ROBUSTNESS OF THE MANN, BRADLEY, HUGHES RECONSTRUCTION OF NORTHERN HEMISPHERE SURFACE TEMPERATURES: EXAMINATION OF CRITICISMS BASED ON THE NATURE AND PROCESSING OF PROXY CLIMATE EVIDENCE

EUGENE R. WAHL[1] and CASPAR M. AMMANN[2]

[1] Environmental Studies and Geology Division, Alfred University, Alfred, New York, U.S.A.
One Saxon Dr., Alfred, NY 14802 *wahle@alfred.edu* 607.871.2604 607.871.2697 (fax)
[2] National Center for Atmospheric Research*, Boulder, Colorado, U.S.A.
The authors contributed equally to the development of the research presented.

Abstract. The Mann et al. (1998) Northern Hemisphere annual temperature reconstruction over 1400-1980 is examined in light of recent criticisms concerning the nature and processing of included climate proxy data. A systematic sequence of analyses is presented that examine issues concerning the proxy evidence, utilizing both indirect analyses via exclusion of proxies and processing steps subject to criticism, and direct analyses of principal component (PC) processing methods in question. Altogether new reconstructions over 1400-1980 are developed in both the indirect and direct analyses, which demonstrate that the Mann et al. reconstruction is robust against the proxy-based criticisms addressed. In particular, reconstructed hemispheric temperatures are demonstrated to be largely unaffected by the use or non-use of PCs to summarize proxy evidence from the data-rich North American region. When proxy PCs are employed, neither the time period used to "center" the data before PC calculation nor the way the PC calculations are performed significantly affects the results, as long as the full extent of the climate information actually in the proxy data is represented by the PC time series. Clear

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

convergence of the resulting climate reconstructions is a strong indicator for achieving this criterion. Also, recent "corrections" to the Mann et al. reconstruction that suggest 15th century temperatures could have been as high as those of the late-20th century are shown to be without statistical and climatological merit. Our examination does suggest that a slight modification to the original Mann et al. reconstruction is justifiable for the first half of the 15th century (~ +0.05°), which leaves entirely unaltered the primary conclusion of Mann et al. (as well as many other reconstructions) that both the 20th century upward trend and high late-20th century hemispheric surface temperatures are anomalous over at least the last 600 years. Our results are also used to evaluate the separate criticism of reduced amplitude in the Mann et al. reconstructions over significant portions of 1400-1900, in relation to some other climate reconstructions and model-based examinations. We find that, from the perspective of the proxy data themselves, such losses probably exist, but they may be smaller than those reported in other recent work.

1. Introduction

The Northern Hemisphere mean annual temperature reconstruction of Mann, Bradley, and Hughes (Mann et al., 1998, 1999; hereafter referred to as "MBH98" and “MBH99”, respectively, or together as "MBH”) is one of a growing set of high-resolution reconstructions of global/hemispheric surface temperatures that cover all or significant portions of the last millennium (e.g., Rutherford et al., 2005; Cook et al., 2004; Mann and Jones, 2003; Crowley et al., 2003; Esper et al., 2002; Briffa et al., 2001; Crowley and Lowery, 2000; Jones et al., 1998; as well as the decadal-resolution reconstruction of Bradley and Jones, 1993). Additionally, Moberg et al. (2005) and Huang (2004) present different approaches in which high frequency (annual)

and low frequency (multidecedal-centennial) information are combined. [See Jones and Mann (2004) for a recent summary of climate reconstructions over the last two millennia.] The MBH reconstruction was the first of this group to assimilate multiple kinds and lengths of climate proxy data sets into eigenvector-based climate field reconstruction (CFR) techniques (cf. Kaplan et al., 1997; Cane et al., 1997; Evans et al., 2002). The ability to reconstruct large-scale climate fields (in this case surface temperature) represents a major breakthrough provided by CFR techniques, which in turn can be used to estimate global/hemispheric mean temperatures. Reconstructions of climate fields allows much more nuanced evaluation of past climate dynamics (e.g., Raible et al., in press) than do simple large-scale averages, a point that has been obscured in the reexaminations of the MBH Northern Hemisphere temperature record considered here. Although the MBH reconstruction quite closely resembles previous (Bradley and Jones, 1993) and also more recent reconstructions (cf. Rutherford et al., 2005; Mann and Jones, 2003), its "hockey stick"-shaped result of a slow cooling trend over the past millennium (the “stick”) followed by anomalous 20th century warming (the “blade”) has been widely cited and featured in international assessments of global climate change, most prominently in the Intergovernmental Panel on Climate Change Third Assessment Report (IPCC-TAR)(Folland et al., 2001). As an important scientific work, the MBH reconstruction has been subjected to significant scrutiny, which can be categorized into three areas.

First, the MBH reconstruction has been examined in light of its agreement/lack of agreement with other long-term annual and combined high/low frequency reconstructions. In particular, the amplitude (difference from a 20th century reference) of the Northern Hemispheric mean surface temperatures in MBH is significantly less at some times during the millennium than the amplitude of some other long-term reconstructions (notably, Moberg et al., 2005; Esper

et al., 2002; Briffa et al., 2001; Harris and Chapman, 2001; Huang et al., 2000). New work by Rutherford et al. (2005) suggests that the extra amplitude of the tree ring-based reconstruction of Esper et al. (2002; cf. Cook et al., 2004) can be explained as the expected result of significantly restricted sampling (14 extra-tropical, continental tree-ring sites) of the spatially variable hemispheric temperature field. In contrast, the MBH99 reconstruction is calibrated using data from a minimum of 36 sites over AD 1000-1399 (12 actual predictands--employing principal component, or PC, summaries of dense proxy networks--including southern tropical and northern high-latitude ice core data along with temperate- and high-latitude tree ring data, with at least one proxy site on each of five continents) and a maximum of 415 proxy and instrumental records over 1820-1980 (112 actual predictands, with greater spatial coverage on all continents except Africa and a small number of tropical oceanic sites that include coral proxies). Similarly, Mann et al. (2005), in a model-based exercise, suggest that the Moberg et al. frequency banding has a tendency to exaggerate reconstruction amplitude, thus overstating the possible amplitude loss in MBH. The comparison of the MBH reconstruction, derived from multi-proxy (particularly tree ring) data sources, with widespread bore-hole-based reconstructions (Harris and Chapman, 2001; Huang et al., 2000) is still at issue in the literature (Chapman et al., 2004; Schmidt and Mann, 2004; Mann and Schmidt, 2003; Mann et al., 2003; Rutherford and Mann, 2004).

Second, a related area of scrutiny of the MBH reconstruction technique arises from an atmosphere-ocean general circulation model (AOGCM) study (von Storch et al., 2004), which also examines the potential for loss of amplitude in the MBH method (and other proxy/instrumental reconstructions that calibrate by using least squares projections of the proxy vectors onto a single- or multi-dimensional surface determined by either the instrumental data or its eigenvectors). This kind of analysis using the virtual climates of AOGCMs allows the long-

term quality of climate reconstructions to be assessed directly, since the modelled reconstruction can be compared to "real" model temperatures (known everywhere in the 2D surface domain). In the real world, such a comparison is only possible, at best, over the recent ~ 150 years and over a restricted spatial domain. However, a number of issues specific to the modelling situation could arise in this context, including: how realistically the AOGCM is able to reproduce the real world patterns of variability and how they respond to various forcings; the magnitude of forcings and the sensitivity of the model that determine the magnitude of temperature fluctuations (see Shindell et al., 2003, 2001; Waple et al., 2002; MBH99); and the extent to which the model is sampled with the same richness of information that is contained in proxy records (not only temperature records, but series that correlate well with the primary patterns of temperature variability--including, for example, precipitation in particular seasons). In addition, the MBH method itself was inaccurately implemented by von Storch et al., leading to erroneously high amplitude losses (Wahl et al., accepted). Another model-based examination paralleling the experimental structure of von Storch et al. has been undertaken by Mann et al. (2005), who find no under-expression of amplitude using the more recent "regularized expectation maximization" (RegEM) CFR technique (Rutherford et al., 2005), which yields reconstructions highly similar to MBH.

A third area of scrutiny has focused on the nature of the proxy data set utilized by MBH, along with the characteristics of pre-processing algorithms used to enhance the climate signal-to-noise characteristics of the proxy data (McIntyre and McKitrick, 2003, 2005a, b; hereafter referred to as "MM03", "MM05a", and "MM05b", respectively, or together as "MM"). In MM03, a reconstruction for Northern Hemisphere mean surface temperature from 1400-1980 is presented that is clearly inconsistent with the hockey-stick result of anomalous 20th century

warming, showing not only relatively high temperatures in the 20th century, but also sustained temperatures in the 15th century that are *higher* than any sustained temperatures in the 20th century--descriptively, a "double-bladed hockey stick" (i.e., an upward trend at either end of the reconstruction period). In MM03, the authors describe this result as being developed using the MBH reconstruction methodology, albeit with elimination of a large number of the proxy data series used by MBH, especially during the 15th century. In MM05b, a second version of the MM reconstruction is presented that the authors describe as virtually identical to the one presented in MM03. The version in MM05b is based on the substitution of a new version of a single tree ring series from northeastern Canada's St. Anne Peninsula (the "Gaspé" series) and on newly-computed PC summary series for North American tree ring data in MBH that are derived from the International Tree Ring Data Base (ITRDB), discussed below [cf. section 2.4(5)]. Although the authors state that the MM03 and MM05b reconstructions are nearly identical, the high excursions of 15th century temperatures in MM03 are clearly larger than those in MM05b (by ~ 0.2°) and are much more clearly differentiated from late 20th century temperatures, especially in relation to instrumental temperatures in the 1990s and the current decade. Thus, the strongest departure from the single-bladed hockey stick depiction of highly anomalous temperatures in the later 20th century is presented in MM03. Associated validation statistics for these reconstructions are not given in either MM03 or MM05b, and thus it is not possible to gauge directly from the published articles the climatological meaningfulness of the time series and thereby evaluate the authors' criticism of the single-bladed hockey stick result.

Neither of the first two areas of scrutiny (congruence with other reconstructions and potential loss of reconstruction amplitude in model virtual worlds) have fundamentally challenged the MBH conclusion of an anomalous rise (in terms of both duration and magnitude)

in surface temperatures during the late 20th century (Jones and Mann, 2004; von Storch et al., 2004; Moberg et al., 2005). The conclusions emphasized by MM, however, warrant a more complete assessment because they question the fundamental structure of climate evolution over time, and thus the corresponding interpretation of cause (external forcing) and effect, over much of the last millennium (e.g., Jones and Mann, 2004; Crowley, 2000; MBH98). In particular, MM present two climate reconstructions, both of which imply that the early 15th century could have experienced the warmest sustained surface temperatures in the Northern Hemisphere in the past six centuries; all other last-millennium and last-two-millennium climate reconstructions do not show temperatures anywhere near the high MM 15th century values over their entire lengths prior to the twentieth century (Jones and Mann, 2004; Moberg et al., 2005). In fact, MBH tends to be among the warmest of these reconstructions in the early 15th century period of greatest contention (Jones and Mann, 2004), although still much lower than MM.

In this paper, we examine the MM results and criticisms of proxy data and methods employed by MBH, based on our own independent emulation and use of the MBH method to reconstruct Northern Hemisphere mean surface temperature over 1400-1980. In addition, we also address aspects of logic in the arguments presented in MM05a and MM05b and the absence of confirmatory validation statistics in MM03 and MM05b, since these issues have bearing on the efficacy of the MM results and have not yet been addressed in the scientific literature. We also use our results to briefly address the issue of loss of amplitude, from the perspective of the information contained in the proxy data themselves.

1.1 DETAILS OF MM CRITICISMS

A major focus of MM has been on two aspects of the primary proxy data used in the MBH reconstruction. First, MM challenge the use of specific proxy records deemed unjustifiable because of suggested lack of temperature information contained in the time series, duplication of series, or improper extrapolation/infilling of missing data (MM03 and MM05b). Second, the authors strongly criticize the method applied by MBH to generate PC summaries (MM05a/b) of spatially-dense proxy networks in North America prior to the actual calibration of the proxies to the instrumental climate field, which is a common technique in high-resolution paleoclimatology. In essence, this technique, derived from the field of dendroclimatology (Fritts, 1976), attempts both to downweight the impact of data-dense regions on statistical calibrations between proxies and instrumental values and to lower the potential impact of calibration "overfitting". In paleoclimatology, the latter situation occurs when a large number of proxy regressors in a least squares projection improves the fit between proxies and climate values during the calibration period, but in the process this fit becomes so specific to the calibration period that it does not work well when applied in the prediction period--including in an independent "verification" period. Reducing the number of individual predictors of data-dense regional proxy networks to a few PCs helps alleviate this problem, and, at the same time, aids in isolating the climate signal in the proxy variations from non-climate noise.

Relative to MBH, MM05a question the method and period over which the proxy values are transformed into standardized anomalies before the PC extraction procedure is performed. MM05a claim that the MBH method as originally applied (standardization and transformation over the calibration period, 1902-1980, before PC extraction) leads to proxy PC summary series that inappropriately weight for hockey stick-bladed shapes for PC1 (the leading component of

variance) in the 20th century. The temporal information captured by PC1 of the North American tree ring network is a crucial source of information in calibration for the 11th-14th century reconstruction (MBH99), and in verification for the early 15th century reconstruction in MBH98, as shown here (cf. section 3.2). A further aspect of this critique is that the single-bladed hockey stick shape in proxy PC summaries for North America is carried disproportionately by a relatively small subset (15) of proxy records derived from bristlecone/foxtail pines in the western United States, which the authors mention as being subject to question in the literature as local/regional temperature proxies after approximately 1850 (cf. MM05a/b; Hughes and Funkhauser, 2003; MBH99; Graybill and Idso, 1993). It is important to note in this context that, because they employ an eigenvector-based CFR technique, MBH do not claim that all proxies used in their reconstruction are closely related to local-site variations in surface temperature. Rather, they invoke a less restrictive assumption that "whatever combination of local meteorological variables influence the proxy record, they find expression in one or more of the largest-scale patterns of annual climate variability" to which the proxy records are calibrated in the reconstruction process (Mann et al., 2000). MM directly note the link between bristlecone/foxtail pines and precipitation (p. 85, MM05b), which is exactly the kind of large-scale pattern registration that the MBH CFR method takes as axiomatic because large portions of this region are known to have important ENSO/precipitation teleconnections (cf. Cole and Cook, 1998; Rajagopalan et al., 2000). Since ENSO has a strong role in modulating global temperatures as well as affecting regional precipitation patterns, a CFR method of temperature reconstruction can effectively exploit regional ENSO/precipitation teleconnections that register in proxy data.

The salient implication of these critiques in MM05a/b is that inappropriate variance is carried into the calibration process by proxy summary records that presumably exhibit an artificial single-bladed shape in the 20th century, and thus these summary data series will be given excessively high weight in the calibration process because eigenvector1/PC1 of the instrumental temperature field is strongly associated with the upward trend in 20th century instrumental temperatures. This relatively high weighting will then be carried by these proxies throughout the reconstruction period, affecting the shape of the reconstruction for the entire 14th-20th century span. Logically, this situation suggests that such a biased weight given to an artificial signal, which is not part of the proxy records themselves, could introduce a climate reconstruction that is not appropriately supported by the original, underlying data. The primary analytical structure of this paper, as outlined in section 2.4, considers these MM proxy criticisms in detail.

In MM05a/b, the authors also examine two issues concerning validation statistics and their application by MBH. The first issue concerns which statistics should be applied as validation measures; the second issue concerns estimating appropriate threshold values for significance for the reduction of error (RE) statistic, which is commonly used as a validation measure in paleoclimatology (Fritts, 1976; Cook et al., 1994). MM discuss the latter issue in the context of their argument concerning methods for deriving PC summary series from proxy data. We examine the issue of appropriate validation statistics in section 2.3 and in greater technical detail in Appendix 1. We consider the issue of appropriate thresholds for the RE statistic in Appendix 2, based on analysis and results reported elsewhere (Ammann, C.M. and E.R. Wahl, 'Comment on “Hockey sticks, principal components, and spurious significance” by S. McIntyre and R. McKitrick', in review with *Geophysical Research Letters*).

2. Methods

2.1 REPRODUCTION OF MBH98 METHODOLOGY

Before addressing issues related to the MM criticisms, we first introduce our own reproduction of the MBH algorithm for reconstructing climate. It is important to note that in this replication we retain the primary MBH assumptions concerning the stationarity of the relationship between proxy climate indicators and eigenvector spatial patterns of temperature anomalies and the relative temporal stability of the large-scale patterns themselves. These fundamental assumptions have not been seriously challenged and are thus not subject in the ongoing debate, although they do represent one level of uncertainty of the overarching approach (see discussion of this issue in Mann et al., 2000). Our implementation of the MBH method was based on published descriptions (MBH; Mann et al, 2000) and on additional material (clarifications and data) available via a "ftp" site maintained by M.E. Mann (http://www.meteo.psu.edu/~mann/shared/research/MANNETAL98/). The same data and supporting material are also available through *Nature* online supplementary information, as updated in 2004 (Mann et al., 2004). Dr. Mann also answered clarifying questions concerning a few details of the method.

The core calibration and reconstruction components of the MBH method were first translated into matrix algebra formulae. The entire algorithm was then coded in the publicly available "R" software (R Development Core Team, 2004), which has particular advantages for implementing matrix algebra equations. The calibration component was simplified to have a consistent annual time step throughout, rather than the temporally heterogeneous method used in MBH of calculating PCs of the instrumental data based on a monthly time step and subsequently

forming annualized PC series from averages of the monthly values (MBH).[*] A second simplification eliminated the differential weights assigned to individual proxy series in MBH, after testing showed that the results are insensitive to such linear scale factors. The MBH technique of specifying 11 separate calibrations between the proxy data and specific numbers of retained PCs from the global instrumental data grid (for different sets of proxy data richness going back in time) was followed exactly; thus, specific calibrations were used for each of the following periods: 1400-1449, 1450-1499, 1500-1599, 1600-1699, 1700-1729, 1730-1749, 1750-1759, 1760-1779, 1780-1799, 1800-1819, and 1820-1980. The standard *calibration* period of 1902-1980 and independent *verification* period of 1854-1901 also followed MBH exactly. For the purposes of testing our emulation of the MBH algorithm, the instrumental and proxy data used (Jones and Briffa, 1992, updated to include 1993) were the same as those employed in the original MBH reconstructions, even though newer versions of instrumental data exist (Jones and Moberg, 2003). Our full code, complete data series (instrumental and proxy), and results for all the scenarios outlined in section 2.4 are accessible in a user-friendly format through our website at the National Center for Atmospheric Research (NCAR), USA (http://www.cgd.ucar.edu/ccr/ammann/millennium/MBH_reevaluation.html). In addition, further technical details are provided in several appendices to this article (see below).

2.2 RETENTION OF INSTRUMENTAL EIGENVECTORS/PC SERIES IN CALIBRATION

[*] MBH noted that use of monthly data *insures* a formally over-determined PCA decomposition. This occurs because the number of monthly time steps is greater than the maximum number of eigenvectors possible from a global instrumental data field with 1082 components (Jones and Briffa, 1992). This stipulation is not a strict necessity statistically (cf. Zorita et al., 2003), and annual mean data can be used if the eigenvalue noise spectrum falls off quickly enough, as happens in this case.

When decomposing a temperature field into its eigenvectors and corresponding PCs, a choice has to be made regarding the number of PCs to be retained. Here, the numbers of PCs from the global instrumental data grid to which the various proxy sets were calibrated were held the same as those used in the corresponding calibrations of the MBH98. We do this for all scenarios (cf. 1, 5 and 6, in section 2.4 below) where the numbers of proxies assimilated into calibration are equal or similar to those used in MBH. For scenarios that assimilate larger numbers of individual proxy series into calibration (especially scenarios 2 and 3, cf. section 2.4 below), the method is justified by experiments in which results for the two 15th century calibrations were replicated by retaining the first four instrumental PCs (rather than the MBH specification of one PC for 1404-1449 and two for 1450-1499). The results of these replications demonstrate robustness of the Northern Hemisphere temperature reconstructions to the number of retained instrumental PCs; the calibration and verification statistics are, at most, only slightly different from those of the original experiments (with no change of meaning), and the actual reconstructed temperature series that result are virtually indistinguishable from those of the original experiments. Thus justified, the use of the same number of instrumental "training" PCs in the scenarios vis-à-vis the corresponding calibrations in the MBH emulation allows the pure effects of eliminating/changing proxy summary PCs and truncations of proxy richness to be isolated, as described in section 2.4.

2.3 CALIBRATION AND VERIFICATION STATISTICS--VALIDATION

Each reconstruction is evaluated regarding its skill in reproducing instrumental data over both the calibration and independent verification periods. Good *validation* performance in both of these periods is considered necessary to infer that a paleoclimate reconstruction is likely to

yield good quality in the target period (i.e., the pre-verification period), during which instrumental data are largely/totally absent (Fritts, 1976; Cook et al., 1994). In choosing validation measures for assessing the calibration and verification performance of the scenarios in this paper, we have sought accurate measures of reproduction quality for both interannual and low-frequency mean behavior that simultaneously support an explicit balance of *jointly* reducing the likelihood of false positive and false negative errors. A *false positive error* occurs when a reconstruction is accepted as being of useful quality, but which in fact is of poor quality (avoidance of this kind of error is the typical logical standard used in validating paleoclimate reconstructions). A *false negative error* occurs when a reconstruction is rejected as of poor quality, but which in fact is of useful quality (Wahl, 2004; Lytle and Wahl, 2005) (avoidance of this kind of error has been typically underappreciated in paleoclimate validation). These two kinds of errors and their corresponding correct outcomes--*true positive* acceptance of a useful quality reconstruction and *true negative* rejection of a poor quality reconstruction--comprise a general set of test outcomes that is used as an assessment tool in a wide variety of scientific disciplines, including medicine, psychology, signal theory, and weather forecasting (formally developed as "receiver operating characteristic" or ROC analysis; cf. Green and Swets, 1988). In the absence of specific criteria that suggest otherwise, a reasonable default criterion in this kind of analysis is to develop a test or tests (here, validation of reconstruction performance) that attempt to jointly minimize both false positive and false negative errors (Wahl, 2004).

In the process of validating paleoclimate reconstructions, the time scale of primary interest is a key parameter in the search for an appropriate balance in the joint reduction of the two errors. Generally, this consideration can be looked at in terms of two contrasting temporal foci--short term (or high frequency) and long term (or low frequency), with more precise

definition of the actual periods involved determined for a specific kind of reconstruction. In high-resolution paleoclimatology, high frequency typically denotes an interannual (even subannual) scale, and one useful definition of low frequency is the span of the verification period--in this case approximately one-half century. We use these definitions here. Three possible situations for evaluating the balance of false positive and false negative errors are available under these definitions: 1) primary focus on high frequency fidelity between climate reconstructions and instrumental data; 2) primary focus on both high and low frequency fidelity between reconstructions and data; and 3) primary focus on low frequency fidelity between reconstructions and data. In (1) a reconstruction that detects interannual variations well would be accepted as useful, even if its ability to detect semi-centennial-scale variations was poor, because of the primary focus on high frequency fidelity. In (2) a reconstruction would need to detect variations well at both time scales to be accepted as useful. In (3) a reconstruction that detects semi-centennial-scale variations well would be accepted as useful, even if its ability to detect interannual variations was poor, because of the primary focus on low frequency fidelity. In addition, it can be reasonable to employ a mixed set of these standards across the calibration and verification periods, if there are particular climate phenomena the reconstructions are intended to register whose time scales of primary interest vary across these periods.

The primary issue in the MM criticisms of the MBH reconstruction concerns Northern Hemisphere surface temperature over the first half of the 15th century; specific years are not the key concern here, but multi-decadal averages are. Thus, in this analysis the most generally appropriate temporal criterion from those listed is (3), a primary focus on low frequency fidelity in validation. It is secondarily appropriate in this case to separate the criteria for the calibration and verification periods; with criterion (2) being used in calibration to be as stringent as possible

over the period in which the actual reconstruction model is fitted, and criterion (3) being used in the verification period in light of the fact that the most important feature in this period is its large downward shift in *mean* from that of the calibration period. In this way, we explicitly seek to avoid inappropriate false negative judgments in the verification period if there are cases of good performance in capturing the mean offset between the calibration and verification periods, but interannual fidelity is poor. This criterion ensures that objectively validated low frequency performance, at the time scale of primary interest, is not rejected because it is coupled with poor high frequency performance. In these judgments, we disagree with MM05a/b, who indirectly insist that criterion (2) be employed as the only valid temporal standard over both the calibration and verification periods. [MM05a/b claim that interannual-oriented statistics such as Pearson's product moment correlation coefficient (r), the sign test, and the coefficient of efficiency (CE) (Cook et al., 1994) should be used as key validation measures along with the temporally multi-valent RE statistic (cf. below and Appendix 1 for further evaluation of these statistics), even though the crux of their criticism of the actual MBH reconstruction is at the semi-centennial scale.] MM05a/b do not consider the need to jointly avoid both false positive and false negative errors, and the generalized temporal validation approach they advocate incurs significant risk of generating a false negative outcome for the frequency band of most interest in the MBH context. We avoid this unbalanced risk by using the temporal validation criteria outlined.

To implement these criteria, we use the RE statistic, and additionally the verification period mean offset (difference between the reconstructed and instrumental means), as measures of validation merit. In the verification period, RE (by design) registers a combination of interannual and mean reconstruction performance (Cook et al., 1994; Appendix 1 here), whereas the verification mean offset specifically registers multi-decadal verification performance. In the

calibration period, RE is equivalent to using empirical R^2 (the ratio of the climate variation explained by the reconstructed (fitted) values to the total climate variation), which is an interannual-oriented measure by definition, but with the important addition that a reconstruction to which it is applied must also register any trend in order to yield high values (strong trends are the case in the 1902-1980 and 1902-1971 calibration periods used here). This combination of measures includes both interannual and low frequency information in a way intended to balance joint minimization of false positive and false negative errors; we apply this balance consistently across all the reconstruction scenarios presented. Numerically, we consider successful validation to have occurred if RE scores are positive, and failed validation to have occurred if RE scores are negative (Ammann and Wahl, in review; Appendix 2). This threshold also has the empirical interpretation that reconstructions with positive RE scores possess skill in relation to substituting the calibration period mean for the reconstructed climate values. We specifically avoid interannual-oriented statistical measures that cannot recognize successful reproduction of differences in mean between the calibration and verification periods (such as Pearson's *r*, the sign test, and CE), or that cannot detect addition to/multiplication of a time series by arbitrary constants (such as Pearson's *r*), as being inappropriate in terms of properly balancing the joint reduction of false positive and false negative errors in this context (cf. Appendix 1).

A more general discussion of validation in paleoclimate reconstruction is provided in Appendix 1, including further consideration of joint reduction of false positive and false negative errors and examination of a suite of statistics (cf. Cook et al., 1994) that might be used for comparison of reconstructed and actual data series. Particular emphasis is given to issues concerning the use of Pearson's *r* for validation of last-millennium climate reconstructions. Although we find Pearson's *r* and CE not germane to use in this paper for decision-making

purposes--for the reasons described above and in Appendix 1, we include the values of these measures in Appendix 1 for all the scenarios we examine so that they are available to the community.

2.4 RECONSTRUCTION SCENARIOS REFLECTING MM CRITICISMS

After testing verified that our algorithm could reproduce the original MBH results (Fig.1), we developed six reconstruction scenarios to examine the different criticisms of the MBH approach made by MM. Scenario 1 examines the MM03 climate reconstruction. Scenarios 2-4 examine the influence of the North American tree ring data, with inclusion/exclusion of the bristlecone/foxtail pine records, by utilizing the proxies *individually* rather than through employing PC summaries. Scenarios 5 and 6 examine the influence resulting from various approaches to generating PC summaries. Scenario 6 also examines the MM05b climate reconstruction.

1) Examination of MM03 Northern Hemisphere temperature reconstruction

Reconstruction was done over 1400-1499 by exclusion of all the North American proxy data from the ITRDB along with additional series from the North American Southwest and North American tree line, with calibration over the standard period. This scenario mimics the reconstruction scenario developed for this period in MM03 (cf. Rutherford et al., 2005). It allows examination of the statistical validity of the MM03 result that Northern Hemisphere surface temperatures in the 15th century were higher than even later-20th century temperatures. Failure to pass validation tests for either the calibration or verification periods would indicate that the MM03 double-bladed hockey stick reconstruction for the 15th century cannot be

ascribed climatological meaning; passing validation tests for both calibration and verification would indicate climatological meaning for the reconstruction.

2) Replacement of North American ITRDB PC summaries with full proxy network

Reconstructions were done for different configurations representing the proxy networks of 1404-1449, 1450-1499, 1700-1729 and 1800-1819 by employment of *all* proxy data series available for each period, i.e., *without* the use of second-order PC summarizations of the proxies in data-rich regions of North America and Eurasia and excluding all data for the years 1400-1403. Calibration was restricted to 1902-1971. (The truncations of the reconstruction and calibration periods account for criticism in MM03 concerning infilling of values in a small number of proxy records by persistence from the earliest year backward and the latest year forward). The purpose of evaluating the reconstructions in the different proxy network configurations is to check whether the use of all the proxy data without PC summarizations is sensitive to the richness of the proxy data set, focusing on four reconstruction periods with specific qualities: a) the key period at issue in MM (15th century) and b) the warmest (18th century) and coolest (19th century) pre-20th century reconstructions when the proxy set is close to its full richness (415 series) and includes a few long instrumental records. The number of proxy records successively increases for the four scenarios (95, 139, 378, 405). The reconstructions were not restricted to the decades for which they represent the optimal network, but all extend to 1971, so that any outcome that might cast doubt on the hockey stick-blade result in the 20th century would be included in the analysis.

The Gaspé proxy series, noted in MM05b as potentially problematic during the first half of the 15th century due to a small sample size in the site chronology over much of this period, is retained in this scenario (unlike in scenarios 5 and 6 below). This retention is based on

"individual proxy" tests over the entire 15th century that yielded highly similar reconstructions whether this series is included over 1404-1449 or not, and whether it occurs once or is replicated over 1450-1499 [occurring both as an individual record and as part of the ITRDB set used to calculate North American PC summaries--as in MBH98 (cf. MM05b)]. The Twisted Tree/Heartrot Hill boreal treeline proxy highlighted as potentially problematic in the early part of the reconstruction period by MM03 (cf. their Fig. 4 and associated text) is not included in the proxy set used in this scenario. These two criteria make the individual proxy set used in this scenario identical to the one that underlies our emulation of the original MBH98 reconstruction.

With this structure, scenario 2 allows examination of the robustness of the climate information in the proxy data set in relation to the use/non-use of PC summaries of the proxies, by determining whether the data *per se* (isolated from any PC pre-processing) carry only a single-bladed hockey stick shape in the 20th century. If no second blade is found in the 15th century, then the issue of whether this overall result is somehow an artifact of the use of second-order PC proxy summarizations becomes essentially moot.

3) Replacement of PC summaries and elimination of bristlecone/foxtail pine proxy records Reconstruction was done over 1404-1971 using the same structure as scenario 2, but *excluding* the 15 bristlecone/foxtail pine records from western North America identified in MM05a/b as carrying little local/regional temperature information during the calibration and verification periods (although note, a close local proxy-temperature correlation is not strictly required; cf. section 1.1).

Scenario 3 allows examination of the contribution of the bristlecone/foxtail pine records as meaningful proxies in relation to the reconstruction of global eigenvector patterns of surface temperature (in contrast to the relationship between these proxies' ring widths and local/regional

surface temperatures), and whether their exclusion has discernible impact on Northern Hemisphere mean surface temperature reconstructions, especially in the 15th century. If excluding these proxies degrades reconstruction performance in either the calibration or verification periods, then they do carry empirically meaningful information at the scale of global/hemispheric surface temperature variance patterns and this result would argue for their continued inclusion in the MBH-style method. However, independent of whether or not the bristlecone/foxtail pine records carry meaningful large-scale temperature information, should their exclusion have relatively little impact on the magnitude and trajectory of Northern Hemisphere temperature reconstructions over the last 600 years, then the question of whether they inappropriately lead to the single-bladed hockey stick result in the 20th century (i.e., an inappropriate diminishing of amplitude in the 15th century) ceases to be of significant import.

4) Replacement of PC summaries with strong truncation of proxy richness in North America Reconstruction was done over 1404-1971 using the same structure as scenario 2, but randomly *excluding* 50 out of 60 proxy data series in the Southwestern United States (defined as 100-123° W, 32-42° N) from the ITRDB; along with deterministically *excluding* four spatially repetitive ITRDB proxy series in Northwestern North America (defined as 100-124° W, 42-50° N), three spatially repetitive ITRDB proxy series in the Southeastern United States (defined as 76-92° W, 32-37° N), and four spatially repetitive non-ITRDB proxy series ("Stahle" series, MBH) in the Southwestern United States and Northwestern Mexico. Two repetitions of this scenario were done based on independent random exclusions. The Sheep Mountain bristlecone pine series, which MM05b characterize as the (inappropriately) top-weighted proxy in PC1 of the North American ITRDB series in MBH98, is not included in either selection to guard against

such a potential influence. Only two other bristlecone/foxtail series are included in each selection.

Scenario 4 allows an important logical check on the use of scenarios 2 and 3, by determining whether the initial intent of employing PC summations of proxy data from the North American region (to reduce the influence of this spatially restricted, but data-rich region in calibration, and to guard against calibration overfitting) is violated by the use of all the individual proxies in calibration. Since this technique is the key method used to examine the robustness of the climate information in the proxy data in relation to the use/non-use of proxy PC summaries, it is crucial that any potential bias caused by overweighting of the North American proxy data in calibration be identified. If there is little or no difference in the reconstruction results whether the full North American ITRDB *cum* "Stahle" set of proxies is used versus employing strongly truncated subsets of these proxies (~83% reductions for the particularly rich Southwestern data, ~66% reductions for the North American data as a whole), then there is no empirical indication of the presence of spurious overemphasis being given to North America in the full "all proxy" scenarios. Repetition with independent random exclusion for the American Southwestern region allows a further check of the existence/non-existence of an impact. [The division of the truncated proxies into three sets follows the general distribution of direction of ENSO-related teleconnections in the three regions (Cole and Cook, 1998; Rajagopalan et al., 2000). Deterministic exclusions of proxies for the North American Northwest and the United States Southeast were used, rather than random exclusions, due to the small number of spatially-repetitive proxies in each of these regions.]

5) Inclusion of PC summaries with restriction of Gaspé series over 1400-1499

Reconstruction was done over 1400-1499 by *including* PC summaries of the North American

proxy data from the ITRDB: replacing the original MBH-calculated PCs with newly calculated PCs based on excluding the Gaspé series (i.e., removing it as an individual proxy over 1400-1449 and as part of the data set used to calculate the North American ITRDB PCs). Calibration was done over 1902-1980. Reconstruction to 1400 and calibration to 1980 (rather than truncating at 1404 and 1971, respectively, as in scenarios 2-4) follows the periods used in MM05b. Three ways of calculating the PC summaries were employed: a) the original MBH reference period for determining standardized anomalies (1902-1980), which were input into the "svd" routine in R (utilizing the singular value decomposition method of PC extraction); b) the MM05a/b reference periods for determining standardized anomalies (1400-1980 and 1450-1980 for the 1400-1449 and 1450-1499 reconstruction periods, respectively), which were input into "svd"; and c) the MM05a/b reference periods for determining anomalies (centered but unstandardized), which were input into the "princomp" routine in R (which extracts PCs using the 'eigen' routine on the variance-covariance matrix of the data [default setting]). From this data pre-processing, we retained 2 through 5 PC series for representation of the underlying proxy data for each of the three distinct calculation methods; providing 12 PC sets altogether that are then used separately in the subsequent climate reconstructions.

6) Inclusion of PC summaries with restriction of the Gaspé series and elimination of bristlecone/foxtail pine records Reconstruction was done over 1400-1499 using the same structure as scenario 5, with the additional elimination of the 15 bristlecone/foxtail pine records from the North American ITRDB data prior to PC calculations. (The 15 records eliminated are the same as those eliminated in scenario 3.)

Scenarios 5 and 6 allow direct examination of the MM05a/b criticism that a single-bladed hockey stick shape in the 20th century (with no second blade in the 15th century) is an

artifact of the use of PC proxy summarizations, especially when done in conjunction with elimination of the bristlecone/foxtail pine records. This direct examination is the logical counterpart of the indirect examination of this issue by exclusion of PC summarizations in scenarios 2-4. Such coupling of indirect and direct tests allows for the possibility of drawing highly robust conclusions (when both methods agree), as demonstrated powerfully in paleoenvironmental studies by Prentice et al. (1991) concerning the millennial-scale synchrony of vegetation and climate change over the last glacial-interglacial transition in North America.

A further motivation for scenarios 5 and 6 arises from the recognition of a mathematical inconsistency in the published replication of the MBH North American proxy PC calculations in MM05a/b, in relation to the method used by MM in their re-calculation of the PCs according to the MM centering convention (cf. Supplementary Information website, MM05b). In MM05b, the authors report that they were able to exactly replicate the MBH North American PC series using "svd" with standardized anomalies of the proxy data formed according to the MBH centering convention. In the R code for the re-calculation of these PCs at the MM Supplemental Information website, the method used for this purpose is the "princomp" routine on the same proxy data matrix, formed into anomalies (but not standardized) using the MM centering convention. The effect of using "princomp" without specifying that the calculation be performed on the correlation matrix (an alternate argument of "princomp") forces the routine to extract eigenvectors and PCs on the variance-covariance matrix of the unstandardized proxy data, which by its nature will capture information in the first one or two eigenvectors/PCs that is primarily related to the absolute magnitude of the numerically largest-scaled variables in the data matrix (Ammann and Wahl, in review). We have demonstrated that this method of PC extraction has the effect of shifting the actual temporal information common to the North American ITRDB

proxy data into higher-order PCs, especially the third and fourth PCs (Ammann and Wahl, in review; cf. MM05a/b where this shifting of information is also noted). Given such a shift, it would be required to reevaluate how many PCs need to be retained in order to still represent the full structure of the underlying data. For example, when "svd" is used on the standardized data, the temporal structure of the data is captured in the first two PCs, and the MBH/MM centering conventions only have the effect of reversing the order of the PC in which a single-bladed 19th-20th century hockey-stick shape occurs (PC1 with the MBH centering convention and PC2 with the MM centering convention). In this situation, when the first two PCs are summed (either arithmetically or as vectors) for both centering conventions, the resulting time series are nearly identical in structure, with a small difference in scale over the pre-calibration period (Ammann and Wahl, in review). Scenarios 5 and 6 allow systematic examination of the effects of these different PC calculation methods on the reconstruction of Northern Hemisphere temperatures over 1400-1499, in the context of the two salient truncations of the 15th century proxy data set proposed in MM05b. The results from scenario 6 further allow examination of the statistical validity of the primary MM05b climate reconstruction, in the same manner that the MM03 reconstruction is examined in scenario 1.

3. Results

3.1 MBH ORIGINAL RECONSTRUCTION

Our emulation of the MBH98 results is shown in Figure 1. The Northern Hemisphere mean surface temperature reconstruction (WA) is nearly identical to the original MBH98 reconstruction (Fig. 1, red and grey lines). The slight differences are related to our methodological simplifications of using a consistent annual time step throughout the

reconstruction process and equal weighting of the proxies. Validation statistics for the WA reconstruction in relation to the original MBH98 results are reported in Table 1, comparable statistics for the various MM-examination scenarios (1-6) are reported in Table 2. Beyond their calibration quality, and just as importantly, the reconstructions also pass temporally independent cross-verification tests over the interval 1854-1901, employing a spatially more restricted instrumental record available during this time (with only 219 out of the 1082 grid cells available globally during calibration). These results confirm that the MBH reconstruction, per se, is closely reproducible if the procedural sequence of MBH is followed and all original proxy data are applied. Additionally, the reconstruction is robust in relation to two significant methodological simplifications--the calculation of instrumental PC series on the annual instrumental data, and the omission of weights imputed to the proxy series. Thus, this new version of the MBH reconstruction algorithm can be confidently employed in tests of the method to various sensitivities. Without such confirmation (using the original data) that the overall procedure has been reliably repeated, which is not reported, e.g., by Zorita et al. (2003) or von Storch et al. (2004), evaluations of issues with the procedure have to be taken with a significant degree of caution.

3.2 EVALUATION OF MM CRITICISMS

The 15th century reconstructions that result from elimination of significant proxy information in scenario 1 (MM03; cf. Rutherford et al., 2005) are also shown in Figure 1 (pink line). Similar to MM03, this scenario yields much warmer NH temperatures for the 15th century than both MBH98 and WA, which are also at odds with 15th century temperatures in other empirical reconstructions (see Jones and Mann, 2004). According to our assessment, however,

this result does not have climatological meaning because the reconstructions clearly fail validation tests, returning negative RE scores for both calibration and verification (-0.42 and -0.57, respectively, for 1400-1449; –0.65 and –2.71, respectively, for 1450-1499) indicative of no reconstruction skill (Table 2). Reconstruction of the verification period mean is also very poor in this scenario for both examined time frames (Table 2, Fig. 5b). The WA reconstructions including the records excluded in scenario 1 perform far better, with positive calibration and verification RE scores (0.39 and 0.48, respectively, for 1400-1449; 0.47 and 0.44, respectively, for 1450-1499) and good reconstruction of the verification period mean (Tables 1 and 2, Fig. 5a). It should be noted that in this experiment, the MBH step of scaling the proxy-reconstructed (i.e. fitted) instrumental PCs--so that those estimated for the calibration period have the same variance as the actual instrumental PCs--was not used, consistent with the method of MM03 (MM Supplemental Information).

Results for scenarios 2 and 3 are shown in Figure 2. The "all proxy" scenarios (2), based on different sets of proxy data richness (blue and light blue curves in Fig. 2), are highly similar to each other over their common interval, and are also generally similar, especially in overall trajectory, to WA, which used proxy PC summaries (smoothed red curve). [The 1700-onwards reconstruction is not shown in Figure 2, since it is highly similar to the 1404-1971 reconstruction (blue) over their common interval.] It is interesting to note that the 1404-1971 reconstruction is also highly similar to the same kind of "all proxy" scenario developed using the entirely independent RegEM reconstruction technique (Schneider, 2001) reported by Rutherford et al. (2005, cf. their Fig. 2). However, there is a noticeable reduction of amplitude relative to the 1902-1980 mean (defined to be zero anomaly) in the "all proxy" scenarios versus WA, in particular on the order of 0.10° over ~1425-1475, ~1680-1750, and ~1815-1835, and on the

order of 0.15° over ~1575-1600 and significant portions of 1850-1900. Actual deviations from the verification period mean average ~ 0.12° for the four reconstructions in this scenario. The calibration statistics and performance are extremely good (and actually above the WA original) for all sets of proxy richness, and the verification statistics show meaningful skill--although with reduction from the original values, especially for the 1700-onwards reconstruction, paralleling the poorer reconstruction of the verification period mean. The increase in calibration skill and reduction of verification skill shown by the 1700-onward reconstruction in scenario 2, in relation to the data-poorer 1404- and 1450-onwards reconstructions, could be an indication of some calibration overfitting, as could the similar relationships of skill statistics between scenario 2 and the WA reconstructions using proxy summaries (Table 2). Overall, the single-bladed hockey stick result of the MBH original reconstruction is robustly retained, although there is suggestion of a slight decrease in the amount of "extension of the blade" from a 15th-19th century baseline. The close fit of the calibration-period reconstructions with the 20th century instrumental record in all the reconstructions indicates that extending the perspective of the hockey stick blade with the instrumental series into the 1990s and the current decade is appropriate.

Results for the exclusion of the bristlecone/foxtail pine series developed according to scenario 3 are shown by the green curve in Figure 2. The exclusion of these proxy records generally results in slightly higher reconstructed temperatures than those derived from inclusion of all the proxy data series, with the greatest differences (averaging ~ +0.10°) over the period 1425-1510. The highest values before the 20th century in this scenario occur in the early 15th century, peaking at 0.17° in relation to the 1902-1980 mean, which are nevertheless far below the +0.40-0.80° values reported for scenario 1. The verification RE scores for this scenario (Table 2) are only slightly above the zero value that indicates the threshold of skill in the independent

verification period, and the verification mean reconstructions are correspondingly poor. These results, which cannot be attributed to calibration overfitting because the number of proxy regressors is *reduced* rather than augmented, suggest that bristlecone/foxtail pine records *do* possess meaningful climate information at the level of the dominant eigenvector patterns of the global instrumental surface temperature grid. This phenomenon is an interesting result in itself, which is not fully addressed by examination of the local/regional relationship between the proxy ring widths and surface temperatures (noted in section 1.1) and which suggests that the "all proxy" scenarios reported in Figure 2 yield a more meaningful comparison to the original MBH results than when the bristlecone/foxtail pine records are excluded. Even in the absence of this argument, the scenario 3 reconstructions in the 15th century do not exhibit large enough excursions in the positive direction (in relation to the 20th century instrumental record) to yield a double-bladed hockey stick result that diminishes the uniqueness of the late 20th century departure from long-term trends.

The results for scenario 4 (not shown), examining potential overrepresentation of North American tree ring data though strong truncation, are generally very close to the reconstructions represented by the blue and light blue curves in Figure 2. The statistical performance of the reconstruction models is shown in Table 2. The longest-duration differences from the full "all proxy" scenarios are three limited periods (decadal length or less) of small positive offsets ($\leq$ +0.12°) in the second half of the 15th century. Over the rest of the reconstruction period, scenario 4 yields interannually varying positive (relatively more) and negative (relatively fewer) anomalies from the "all proxy" scenarios, with maximum absolute differences on the order of ~ 0.20°. (The exact nature and timing of anomalies from the "all proxy" scenarios varies according to the random selection of Southwestern United States ITRDB proxies excluded from

reconstruction, but within the boundaries mentioned.) These results, based on strong truncations of the North American proxy data used in reconstruction, indicate that the full "all proxy" scenarios show very little, if any, bias due to over-representation of North American proxy records. Thus, the logical appropriateness of using the "all proxy" scenarios as a valid "test bed" for examining the impact of inclusion/non-inclusion of proxy PC summaries is established.

The results for scenario 5 are shown in Figure 3 for the representative cases indicated by 5a-d in Table 2. The scenario 5a-c reconstructions are all close to each other; the range of their variation is shown by the thick blue curve in Fig. 3. The pure effect of removing the Gaspé series as an individual proxy over 1400-1449 and as part of the North American ITRDB network used to calculate proxy PCs over 1450-1499 (scenario 5a) leads to a reconstruction that is higher than WA by ~ 0.05° on average, with a maximum of ~ +0.15° (1402). Over 1450-1499, this reconstruction (the low end of the blue range in this period) is virtually identical to WA. The divergence of scenario 5a from WA in the early 15th century is comparable to the difference shown in MM05b (their Figures 1a and 1b). The highest absolute reconstructed temperature in scenario 5a is 0.21° in 1406 (near the high end of the blue range for this year), contrasted with 0.12° in the same year in WA. This value is below the maximum high excursions of the mid-20th century (0.34°) and ~ 0.6° lower than the highest late-20th century instrumental temperatures (Jones and Moberg, 2003, updated). Scenario 5b represents the additional effect of changing from MBH (5a) to MM (5b) centering (both based on standardized data) in the proxy PC calculations. The reconstructed temperatures in 5b are slightly higher on average (~ 0.05°) than those in scenario 5a over the 15th century, with the same difference in terms of peak high excursions (0.26° in 1402, the high end of the blue range for this year). This slight shift gauges the differential impact on reconstructed climate in scenario 5 related to using the full series

length for proxy centering (MM) as opposed to a centering in relation to the calibration period (MBH convention) before performing PC calculations. The overall effect is far smaller than the impact for this methodological difference suggested in MM05a/b.

In scenarios 5a/b, proxy PC calculations differ only by being based on series that were centered over a different period, and thus do not include the fact that MM05a/b omit the step of standardizing the individual records. As shown in Ammann and Wahl (in review), the method used in MM05a/b causes the PC analysis to capture the variance in the data in a very different way, with the first PCs mostly picking up time series with the largest amplitudes, but not what is common among the series. Only subsequent PCs (after the series have been indirectly standardized to the same variance level) capture variability that is in most of the individual series (Ammann and Wahl, in review). Thus, the number of PCs required to summarize the underlying proxy data changes depending on the approach chosen. Here we verify the impact of the choice of different numbers of PCs that are included in the climate reconstruction procedure. Systematic examination of the Gaspé-restricted reconstructions using 2-5 proxy PCs derived from MM-centered but unstandardized data demonstrates changes in reconstruction as more PCs are added, indicating a significant change in information provided by the PC series. When two or three PCs are used, the resulting reconstructions (represented by scenario 5d, the pink (1400-1449) and green (1450-1499) curve in Fig. 3) are highly similar (supplemental information). As reported below, these reconstructions are functionally equivalent to reconstructions in which the bristlecone/foxtail pine records are directly excluded (cf. pink/blue curve for scenarios 6a/b in Fig. 4). When four or five PCs are used, the resulting reconstructions (represented by scenario 5c, within the thick blue range in Fig. 3) are virtually indistinguishable (supplemental information) and are very similar to scenario 5b. The convergence of results obtained using four

or five PCs, coupled with the closeness of 5c to 5b, indicates that information relevant to the global eigenvector patterns being reconstructed is no longer added by higher-order PCs beyond the level necessary to capture the temporal information structure of the data (four PCs using unstandardized data, or two PCs using standardized data). More generally, the overall similarity of scenarios 5a-c demonstrates that when the full information in the proxy data is represented by the PC series, regardless of the precise ordering of the PCs and which centering convention is used, the impact of PC calculation methods on climate reconstruction in the MBH method is extremely small.

It is important to note that not all of the reconstructions developed in scenarios 5a-d are equivalent in terms of climatological meaning. Applying PC calculations over 1400-1449 as used in scenario 5d yields *negative RE scores in verification* and very poor performance in reconstructing the verification period mean (Table 2; similarly when three proxy PCs are used, supplemental information), indicating that these reconstructions *cannot be ascribed climatological meaning*. Thus, only the thick blue curve in Figure 3 (scenarios 5a-c) represents potentially appropriate adjustments to the original MBH results in the early 15th century, based on question of adequate replication of the individual tree chronologies included in the Gaspé record during this time (MM05b). These scenarios pass both calibration and verification tests, although with somewhat lower RE scores and distinctly poorer reconstruction of the verification period mean than WA (Table 2). In particular, verification performance is differentially poorer when the MM centering convention is used (5b/c) rather than the MBH convention. The failure to verify by scenario 5d, including only two PCs derived from unstandardized data in the MM centering convention, demonstrates that the bristlecone/foxtail pine records are important for the 1400-1449 network--the information they add to PC4 in this way of calculating the PCs is

necessary for a meaningful climate reconstruction. Restricting the PCs in MM05a/b to only the first two (5d) *indirectly* omits the information carried by the bristlecone/foxtail pine records and thereby leads to a non-meaningful reconstruction.

Scenario 6 illustrates this observation by *direct* exclusion of the bristlecone/foxtail pine records from the data set used to calculate North American ITRDB proxy PCs (note, the Gaspé data are also removed as in scenario 5). As shown in Figure 4 and Table 2, such direct exclusion leads to functionally and statistically equivalent results to scenario 5d. The 6a (using the MBH centering convention) and 6b (using the MM centering convention) reconstructions are both nearly identical to the 5d reconstruction and, again, *cannot be ascribed climatological meaning* for the period 1400-1449, based on negative RE scores in verification and corresponding very poor performance in reconstructing the verification period mean. For the period 1450-1499, scenarios 6a/b (and 5d) yield performance measures very much like those of the WA reconstruction (Table 2), thus, from a strictly statistical perspective inclusion of the bristlecone/foxtail pine data in the proxy PC calculations neither enhances nor degrades reconstruction performance during the second half of the 15th century. Over 1450-1499, scenarios 6a/b (and 5d) have maximum high reconstructed temperatures of ~ 0.20°, virtually identical to average mid-20th century values, but well below temperatures reached in the later part of the 20th century. The average difference between the WA reconstruction and scenarios 6a/b over this period is ~ +0.13° (Fig. 4). Given these observations, from a climate reconstruction point of view one can argue that, in general, the bristlecone/foxtail pine records do not introduce spurious noise and their inclusion is justifiable; or said more strongly, their elimination is not objectively justifiable. Their inclusion by standardization of the individual proxy records (independent of the centering convention) or, even if non-standardized series are

applied, by using at least four PCs (until the resulting climate reconstructions converge), leads to reconstruction models that demonstrate skill in both calibration and independent verification.

Scenario 6c (purple/green curve in Fig. 4) parallels scenario 1 (fitted instrumental PCs not rescaled), and is comparable to Figure 1c in MM05b. Like scenario 1, scenario 6c fails both calibration and verification tests over 1400-1449 (RE values for calibration and verification are -0.34 and -0.56, respectively and the verification mean is very poorly reconstructed; Table 2). Although the highest temperatures in this scenario for the early 15th century are similar to those reported in MM05b (max 0.53°), which would, on face value, suggest the possibility of a double-bladed hockey stick result, these values once again *cannot be ascribed climatological meaning* and thus *cannot represent a correction to the original MBH reconstructions*. Over 1450-1499, the scenario 6c reconstructions do pass calibration and verification tests, with a maximum reconstructed temperature of 0.27°. However, the RE scores in both verification and calibration at the grid-cell level for this reconstruction are far lower than those exhibited by scenarios 6a/b for this period (Table 2).

The results for the other combinations of proxy PC calculation method and retained proxy PCs not shown here conform closely to the representative cases reported in scenarios 5 and 6, and are archived at the supplemental information website. A similar, but less elaborated, examination of differing scenarios for possible reconstructions using the MBH method is provided in MM05b. Specific reconstructions and confirmatory validation statistics are not reported in this examination, and thus it does not contain the detailed examinations of reconstruction results and climatological meaningfulness provided here.

4. Discussion and Conclusions

4.1 PRINCIPAL COMPONENT SUMMARIES OF PROXY CLIMATE DATA AND THE USE/NON-USE OF SPECIFIC PROXY RECORDS

Our results show that the MBH climate reconstruction method applied to the original proxy data is not only reproducible, but also proves robust against important simplifications and modifications. The results of this study demonstrate that the primary climatological claim described in MM05a--that the method used by MBH to form PC summaries of climate proxies from data-rich regions results in calibrations that inappropriately weight proxies with a single-bladed hockey stick-like shape in the 20th century--cannot be upheld, and leaves unchanged the overall MBH result of uniquely high Northern Hemisphere temperatures in the late 20th century (relative to the entire 15th-20th century period). Indirect examination of this issue by use of all the continuous individual proxy data over this period, without any form of summarization (PC or otherwise), results in a reconstruction that is similar to the MBH original. Such an approach produces a small reduction of amplitude over the pre-20th-century period (possibly due to calibration overfitting), but the temporal structure exhibits a clear single-bladed hockey stick shape in the 20th century (Fig. 2). A parallel result is derived from direct examination of the impact on reconstruction quality of the *methods* used to calculate proxy PC series based on the North American ITRDB data set in the 15th century (effects on PCs shown in Ammann and Wahl, in review; climatic implications shown here). *Thus, it is the information content of the proxies themselves that drives the shape of the MBH reconstruction, not methodological issues concerning PC summarizations of the proxy series*. This conclusion is robust to several forms of assessment, most powerfully the agreement of the indirect and direct examinations, and also to strong truncations of the individual North American proxy data assimilated into reconstruction (especially from the data-rich Southwestern United States). The latter assessment demonstrates

that the "all proxy" scenario is robust in calibration to the relative density of North American data used in these reconstructions.

The bristlecone/foxtail pine proxies from the Southwestern United States are shown to add necessary verification skill to the climate reconstructions for 1400-1449 when PC summaries are used and significantly greater verification skill to the reconstructions for 1400-1499 when no PC summaries are used--indicating that in these cases the records carry important climate information at the level of eigenvector patterns in global surface temperatures. These results are valid notwithstanding issues concerning these proxies' empirical relationship to local/regional surface temperatures after 1850, noted by MBH in previous work (MBH99; cf. MM05a/b; Hughes and Funkhouser, 2003; Graybill and Idso, 1993). These results enhance the validity of the MBH assumption that proxies used in the reconstruction process do not necessarily need to be closely related to local/regional surface temperatures, as long as they register climatic variations that are linked to the empirical patterns of the global temperature field that the MBH method (and other climate field reconstructions) target. The fact that these additions of skill occur *only* in the verification period (1854-1901) for scenarios 2 and 3 (Table 2), and are far more pronounced in verification for scenarios 5 and 6 (Table 2; supplementary information), leads to the further conclusion that these proxies do not generate spurious increases in calibration fit that thereby downweight the value of other (presumably more locally climate-correlated) proxy series. Over 1450-1499, the bristlecone/foxtail pine proxies neither enhance nor degrade reconstruction performance when PC summaries are used. Thus, in this situation, it is logically appropriate to retain these proxies over the entire 15th century, since they are necessary for verification skill in the first half of this period and have no impact on calibration and verification performance in the later half.

Removal of the Gaspé record from the MBH reconstruction during the early 15th century when proxy PC summaries are used, as mentioned in MM05b, represents a potentially useful adjustment of the original MBH results. With this correction, the adjustment of the MBH time series over 1400-1449 averages ~ + 0.05-0.10°, depending on the centering convention used for the proxy PC calculations, as shown by the blue curve in Figure 3. This adjustment yields a maximum high excursion of 0.21-0.26° over the entire 15th century, which is similar to the mean of mid-20th century temperatures, but ~ 0.6° below late-20th century temperatures. Since the calibration and verification statistical performance for these adjusted reconstructions is somewhat poorer than that achieved by the WA reconstruction, it can be argued empirically that WA represents the best reconstruction achieved in our analysis for the 15th century, and that no change to the MBH reconstruction is conceivably an appropriate judgment. On the other hand, if adjustment is considered appropriate, then the MBH centering convention for proxy PC calculation might be favored over the MM centering convention, based on better statistical performance in verification (associated with the lower values for adjustment given above). Our assessment is that these two possibilities remain as the sole legitimate points of discussion concerning the early 15th century from the MM criticisms; and, in the extreme case, would suggest that the 1400-1449 reconstruction of MBH could be adjusted by + 0.05-0.10°, a minute amount compared to the uncertainty inherent in the reconstruction (MBH99).

4.2 ROBUSTNESS OF MBH98 RESULTS IN RELATION TO MM "CORRECTED" RECONSTRUCTIONS

Our results do not support the large upward “correction” to MBH reconstructed 15th century temperatures described in MM03 (p. 766) and MM05b (p. 71), and leave unaltered the

single-bladed hockey stick conclusion of strongly anomalous late 20th century temperatures. The conclusion of strongly anomalous late 20th century temperatures is retained even if the bristlecone/foxtail pine records were (inappropriately) eliminated for the 15th century, because the maximum high excursion when following the MBH method (rescaling of RPCs) would be ~ 0.35° during the entire 15th century, which is essentially the same as the highest values that occurred during the mid-20th century, but still well below late-20th century temperatures (scenarios 5d, 6a-b, Figs. 3 and 4). Since these scenarios do not pass validation testing for the time frame in which this high excursion occurs (1400-1449), this consideration is used only to show that, even from the standpoint of simple computability, exclusion of these records cannot yield a MM-style double-bladed hockey stick result within the framework of the actual MBH algorithm.

More generally, our results highlight the necessity of reporting skill tests for each reconstruction model. Taking this consideration into account, there is strong reason to conclude that the 15th century reconstructions reported in MM03, which show hemispheric temperatures much higher than those of the mid 20th century, do not have climatological meaning. This double-bladed hockey stick result, while computable using the MBH algorithm, does not pass standard validation tests (RE scores <0 for both calibration and verification). These validation results indicate that the annual climatological mean of the 1902-1980 calibration period would be a *better* predictor over 1854-1980 (the period of available instrumental values that can be used for comparison) than the reconstruction models of this scenario (cf. Cook et al., 1994). The same result holds for the somewhat lower-amplitude double-bladed hockey stick reconstruction reported in MM05b over 1400-1449, which also fails to pass calibration and verification tests for this period. Thus, the primary climatological argument offered by MM for rejecting the

uniqueness of high late-20th century temperatures is found to be without merit, based on examination of the empirical quality of the reconstruction models these authors report.

Overall, the primary outcome from our results is that the work reported in MM03, MM05a, and MM05b does not provide substantive reason to invalidate the general conclusion of anomalous warmth in the later 20th century derived from the MBH reconstruction method and proxy data framework. We find that this result is neither an artifact of selection of the proxy series nor the result of formation or application of PC summaries in the reconstruction procedure. With only a slight upward adjustment (on average ~ +0.05°) to the original MBH reconstruction, which is potentially warranted by removal of the Gaspé record over 1400-1449, the MBH result of anomalous warmth in the later 20th century remains consistent with other paleoclimate reconstructions developed for the last 1-2 millennia (Osborn and Briffa, 2006; Rutherford et al., 2005; Moberg et al., 2005; Oerlemans, 2005; Cook et al., 2004; Huang, 2004; Jones and Mann, 2004; Mann and Jones, 2003; Esper et al., 2002; Briffa et al., 2001; Huang et al., 2000; Crowley and Lowery, 2000; Jones et al., 1998; Bradley and Jones, 1993), especially in light of the recent reconciliation of the Esper et al. (2002; cf. Cook et al., 2004) and MBH reconstructions reported by Rutherford et al. (2005).

4.3 POTENTIAL AMPLITUDE LOSS IN MBH RECONSTRUCTION

Issues concerning possible loss of amplitude in the MBH reconstruction (Moberg et al., 2005; Huang, 2004; von Storch et al., 2004) remain difficult to assess within the proxy framework alone (cf. MBH99, for initial examination of this issue in the MBH reconstructions). Empirically, one indication for such a loss could be derived from positive anomalies for the reconstructed versus instrumental means in the verification period (Table 1). Positive anomalies

indicate that reductions of low frequency amplitude during the verification period (by values shown in Table 1) might reasonably be expected for the original MBH reconstructions between 1400-1449 (~0.04°), 1600-1699 (~0.04°), 1700-1749 (~0.02°), 1750-1759 (~0.03°), and 1760-1779 (~0.01°). These values for amplitude loss are much less than those suggested by Moberg et al. (2005), von Storch et al. (2004), and Huang (2004), and may represent a lower bound on the actual amount of loss, particularly because the verification period is among the coldest half-centuries in MBH. However, the results of Moberg et al. and von Storch et al. are themselves subject to criticism as overstating the possible amplitude loss of the MBH reconstruction (cf. discussion in the "Introduction" here, Mann et al. 2005, and Wahl et al., accepted), and thus the "lower bound" we identify could potentially be somewhat closer to the actual loss. The possibility of significant amplitude loss in general, and specifically in relation to borehole-based reconstructions (cf. Huang et al., 2000; Harris and Chapman, 2001; Huang, 2004), remains at issue in the literature.

The anomalies of reconstructed versus instrumental verification means for scenarios 2-4 indicate similar, and enhanced, tendencies for reconstruction amplitude loss (Table 2). These results carry a final implication in relation to the arguments made by MM. The enhanced loss of verification-period amplitude by the individual proxy scenarios coupled with the conclusion of support for the uniqueness of late-20th century temperatures indicate that, rather than being problematic, PC summaries as developed in the original MBH method actually enhance the quality of the resulting climate reconstructions. This conclusion, focused on the verification period results, demonstrates the usefulness of the PC summaries in a framework that is independent of appropriate/inappropriate variation carried by the PC summaries in the calibration period. It might additionally illustrate that, by reducing the number of predictors

concentrated in a relatively small area (as compared to the rest of the records spread over the globe), the potential for overfitting in the calibration period is reduced and thus the predictive skills of the reconstruction models are enhanced.

5. SUMMARY

Figure 5 gives an overview of our primary results and conclusions. In panel (a) the WA emulation of the MBH reconstruction is compared with the original, without the MM03 emulation (scenario 1) included in Figure 1. The Northern Hemisphere instrumental surface temperature data used in calibration (Jones and Briffa, 1992, updated) are shown, as is the mean from the same data set for the independent verification period (1854-1901). The instrumental series of Jones and Moberg (2003, updated) is also shown to extend the instrumental data through 2005. The nearly identical nature of the reconstructions in terms of decadal-to-multi-decadal structure and variance is evident, as is the very good performance of the reconstructions at capturing both the 20th century trends and the verification period mean.

In panel (b) the WA emulation of MBH is compared with the scenario 1 results, emulating the MM03 reconstruction. The reconstructions based on the 1450-1499 proxy network are continued through 1980, in order to visually demonstrate the poor calibration and verification performance of the scenario 1 results, in parallel with their statistical performance reported in Table 2. The very poor performance of the MM03 emulation is particularly clear in the verification period, averaging very close to the *calibration* mean rather than the verification mean, indicating no skill in detecting the major offset in means between these two periods. The MM03 emulation results also show much greater variance than WA.

In panel (c) the WA emulation of MBH is compared with the scenario 6 results, emulating the MM05b reconstruction. Here, the 1400 and 1450 proxy network results are both continued through 1980, again to graphically show the calibration and verification performance of the scenario 6 results in parallel with their statistical performance in Table 2. The very poor performance of the reconstructions based on the 1400 proxy network when the bristlecone/foxtail pine records in North America are excluded (in line with MM05b) is again evident, actually averaging slightly *above* the calibration mean in the verification period and showing much greater variance than WA. In contrast, the climatologically meaningful performance of the 1450 proxy network in this scenario is evident, albeit with over-represented variance related to the lack of rescaling the reconstructed instrumental PCs in the fitting process. These results graphically highlight the evidence for reasoning that inclusion of the bristlecone/foxtail pine data is empirically warranted during the first half of the 15th century (based almost solely on *verification* performance) and empirically neutral during the second half of the century--with the overall conclusion that including these records in the 15th century is thus logically appropriate and does *not* lead to inappropriate downweighting of information contained by other proxies in the calibration process.

In panel (d) the WA MBH emulation is compared with the scenario 5 result showing the impact of excluding the Gaspé record over the first half of the 15th century along with using the MM centering convention for the formation of proxy PC summaries, both modifications that conceivably could be implemented (despite somewhat poorer verification scores in relation to WA). The 1450 proxy network results are again extended through 1980. Overall, this reconstruction is very much like the WA MBH emulation, with a slight increase averaging ~ 0.10° over 1400-1449. Otherwise this result closely tracks the decadal-to-multi-decadal structure

and variance of the WA MBH emulation. This reconstruction represents the extreme case in our sensitivity tests for a potentially appropriate alteration to the original MBH result, and clearly shows how restricted the impact of such a modification would be. There is hardly any effect on the "hockey stick" shape of the original MBH reconstruction, *at most* adding a small "knob" to the "stick" in the early 15th century. The overall trajectory and conclusions of MBH are completely unaffected by this result.

The results presented here show no evidence for removing the MBH Northern Hemisphere temperature reconstruction from the list of important climate reconstructions of the past six centuries, on the basis of alleged “flaws” in its use of proxy data or underlying methodology. Indeed, our analyses act as an overall indication of the robustness of the MBH reconstruction to a variety of issues raised concerning its methods of assimilating proxy data, and also to two significant simplifications of the MBH method that we have introduced. The shape of a single-bladed “hockey stick”-like evolution of Northern Hemisphere temperature over the last 600 years is strongly confirmed within the MBH reconstruction framework (general algorithm and proxy data). Questions of potential loss of downward amplitude in the MBH method remain, but the evidence developed here from the perspective of the proxy data themselves suggests such losses may be smaller than those shown in other recent work.

Acknowledgments

We are grateful to L. Mearns and the NCAR Weather and Climate Impact Assessment Science Initiative for support, and acknowledge D. Nychka and C. Tebaldi of NCAR for statistical and programming aid. We are also thankful for comments by four referees, which

have greatly helped in improving the thoroughness and clarity of the text. The National Center for Atmospheric Research is sponsored by the National Science Foundation, USA.

Appendix 1 -- Validation in Paleoclimate Reconstruction

The primary purpose of validation in climate reconstruction is to help evaluate the accuracy and usefulness of reconstructions at particular time scales of interest. As generally practiced, validation involves two parallel processes. Prior to validation, a *calibration* is determined between climate proxy data and instrumental data, employing appropriate mathematical fitting procedures tailored to the data being used. The MBH eigenvector-based climate field reconstruction technique is a relatively sophisticated form of such fitting. Other forms include regression of proxy data against climate data of nearby or teleconnected regions, and, at larger spatial scales, simple averaging of individual-site reconstructions to form regional/hemispheric/global composites. [A variant of the latter technique scales composited averages of normalized proxy data per se (e.g. tree ring index values) against the target instrumental data (e.g. Northern Hemisphere mean temperature) over a chosen calibration period.] The first component of validation is to examine how well the calibration fitted values approximate the calibration instrumental data. Second, an independent cross-validation, typically called *verification*, is done by applying the calibration fitting relationships to a set of independent proxy data not used in establishing the calibration relationships. This kind of verification utilizes an independent time period, known as the *verification period*, for which instrumental data are also available. The second component of validation examines how well the verification fitted values approximate the verification instrumental data.

The typical focus of judgment in both calibration and verification is to enhance the likelihood of avoiding *false positive* outcomes (Fritts, 1976; Cook et al., 1994; cf. Wahl, 2004). An *actual* false positive outcome would be acceptance of a reconstruction of poor accuracy (and possibly precision) according to statistical measures employed to gauge these qualities, *with particular focus on the pre-verification period* during which instrumental data are not available. Because this determination is generally impossible to make, due to the lack of pre-verification instrumental data, the statistical tests are done during the calibration and verification periods and their results are employed to *infer* the possible quality of pre-verification reconstructions. Often, these examinations are formalized by the use of null-hypothesis testing, in which a threshold of a selected validation measure is established representing a low likelihood that a value at or above the threshold would occur in the reconstruction process purely by chance. When theoretical distributions are not available for this purpose, Monte Carlo experiments with randomly-created data containing no climatic information have been used to generate approximations of the true threshold values (Fritts, 1976; cf. MM05a; Huybers, 2005; Ammann and Wahl, in review--note that the latter two references correct errors in implementation and results in MM05a). Generally, the passing of statistical tests in *both* calibration and verification is considered necessary to reject the hypothesis of significant possibility of a false positive reconstruction in the pre-verification period (Fritts, 1976; Cook et al., 1994). Because such a result is inferential, it is not, strictly speaking, logically sufficient, but it is the best objective empirical assessment that can be made given instrumental data limitations.

While the concern for avoiding false positive outcomes is a critical component of validation, the simultaneous need to avoid *false negative* outcomes is also important to consider (Wahl, 2004). An *actual* false negative outcome would be rejection of a reconstruction of useful

accuracy (and possibly precision) according to statistical measures employed to gauge these qualities, *again with (inferential) focus on the pre-verification period*. False positive and false negative errors are co-occurring, and generally are inversely related so that reduction of one error leads to increase of the other and vice-versa. It is important to note that the inherent tradeoff between false positive and false negative errors is often not adequately recognized in paleoclimate validation. In recent analyses, a previously underappreciated bias towards reducing false positives has been demonstrated in the fields of microfossil-based climate and ecological reconstructions (Wahl, 2004, Lytle and Wahl, 2005). In that arena, highly conservative criteria for identification of useable proxy reconstruction information, explicitly oriented towards strong reduction of false positive error, have been shown to result in significant, even quite large, under-utilization of valuable climate information in the proxies (Lytle and Wahl, 2005).

We have made *joint* reduction of both false positive and false negative errors our explicit validation criterion in this paper, and use the RE statistic and the verification period mean offset (difference between reconstructed and instrumental means--based on the reduced verification grid where instrumental data are available) as the key validation measures to achieve this purpose, as explained in section 2.3. MM have criticized the use of RE verification statistics in MBH as insufficient, including an alleged improper determination of the threshold value, and have stressed the need to use other, interannual-oriented, verification measures. Below we consider various measures of merit summarized in Cook et al. (1994), to better examine their relevance to the joint reduction of false positive and false negative outcomes. The first four measures generally isolate high-frequency (interannual) information on reconstruction performance. The RE statistic is then evaluated in relation to the other statistics, in order to demonstrate the potential for significant loss of valuable climate information that would arise

from false negative errors if the strict MM validation procedures were to be followed (cf. section 2.3).

MEASURES OF RECONSTRUCTION PERFORMANCE ORIENTED TO THE INTERANNUAL TIME SCALE

General comment As mentioned in section 2.3, validation statistics that isolate interannual information are considered inappropriate for use in the verification period in this paper. This judgment arises because the analysis presented has the primary purpose of gauging multi-decadal reconstruction performance--especially in relation to verification period performance as the sole objective tool for assessing possible reconstruction quality outside the calibration period. In this context, setting a decision criterion that reconstructions must independently pass *both* high and low frequency verification tests introduces significant risk of making a false negative judgment concerning the usefulness of reconstructions that demonstrate good fidelity in the multi-decadal frequency range of primary interest, if they do not also exhibit good fidelity in verification at high (interannual) frequency. [The low frequency fidelity of particular interest in this examination is the ability of the reconstructions to detect changes in mean climate between the calibration and verification periods.]

1) The product moment correlation coefficient (Pearson's r) This measure, although commonly used in comparison of fitted and actual values, presents specific problems in the context of the purposes of this paper. In particular, *r* has the two properties that: a) it is invariant in the face of arbitrary offsets in mean between the two data series being compared; and b) it is invariant in terms of multiplication of either series by an arbitrary scalar. In terms of vector geometry, these properties are important, in that they ensure *r* measures only the cosine of the

angle between the two vectors determined by two equal-element data series (Wonnacott and Wonnacott, 1979)--regardless of their (implied if not actual) point of intersection in the space they span and their relative lengths. In this way *r* focuses uniquely on the closeness with which the two data vectors are *oriented* in space.

For use in gauging the quality of climate reconstructions, however, these angle-retaining properties--which express themselves empirically in terms of highest frequency (in this case interannual) tracking between two time series--can lead to incorrect assessments of reconstruction fidelity over *both* multi-decadal and interannual time periods. In this context, the offset or lack thereof in mean state gives important information about lowest-frequency closeness between a reconstructed time series and corresponding instrumental data. The relative scaling between the two vectors can give useful information in terms of relative amplitude offsets between the two time series, even when they are tracking perfectly in terms of the relative signs of their interannual changes.

Examples of these situations are given in Figure 1S (adjusted from Rutherford et al., 2005). Panel "b" illustrates the potential for false negative rejection of climate reconstructions based on verification period performance at the interannual time scale measured by *r*, even though the reconstructions accurately represent multi-decadal mean information. Panel "a" illustrates the symmetric potential for false positive acceptance of the reconstructions based on verification period performance at the interannual time scale measured by *r*, even though the reconstructions fail entirely to represent the multi-decadal shift in mean between the calibration and verification periods. Panel "c" illustrates the additional potential for *r* to suggest an inappropriately positive perspective concerning verification period performance, due to *r*'s

inability to recognize relative amplitude differences between time series with the same mean and perfect tracking in terms of the relative signs of their interannual changes.

2) The sign test The sign test measures year-to-year changes in the direction of sign of the reconstructions in relation to those of the instrumental values. It is thus also incapable of recognizing that a reconstruction detects changes in instrumental mean values between the calibration and verification periods.

3) The product means test The product means test is likewise insensitive to detection of changes in mean climate state, since it is based on the "cross-products of the actual and estimated yearly departures from their respective mean values" (Cook et al., 1994). Because these means are both relative to the verification period, this statistic becomes a high-frequency-only measure.

4) The coefficient of efficiency (CE) CE subtracts from one the ratio of the sum of squared residuals of reconstruction to the total sum of squared deviations of the instrumental values from their mean. During the calibration period, CE is identical to RE by definition and both are also identical to empirical R^2. During the verification period, CE and RE can differ because CE uses the *verification* period mean to calculate the total sum of instrumental squared deviations, whereas RE continues to use the *calibration* period mean for this calculation. Thus, by construction, CE is incapable of recognizing the ability of a reconstruction to *successfully* detect changes in mean state between the calibration and verification periods.

It should be noted that CE *is* able to recognize when a reconstruction *fails* to detect changes in instrumental mean state between the calibration and verification periods (as in Fig. 1Sa), because the sum of squared verification period residuals will be very large in relation to the sum of squared deviations of the instrumental values from their verification mean--driving CE to relatively large negative values. However, the closer the verification period reconstructions

come to *correctly* detecting a change in mean between the calibration and verification periods--but miss interannual variation in verification (Fig. 1Sb)--the smaller the sum of squared verification residuals becomes, while the sum of squared deviations of the instrumental values from the verification mean remains constant. In the limit of a situation such as Figure 1Sb, these two sums of squares become nearly equal and CE goes towards zero. This asymmetric inability to recognize that a reconstruction is successfully detecting a low frequency change in mean state (and consequent reduction of CE to recognizing only interannual behavior in this case) is what makes CE a poor validation measure in the context of the examinations presented in this paper.

THE REDUCTION OF ERROR STATISTIC (RE)

With one exception, RE is identical to CE as described above. The fact that *verification* RE uses the calibration period mean as the standard of reference for calculating the total sum of instrumental squared deviations allows it to detect changes in the mean of the reconstructed values from the calibration period mean. RE rewards this skill (Fig. 1Sb) and symmetrically penalizes the lack of detection of a shift in mean between the calibration and verification periods (Fig. 1Sa). Thus, RE can register as a valid reconstruction one that has poor high-frequency fidelity in the verification period, but which retains useful low-frequency fidelity in relation to offsets from the calibration period mean, as is the case in the MBH/WA context for a number of the separate calibrations. Cook et al. discuss this "odd behavior" that a high-frequency test (they mention r^2) can show poorer performance than RE in such a situation. However, the concern of the Cook et al. discussion is focused on high-frequency reconstruction fidelity as the target of interest. Conversely, and most importantly in relation to the concerns of this paper, Cook et al. do not consider the detection of differences of mean between the calibration and verification

periods as a target of examination. Thus, using what these authors describe as "odd behavior" as a basis for making validation judgments can readily lead to inappropriate false negative outcomes--if this behavior is used to characterize as without merit a reconstruction that objectively shows skill in reconstructing changes in mean between the calibration and verification periods, although it's interannual verification performance shows little skill.

EXAMINATION AND CONTEXTUALIZATION OF INTERANNUAL RECONSTRUCTION PERFORMANCE IN MBH

Tables 1S and 2S give r^2 and CE values for the WA emulation of MBH and the MM-motivated scenario sets, which parallel the RE and verification mean offset performance reported in Tables 1 and 2. These data suggest (but see below) that a number of the MBH calibration exercises exhibit poor/very little skill at the interannual temporal scale in the verification period, although they all show very good skill at the multi-decadal scale of reconstructing the shift between the verification and calibration period means. All of the MM-motivated scenarios exhibit essentially no skill at the interannual scale in the verification period, yet at the same time scenarios 5a-d and 6a-b exhibit very good skill for the 1450 calibrations at the multi-decadal scale of the verification/calibration mean offset.

These results highlight the need to be very careful about the logical framework for determining the kinds of errors for which minimization is being sought in validation. To use, for example, just the interannual information available from r^2 would, under the criterion of minimizing the risk of committing a false positive error, lead to verification rejection of most of the MBH/WA emulations and all of the MM-motivated scenarios reported. However, this judgment would entail committing large proportions of false negative errors for these

reconstructions at the low frequency scale of the overall verification period, whose multi-decadal perspective is the primary temporal focus of this paper. Our assessment is that such a rejection of validated performance at the low frequency scale would be a waste of *objectively* useful information, similar to that documented for micro-fossil based paleoclimate reconstructions that use highly conservative criteria focused on strong reduction of false positive errors (Lytle and Wahl, 2005). Significant attention to appropriate balancing of these errors is now being sought in paleoclimatology and paleoecology (cf. Lytle and Wahl, 2005; Wahl, 2004), and remains an important venue for further research targeted at recovering maximal information in paleoenvironmental reconstructions.

It also must be noted that indirect verifications of the MBH reconstruction actually suggest quite good interannual performance, potentially raising the question of the quality of the truncated-grid instrumental values used for validation testing in the verification period (219 grid points versus 1082 grid points in the calibration data set). A spatial subset of the MBH annual temperature reconstruction for European land areas (25°W-40°E, 35°N-70°N) compares very well with an independent CFR reconstruction for that region, using a regionally much richer, fully independent set of different instrumental series in combination with documentary proxy evidence (Luterbacher et al. 2004; Xoplaki et al. 2005). Over the 1760-1900 period of this comparison, the r^2 between the regional annual temperature averages of these two reconstructions is 0.67, corresponding with excellent visual temporal tracking of these time series (not shown) at interannual, decadal, and centennial scales. [The interannual amplitude in the European-based annual reconstruction is slightly greater compared to MBH.] Additionally, von Storch et al. (2004) show that the higher-frequency (interannual to decadal) tracking of MBH-style reconstructed temperatures with "actual" temperatures in an AOGCM context is very

good, although an implementation error in the von Storch et al. analysis incorrectly showed very large amplitude losses for the MBH method at lower-frequency (approximately centennial) scales (Wahl et al., accepted). Good interannual/decadal tracking is robust to correction of the implementation error (Wahl et al., accepted). These indirect tests of the MBH reconstruction add a significant caveat to the indications of poor interannual performance based on the verification instrumental data used by MBH (and by us here). It will be an important area of further testing of the MBH/WA reconstructions to attempt to resolve this inconsistency of verification validation results for the interannual frequency band, which is an issue that is potentially relevant to all high-resolution proxy-based climate reconstructions because of the limited spatial representation of pre-20th century instrumental data.

Appendix 2 -- Benchmarking Significance for the RE Statistic

We downloaded and used the MM05a code for benchmarking the RE statistic, and have replicated their results to the degree of agreement one can expect using separate realizations of a Monte Carlo process. [Note: we used 1000 iterations vs. the 10,000 in MM05a, but this has no bearing on the results.] In implementing this procedure, we found a technical problem that we reported in Ammann and Wahl (in review, and supplemental material there referenced). The method presented in MM05a generates apparently realistic pseudo tree ring series with autocorrelation (AC) structures like those of the original MBH proxy data (focusing on the 1400-onward set of proxy tree ring data), using red noise series generated by employing the original proxies' complete AC structure. However, one byproduct of the approach is that these time series have nearly uniform variances, unlike those of the original proxies, and the PCs derived from them generally have AC structures *unlike* those of the original proxies' PCs. Generally, the

simulated PCs (we examined PCs 1-5) have significant spurious power on the order of 100 years and approximate harmonics of this period. When the original relative variances are restored to the pseudoproxies before PC extraction, the AC structures of the resultant PCs are much like those of the original proxy PCs. Following MM05a, the first PCs of this process were then used as regressors in a calibration with the Northern Hemisphere mean from the MBH verification data grid and the RE of verification determined, for each Monte Carlo iteration. Approximate RE significance levels can then be determined, assuming this process represents an appropriate null hypothesis model. Using the AC-correct PC1s in the RE benchmarking algorithm had little effect on the original MM benchmark results, but does significantly improve the realism of the method's representation of the real-world proxy-PC AC structure.

In implementing this analysis, we considered the criticism by Huybers (2005) that MM05a do not equate the variances of the calibration-period fitted and observed values for Northern Hemisphere mean temperature. The factor Huybers applies to the fitted values to ensure this equality of calibration variances is also then applied to the verification period fitted values. Huybers applies this correction from a theoretical perspective. Possibly more importantly, *variance re-scaling is appropriately applied in this case because it more accurately mimics the actual MBH procedure,* which applies a parallel rescaling to the fitted instrumental PCs that drive the MBH climate field reconstruction process. Applying this variance rescaling in the MM code with AC-incorrect PC1s, Huybers determined a 99% significance RE benchmark of 0.0 in verification (significant digits as reported by Huybers). We also replicated this result at the level of precision available from different Monte Carlo realizations of the same process. When we applied the Huybers' variance rescaled RE calculation to our AC-correct pseudoproxy PC1s, we generated a 98.5% significance RE benchmark of 0.0. We find that the combination of

AC-correct pseudoproxy PC series with the variance-rescaled RE calculation provides the most appropriate mimicking of MBH possible in this simple case of examining the potential verification skill available merely from the non-climatic persistence contained in PC1 of the ITRDB N. American data.

Appendix 3 -- Areal Weighting of Instrumental Temperature Data

We followed the MBH convention of weighting the instrumental temperature data input into the reconstruction algorithm by the cosine of latitude of each grid cell. We recognize that square root of cosine (latitude) weighting can be justified as appropriate when eigenvector values are the actual mathematical units of analysis, since these are often determined from covariance or correlation matrices of original values of interest. However, in the case of reconstructing Northern Hemisphere temperature, the units of interest are the grid cell level temperatures themselves, not eigenvector values, and thus areal weighting by cosine (latitude) is appropriate. Additionally, tests using the 1450 and 1820 proxy data sets showed little difference in terms of both statistics and actual reconstructions across both kinds of areal weighting, and thus we find that there is no substantive issue in this regard in practical terms.

Appendix 4 -- WA and MBH Verification RE Scores

The reductions of verification RE scores for WA versus MBH in Table 1 are possibly artifacts of using a spatially restricted instrumental grid to calculate the verification statistics. As mentioned in the "Results" section, the number of grid cells available over the 1854-1901 verification period is 219, versus the 1082 cells used in calibration. Of the 219 global total, 172 cells are in the Northern Hemisphere, versus 707 Northern Hemisphere cells that are actually

reconstructed in calibration (and thereby throughout the entire length of the reconstruction period). For the 1820-1980 proxy network for which comparable "sparse grid" verification reconstructions are available, the WA average reconstructed Northern Hemisphere mean temperature (entire verification period) is highly similar to that reported by MBH (-0.224° and -0.229°, respectively), while the corresponding standard deviations show a reduced variation in WA compared to MBH (0.111° and 0.146°, respectively). The corresponding MBH instrumental standard deviation is 0.165°. [MBH sparse grid data are from "Global Temperature Patterns in Past Centuries: an Interactive Presentation"; NOAA Paleoclimatology Program.] Thus, there is appropriate variation in the MBH sparse grid reconstructions that is underrepresented in the parallel WA reconstructions.

However, this loss of variation by WA does *not* occur in the reported verification period annual mean reconstructions, calculated from the full 707 Northern Hemisphere grid-cell set (standard deviations of 0.080° for WA and 0.084° for MBH). Thus, there is apparently a slight differential impact of the two reconstruction procedures in terms of the variability present in the sparse verification grid, which does not "pass through" at the scale of the full reported reconstruction grid. This observation suggests that the differences between the WA and MBH verification RE values shown could be overestimates of the true (but unmeasurable) differences occurring at the level of the full Northern Hemisphere reconstruction grid, driven by the non-random sampling of the sparse sub-grid used to calculate verification performance statistics.

References

Ammann, C.M. and Wahl, E.R.: in review, 'Comment on "Hockey sticks, principal components, and spurious significance" by S. McIntyre and R. McKitrick', *Geophys. Res. Lett.*

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

Briffa, K. R., Osborn, T. J., Schweingruber, F. H., Harris, I. C., Jones, P. D., Shiyatov, S. G., and Vaganov, E. A.: 2001, 'Low-frequency temperature variations from a northern tree ring density network', *J. Geophys. Res.* **106**, 2929-2941.

Cane, M. A., Clement, A. C., Kaplan, A., Kushnir, Y., Pozdnyakov, D., Seager, R., Zebiak, S. E., and Murtugudde, R.: 1997, 'Twentieth-century sea surface temperature trends', *Science* **275**, 957-960.

Chapman, D. S., Bartlett, M. G., and Harris, R. N.: 2004, 'Comment on "Ground vs. surface air temperature trends: Implications for borehole surface temperature reconstructions", by M. Mann and G. Schmidt', *Geophys. Res. Lett* **31**, doi:10.1029/2003GL019054.

Cole, J. E. and Cook, E. R.: 1998, 'The changing relationship between ENSO variability and moisture balance in the continental United States', *Geophys. Res. Lett.* **25(24)**, 4529-4532.

Cook, E. R., Briffa, K. R., and Jones, P. D.: 1994, 'Spatial regression methods in dendroclimatology: a review and comparison of two techniques', *Int. J. Clim.* **14**, 379-402.

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

Crowley, T. J. and Lowery, T.: 2000. 'How Warm Was the Medieval Warm Period? A comment on "Man-made versus natural climate change"', *Ambio* **29**, 51-54.

Crowley, T. J., Baum, S. K., Kim, K.-Y., Hegerl, G. C. and Hyde, W. T.: 2003, 'Modeling ocean heat content changes during the last millennium', *Geophys. Res. Lett.* **30(18)**, 1932, doi:10.1029/2003GL017801.

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

Evans, M. N., Kaplan, A., and Cane, M. A.: 2002, 'Pacific sea surface temperature field reconstruction from coral $d^{18}O$ data using reduced space objective analysis', *Paleoceanography* **17**, 7,1-7,13.

Folland, C. K., Karl, T. R., Christy, J. R., Clarke, R. A., Gruza, G. V., Jouzel, J., Mann, M. E., Oerlemans, J., Salinger, M. J., and Wang, S.-W.: 2001, 'Observed climate variability and change', in Houghton, J. T., et al. (eds.), *Climate Change 2001: The Scientific Basis*, Cambridge Univ. Press, Cambridge, 99-181.

Fritts, H.C.: 1976, *Tree Rings and Climate*, Academic Press London,.

Goosse, H., Crowley, T. J., Zorita, E., Ammann, C. M., Renssen, H., and Driesschaert, E.: 2005, 'Modeling the climate of the last millennium: What causes the differences between simulations?', *Geophys. Res. Lett.* 32, L06710, doi:10.1029/2005GL022368.

Graybill, D. and Idso, S.: 1993, 'Detecting the aerial fertilization effect of atmospheric CO_2 enrichment in tree-ring cohronologies', *Global Biogeochemical Cycles* **7**, 81-95.

Green, D.M. and Swets, J.A.: 1988. *Signal Detection Theory and Psychophysics*, Peninsula Publishing, Los Altos, CA.

Harris, R. N. and Chapman, D. S.: 2001, 'Mid-latitude (30-60N) climatic warming inferred by combining borehole temperatures with surface air temperatures', *Geophys. Res. Lett.* **28**, 747-750.

Huang, S., Pollack, H.N., and Shen, P.-Y.: 2000, 'Temperature trends over the past five centuries reconstructed from borehole temperatures'. *Nature* **403**, 756-758.

Huang, S.: 2004, 'Merging information from different resources for new insight into climate change in the past and future', *Geophys. Res. Lett*. **31**, L13205.

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

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

Jones P. D. and Briffa, K. R.: 1992, 'Global surface temperature variations during the 20th century: Part 1: Spatial, temporal and seasonal details', *Holocene* **1**, 165-179.

Jones, P. D., Briffa, K. R., Barnett, T. P., and Tett, S. F. B.: 1998, 'High-resolution paleoclimatic records for the last millennium: interpretation, integration and comparison with circulation model control-run temperatures', *Holocene* **8**, 455-471.

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

Jones, P.D. and Moberg, A.: 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., Osborn, T. J. ,and Briffa, K. B.: 2001, 'The evolution of climate over the last millennium', *Science* **292**, 662-667.

Kaplan, A., Kushnir, Y., Cane, M. A., and Blumenthal, M. B.: 1997, 'Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperatures', *J. Geophys. Res.* **102(C13)**, 27,835-27,860.

Lytle, D. and Wahl, E.: 2005, 'Paleoenvironmental reconstructions using the modern analog technique: the effects of sample size and decision rules', *Holocene* **15**, 554-566.

Luterbacher, J., Dietrich, D., Xoplaki, E., Grosjean, M., and Wanner, H.: 2004, 'European seasonal and annual temperature variability, trends and extremes since1500', *Science* **303**, 1499-1503.

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

Mann, M. E., Bradley, R. S., and Hughes, M. K.: 1999, 'Northern Hemisphere temperatures during the past millennium: inferences, uncertainties, and limitations', *Geophys. Res. Lett.* **26**, 759-762.

Mann, M. E., Bradley, R. S., and Hughes, M. K.: 2000, 'Long-term variability in the El Nino Southern Oscillation and associated teleconnections', in Diaz, H. F. and 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., Bradley, R.S., Hughes, M.K.: 2004, 'Corrigendum: Global-scale temperature patterns and climate forcing over the past six centuries', *Nature* **430**, 105.

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

Mann, M. E. and Rutherford, S.: 2002, 'Climate reconstruction using "pseudoproxies"', *Geophys. Res. Lett.* **29**, 139,1-139,4.

Mann, M. E., Rutherford, S., Bradley, R. S., Hughes, M. K., Keimig, F. T.: 2004 'Optimal surface temperature reconstructions using terrestrial borehole data', *J. Geophys. Res.* **108(D7)**, 4203, doi: 10.1029/2002JD002532.

Mann, M. E., Rutherford, S., Wahl, E., and Ammann, C.: 2005, 'Testing the fidelity of methods used in "proxy"-based reconstructions of past climate', *J. Climate* **18**, 4097-4107.

Mann, M. E. and Schmidt, G. A.: 2003, 'Ground vs. surface air temperature trends: implications for borehole surface temperature reconstructions', *Geophys. Res. Lett.* **30(12)**, 1607, doi: 10.1029/2003GL017170.

McIntyre, S. and McKitrick, R.: 2003, 'Corrections to the Mann et al (1998) proxy data base and Northern Hemispheric average temperature series', *Energy and Environment* **14**, 751-771.

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

McIntyre, S. and McKitrick, R.: 2005b, 'The M&M critique of the MBH98 Northern Hemisphere climate index: update and implications', *Energy and Environment* **16**, 69-100.

Moberg, A., Sonechkin, D., Holmgren, K., Datsenko, N., and Karlén, W.: 2005, 'Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data', *Nature* **433**, 613-617.

Oerlemans, J.: 2005, 'Extracting a climate signal from 169 glacier records', *Science* **308**, 675-677.

Osborn, T.J. and Briffa, K.R.: 2006, 'The spatial extent of 20th-Century warmth in the context of the past 1200 years', *Science* **311**, 841-844.

Prentice, I. C., Bartlein, P. J., and Webb III, T. I.: 1991, 'Vegetation and climate change in eastern North America since the Last Glacial Maximum', *Ecology* **72**, 2038-2056.

Raible, C., Casty, C., Luterbacher, J., Pauling, A., Esper, J., Frank, D., Büntgen, U., Roesch, A., Tschuck, P., Wild, M., Vidale, P-L., Schar, C., and Wanner, H.: in press, "Climate variability - observations, reconstructions, and model simulations for the Atlantic-European and Alpine region from 1500-2100 AD", *Clim. Change*.

Rajagopalan, B., Cook, E.R., Lall, U., and Ray, B.K.: 2000, 'Spatiotemporal variability of ENSO and SST teleconnections to summer drought over the United States during the twentieth century', *J. Climate* **13**, 4244-4255.

Rutherford, S., Mann, M. E., Osborn, T. J., Bradley, R. S., Briffa, K. R., Hughes, M. K., and Jones, P. D.: 2005. 'Proxy-based Northern Hemisphere surface temperature reconstructions: sensitivity to method, predictor network, target season, and target domain', *J. Climate* **18**, 2308-2329.

Rutherford S. and Mann, M. E.: 2004, 'Correction to "Optimal surface temperature reconstructions using terrestrial borehole data"', *J. Geophys. Res.* **109(D11107)** doi:10.1029/2003JD004290.

Schmidt, G. A. and Mann, M. E.: 2004, 'Reply to comment on "Ground vs. surface air temperature trends: implications for borehole surface temperature reconstructions" by D. Chapman et al.', *Geophys. Res. Lett.* **31**, L07206, doi: 10.1029/2003GL0119144.

Shindell, D. T., Schmidt, G. A., Mann, M. E., Rind, D., and Waple, A.: 2001, 'Solar forcing of regional climate change during the Maunder Minimum', Science 294, 2149-2152.

Shindell, D. T., Schmidt, G. A., Miller, R. L. and Mann, M. E.: 2003, 'Volcanic and solar forcing of climate change during the preindustrial era', *J. Climate* **16**, 4094-4107.

Schneider, T.: 2001, 'Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values', *J. Climate **14***, 853-887.

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

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

Wahl, E.: 2004, 'A general framework for determining cut-off values to select pollen analogues with dissimilarity metrics in the modern analogue technique', *Rev. Palaeobot. Palyno.* **128**, 263-80.

Wahl, E., Ritson, D., and Ammann, C.: accepted, 'Reconstruction of century-scale temperature variations', *Science*

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

Wonnacott, R. and Wonnacott, T.: 1979, *Econometrics* (2nd ed.), John Wiley and Sons, New York.

Zorita, E., Gonzalez-Rouco, F., and Legutke, S.,: 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, 2003.

Figure Captions

Figure 1 Comparison of Wahl/Ammann and Mann/Bradley/Hughes (1998) Northern Hemisphere annual mean temperature reconstructions with scenario 1 reconstruction [emulation of McIntyre/McKitrick (2003) results]. Wahl/Ammann reconstruction is shown in red. Mann/Bradley/Hughes reconstruction is shown in grey. Scenario 1 reconstruction for 15th century is shown in pink. Zero reference level is mean value for 1902-1980 instrumental data. Instrumental data used in calibration and verification are shown in black: annual data for full Northern Hemisphere grid over 1902-1993, and the mean of the spatially-restricted Northern Hemisphere grid over 1854-1901 (Jones and Briffa, 1992, updated). Instrumental data for 1902-2005 from Jones and Moberg (2003, updated) are also plotted, in dark blue. Pink coding of Scenario 1 shows validation failure according to criteria described in section 2.3.

Figure 2 Northern Hemisphere annual mean temperature reconstructions using only individual proxy records--without principal component summaries of proxy data (scenarios 2 and 3, described in text; cf. Table 2). Blue--all individual records for 1404-1449 and 1450-1499 Mann/Bradley/Hughes (1998) proxy networks (1450-1499 network used to reconstructed entire period from 1450-1971) (scenario 2). Light blue--all individual records for 1800-1820 Mann/Bradley/Hughes (1998) proxy network used to reconstruct 1800-1971 (scenario 2). Green--same as for blue curve, except that 15 bristlecone/foxtail pine records questioned in McIntyre/McKitrick (2005a/b) are removed from proxy roster (scenario 3). WA (red line), zero reference level, and instrumental data same as in Figure 1.

Figure 3 Northern Hemisphere annual mean temperature reconstructions for 15th century with proxy principal component (PC) summaries retained, excluding the St. Anne's River (Gaspé)

series as an individual proxy over 1400-1449 and as part of data set used to calculate PCs of North American proxies from the International Tree Ring Data Base (ITRDB) (scenario 5, described in text; cf. Table 2). Thick blue line includes range associated with following scenarios: reconstructions using standardized anomalies for ITRDB proxies (for input into PC extraction) referenced to 1902-1980 mean values (scenario 5a); reconstructions using standardized anomalies for ITRDB proxies referenced to mean values over 1400-1980 and 1450-1980, for 1400-1449 and 1450-1499 reconstructions, respectively (scenario 5b); and reconstructions using *non*-standardized anomalies for ITRDB proxies referenced to mean values over 1400-1980 and 1450-1980, for 1400-1449 and 1450-1499 reconstructions, respectively (with retention of sufficient PC series to capture temporal information structure of ITRDB data) (scenario 5c). Pink (1400-1449) and green (1450-1499) reconstruction is same as scenario 5c, *except* with too few PC series retained to capture information dynamic structure of ITRDB data (acting in effect as *exclusion* of bristlecone/foxtail pine records from PC calculations) (scenario 5d). WA (red line) and zero reference level same as in Figure 1. Dark blue line is Jones and Moberg (2003, updated) instrumental data series through 2005, showing latest 20th century and earliest 21st century hemispheric temperatures for comparative purposes. Pink-coded portion of scenario 5d shows validation failure according to criteria described in section 2.3.

Figure 4 Gaspé proxy restrictions as in Figure 3, with additional exclusion of 15 bristlecone/foxtail pine records from data set used to calculate principal components (PCs) of North American proxies from the International Tree Ring Data Base (ITRDB) (scenario 6, described in text; cf. Table 2). Thick pink (1400-1449) and blue (1450-1499) line includes range associated with following scenarios: reconstructions using standardized anomalies for ITRDB

proxies (for input into PC extraction) referenced to 1902-1980 mean values (scenario 6a) and reconstructions using standardized anomalies for ITRDB proxies referenced to mean values over 1400-1980 and 1450-1980, for 1400-1449 and 1450-1499 reconstructions, respectively (scenario 6b). Purple (1400-1449) and green (1450-1499) reconstruction uses *non*-standardized anomalies for ITRDB proxies referenced to mean values over 1400-1980 and 1450-1980, for 1400-1449 and 1450-1499 reconstructions, respectively (with fitted instrumental PCs *not* rescaled by factor which equates variances of fitted and instrumental PCs over calibration period) (scenario 6c). WA (red line), zero reference level, and instrumental data same as in Figure 3. Pink- and purple-coded portions of Scenarios 6a-c show validation failures according to criteria described in section 2.3.

Figure 5 Summary of results. Panel (a) compares the Wahl-Ammann (WA) emulation of the MBH reconstruction (red) with the original (grey). Panel (b) compares the WA reconstruction (red) with an emulation of the MM03 *Energy and Environment* reconstruction (pink). The MM03 emulation for 1400-1449 uses the MBH 1400 proxy network as adjusted by MM03; the MM03 emulation for 1450-1980 uses the MBH 1450 proxy network as adjusted by MM03. Panel (c) compares the WA reconstruction (red) with emulations of the MM05b *Energy and Environment* reconstruction. The emulations directly exclude the bristlecone/foxtail pine records from calculation of PC summaries of N. American tree ring data (which are indirectly excluded by MM05a/b, cf. "Results" in text). The MM05b emulation using the 1400 proxy network is continued through 1980 (pink), as is the MM05b emulation using the 1450 proxy network (green). Panel (d) compares the WA reconstruction (red) with a reconstruction based on exclusion of the Gaspé record over 1400-1449 and use of the MM centering convention for

forming PC summaries of North American tree ring data (dark magenta). Pink-coded reconstructions show validation failure according to criteria described in section 2.3.

Zero reference level in each panel is mean value for 1902-1980 instrumental data. Instrumental data in all panels are indicated as follows. Instrumental data used in calibration and verification are shown in black: annual data for full Northern Hemisphere grid over 1902-1993, and the mean of the spatially-restricted Northern Hemisphere grid over 1854-1901 (Jones and Briffa, 1992, updated). Instrumental data for 1902-2005 from Jones and Moberg (2003, updated) are also plotted, in dark blue.

Table 1 RE Scores and Verification Mean Anomalies (deg. C) for MBH Reconstruction Emulations

Proxy Network MBH – periods	NH Mean RE Calibration-period MBH / WA	All Grid Cells RE Calibration-period MBH / WA	NH Mean RE Verification-period MBH / WA	Reconstructed *minus* Instrumental Means Verification-period WA
1400-1449	0.42 / 0.39	0.08 / 0.06	0.51 / 0.48	0.039 *
1450-1499	0.41 / 0.47	0.09 / 0.09	0.51 / 0.44	-0.040
1500-1599	0.42 / 0.48	0.10 / 0.10	0.49 / 0.47	-0.035
1600-1699	0.67 / 0.64	0.14 / 0.14	0.53 / 0.46	0.038
1700-1729	0.71 / 0.69	0.14 / 0.18	0.57 / 0.50	0.019
1730-1749	0.71 / 0.69	0.15 / 0.18	0.61 / 0.55	0.015
1750-1759	0.74 / 0.71	0.18 / 0.24	0.57 / 0.61	0.028
1760-1779	0.74 / 0.73	0.26 / 0.32	0.70 / 0.54	0.009
1780-1799	0.74 / 0.74	0.27 / 0.35	0.69 / 0.59	-0.017
1800-1819	0.75 / 0.75	0.27 / 0.35	0.68 / 0.60	-0.026
1820-1980	0.76 / 0.75	0.30 / 0.37	0.69 / 0.62	-0.031

* Verification-period instrumental mean is -0.193°, relative to mean for period 1902-1980

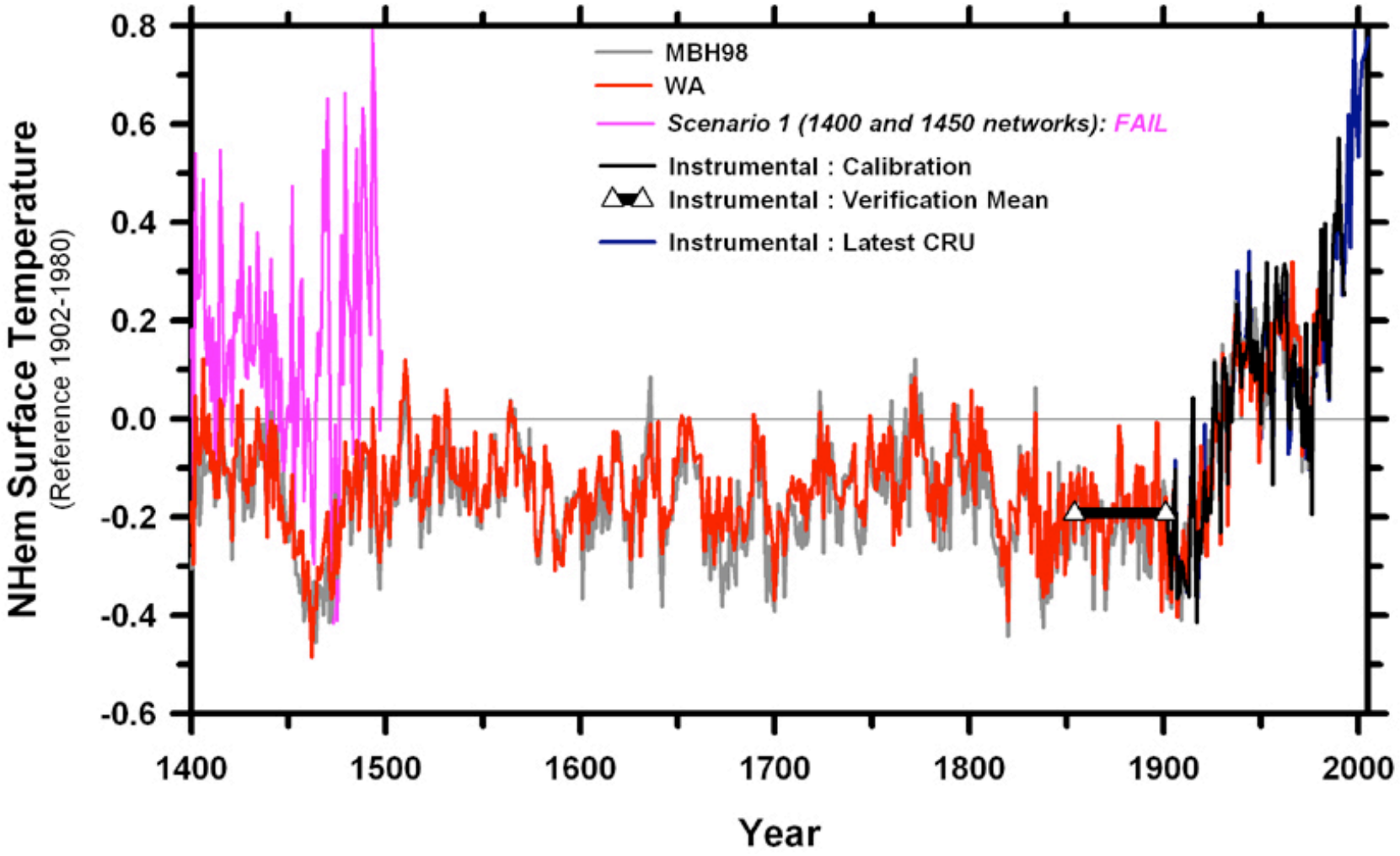
Note: The reductions of verification RE scores for WA versus MBH in Table 1 are possibly artifacts of using a spatially restricted instrumental grid to calculate the verification statistics. Cf. Appendix 4.

Table 2 RE Scores and Verification Mean Anomalies (deg. C) for MM Reconstruction Scenarios

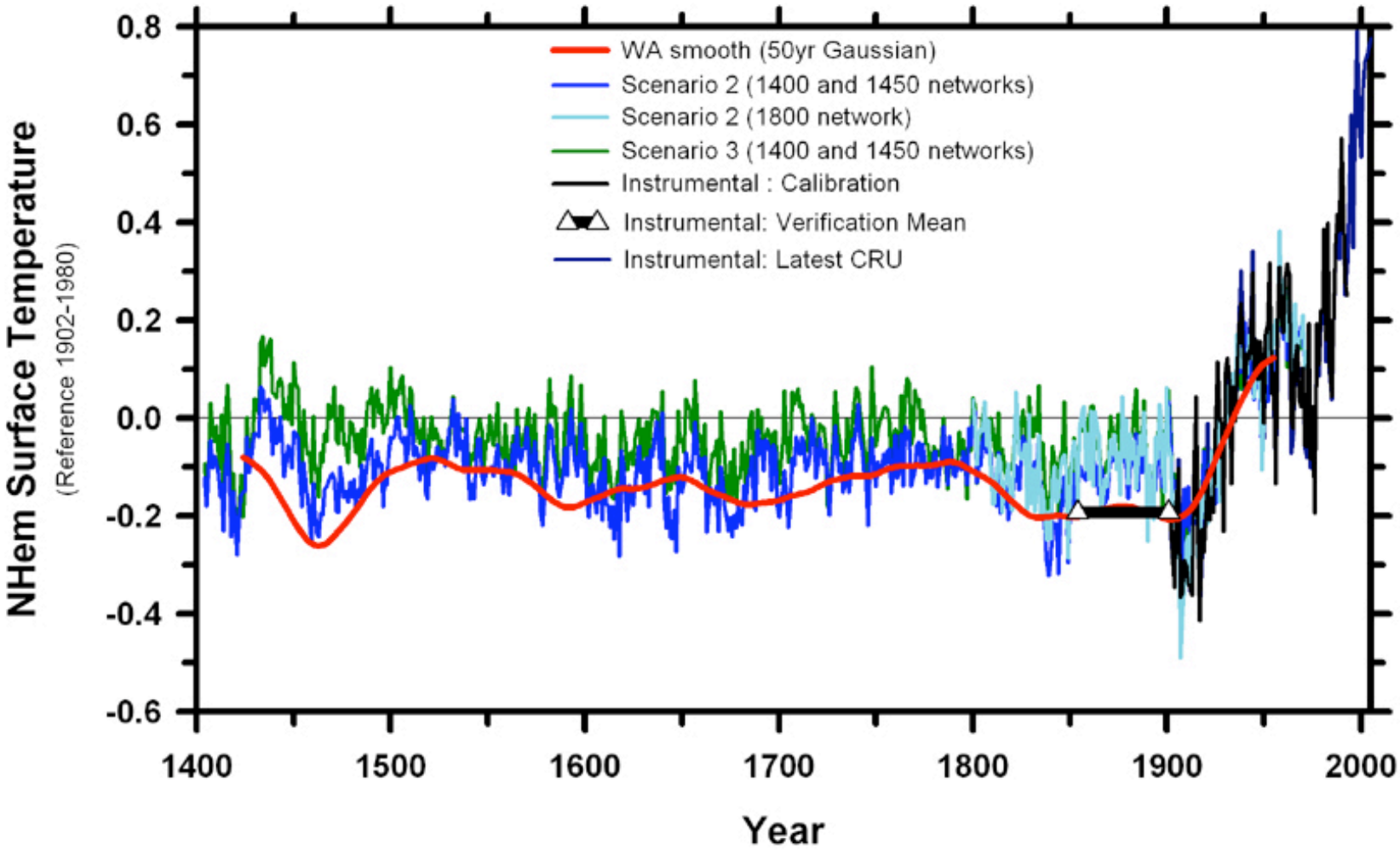
Scenarios with MBH Proxy networks	NH Mean RE Calibration period	All Grid Cells RE Calibration period	NH Mean RE Verification period	NH Offset in Verification period *Instrumental mean (-0.193° C)*
Reference : Wahl-Ammann Emulation of MBH				
1400-1449	0.39	0.06	0.48	0.039
1450-1499	0.47	0.09	0.44	-0.040
1700-1729	0.69	0.18	0.50	0.019
1800-1819	0.75	0.35	0.61	-0.026
Scenario 1 Omission ITRDB & other N. Am. Data *(reconstructed instrumental PCs not re-scaled)*				
1400-1449	*-0.42*	*-0.09*	*-0.57*	0.194
1450-1499	*-0.65*	*-0.20*	*-2.71*	0.170
Scenario 2 no PC for ITRDB Data				
1404-1449	0.60*	0.06*	0.36*	0.113*
1450-1499	0.67	0.12	0.28	0.107
1700-1729	0.73	0.17	0.19	0.140
1800-1819	0.76	0.36	0.35	0.116
Scenario 3 no PC for ITRDB and no Bristlecone/Foxtail				
1404-1449	0.62*	0.06*	0.06*	0.169*
1450-1499	0.71	0.12	0.10	0.138
Scenario 4 no PC for ITRDB and Limitation of Records				
(a) 1404-1449	0.57*	0.06*	0.03*	0.167*
1450-1499	0.70	0.13	0.15	0.115
(b) 1404-1449	0.56*	0.06*	0.20*	0.139*
1450-1499	0.68	0.13	0.30	0.081
Senario 5 No Gaspé**				
(a) MBH centered and standardized 2 PCs (svd):				
1400-1449	0.28	0.04	0.34	0.093
1450-1499	0.47	0.09	0.44	-0.030
(b) MM centered and standardized 2 PCs (svd):				
1400-1449	0.32	0.04	0.18	0.135
1450-1499	0.49	0.09	0.46	-0.007
(c) MM centered, not-standardized 4 PCs (princomp):				
1400-1449	0.29	0.04	014	0.144
1450-1499	0.48	0.09	0.43	0.001
(d) MM centered, not-standardized 2 PCs (princomp): *(indirect Bristlecone/ Foxtail exclusion)*				
1400-1449	0.19	0.02	*-0.18*	0.190
1450-1499	0.46	0.09	0.42	0.008
Scenario 6 No Gaspé and no Bristlecone/Foxtail**				
(a) MBH-centered, standardized, 2 PCs (svd):				
1400-1449	0.24	0.03	*-0.13*	0.182
1450-1499	0.48	0.09	0.43	0.008
(b) MM-centered, standardized, 2 PCs (svd):				
1400-1449	0.19	0.02	*-0.20*	0.193
1450-1499	0.47	0.09	0.42	0.008
(c) MM-centered, *not*-standardized, 2 PCs (princomp): *(reconstructed instrumental PCs not re-scaled)*				
1400-1449	*-0.34*	*-0.08*	*-0.56*	0.196
1450-1499	0.32	0.005	0.14	-0.062

* note slight difference in data coverage with scenario limited to 1404-1449

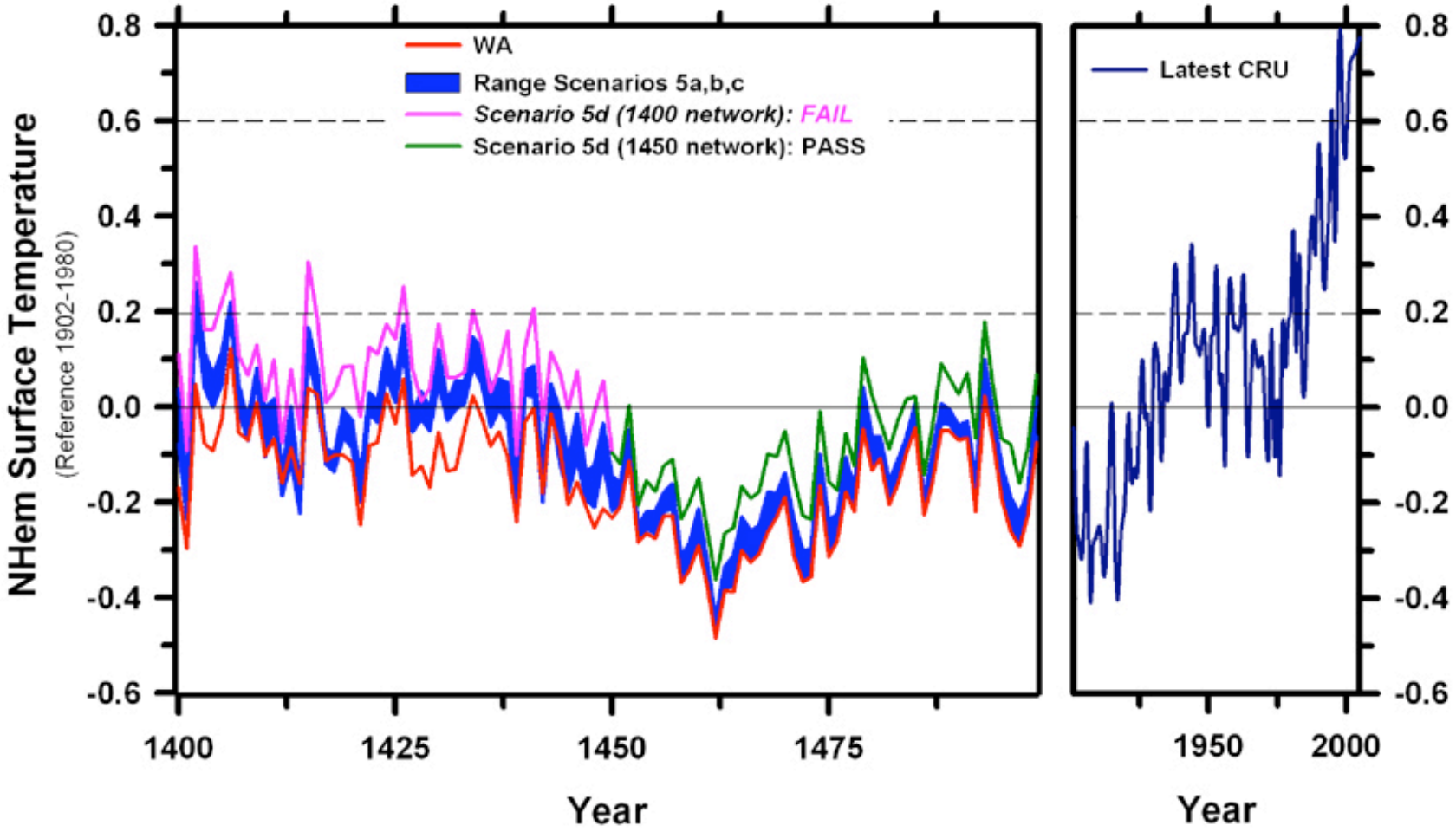
** statistics given for scenario iterations described in "Results"; statistics for *all* iterations of each scenario available at supplemental information website



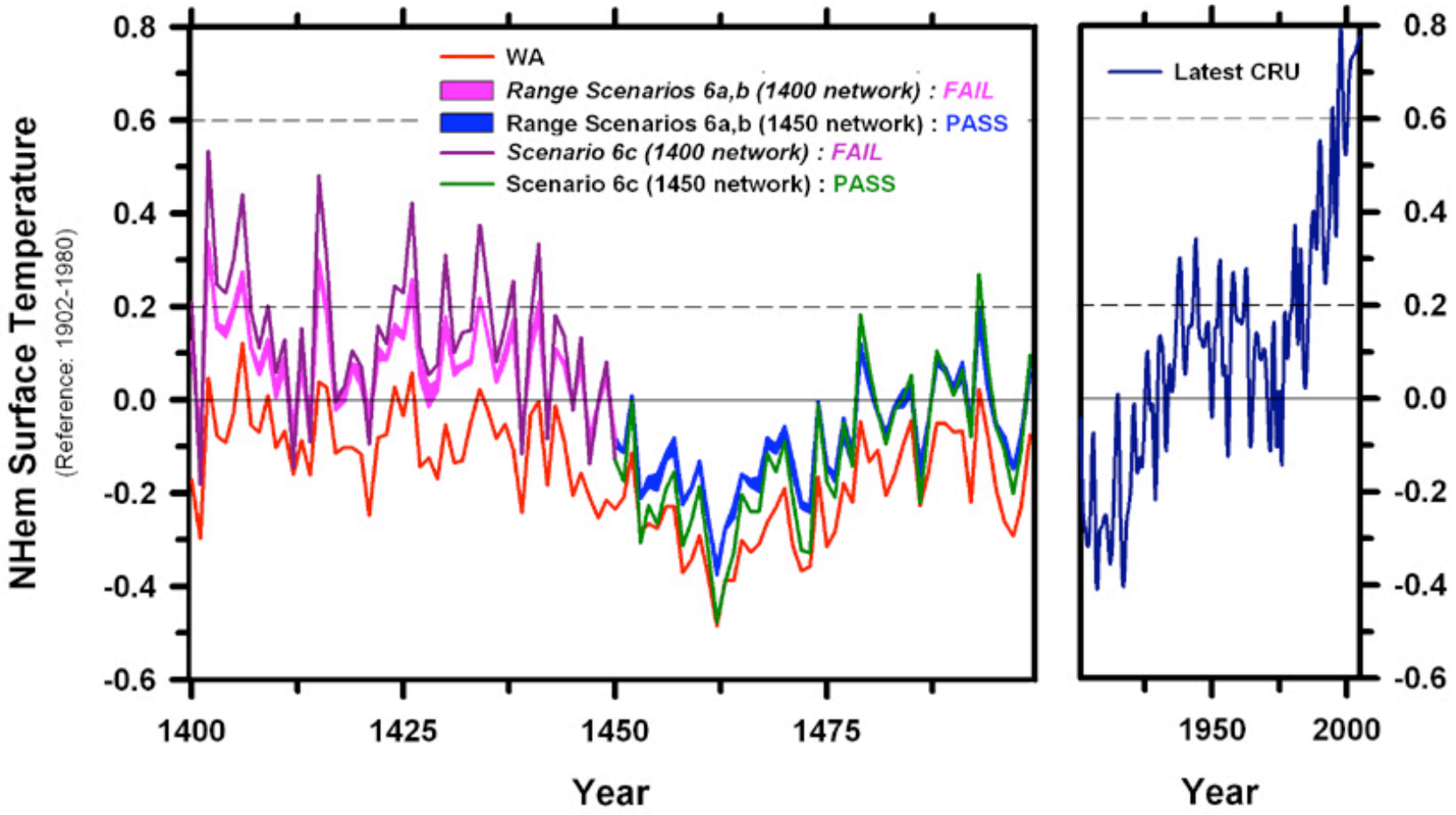
Wahl-Ammann Climatic Change: Figure 1



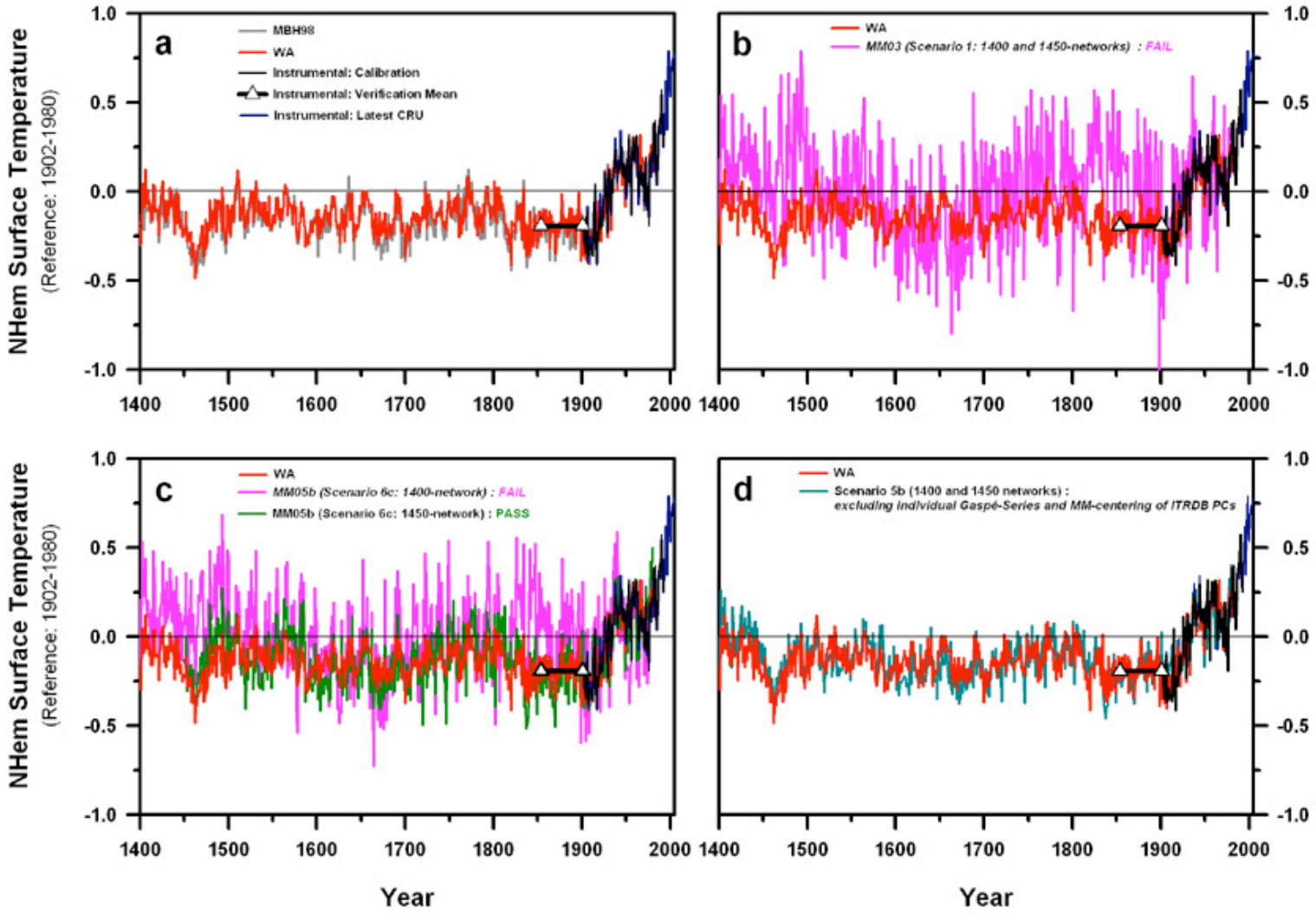
Wahl-Ammann Climatic Change: Figure 2



Wahl-Ammann Climatic Change: Figure 3



Wahl-Ammann Climatic Change: Figure 4



Wahl-Ammann Climatic Change: Figure 5

Figure 1S Relationships of r^2 and the "Reduction of Error" statistic (RE) to reconstruction performance, highlighting: (a and b) how r^2 is sensitive to the interannual tracking of two time series, yet is insensitive to the relationship between the means of the same series; and (c) how r^2 is insensitive to amplitude scaling. RE, on the other hand rewards accurate performance of capturing changes in mean state even when interannual tracking is quite poor (b), while it penalizes lack of capturing changes in mean state when interannual tracking is perfect (a). RE also does not inappropriately reward insensitivity to amplitude scaling even when the mean state is accurately captured (c). The time series presented are arbitrary. Statistics are calculated over years 0-49 only. (Adapted from Rutherford et al., 2005).

Table 1S Pearson's r^2 and CE Scores for MBH Reconstruction Emulations

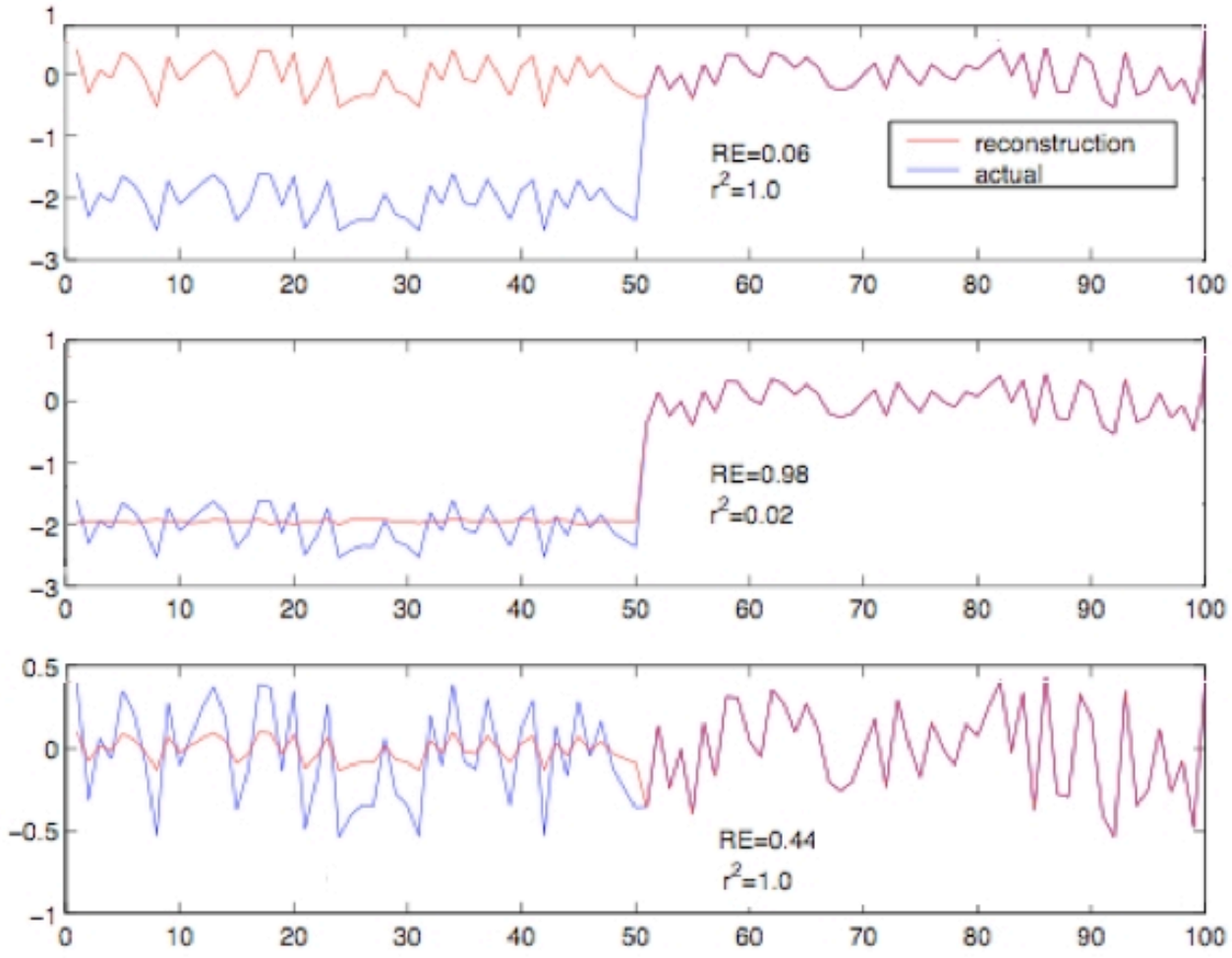
Proxy Network MBH – periods	NH Mean r^2 Calibration-period	NH Mean r^2 Verification-period	NH Mean CE Verification-period
1400-1449	0.414	0.018	-0.215
1450-1499	0.483	0.010	-0.314
1500-1599	0.487	0.006	-0.253
1600-1699	0.643	0.004	-0.259
1700-1729	0.688	0.00003	-0.161
1730-1749	0.691	0.013	-0.063
1750-1759	0.714	0.156	0.077
1760-1779	0.734	0.050	-0.070
1780-1799	0.750	0.122	0.040
1800-1819	0.752	0.154	0.069
1820-1980	0.759	0.189	0.103

Table 2S Pearson's r^2 and CE Scores for MM Reconstruction Scenarios

Scenario with MBH Proxy networks	NH Mean r^2 Calibration-period	NH Mean r^2 Verification-period	NH Mean CE Verification-period
1 Omission ITRDB & other N. Am. Data			
(reconstructed instrumental PCs not re-scaled)			
1400-1449	0.249	0.000008	-2.676
1450-1499	0.261	0.017	-7.686
2 no PC for ITRDB Data			
1404-1449	0.610*	0.003*	-0.508*
1450-1499	0.676	0.005	-0.684
1700-1729	0.742	0.0001	-0.891
1800-1819	0.770	0.068	-0.518
3 no PC for ITRDB and no Bristlecone/Foxtail			
1404-1449	0.633*	0.012*	-1.202*
1450-1499	0.716	0.016	-1.101
4 no PC for ITRDB and Limitation of Records			
(a) 1404-1449	0.574*	0.041*	-1.264*
1450-1499	0.704	0.110	-0.985
(b) 1404-1449	0.560*	0.008*	-0.863*
1450-1499	0.689	0.084	-0.632
5 No Gaspé**			
(a) MBH centered and standardized 2 PCs (svd):			
1400-1449	0.327	0.014	-0.538
1450-1499	0.485	0.008	-0.308
(b) MM centered and standardized 2 PCs (svd):			
1400-1449	0.353	0.007	-0.928
1450-1499	0.499	0.010	-0.272
(c) MM centered, not-standardized 4 PCs (princomp):			
1400-1449	0.333	0.003	-1.014
1450-1499	0.489	0.002	-0.345
(d) MM centered, not-standardized 2 PCs (princomp):			
***(indirect Bristlecone/ Foxtail exclusion*)**			
1400-1449	0.262	0.0001	-1.761
1450-1499	0.476	0.002	-0.358
6 No Gaspé and no Bristlecone/Foxtail**			
(a) MBH-centered, standardized, 2 PCs (svd):			
1400-1449	0.297	0.0001	-1.647
1450-1499	0.490	0.006	-0.338
(b) MM-centered, standardized, 2 PCs (svd):			
1400-1449	0.259	0.00003	-1.819
1450-1499	0.477	0.0006	-0.369
(c) MM-centered, *not*-standardized, 2 PCs (princomp):			
(reconstructed instrumental PCs not re-scaled)			
1400-1449	0.267	0.0002	-2.666
1450-1499	0.475	0.0002	-1.016

* note slight difference in data coverage with scenario limited to 1404-1449

** statistics given for scenario iterations described in "Results"; statistics for *all* iterations of each scenario available at supplemental information website



Wahl-Ammann Climatic Change: Figure S1