Clim. Past Discuss., 2, S139–S152, 2006
www.clim-past-discuss.net/2/S139/2006/

Climate
of the Past
Discussions

Interactive comment on "On the verification of climate reconstructions" *by* G. Bürger and U. Cubasch

Anonymous Referee #2

Received and published: 4 July 2006

This is a deeply flawed manuscript, and it's publication would damage the reputation of this promising new journal. The authors display a disregard for existing peer-reviewed literature that unambiguously refutes their main claims. Each of their primary claims is false or misleading, as detailed below in this review. Moreover, the focus the paper is now plainly inappropriate, focusing on nearly decade-old work, the details of which and the key conclusions of which have now been independently validated by numerous other studies. The manuscript is backward-looking, invoking flawed criticisms of now very old work, while current studies have moved well beyond this spurious debate about statistical minutia, focusing instead on real scientific issues: the reconstruction of spatial patterns of climate, and the elucidation of mechanisms of variability that can inform our understanding of climate and/or climate sensitivity (e.g. Mann et al, 2000; Delworth and Mann, 2000; Shindell et al, 2001; 2003; 2004; Waple et al, 2002; Braganza et al, 2003; Adams et al, 2003; Shindell et al, 2003;2004; Jones and Mann, 2004; Osborn and Briffa, 2006; Goosse et al, 2006; Hegerl et al, 2006).

This represents a 2nd attempt by the authors to publish this flawed manucript, as it has been rejected once already from the peer-reviewed climate literature. The problems with this paper are fundamental, and it would take far more time and effort than the paper merits, to detail all of the problems with the paper. In my review, I will focus on the few most serious problems that undermine the essence of the authors' claims.

It should be noted that comment recently published by Wahl et al in *Science*, it is revealed that there was an erroneous undisclosed procedure used first by Von Storch et al (2004) and subsequently adopted by Von Storch associate Cubasch in Burger and Cubasch (2005) and Burger et al (2006), which involves the detrending of data prior to calibration. Wahl et al (2006) note that use of this inappropriate procedure invalidates any conclusions drawn in studies that employ that procedure, undermining the conclusions of Burger and Cubasch (2005) and Burger et al (2006). These studies greatly misrepresented the actual method used by Mann et al (1998). The present submitted manuscript similarly misrepresents the RegEM method used more recently by Rutherford, Mann and colleagues in climate reconstruction.

The authors' main assertions are summarized in their abstract as follows:

Using a rigorous verification method, we show that previous estimates of skill approaching 60% mainly reflect a sampling bias, and that more realistic values vary about 25%. The bias results from the strong trends in the instrumental period, together with the special partitioning into calibration and validation parts. This setting is characterized by very few degrees of freedom and leaves the regression susceptible to nonsense predictors.

These assertions are shown below to be false, resulting from (1) erroneous alteration of the methodologies Burger and Cubasch purport to be testing, (2) a demonstrably untrue claim that verification skill estimates in past work was artificially inflated by spatial and/or temporal sampling bias, (3) use of statistical significance estimates that have previously been demonstrated as erroneous, and (4) an already refuted claim by the authors that Climate Field Reconstruction ("CFR") methods used by Mann, Rutherford, and collaborators overfit with "nonsense predictors" in the presence of calibration period trends. Each of these issues is dealt with in detail below:

1. Erroneous Alteration of RegEM method

Rutherford et al (2005) use the specific method of regularization of the classical Expectation

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

Maximization (EM) algorithm that is described by Schneider (2001). We henceforth refer to this specific implementation as “RegEM”. Applied to the same proxy data set as Mann et al, ‘98, RegEM produces a nearly indistinguishable reconstruction from Mann et al (1998), indicating that the overall reconstruction is robust with respect to the particular CFR method used.

The so-called “flavors” entertained by Burger and Cubasch are inconsistent with the correctly implemented RegEM method (much as the various “flavors” they entertained in Burger and Cubasch ’05 –in particular the erroneous aforementioned detrending procedure–were clearly inconsistent with correct method used by Mann et al ’98). This is explained in more detail below. The authors’ discussion and use of RegEM indicates a lack of understanding of the method, and the past implementation of that method in paleoclimate reconstruction. The various subjective steps they attempt to force into the method are inappropriate, and distort the otherwise objective RegEM method of Schneider (2001). These distortions are certain to degrade the performance of the method.

In RegEM, as in conventional expectation maximization (EM), an explicit statistical model is provided for both the regression coefficients and residual (error) term in relating missing “m” and available “a” values. For a given record **x** (a row vector of missing and available values), missing values are related to available values through

$$x_m = \mu_m + (xa - \mu_a)B + e$$

where **B** is a matrix of regression coefficients, and the residual vector **e** is an assumed random “error” vector with mean zero and covariance matrix **C** to be determined. In each iteration of the EM algorithm, estimates of the mean μ and of the covariance matrix Σ of the data **x** are taken as given, and from these, the conditional maximum likelihood estimates of the matrix of regression coefficients **B** and of the residual covariance matrix **C** are computed for each record with missing values. In the conventional EM algorithm, the estimate of **B** is the conditional maximum likelihood estimate given the estimates of μ and Σ. To insure that Σ be well conditioned, it is necessary to “regularize” the EM algorithm, through the use of a regularized estimate of **B**. The regression model (1) with the estimated regression coefficients B is then used to estimate missing values, and using the estimated missing values and the residual covariance matrix, the estimates of μ and Σ are updated. It should be noted that Σ in this context contains not just the sample covariance matrix of the completed dataset, but also a contribution due to the residual covariance **C**. The above steps are iterated until convergence.

The RegEM method of Schneider (2001) accomplishes regularization through the use of ridge regression, introducing a “regularization parameter” h that specifies the degree of inflation $(1 + h^2)$ of the main diagonal of the covariance matrix Σ. The parameter h determines the degree of smoothing of the estimated missing values. Schneider (2001) uses Generalized Cross Validation (“GCV”) to determinate an objective estimate for h.

The RegEM approach, as applied by Mann and coworkers to proxy-based CFR [Mann and Rutherford, 2002; Rutherford et al, 2003; 2005; Zhang et al, 2004; Mann et al, 2005;2006], simultaneously estimates the covariance structure both within and between the proxy (“predictors”) and instrumental (“predictand”) data sets, through an objective smoothing of missing values in the joint (proxy+instrumental) data matrix. The surface temperature field to be reconstructed is treated as a set of missing values in an incomplete data matrix consisting of the combined available standardized proxy and instrumental data. The approach, in this manner, also readily accommodates missing values in either proxy or instrumental data.

Burger and Cubasch (2005) have made the following criticism of the Mann et al (1998) truncated EOF-based method: that there is putative absence of either a “sound mathematical derivation of the model error” or “sophisticated regularization schemes that can keep this error small”. It is especially easy to see that these criticisms cannot apply to the RegEM procedure used by Rutherford et al (2005) and Mann et al (2005): As is clear from the discussion above RegEM both employs a rigorous, objective regularization scheme, and an explicit statistical modeling of the error term. The RegEM method as implemented by Schneider (2001) and used in paleoclimate reconstruction by Rutherford et al (2005) and Mann et al (2005) is therefore obviously immune to the Burger and Cubasch (2005) criticisms. And, as shown by Rutherford et al (2005), moreover, RegEM yields the same result as the original Mann et al (1998) approach, applied to the same proxy dataset. In other word, the central claim of Burger and Cubasch (2005) that the Mann et al (1998) reconstruction is not robust with regard to methodological issues, is clearly false, a point that is clear from a reading of Rutherford et al (2005). Burger and Cubasch similarly here, now misrepresent the RegEM method used by Rutherford et al (2005) and Mann et al (2005), introducing subjective, indefensible procedures that simply distort that method.

It is easy to understand why the 48 so-called “flavors” entertained by Burger and Cubasch have no relevance to the performance of the RegEM procedure described by Schneider (2001) and applied by Rutherford et al (2005) to the problem of proxy-based climate reconstruction. First, consider the issue of the target variable, what Burger and Cubasch call “GLB”. In fact, there is only one legitimate target variable in the RegEM approach, the set of unknown values of the spatial field of interest (in this case, the surface temperature field). Any other quantities of interest (e.g. the NH mean temperature) are simply diagnosed from the reconstructed field, just as they would be from the modern available instrumental temperature field. It makes particularly little sense to define an “EOF” as a “target” with RegEM, since RegEM does not employ an EOF-based representation of the data covariance. To define an EOF as a target variable would only make sense in the context of Principal Components Regression (PCR). But this represents an entirely different type of regularization procedure than RegEM, and is thus irrelevant to the issue at hand.

Now, let us consider their variable “MDL”. Again, the implementation of RegEM as defined by Schneider (2001) and employed by Rutherford et al (2005) involves only one statistical model, the solution of the above equation above using ridge regression for regularization, using GCV to select the regularization parameter h. It is true that there are a number of other possible ways to regularize the EM algorithm, including Principal Components Regression (PCR), truncated total least squares regression, and ridge regression. Schneider (2001) however specifically favors one unique regularization approach, ridge regression, since it arises as a regularization method when the observational error in the available data is taken into account. Ridge regression regularizes a total least squares regression, provided the relative variance of the observational error is homogeneous. This assumption is appropriate when, as in applications to paleoclimate reconstruction (e.g. Rutherford et al, 2005; Mann et al, 2005;2006), available data series are standardized prior to their use in CFR. Even if this assumption is not met, the true regularized estimates are close to the estimates provided by ridge regression. The alternative models proposed by Burger and Cubasch are likely to provide estimates with greater variance. Consider, for example, their use of truncated “Total Least Squares” (TLS). This is not an optimal approach to regularizing TLS. Under the assumption of homogeneous relative errors in the standardized data as discussed above, an optimal regularization of TLS leads directly to ridge regression, and not truncated TLS, which is indeed the reason Schneider(2001) employed ridge regression in the RegEM algorithm.

Now, let us finally consider the final “flavor” variable “RSC” proposed by the authors, that is, whether or not estimated values should be rescaled by their calibration period variance. The introduction of this step makes absolutely no sense at all in the context of RegEM. Not only is there no reason to rescale the values estimated by RegEM, but doing so would insure that the procedure no longer yield optimal (in the sense of minimum GCV) estimates of the missing data and and unresolved variance.

In short there are no legitimate “flavors” of the sort argued by Burger and Cubasch in the RegEM procedure as described by Schneider (2001) and as applied to the problem of paleoclimate reconstruction by Rutherford, Mann, and colleagues in numerous papers. The introduction of various subjective “flavors” by Burger and Cubasch into the RegEM framework simply serves to distort the procedure, and degrade its likely performance.

2. False assertion of sampling bias in verification skill estimates

Let us now consider the authors’ claim that the skill estimates of Mann et al (1998) based on use of the standard metric RE applied the NH mean reconstruction are somehow artificially inflated by a supposed spatial or temporal sampling bias resulting from the specific calibration and validation intervals used. The findings of Rutherford et al (2005) already demonstrate the falsehood of this claim, as described below.

Mann et al (1998) estimated statistical skill of RE=69% for the full network (i.e., available back to AD 1820), and RE=51% (note that Burger and Cubasch incorrectly cite an estimate 57%, which is for the global domain mean„ and not the Northern Hemisphere mean) for the sparsest network (available back to AD 1400), using a calibration interval of 1902-1980 with 1082 nearly continuous global surface temperature gridpoints, and a verification interval of 1854-1901 with 219 nearly continuous global surface temperature gridpoints. Small gaps were linearly interpolated. Rutherford et al (2005) made use of Northern Hemisphere only instrumental surface temperature data to calibrate the same multiproxy network as Mann et al (1998). For calibration, they employed 1002 mostly continuous Northern Hemisphere gridboxes (requiring 70% completeness and single gaps no longer than 6 months). They used RegEM to impute any temporal gaps and produce a temporally complete annual mean surface temperature dataset over the 1856-1971 interval, with a greater proportion of the early (sparser) data were based on infilled values than of the later (more widespread). They used a split calibration/validation procedure, so that the early and late portions of data were alternatively used for calibration and cross-validation. In the “early verification” analysis (1900-1971 calibration interval and 1856-1899 verification interval), only 210 Northern Hemisphere gridboxes that are nearly complete over the 1856-1899 interval were used for validation while in the “late verification” experiments (1856-1926 calibration interval and 1927-1971 verification interval), both this sparser set and the entire 1002 Northern Hemisphere gridboxes, were alternatively used for cross-validation. Imputed (i.e., originally missing) values were not used in the calculation of verification scores.

Similar validation scores to those estimated by Mann et al (1998) were found by Rutherford et al (2005) using both the “early verification” and “late verification” period. For the “early verification” results as provided in the paper, the results for the NH mean series were RE=72% for the full (1820) network, and RE=46% for the 1400 network, remarkably close to those estimated by Mann et al (1998). Using the “late verification” experiments (see online supplementary information provided for the paper) the validation scores were higher, not lower: RE=84% for the full network, and RE=66% for the 1400 network using the sparse (210 gridbox) verification grid, and RE=83% for the full network, and RE=71% for the 1400 network using the full (1002 gridbox) verification grid. So, Rutherford et al (2005) find similar verification results using both the early verification period when instrumental observations were sparse and there is a substantial trend in the NH mean series over the (late) calibration interval, and for the late verification period when instrumental observations are widespread and there is little trend over the (early) calibration interval. The fact that the validation scores of Rutherford et al (2005) are even higher that those of Mann et al (1998) in the ‘late verification’ period which employs a calibration over a relatively trend-free interval, directly contradicts the claim by Burger and Cubasch that the RE scores of Mann et al (1998) are somehow artificially inflated by the presence of trend over the calibration interval.

3. Erroneous Statistical Significance Estimation

Let us next consider the erroneous discussion by the authors of statistical significance of verification scores. The authors state the following:

The nonsense predictor mentioned above (number of available grid points) scores RE=46% (and CE=-23%), which is more than any of the flavors ever approaches in the 100 random samples. And it is not unlikely that other nonsense predictors score even higher. On this background, the originally reported 57% of verification are hardly significant. This has already been claimed by McIntyre and McKitrick 2005b in a slightly different context.

There are several problems here. First, the authors’ statistical significance estimates are meaningless, ass they are based on random permutations of subjective “flavors” that are simply erroneous in the context of the correctly implemented RegEM procedure, as detailed in section “1” above. Mann et al (2005) provide a true rigorous significance estimation procedure and their code is available as supplementary information to that paper. The procedure is based on the null hypothesis of AR(1) red noise predictions over the validation interval, using the variance and lag-one autocorrelation coefficient of the actual NH series over the calibration interval to provide surrogate AR(1) red noise reconstructions. From the ensemble of surrogates, a null distribution for RE and CE is developed. In other words, the appropriate significance estimation procedure requires the use appropriate AR(1) red noise surrogates against which the performance of the correct RegEM procedure as implemented by Schneider (2001)/Rutherford et al (2005)/Mann et al (2005) can be diagnosed. Instead, Burger and Cubasch have simply analyzed the sensitivity of the procedure to the introduction of subjective, and erroneous alterations. Their results are consequently statistically meaningless. This view was recently endorsed by the U.S. National Academy of Sciences in their report "Surface Temperature Reconstructions for the Last 2000 Years" which specifically took note of the inappropriateness of the putative significance estimation procedures used by Cubasch and collaborators in their recent work.

Burger and Cubasch (2005) attempt to bolster their claims based on a reference to claims by "McIntyre and McKitrick (2005b)", published in a social science journal “Energy and Environment” that is not even in the ISI database. McIntyre and McKitrick made essentially the same claim in a 2005 GRL article. Huybers (2005), in a comment on that article, demonstrated that McIntyre and McKitrick’s unusually high claimed thresholds for significant were purely an artifact of an error in their time series standardization. Huybers (2005), after correcting the mistake by McIntyre and McKitrick, verified the original RE significance thresholds indicated by Mann et al (1998). It is extremely surprising that Burger and Cubasch appear unaware of all of this.

The RE (and CE) values obtained by Rutherford et al (2005) are easily shown to be statistically significant against the null hypothesis of AR(1) red noise, using the Mann et al (2005) Monte Carlo procedure discussed above. Indeed, in tests with “pseudoproxy” data derived from long model simulations, Mann et al (2005) have demonstrated that highly skillful reconstructions (i.e., those which correctly predict the true long-term model history–see more discussion below) often produce negative CE scores (which Burger and Cubasch erroneously conclude to be statistically insignificant) applied to the short validation intervals available for actual proxy reconstructions. In such cases, both the RE scores and the moderately negative CE scores are statistically significant against the null hypothesis of red noise (see e.g. Table 1 of Mann et al, 2005). Burger and Cubasch are correct in stating that r^2 scores are not an appropriate metric of reconstruction skill. As discussed by Wahl and Ammann (2006) and explicitly demonstrated by Mann et al (2006), this statistic yields unacceptably high type I and type II errors of statistical inference due to its insensitivity to changes in mean or variance outside the calibration interval. The aforementioned National Academy of Sciences report also noted the inappropriateness of r^2 as a metric of goodness of fit in paleoclimate reconstruction validation.

4. Previously falsified claims of methodological tendency to overfit with "nonsense predictors"

Finally we come to "nonsense variables" argument made by Burger and Cubasch. Their basic argument here appears to be that CFR methods such as those used by Mann et al (1998), Rutherford et al (2005) and dozens of other groups of climate and paleoclimate researchers, are somehow prone to statistical overfitting in the presence of trends, which leads to the calibration of false relationships between predictors and predictand. This argument misunderstands the basic underlying assumptions of state-space based CFR methods, which simply assume that the calibration period capture the basic spatial patterns of the large-scale field, regardless of the temporal characteristics of those patterns (be they random, oscillatory, trended, etc).

Their argument is most readily seen to be false by demonstration. Burger and Cubasch (see e.g. Burger et al, 2006) argue that the signal-to-noise amplitude ratio ("SNR") in the multiproxy dataset used by both Mann et al (1998) and Rutherford et al (2005) is SNR=0.4 (or "86% noise" using the terminology Burger and Cubasch). So their above claim can be restated as follows: calibration using CFR methods such as Rutherford et al (2005) and multiproxy networks with the attributes (e.g. spatial locations and signal-vs-noise characteristics) similar to those of the proxy network used by Rutherford et al (2005) will produce unreliable reconstructions if the calibration period includes a substantial trend. They believe that such conditions will necessarily lead to the selection of "nonsense predictors" in reconstructing past variations from the available predictor network.

The real "nonsense" however is associated with their claim, which has already been tested and falsified by Mann et al (2005). Burger and Cubasch (2005) cite and discuss this paper, yet they appear unfamiliar with its key conclusions. Mann et al (2005) tested the RegEM CFR method using synthetic "pseudoproxy" proxy datasets derived from a 1150 year forced simulation of the

Full Screen / Esc
Printer-friendly Version
Interactive Discussion
Discussion Paper

NCAR CSM1.4 coupled model, with SNR values even lower (SNR=0.25 or “94% noise”) than Burger and Cubasch estimate for the actual Mann et al (1998)/Rutherford et al (2005) proxy data network. Mann et al (2005) demonstrate that even at these very low SNR values, and calibrating over precisely the same interval (1856-1980) as Rutherford et al (2005) (over which the model contains an even greater warming trend than the actual surface temperature observations), a highly skillful and unbiased reconstruction of the prior history of the model surface temperature field (and NH mean series) is produced. In other words, the RegEM CFR method applied to multiproxy networks that are even “noisier” than Burger and Cubasch estimate for the actual Mann et al (1998)/Rutherford et al (2005) proxy network, and calibrated over intervals in which the surface temperature field exhibits even greater trends than in the actual data, yields faithful long-term reconstructions of the past climate history (this can be confirmed in the model, because the 1000 year evolution of the temperature field prior to the calibration interval is precisely known). Certainly, if the method were–as Burger and Cubasch claim–prone to using “nonsense predictors” when applied to multiproxy network with the signal-to-noise attributes of the Rutherford et al (2005) proxy network, and calibrating ove intervals with trends similar to those in the actual observations, there would be at least some hint of this in tests using networks with even lower signal-to-noise ratios, and calibrated over intervals with even larger trends? But there is no such evidence at all in the experiments detailed by Mann et al (2005). In more recent work, Mann et al (2006) show that these conclusions are insensitive to what metric of reconstruction skill is used, whether the “noise” component of the pseudo-proxies is white or substantially red, whether predictors are used individually or represented by their leading Principal Component (PC) summaries, or any of the other issues yet raised by Cubasch, Von Storch, Zorita and their collaborators (Von Storch et al, 2004; Von Storch and Zorita, 2005; Burger and Cubasch, 2005; Burger et al, 2006). Moreover, Mann et al (2006) show that the standard skill metrics (RE and CE) evaluated in these experiments over similarly short validation intervals (e.g. 1856-1899) are similar to those found by Mann et al (1998) (including moderately negative CE scores), and are in all cases statistically significant relative to the properly tested null hypothesis of AR(1) red noise. By contrast, they find that r^2 scores over short (i.e., roughly 50 year intervals) do not correctly diagnose long-term reconstructive skill, and typically reject what can readily be seen over the full (1000 year) interval to be reliable, highly skillful long-term reconstructions. This is associated with the fact that this statistic is prone to high rates of tape I and type II statistical inference errors, as discussed previously.

Full Screen / Esc
Printer-friendly Version
Interactive Discussion
Discussion Paper

Mann et al (2005) have made all of the data and all (Matlab) source codes for implementation of their procedures available through the supplementary web site provided in their paper. Any researchers who feel they have legitimate criticisms of their findings are therefore free to download the code and data and demonstrate, for example, any proneness of the method to the sort of biases argued by Burger and Cubasch at the signal-to-noise ratios in question, and using the types of calibration intervals in question. The *Journal of Climate* would of course publish any legitimate criticisms of the Rutherford e al (2005) or Mann et al (2005) findings.

Some Other Serious Problems:

a. The title of the paper is deeply misleading.

The authors would apparently have their readers believe from the very general-sounding title of the paper that they are provide a meaningful discussion of general issues involved in the validation of paleoclimate reconstructions, when in fact what is provided is simply an invalid criticism of a specific group of researchers (Mann and collaborators), focusing largely on details (e.g. verification statistic estimates) of work done nearly a decade ago.

b. The authors' conclusions wouldn't follow even if their arguments were valid.

The authors attempt to extrapolate sweeping conclusions about the reliability of all millennial reconstructions from their (fundamentally flawed) analysis of one particular set of reconstructions (Mann et al, 1998; Rutherford et al, 2005):

Are 25% RE enough to decide the millennial NHT controversy? This is the crucial question. 25% RE translates to an amplitude error of ... 85%. If one were to focus the controversy into the single question: Was there a Medieval Warm Period (MWP) and was it possibly warmer than recent decades? - we doubt that question can be decided based on current reconstructions alone.

As described above, the "25%" estimate is based on an erroneous set of calculations which disagree with independent confirmations (e.g. Rutherford et al, 2005; Wahl and Ammann, 2006) of the original verification skill estimates of Mann et al (1998).

This point aside, their very sweeping statement about the "MWP" belies the remarkably narrow scope of their analysis, which focuses on one specific, now decade-old study (Mann et al, 1998;1999) that has since been supplanted by more than a dozen independent analyses, using independent proxy data and different and increasingly more robust statistical methodologies, which all come to this same basic conclusion: that the hemispheric-scale warmth of the past few decades is anomalous in the context of at least the past 1000 years. The authors appear to be unfamiliar with all of these studies. The authors moreover appear to be unaware of the large number of model simulations that have been performed that independently agree with these various empirical reconstructions within the estimates uncertainties, reinforcing conclusions regarding the anomalous nature of recent warmth. Indeed, the most recent independent work (see review by Jones and Mann, 2004; also Cook et al, 2004; Moberg et al, 2005; Osborn and Briffa, 2006; Hegerl et al, 2006) comes to the even stronger conclusion that the hemispheric-scale warmth of the past few decades is likely anomalous in the context of at least the past 2000 years.

References:

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

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

Burger, G., Fast, I., and U. Cubasch, Climate Reconstruction by Regression, Tellus, 58A, 227-235, 2006.

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

Goosse, H., Renssen, H., Timmermann, A., Bradley, R.S., Mann, M.E., Using paleoclimate proxy-data to select optimal realisations in an ensemble of simulations of the climate of the past millennium, Climate Dynamics, 27, 165-184, 2006.

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

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

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

Mann, M.E., Rutherford, S., Climate Reconstruction Using ’Pseudoproxies’, Geophysical Research Letters, 29 (10), 1501, doi: 10.1029/2001GL014554, 2002.

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

Mann, M.E., Rutherford, S., Wahl, E., Ammann, C., Testing the Fidelity of Methods Used in Proxy-based Reconstructions of Past Climate, Journal of Climate, 18, 4097-4107, 2005.

Mann, M.E. et al, Robustness of proxy-based climate field reconstruction methods, 2006 (accepted).

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

Osborn, T.J., and Briffa, K.R., The Spatial Extent of 20th-century Warmth in the context of the past 1200 years, Science, 311, 841-844, 2006.

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

Rutherford, S., Mann, M.E., Osborn, T.J., Bradley, R.S., Briffa, K.R., Hughes, M.K., Jones, P.D., Proxy-based Northern Hemisphere Surface Temperature Reconstructions: Sensitivity to Methodology, Predictor Network, Target Season and Target Domain, Journal of Climate, 18, 2308-2329, 2005.

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

Shindell, D.T., Schmidt, G.A., Miller, R.L., Mann, M.E., Volcanic and Solar Forcing of Climate Change during the Preindustrial Era, Journal of Climate, 16, 4094-4107, 2003.

Shindell, D.T., Schmidt, G.A., Mann, M.E., Faluvegi, G., Dynamic winter climate response to large tropical volcanic eruptions since 1600, Journal of Geophysical Research, 109, D05104, doi: 10.1029/2003JD004151, 2004.

Schneider, T., Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values, Journal of Climate, 14, 853-871, 2001.

Wahl, E.R., D.M. Ritson, and C.M. Ammann, C.M. (2006), Comment on ‘Reconstructing Past Climate from Noisy Data’. Science, 312, 529b.

Wahl, E.R. and C.M. Ammann, Robustness of the Mann, Bradley, Hughes Reconstruction of Surface Temperatures: Examination of Criticisms Based on the Nature and Processing of Proxy Climate Evidence, Climatic Change (in press).

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

Zhang, Z., Mann, M.E., Cook, E.R., Alternative Methods of Proxy-Based Climate Field Reconstruction: Application to the Reconstruction of Summer Drought Over the Conterminous United States back to 1700 From Drought-Sensitive Tree Ring Data, Holocene, 14, 502-516, 2004.

Zorita, E. and Von Storch, H., Methodical aspects of reconstructing non-local historical temperatures, Mem. S.A. It. 76, 794, 2005.

Interactive comment on Clim. Past Discuss., 2, 357, 2006.