction domain), the RegEM MXD warm-season NH reconstruction shows a remarkably similar character to the Esper et al. (2002) reconstruction (Figure 4c). This result suggests that the greater variability evident in the Esper et al. reconstruction likely results from the restricted sampling provided by

the network used, though some residual differences may be due to different methods of tree-ring standardization (Esper et al., 2002; Briffa and Osborn, 2002; Mann and Hughes, 2002; Cook et al., 2004) and differences in reconstruction method.

From these comparisons we can draw an important conclusion that might have been anticipated from spatial sampling considerations alone: reconstructions of full hemispheric means are likely to exhibit lower amplitude variability that those based on a more restricted sub-domain of the field, due to the tendency for the cancellation of anomalies of different signs and magnitudes in different regions (see e.g. Mann et al., 2003b).

D. Comparisons Of Hemispheric Mean Series

Figure 5 shows the annual, warm season and cold season NH mean reconstructions produced using the different predictor networks. The annual (Figure 5a) reconstructions are quite similar for all three predictor networks back to approximately AD 1700, and are largely within the statistical uncertainties of each other back through AD 1400. A similar statement holds for the cold-season reconstructions, though the differences are slightly greater during certain time intervals. For the warm season, only the MXD network provides a skillful reconstruction back to AD 1400, but all reconstructions are similar over the interval in which the cross validation experiments indicate a skillful reconstruction (1750 for the Multiproxy/PC network and 1500 for the Combined network). It is clear from the similarity of the MXD and Combined network warm-season reconstructions that the Combined network reconstruction is dominated by the MXD predictors, as one would expect based on the verification scores.

Finally, we compare (Figure 5d-e) the RegEM NH reconstructions based on what appears to be optimal apparent predictor network for each season (see section 4b: MXD-warm season, Multiproxy/PC-cold season, Combined-annual). These reconstructions show that the cold-season mean reconstruction generally exhibits the greatest interannual variability (particularly when a larger number of indicators are available). The warm-season reconstructions often show stronger cooling events, in many cases associated with large explosive volcanic events (e.g. after the AD 1600 eruption). As discussed further below, this observation is consistent with the modeled response to volcanic forcing, which shows cooling to dominate during the warm season, particularly over continental regions (Kirchner et al., 1999; Shindell et al., 2003). Reconstructions which emphasize the warm-season and continental regions are thus likely to exhibit greater summer cooling during periods of intense explosive volcanic activity.

E. Spatial Patterns

It is instructive to examine the spatial and seasonal details evident in the actual reconstructed patterns. We thus focus on the reconstructed temperature patterns for some selected years (Figure 6), using reconstructions based on the optimal networks for each season as discussed above. We consider the year 1601 (cold-season 1600/1601) following the Huaynaputina (Peru) eruption (February 1600); 1783, the year of the Laki eruption in Iceland and an exceptionally cold winter in parts of North America and Europe; 1791, an established unusually strong El Nino year (Quinn and Neal 1992); 1816, the "Year Without a Summer" one year after the explosive Tambora eruption of April 1815, 1817, two years after the eruption, and finally, 1834, an exceptionally warm

year in Europe as evidenced by the central England temperature record (Manley, 1974—see also Mann et al., 2000b and Briffa et al., 1998a,2002b).

There is a tendency for opposite seasonal surface temperature responses to tropical volcanic forcing. Strong warm-season continental cooling is apparent in the summers after the volcanic years 1600, and 1815 contrasting with the tendency for an offsetting pattern of continental warming during the winter following those eruptions (or even two winters, following the 1815 eruption). This pattern has been observed in model simulations of the dynamical response to an explosive tropical eruption (Groisman, 1992; Graf et al., 1993; Robock and Mao, 1995; Kirchner et al., 1999; Shindel et al., 2003). The tendency for cooler summers and warmer winters appears to be responsible for the reduced annual mean cooling response to volcanic forcing (Shindell et al., 2003) evident in the annual mean reconstructions. Large-scale warmth both in the tropical Pacific, and in the extratropics, is clearly evident for the El Nino year of 1791 during all seasons, but is particularly evident in the cold-season (i.e., 1791/1792) pattern.

5. Conclusions

Comparisons both within the suite of reconstructions presented in this study, and between these reconstructions and others previously developed (Mann et al., 1998; Osborn et al., 2004; Esper et al., 2002) allow us to evaluate the impacts of method, target season, target region, and underlying proxy data network on large-scale surface temperature reconstructions. (Note: The reconstructions performed in this study are available at http://fox.rwu.edu/~rutherfo/supplements/jclim2003a.) These evaluations suggest that differing methods of reconstruction (e.g. different CFR techniques or local calibration

approaches) yield nearly indistinguishable results if differences in underlying proxy network, target season and target region are controlled for. We conclude that proxy-based temperature reconstructions are robust with respect to a wide array of alternative statistical approaches. Differences in target region appear to lead to more substantial differences. Spatial averages over increasingly sparse domains (e.g., extratropical continents only, or a small number of isolated regions of the extratropical continents only) yield "hemispheric mean" estimates with increasingly greater variability. Meaningful comparisons of different hemispheric mean estimates are thus only possible when differences in target spatial domain are taken into account e.g. through an appropriate spatial masking of the surface temperature field reconstructions. Differences in target seasonal window are also important, with different predictor networks (e.g. the Multiproxy/PC vs. MXD vs. Combined network) each indicating preferential reconstructive skill for different seasonal windows. The resulting optimal seasonal (cold-season, warm-season, and annual mean) reconstructions indicate modest differences for the main hemispheric mean temperature changes, and more substantial differences spatially, consistent with the distinct spatial and seasonal features typically associated with climate signals such as El Niño or the response to volcanic radiative forcing.

In addition, we find that the number of proxies can be less important than the quality of the proxy and its suitability for reconstructing a particular season. As an example, the MXD network alone clearly outperforms the Combined network in warm-season verification tests. Furthermore, it is not always easy to determine the best network, lag and method to use in every situation because differences in verification

scores can be small. In the situations we examined, however, differences in the reconstructions are also small.

Finally, the evidence for exceptional late 20th century warmth in the context of the period since AD 1400 (in warm, cold and annual temperatures) is a robust conclusion with respect to all of the factors considered.

Acknowledgements

This work was supported by the NSF- and NOAA-funded "Earth Systems History" program (M.E.M., S.R., R.S.B., and M.K.H.-NOAA award NA16GP2913), the Office of Science (BER), U.S. Dept. of Energy, Grants No. DE-FG02-98ER62601 (P.D.J.) and DE-FG02-98ER62604 (R.S.B.) and the European Community under the SOAP project: EVK2-CT-2002-00160 (T.J.O. and K.R.B).

References:

Appenzeller, C., T.F. Stocker, and M. Anktin, 1998: North Atlantic Oscillation Dynamics Recorded in Greenland Ice Cores, *Science*, *282*, 446-449.

Bauer, E., M. Claussen, and V. Brovkin, 2003: Assessing climate forcings of the earth system for the past millennium, *Geophys. Res. Lett.*, 30 (6), 1276, doi: 10.1029/2002GL016639.

Bertrand C., M. F. Loutre, M. Crucifix, and A. Berger, 2002: Climate of the Last millennium: a sensitivity study. *Tellus*, 54(A), 221-244.

Biondi, F., A. Gershunov, and D.R. Cayan, 2001: North Pacific Decadal Climate Variability since 1661, *Journal of Climate*, *14*, 5-10.

Bradley, R.S., 1996: Are there optimum sites for global paleotemperature reconstruction?, in *NATO ASI Series Climatic Variations and Forcing Mechanisms of the last 2000 Years*, edited by P.D. Jones, R.S. Bradley, and J. Jouzel, Springer-Verlag Berlin Heidelberg. pp. 603-624,

Bradley, R.S., 1999: Paleoclimatology: reconstructing climates of the Quaternary Harcourt. Academic Press, San Diego, 610 pp.

Bradley, R.S., K. R. Briffa, T. J. Crowley, M. K. Hughes, P.D. Jones, and M.E. Mann, 2001: Scope of Medieval Warming, *Science*, 292, 2011-2012.

Bradley, R.S., and P.D. Jones, 1993: "Little Ice Age" summer temperature variations: their nature and relevance to recent global warming trends, *The Holocene*, *3* (4), 367-376, 1993.

Bradley, R.S. and P.D. Jones (eds.), 1995: Climate Since A.D. 1500. (Revised edition) Routledge, London, 706 pp.

Braganza, K., D. Karoly, T. Hirst, M.E. Mann, P. Stott, R. J. Stouffer, and S. Tett: 2002. Indices of Global Climate Variability and Change: Part I - Variability and Correlation Structure, *Climate Dynamics*, in press.

Briffa, K.R., P.D. Jones, and F.H. Schweingruber, 1994: Summer temperatures across northern North America: Regional reconstructions from 1760 using tree-ring densities, *Journal of Geophysical Research*, *99*, 25,835-25,844.

Briffa, K.R., P.D. Jones, F.H. Schweingruber, W. Karlen and S. G. Shiyatov, 1996: in *NATO ASI Series Climatic Variations and Forcing Mechanisms of the last 2000 Years*, edited by P.D. Jones, R.S. Bradley, and J. Jouzel, pp. 9-41, Springer-Verlag Berlin Heidelberg.

Briffa, K.R., P.D. Jones, F.H. Schweingruber, and T.J. Osborn, 1998a: Influence of volcanic eruptions on Northern Hemisphere summer temperatures over the past 600 years, *Nature*, *393*, 450-454.

Briffa, K.R., F.H. Schweingruber, P.D. Jones, T.J. Osborn, S.G. Shiyatov, and E.A. Vaganov, 1998b: Reduced sensitivity of recent tree-growth to temperature at high northern latitudes, *Nature*, 678-682.

Briffa, K.R., and T. J. Osborn, 2002: Blowing Hot and Cold, *Science*, 295, 2227-2228.

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

Briffa, K. R., T. J. Osborn, F. H. Schweingruber, P. D. Jones, S. G. Shiyatov, and E. A. Vaganov, 2002a: Tree-ring width and density data around the Northern Hemisphere: Part 1, local and regional climate signals. *The Holocene*, **12,** 737-757.

Briffa, K. R., T. J. Osborn, F. H. Schweingruber, P. D. Jones, S. G. Shiyatov, and E. A. Vaganov, 2002b: Tree-ring width and density data around the Northern Hemisphere: Part 2, spatio-temporal variability and associated climate patterns. *The Holocene*, **12,** 759-789.

Briffa KR, T. J. Osborn and F.H Schweingruber, 2003: Large-scale temperature inferences from tree rings: a review. *Global and planetary change* **40**, 11-26 (doi:10.1016/S0921-8181(03)00095-X).

Bush, A.B.G., 1999: Assessing the impact of Mid-Holocene insolation on the atmosphere-ocean system, *Geophysical Research Letters*, *26* (1), 99-102.

Clement, A.C., R. Seager, and M.A. Cane, 2000: Suppression of El Nino during the mid-Holocene by changes in the Earth's orbit, *Paleoceanography*, *15*, 731-737.

Cobb, K.M., C.D. Charles, H. Cheng, and R.L. Edwards, 2003: El Niño-Southern Oscillation and tropical Pacific climate during the last millennium, *Nature,* 424, 271-276.

Cook, E.R., 2002: Multi-Proxy Reconstructions of the North Atlantic Oscillation (NAO) Index: A Critical Review and a New Well-Verified Winter NAO Index Reconstruction Back to AD 1400. In *North Atlantic Oscillation* (J.W. Hurrell, Y. Kushnir, G. Ottersen and M. Visbeck, Eds.). American Geophysical Union, Washington D.C., 63-81.

Cook, E. R., K.R. Briffa, and P.D. Jones, 1994: Spatial Regression Methods in Dendroclimatology: A Review and Comparison of Two Techniques. *International Journal of Climatology* 14, 379-402.

Cook, E.R., K.R. Briffa, D.M. Meko, D.A. Graybill, and G. Funkhouser, 1995: The "segment length curse" in long tree-ring chronology development for palaeoclimatic studies, *The Holocene*, *5*, 229-237.

Cook, E. R., D. M. Meko, and C. W. Stockton, 1997: A new assessment of possible solar and lunar forcing of the bidecadal drought rhythm in the western United States. J. Climate, 10, 1343-1356.

Cook, E.R., D. M. Meko, D. W. Stahle, and M.K. Cleaveland, 1999: Drought Reconstructions for the Continental United States, *J. Climate*, *12*, 1145-1162.

Cook, E.R., R.D. D'Arrigo, M.E. Mann, 2002: A Well-Verified, Multi-Proxy Reconstruction of the Winter North Atlantic Oscillation Since AD 1400, *J. Climate*, 15, 1754-1764.

Cook, E.R., J. Esper, and R.D. D'Arrigo, 2004: Extra-tropical Northern Hemisphere land temperature variability over the past 1000 years, *Quaternary Science Reviews*, in press.

Cullen, H., R.D. D'Arrigo, E. Cook, M.E. Mann, 2001: Multiproxy-based reconstructions of the North Atlantic Oscillation over the past three centuries, *Paleoceanography*, 15, 27-39.

Crowley, T. J., 2000: Causes of Climate Change Over the Past 1000 Years. *Science* 289, 270-277.

Crowley, T. J. and T. Lowery, 2000: How Warm Was the Medieval Warm Period? *Ambio 29*, 51-54.

Crowley, T.J. and G.R. North, 1991: Paleoclimatology. (Revised edition) Oxford University Press, New York, 349 pp.

D'Arrigo, R.D., E.R. Cook, G.C. Jacoby, and K.R. Briffa, 1993: NAO and Sea Surface Temperature Signatures in Tree-Ring Records From the North Atlantic Sector, *Quaternary Science Reviews*, *12*, 431-440.

Delworth, T.L., and M.E. Mann, 2000: Observed and Simulated Multidecadal Variability in the Northern Hemisphere, *Climate Dynamics,* 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, 1998: Optimal Sites for Coral-based Reconstruction of Global Sea Surface Temperature, *Paleoceanography*, *13*, 502-516.

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, 71.

Fisher, D. A., R. M. Koerner, K. Kuivinen, H. B. Clausen, S. J. Johnsen, J.-P. Steffensen, N. Gundestrup, and C. U. Hammer, 1996: Intercomparison of ice core d (O-18) and precipitation records from sites in Canada and Greenland over the last 3500 years and over the last few centuries in detail using EOF techniques. *Climate variations and forcing mechanisms of the past 2000 years*, P. D. Jones, R. S. Bradley, and J. Jouzel, Eds., Springer, 297-328.

Folland, C.K., T.R. Karl, J.R. Christy, R.A. Clarke, G.V. Gruza, J. Jouzel, M.E. Mann, J. Oerlemans, M.J. Salinger, and S.-W. Wang, 2001: Observed Climate Variability and Change, in *Climate Change 2001: The Scientific Basis,* Houghton, J.T., et al. (eds.), Cambridge Univ. Press, Cambridge, 99-181.

Free, M. and A. Robock, 1999: Global Warming in the Context of the Little Ice Age. *J Geophys. Res. 104*, 19,057-19,070.

Fritts, H.C., 1976: Tree Rings and Climate. London, Academic Press: xii + 567 p.

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. *Journal of Applied Meteorology* 10(5), 845-864.

Fritts, H.C. 1991: *Reconstructing Large-scale Climatic Patterns from Tree-Ring Data*. Tucson & London, The University of Arizona Press: 286 p.

Gedalof, Z., N.J. Mantua, and D.L Peterson, 2002:A multi-century perspective of variability in the Pacific Decadal Oscillation: new insights from tree rings and coral. *Geophysical Research Letters* 29(24).

Gerber, S., F. Joos, P.P. Bruegger, T.F. Stocker, M.E. Mann, and S. Sitch, 2003: Constraining Temperature Variations over the last Millennium by Comparing Simulated and Observed Atmospheric CO2, *Climate Dynamics*, 20, 281-299.

Gonzalez-Rouco, F., H. von Storch, and E. Zorita, 2003: Deep soil temperature as proxy for surface air-temperature in a coupled model simulation of the last thousand years, *Geophys. Res. Lett*., 30, 2116, doi:10.1029/2003GL018,264.

Graf, H.-F., I. Kirchner, A. Robock, and I. Schult, 1993: Pinatubo eruption winter climate effects: Model versus observations, *Clim. Dyn*., *9*, 81–93.

Groisman, P. Y., 1992: Possible regional climate consequences of the Pinatubo eruption: An empirical approach, *Geophys. Res. Lett*., *19*, 1603–1606.

Guiot, J., 1985: The extrapolation of recent climatological series with spectral canonical regression. Journal of Climatology, 5, 325-335.

Guiot, J., 1988: The combination of historical documents and biological data in the reconstruction of climate variations in space and time, *Palaeoclimatforschung*, 7, 93-104.

Hendy**,** E. J., M. K. Gagan, C. A. Alibert, M. T. McCulloch, J. M. Lough, & P.J. Isdale, 2002: Abrupt Decrease in Tropical Pacific Sea Surface Salinity at End of Little Ice Age, *Science* 295, 1511.

Hughen, K.A., J.T. Overpeck and R. Anderson, 2000: Recent warming in a 500-year paleoclimate record from Upper Soper Lake, Baffin Island, Canada. *The Holocene*, 10, 9-19

Hughes, M.K. and G. Funkhouser, 2003: Frequency-dependent climate signal in upper and lower forest border trees in the mountains of the Great Basin. *Climatic Change,* 59, 233-244

Jacoby, G.C., and R. D'Arrigo, 1989: Reconstructed Northern Hemisphere Annual Temperature Since 1671 Based on High-Latitude Tree-Ring Data From North America, *Climatic Change*, *14*, 39-59.

Jones, P.D., 1994: Hemispheric Surface Air Temperature Variations: A Reanalysis and an Update to 1993, *Journal of Climate*, *7*, 1794-1802.

Jones, P.D., and M.E. Mann, 2004: Climate Over Past Millennia, *Reviews of Geophysics*, 42, RG2002, doi: 10.1029/2003RG000143.

Jones, P.D. and A. Moberg, 2003: Hemispheric and large-scale surface air temperature variations: An extensive revision and an update to 2001. *J. Climate* **16**, 206-223.

Jones, P.D., K.R. Briffa, T.P. Barnett and S.F.B. Tett, 1998: High-Resolution Paleoclimatic Records for the Last Millennium: Interpretation, Integration and Comparison with Circulation Model Control-Run Temperatures. *The Holocene 8*, 455-471.

Jones, P.D., K.R. Briffa, and T.J. Osborn, 2003a: Changes in the Northern Hemisphere annual cycle – implications for paleoclimatology?, *J. Geophys. Res., 108*, 4588, doi:10.1029/2003JD003695.

Jones, P.D., T.J. Osborn, and K.R. Briffa, 2003b: Pressure-based measures of the NAO: A comparison and an assessment of changes in the strength of the NAO and in its influence on surface climate parameters. In *North Atlantic Oscillation* (J.W. Hurrell, Y. Kushnir, G. Ottersen and M. Visbeck, Eds.). American Geophysical Union, Washington D.C., 51-62.

Jones, P.D., M. New, D.E. Parker, S. Martin and J.G. Rigor, 1999: Surface Air Temperature and its Changes over the Past 150 Years. *Rev. Geophys. 37*, 173-199.

Jones, P.D., T.J. Osborn, and K.R. Briffa, 2001a: The Evolution of Climate Over the Last Millennium, *Science*, *292* (April 27), 662-667.

Jones, P.D., T.J. Osborn, K.R. Briffa, C.K. Folland, B. Horton, L.V. Alexander, D.E. Parker, and N.A. Rayner, 2001b: Adjusting for sample density in grid-box land and ocean surface temperature time series. *J. Geophys. Res.* **106**, 3371-3380.

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. Journal of Geophysical Research, 102(C13): 27,835-27,860.

Kirchner, I., G.L. Stenchikov, H.F. Graf, A. Robock, J.C. Antuña, 1999: Climate model simulation of winter warming and summer cooling following the 1991 Mount Pinatubo volcanic eruption. *J. Geophys. Res.*, 104, 19039.

Knutson, T.R., T.L. Delworth, K. Dixon and R.J. Stouffer, 2000: Model assessment of regional surface temperature trends (1949-97), J. Geophys. Res., 104, 30981-30996.

LaMarche, V. C., 1974: Frequency-Dependent Relationships Between Tree-Ring Series Along an Ecological Gradient and Some Dendroclimatic Implications, *Tree-Ring Bulletin,* 34, 1-20.

Le Roy Ladurie, E., 1971: *Times of Feast, Times of Famine, A History of Climate since the year 1000*, Doubleday.

Lorenz, E.N., 1956: Empirical orthogonal functions and statistical weather prediction. *M.I.T. Stat. Forecasting Proj. Sci. Rep*. 1, Contract No. AF 19 (604) - 1566

Luterbacher, J., C. Schmutz, D. Gyalistras, E. Xoplaki, and H. Wanner, 1999: Reconstruction of monthly NAO and EU indices back to AD 1675, *Geophysical Research Letters*, *26* (17), 2745-2748.

Luterbacher, J., E. Xoplaki, D. Dietrich, R. Rickli, J. Jacobeit, C. Beck, D. Gyalistras, C. Schmutz, and H. Wanner, 2002a: Reconstruction of sea level pressure fields over the Eastern North Atlantic and Europe back to 1500, *Climate Dynamics*, *18*, 545-561.

Luterbacher, J., E. Xoplaki, D. Dietrich, P.D. Jones, T.D. Davies, D. Portis, J.F. Gonzalez-Rouco, H. von Storch, D. Gaylistras, C. Casty, and H. Wanner, 2002b: Extending North Atlantic Oscillation reconstructions back to 1500. *Atmospheric Science Letters*, doi:10.1006/asle.2001.0044.

Luterbacher, J., D. Dietrich, E. Xoplaki, M. Grosjean and H. Wanner, 2004: European annual and seasonal temperature variability, trends and extremes since 1500. *Science*, 303, 1499-1503.

Manley, G., 1974: Central England Temperatures: monthly means 1659 to 1973. *Quarterly Journal of the Royal Meteorological Society* 100: 389-405.

Mann, M.E., 2002a: The Value of Multiple Proxies, *Science,* 297, 1481-1482.

Mann, M.E., 2002b: Large-scale climate variability and connections with the Middle East in past centuries, *Climatic Change,* 55, 287-314.

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., R.S. Bradley, and M.K. Hughes, 2000a: Long-term variability in the El Nino Southern Oscillation and associated teleconnections, Diaz, H.F. & Markgraf, V. (eds) *El Nino and the Southern Oscillation: Multiscale Variability and its Impacts on Natural Ecosystems and Society*, Cambridge University Press, Cambridge, UK, 357-412.

Mann, M.E., E. Gille, R.S. Bradley, M.K. Hughes, J.T. Overpeck, F.T. Keimig, F.T. and W. Gross, 2000b: Global Temperature Patterns in Past Centuries: An interactive presentation, *Earth Interactions*, *4-4*, 1-29.

Mann, M.E., M.K. Hughes, 2002: Tree-Ring Chronologies and Climate Variability, *Science*, 296, 848.

Mann, M.E. and J.M. Lees, 1996: Robust Estimation of Background Noise and Signal Detection in Climatic Time Series. *Climate Change 33*, 409-445.

Mann, M.E., and S. Rutherford, 2002: Climate Reconstruction Using 'Pseudoproxies', *Geophysical Research Letters*, 29, 139-1-139-4.

Mann, M.E., S. Rutherford, R.S. Bradley, M.K. Hughes, F.T. Keimig, 2003a: Optimal Surface Temperature Reconstructions Using Terrestrial Borehole Data, *Journal of Geophysical Research,* 108 (D7), 4203, doi: 10.1029/2002JD002532.

Mann, M.E., C.M. Ammann, R.S. Bradley, K.R. Briffa, T.J. Crowley, M.K. Hughes, P.D. Jones, M. Oppenheimer, T.J. Osborn, J.T. Overpeck, S. Rutherford, K.E. Trenberth, and T.M.L. Wigley, 2003b: On past temperatures and anomalous late-20th century warmth, *Eos*, **84**, 256-258.

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

McIntyre, S., and R. McKitrick, 2003: Corrections to the Mann et al. (1998) Proxy Data Based and Northern Hemispheric Average Temperature Series, *Energy and Environment*, **14**, 751-771.

Meeker, L.D., Mayewski, P.A., 2002: A 1400-year high-resolution record of atmospheric circulation over the North Atlantic and Asia, *Holocene*, 12, 257-266.

O'Brien, S.R., P.A. Mayewski, L.D. Meeker, D.A. Meese, M.S. Twickler, and S.I. Whitlow, 1995: Complexity of Holocene Climate as Reconstructed from a Greenland Ice Core, *Science*, *270*, 1962-1964.

Osborn T.J., K.R. Briffa, F.H. Schweingruber and P.D. Jones, 2004: Annually-resolved patterns of summer temperature over the Northern Hemisphere since AD1400 from a tree-ring network. Submitted to *Global and Planetary Change*.

Osborn T.J. and K.R. Briffa, 2000: Revisiting timescale-dependent reconstruction of climate from tree-ring chronologies. *Dendrochronologia* **18** , 9-26.

Osborn, T.J., K.R. Briffa, S.F.B. Tett, P.D. Jones, and R.M. Trigo, 1999: Evaluation of the North Atlantic Oscillation as simulated by a coupled climate model, *Climate Dynamics*, *15*, 685-702.

Overpeck, J., K. Hughen, D. Hardy, R. Bradley, R. Case, M. Douglas, B. Finney, K. Gajewski, G. Jacoby, A. Jennings, S. Lamoureux, A. Lasca, G.M.J. Moore, M. Retelle, S. Smith, A. Wolfe, and G. Zielinski, 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, Geophysical Research Letters, VOL. 30, NO. 15, 1787, doi:10.1029/2003GL017589.

Pfister, C., J. Luterbacher, G. Schwarz-Zanetti and M. Wegmann, 1998: Winter air temperature variations in Central Europe during the Early and High Middle Ages (A.D. 750-1300). Holocene, 8, 547-564.

Quinn, W. H. and V. T. Neal, 1992: The Historical Record of El Niño Events. *Climate Since A.D. 1500*, R. S. Bradley and P. D. Jones, Eds., Routledge, 623-648.

Robock, A., and J. Mao, 1995: The volcanic signal in surface temperature observations, *J. Clim*., *8*, 1086–1103.

Rutherford, S., M.E. Mann, T.L. Delworth, and R.J. Stouffer, 2003. Climate Field Reconstruction Under Stationary and Nonstationary Forcing, *J. Clim.*, 16, 462-479.

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*, 294, 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, *Journal of Climate*, 16, 4094-4107.

Schneider, T., 2001: Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values. *Journal of Climate 14*, 853-887.

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 of Climate, 9, 1403-1420.

Stahle, D.W., R.D. D'Arrigo, P.J. Krusic, M.K. Cleaveland, E.R. Cook, R.J. Allan, J.E. Cole, R.B. Dunbar, M.D. Therrell, D.A. Gay, M.D. Moore, M.A. Stokes, B.T. Burns, J. Villanueva-Diaz, and L.G. Thompson, 1998: Experimental Dendroclimatic Reconstruction of the Southern Oscillation, *Bulletin of the American Meteorological Society*, 79 (10), 2137-2152.

Vaganov, E.A., M.K. Hughes, A.V. Kirdyanov, F.H. Schweingruber, and P.P. Silkin, 1999: Influence of snowfall and melt timing on tree growth in subarctic Eurasia. *Nature* 400:149-151

Waple, A., M.E. Mann, R.S. Bradley, 2002: Long-term Patterns of Solar Irradiance Forcing in Model Experiments and Proxy-based Surface Temperature Reconstructions, *Climate Dynamics,* 18, 563-578.

Wigley, T.M.L., M.J. Ingram, and G. Farmer, 1981: Past Climates and their impact on Man: a review. in *Climate and History*, Cambridge University Press, eds. T.M.L. Wigley, M.J. Ingram, G. Farmer, pp. 3-50.

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences*, Academic Press, San Diego, CA, 467pp.

Zhang, Z., M.E. Mann, 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., 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.

Analysis	Verif.	Multiproxy/PC				MXD			Combined			
Season	**Period**	**Ann.**	**No inst.**	**Cold**	**Warm**	**Ann.**	**Cold**	**Warm**	**Ann.**	**No inst.**	**Cold**	**Warm**
non-hybrid	1856-1900	0.22 0.69	0.07 0.57	0.05 0.53	0.06 0.57	0.14 0.56	0.06 0.45	0.14 0.50	0.21 0.69	0.12 0.61	0.03 0.55	0.11 0.35
	1755-1900		0.13 0.24			0.24 0.26				0.21 0.26		
20-year hybrid	1856-1900	0.20 0.72	0.07 0.59	0.05 0.56	0.02 0.15	0.14 0.55	0.08 0.56	0.12 0.38	0.13 0.58	0.12 0.61	-0.04 0.47	0.00 0.14
	1755-1900		0.13 0.20			0.26 0.29				0.19 0.30		

Table 1

Table 1. RegEM-method Reduction of Error (RE) verification scores based on full network that is available back to 1820 (112 predictors in the multiproxy/PC case and 111 grid-box indicators in the MXD case—4 MXD gridboxes are not available prior to 1856 and therefore are never used in reconstructions) for different seasons and proxy networks. For table cells with two rows of verification scores, those scores are 1856-1900 multivariate (calculated over verification gridboxes and years) and 1856-1900 NH mean (calculated for the mean NH time series over the verification years). Where applicable, cells with additional rows include 1755-1900 multivariate and 1755-1900 mean verification scores. "No inst." indicates the long instrumental records present in the proxy network have been withheld from the network. In all cases the calibration period is 1901-1971

Analysis	Multiproxy/PC			MXD			Combined		
Season	**Ann.**	**Cold**	**Warm**	**Ann.**	**Cold**	**Warm**	**Ann.**	**Cold**	**Warm**
Network back to 1750, proxies only*									
nohybrid	0.05	0.02	0.03	0.14	0.06	0.14	0.10	0.00	0.09
	0.52	0.47	0.21	0.56	0.45	0.50	0.64	0.54	0.36
20-year	0.04	0.00	0.06	0.14	0.08	0.12	0.60	-0.09	0.09
hybrid	0.52	0.47	0.20	0.55	0.56	0.38	0.51	0.49	0.13
Network back to 1700									
nohybrid	0.04	0.02	0.01	0.13	0.06	0.12	0.09	-0.02	0.50
	0.51	0.51	0.03	0.56	0.45	0.41	0.61	0.57	0.17
20-year	0.04	0.01	0.00	0.13	0.04	0.11	0.07	-0.08	0.02
hybrid	0.53	0.52	0.01	0.55	0.53	0.33	0.52	0.50	0.08
Network back to 1600									
nohybrid	0.05	0.04	0.01	0.12	0.07	0.10	0.09	0.02	0.04
	0.53	0.53	0.06	0.48	0.40	0.31	0.58	0.54	0.13
20-year	0.06	0.02	-0.02	0.13	0.05	0.09	0.09	-0.02	-0.06
hybrid	0.56	0.52	0.00	0.52	0.53	0.27	0.57	0.50	0.00
Network back to 1500									
nohybrid	0.07	0.08	0.00	0.11	0.07	0.10	0.08	0.02	0.03
	0.39	0.56	-0.42	0.46	0.39	0.36	0.52	0.51	-0.12
20-year	0.08	0.06	-0.01	0.12	0.07	0.09	0.10	0.00	0.04
hybrid	0.47	0.55	-0.36	0.50	0.54	0.34	0.59	0.53	0.11
Network back to 1400									
nohybrid	0.06	0.08	0.00	0.11	0.08	0.09	0.07	0.04	0.00
	0.40	0.55	-0.34	0.49	0.42	0.36	0.46	0.54	-0.40
20-year	0.06	0.08	-0.01	0.10	0.06	0.05	0.09	0.04	0.02
hybrid	0.46	0.54	-0.36	0.48	0.53	0.20	0.57	0.56	-0.05

Table 2

Table 2. Verification Reduction of Error (RE) scores for the different proxy networks and seasons using only the proxies available over different centuries. The verifications scores are presented for the multivariate (top row in each table cell) and NH mean (bottom row in each table cell) cases. In all cases the calibration period is 1901-1971 and the verification period is 1856-1900.

* The MXD proxy network is nearly complete by 1750 so the results for this verification experiment are the same as those for the full network, within round off errors.

Ann.: Annual

	RegEM			Mann et al. (1998)		
Analysis	**mult RE 1820-1900**	**mult RE 1856-1900**	**globe RE 1856-1900**	**mult RE 1820-1900**	**mult RE 1856-1900**	**globe RE 1856-1900**
no hybrid	N/A	0.25(0.21)	0.78(0.74)	N/A	0.22	0.76
hybrid 20	N/A	0.20(0.20)	0.73(0.76)	N/A	N/A	N/A
No long instrumental or historical records as proxies						
nohybrid	0.10(0.14)	0.05(0.02)	0.55(0.46)	0.19	0.11	0.65
hybrid20	0.12(0.14)	0.04(0.01)	0.53(0.55)	N/A	N/A	N/A

Table 3

Table 3: Comparison of verification scores between the Mann et al. (1998) method and the RegEM method used here, based on annual-mean temperature reconstructions using the full network of 112 predictors. The same grid boxes used by Mann et al. (1998) for calibration and for verification were used in the RegEM cases. The RE scores are presented for both the multivariate (mult) and global mean (globe) cases. Here, the global mean is the average of all the available Mann et al. (1998) verification grid boxes, not just those in the NH. (N/A indicates the scores are not relevant). Numbers in parentheses indicate the calibration period ends in 1971 (see text for details) all other calibration periods end at 1980.

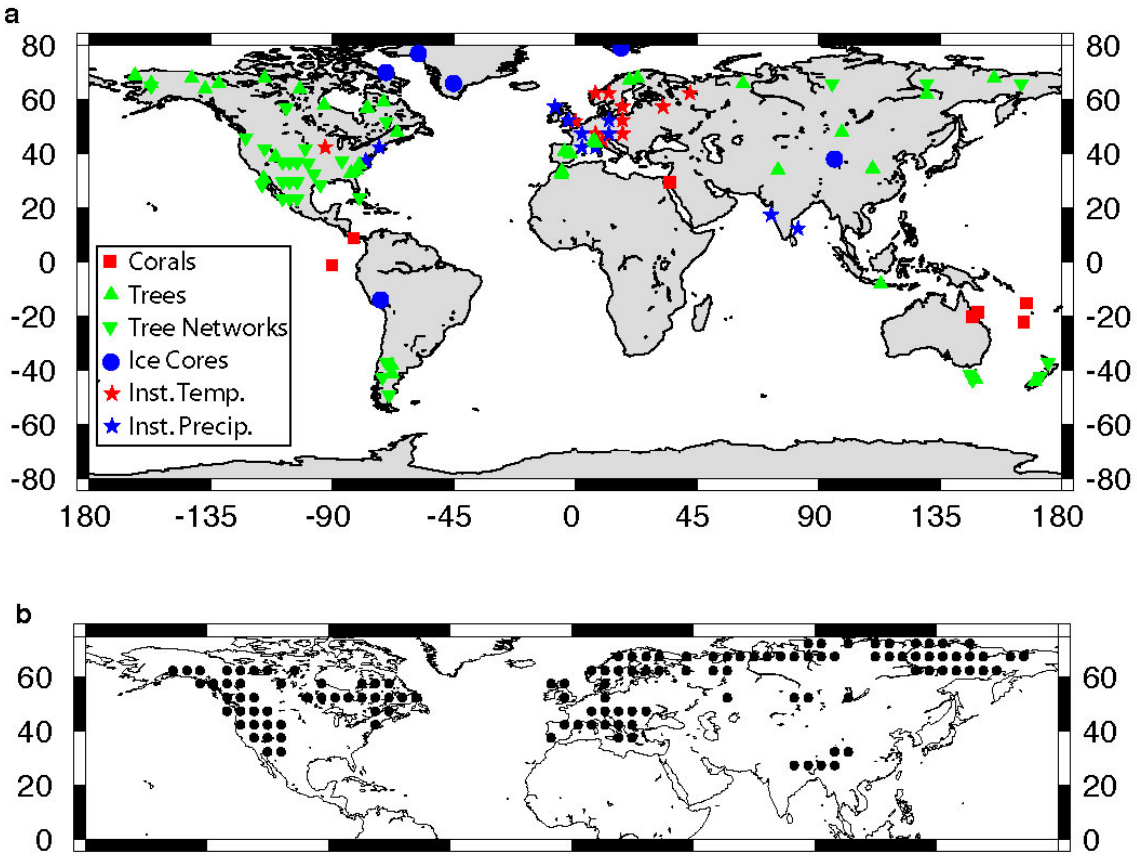
a
80
60
40
20
0
-20
-40
-60
-80
180
-135
-90
-45
0
45
90
135
180
Corals
Trees
Tree Networks
Ice Cores
Inst. Temp.
Inst. Precip.
b
60
40
20
0

Figure 1

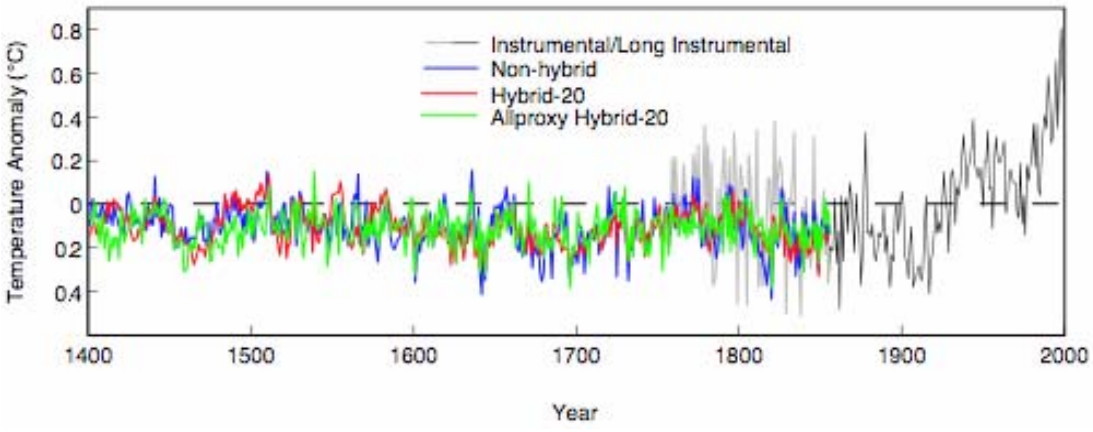
Instrumental/Long Instrumental
Non-hybrid
Hybrid-20
Allproxy Hybrid-20
Temperature Anomaly (°C)
0.8
0.6
0.4
0.2
0
0.2
0.4
1400
1500
1600
1700
1800
1900
2000
Year

Figure 2

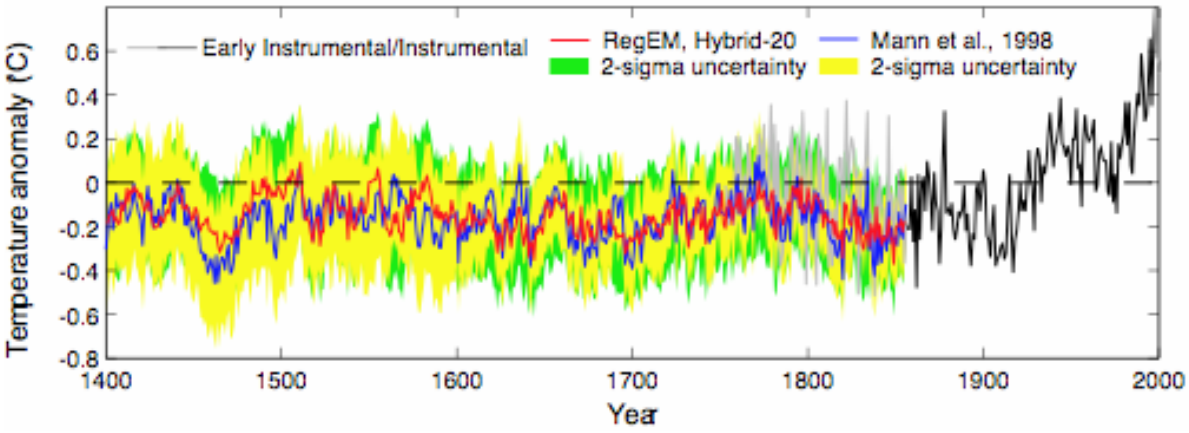
Early Instrumental/Instrumental
RegEM, Hybrid-20
2-sigma uncertainty
Mann et al., 1998
2-sigma uncertainty
Temperature anomaly (C)
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
1400
1500
1600
1700
1800
1900
2000
Year

Figure 3

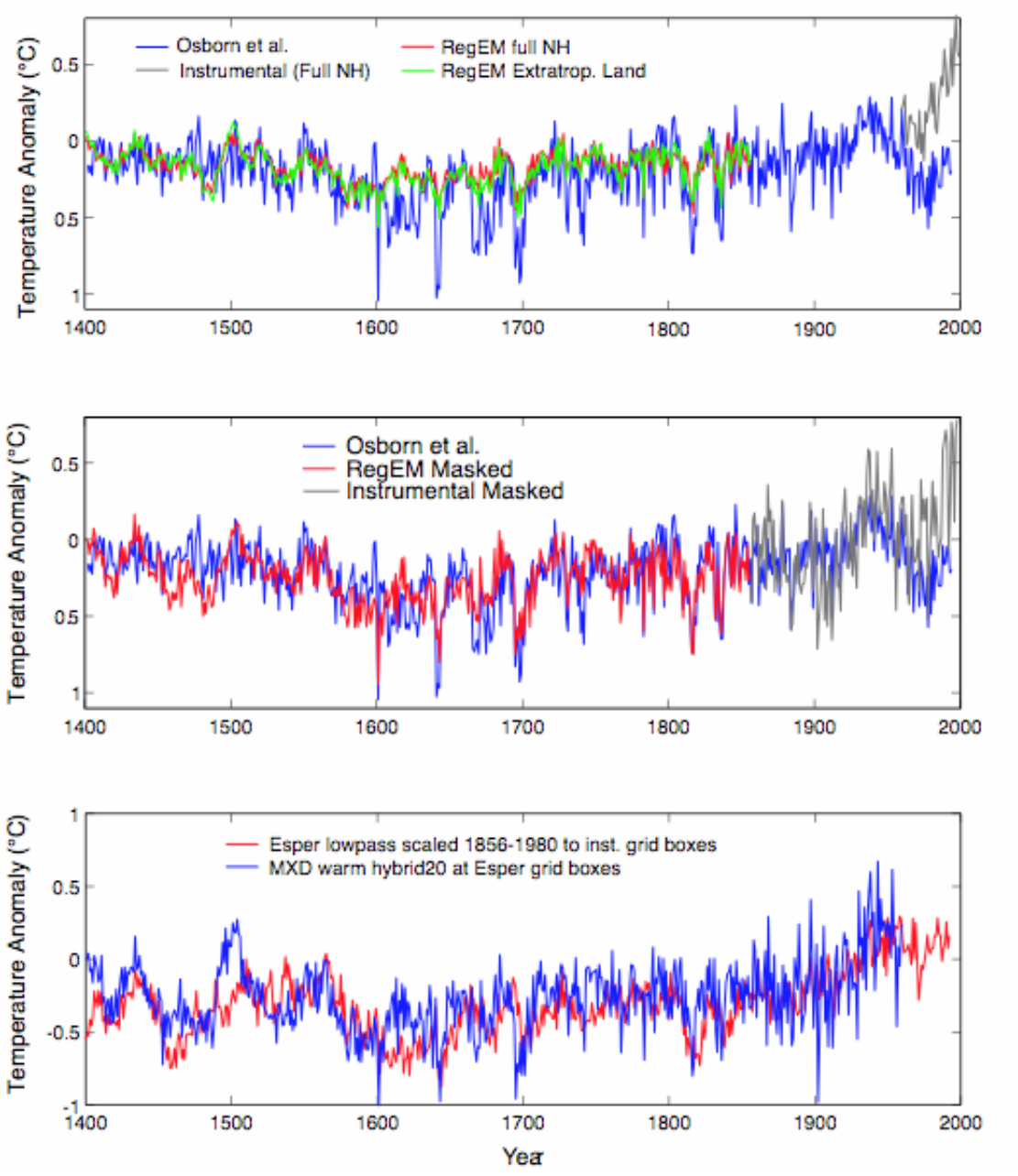
Osborn et al.
Instrumental (Full NH)
RegEM full NH
RegEM Extratrop. Land
Temperature Anomaly (°C)
Osborn et al.
RegEM Masked
Instrumental Masked
Temperature Anomaly (°C)
Esper lowpass scaled 1856-1980 to inst. grid boxes
MXD warm hybrid20 at Esper grid boxes
Temperature Anomaly (°C)
Year
1400
1500
1600
1700
1800
1900
2000

Figure 4

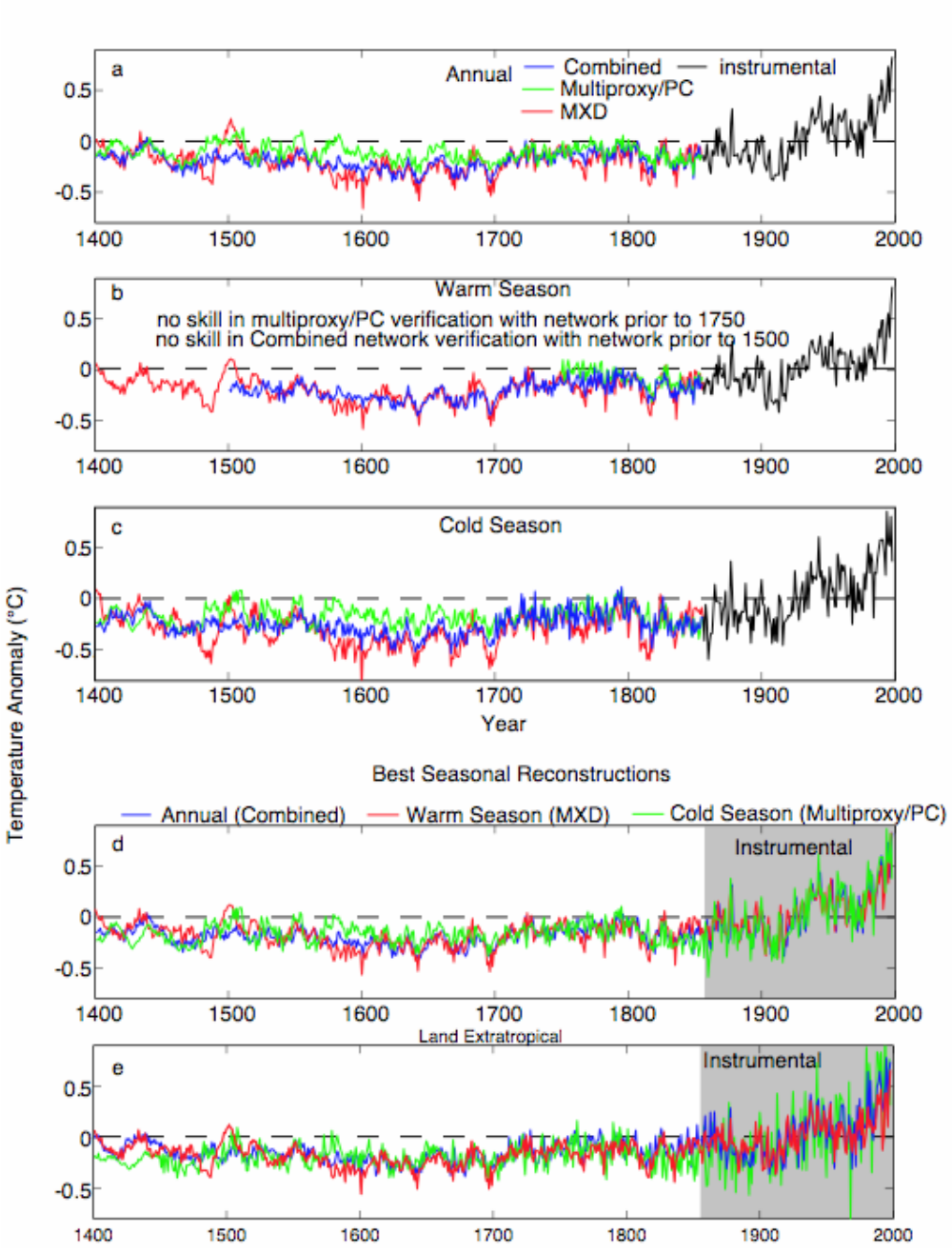
a
Annual
Combined
instrumental
Multiproxy/PC
MXD
b
Warm Season
no skill in multiproxy/PC verification with network prior to 1750
no skill in Combined network verification with network prior to 1500
c
Cold Season
Temperature Anomaly (°C)
Year
Best Seasonal Reconstructions
Annual (Combined)
Warm Season (MXD)
Cold Season (Multiproxy/PC)
d
Instrumental
Land Extratropical
e
Instrumental
0.5
0
-0.5
1400
1500
1600
1700
1800
1900
2000

Figure 5

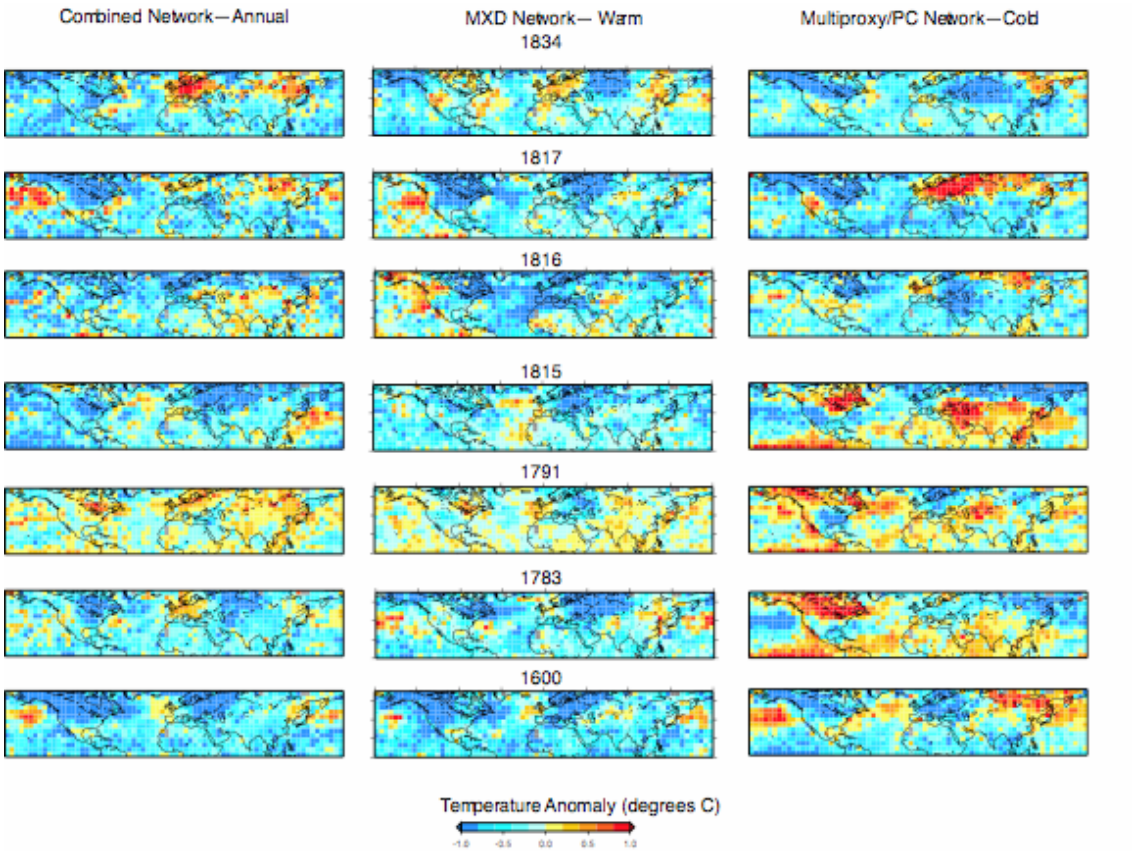
Combined Network—Annual
MXD Network— Warm
Multiproxy/PC Network—Cold
1834
1817
1816
1815
1791
1783
1600
Temperature Anomaly (degrees C)

Figure 6

Figure 1: Distribution of proxies for the two networks used in this study. (**a**) the multiproxy/PC network of Mann et al. (1998). (**b**) the age-banded maximum latewood density (MXD) network of Briffa et al. (2001) where each dot corresponds to the center of one 5° by 5° grid box.

Figure 2. Comparisons of stepwise non-hybrid and hybrid-20 reconstructions for annual mean using the multiproxy/PC network and the hybrid-20 reconstruction using the full multiproxy network. Also shown is the instrumental record consisting of the few available data from 1750-1856 (Mann, 2002a) and the Jones et al. (1999) data for 1856-2000.

Figure 3. Comparison between the RegEM-based hybrid-20 annual mean reconstruction (using the Mann et al. multiproxy/PC network) and Mann et al. (1998) showing overlapping uncertainties in both estimates. Also shown is the long instrumental record (see caption for Figure 3) for comparison.

Figure 4. Comparison of summer mean temperatures based on the MXD network (Briffa et al., 2001, 2002a, 2002b) using the RegEM hybrid-20 method and that of Osborn et al. (2004). (a) The RegEM full NH mean and extratropical land only mean versus Osborn et al. (submitted) (see text for details) (b) Comparison using the RegEM reconstructed grid boxes that coincide with those reconstructed by Osborn et al. (2004) and Briffa et al. (2002b). (c) MXD summer mean reconstruction based on the restricted set of grid boxes sampled in an independent warm-season extratropical temperature reconstruction by Esper et al. (2002) (see text for details).

Figure 5. Comparisons of the RegEM-based results for different seasons and proxy networks. (a) annual mean reconstruction using the three proxy networks (multiproxy/PC, MXD and Combined). (b) warm season comparison and (c) cold season comparison. (d) shows the three seasonal reconstructions produced by the best network for each season based on the verification scores, while (e) is the same comparison restricted to the extratropical continents.

Figure 6. Maps of the full field reconstruction for interesting years using the three different proxy networks (multiproxy/PC, MXD and Combined) to reconstruct three seasonal targets. The years shown are the following: 1600, the year of the Huaynaputina (Peru) eruption; 1783, the year of the Laki eruption in Iceland; 1791, an El Nino year (Quinn and Neal 1992); 1816, the "Year Without a Summer" following the Tambora eruption in 1815; and 1834, an exceptionally warm year in Europe (Mann et al., 1998).