*Tellus (2007), 59A, 591–598*
*Printed in Singapore. All rights reserved*



**TELLUS**

# The 'hockey stick' and the 1990s: a statistical perspective on reconstructing hemispheric temperatures

*By* BO LI[1]*, DOUGLAS W. NYCHKA[1] and CASPAR M. AMMANN[2], [1]*Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, Boulder, Colorado, USA;* [2]*Climate and Global Dynamics Division, National Center for Atmospheric Research, Boulder, Colorado, USA*

(Manuscript received 22 March 2007; in final form 28 June 2007)

ABSTRACT

The short instrumental record of about 100–150 yr forces us to use proxy indicators to study climate over long timescales. The climate information in these indirect data is embedded in considerable noise, and the past temperature reconstructions are therefore full of uncertainty, which blurs the understanding of the temperature evolution. To date, the characterization and quantification of uncertainty have not been a high priority in reconstruction procedures. Here we propose a new statistical methodology to explicitly account for three types of uncertainties in the reconstruction process. Via ensemble reconstruction, we directly obtain the distribution of decadal maximum as well as annual maximum. Our method is an integration of linear regression, bootstrapping and cross-validation techniques, and it (1) accounts for the effects of temporal correlation of temperature; (2) identifies the variability of the estimated statistical model and (3) adjusts the effects of potential overfitting. We apply our method to the Northern Hemisphere (NH) average temperature reconstruction. Our results indicate that the recent decadal temperature increase is rapidly overwhelming previous maxima, even with uncertainty taken into account, and the last decade is highly likely to be the warmest in the last millennium.

## 1. Introduction

The study of the Earth's past climate provides insight into how the Earth system varied over longer timescales and how it responds to various forcings. Knowledge of the natural background is crucial to understand the dynamics of current climate change. In particular, the temperature reconstruction for the last one to two millennia can shed light on how the warming of recent decades compares to the range of natural fluctuation. Over the past 10 yr the quality of regional and hemispheric temperature reconstruction for the past millennium has increased substantially, due to both methodological development as well as to improved data availability (Bradley and Jones, 1993; Jones et al., 1998; Mann et al., 1998; Crowley and Lowery, 2000; Briffa et al., 2001; Esper et al., 2002; Mann and Jones, 2003; Cook et al., 2004; Luterbacher et al., 2004; Moberg et al., 2005; Rutherford et al., 2005; Xoplaki et al., 2005; D'Arrigo et al., 2006; Hegerl et al., 2006). However, it is widely acknowledged that these reconstructions may contain substantial uncertainty that is difficult to quantify in its full complexity. The recent debate about the 'hockey stick' temperature reconstruction (Mann et al., 1998, 1999, subsequently referred to as MBH98 and MBH99)—a time evolution of past NH temperatures showing a slightly decreasing but relatively stable millennium (the shaft) followed by a comparatively rapid warming over the last century (the blade)—has highlighted not only the importance of these past temperature estimates but also the fact that the community still struggles with uncertainty inherent in several aspects of the reconstruction procedures (MBH99; North et al., 2006). Although many specific criticisms on MBH98 have been examined and only minor corrections were found to be necessary to address many of the concerns debated in the literature (Mann and Jones, 2003; von Storch et al., 2004; Bürger and Cubasch, 2005; North et al., 2006; Wahl and Ammann, 2007; Ammann and Wahl, 2007), there are some basic aspects of the statistical uncertainty in hemispheric reconstructions that have not been properly developed and implemented. In any event, our paper does not attempt to investigate the validity of the underlying assumptions of MBH98 nor do we include uncertainty from the regression models (Bürger and Cubasch, 2005), though the Monte Carlo idea can be extended to account for those types of uncertainties.

Traditionally, most large-scale temperature reconstructions are presented as single time series without error bars or

*Corresponding author.
e-mail: boli@ucar.edu.
DOI: 10.1111/j.1600-0870.2007.00270.x

confidence ranges representing the direct result of a regression procedure of proxies against instrumental data. Although it is implicitly or explicitly recognized that the proxies contain noise, the reconstruction outcomes are often used as 'unique'. This, however, does not take into account that the noise component in all proxies (and to some degree, albeit smaller, in the instrumental 'truth', see Brohan et al., 2006) is but a single representation of many possible noise realizations. Should hypothetically the exact same climate occur again, the noise realization would be different. Therefore, it would be more appropriate to recognize each proxy based climate reconstruction as an individual member of a family of possible realizations; in a similar way, the different climate reconstructions with their underlying methodologies and data are forming a family of estimates of past climate, each with their own properties.

Of course, uncertainty has not completely been ignored in the past temperature reconstruction. In the literature some (MBH98; MBH99; Briffa et al., 2001; Mann and Jones, 2003; D'Arrigo et al., 2006) have published their reconstructions with an associated range of uncertainty. Such ranges have generally been determined from the unresolved variance in the calibration period, and the 1- or 2-standard deviations have then been simply applied uniformly to each annual or decadal temperature estimate. Following a different concept, Esper et al. (2002) and Cook et al. (2004) provided bootstrap confidence intervals for low frequency temperature by resampling the tree ring chronology sites. The uncertainty range shown in Moberg et al. (2005) is of a slightly different nature because it is primarily the result of dating uncertainty in the applied low-frequency proxies, but not calibration uncertainty due to noise.

Here, we develop a statistical method to reconstruct past temperatures together with its confidence ranges by keeping track of different sources of uncertainties (Section 2). We illustrate our approach with relatively simple, direct multivariate reconstruction of hemispheric mean temperatures using a limited proxy set (Section 3); however, the same protocol can easily be applied in other applications using different regression methods (Hegerl et al., 2006) or in more involved climate field reconstructions (MBH98; Rutherford et al., 2005). The benefit of our approach is demonstrated in the evaluation of the question about past decadal mean temperature maxima and how they compare to the present (Section 3.2). Such an analysis is not directly achievable from previous methods of uncertainty estimates because they were based on the uncertainty ranges developed at the interannual timescale. Using Monte Carlo ensemble simulations of past temperature evolution, we approximate a probability distribution of NH decadal means over the period (1000–1849 AD). Although still based on the assumption that the relationship between the temperature and proxies remains the same over time (Principle of Uniformitarianism), our method offers, for the first time, an explicit answer to the question on decadal average temperatures because we are able to take serial correlation of uncertainty into account.

## 2. Data and methods

### 2.1. Instrumental and proxy data

Motivated by the recent discussion of uncertainty in the MBH99 reconstruction (North et al., 2006), we illustrate our statistical procedures for the purpose of this article by restricting our network of proxy records to the 14 series originally used in MBH99 for the period back to the year 1000 (see table 1 in MBH99). Again because we mainly focus on illustrating how to investigate some uncertainties that cannot be solved by MBH99, we do not critically evaluate these proxies but simply apply them as the best available set at the time of 1999, despite the fact that many more, and possibly better, series are available today. We also assume the linear and stationary relationship between these 14 proxies and the temperature evolution following MBH98, MBH99. We use this network to compare our results with the detailed discussion of maximal temperature and uncertainty in MBH99.

Instead of performing a full field reconstruction, we simply focus on the NH average temperature as the target series. We use the latest HadCRUT3v series available at *http://www.cru.uea.ac.uk/cru/data/temperature/* (Brohan et al., 2006), which is significantly updated from the instrumental series used in MBH99. As described below, we use the maximum possible length for calibration (1850–1980, bounded by the proxies ending in 1980), and apply the 1961–1990 mean as reference period for all results.

### 2.2. Statistical uncertainty concept

Let $T_t$ and $\mathbf{p}_t$ denote the temperature and proxies at time $t$, respectively. Our basic statistical model is:

$$T_t = \mathbf{p}_t'\boldsymbol{\beta} + e_t, \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients, and the vector of error $\mathbf{e} = (e_1, \ldots, e_t)' \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma})$, which means $\mathbf{e}$ has a joint normal distribution with zero mean and some covariance matrix $\boldsymbol{\Sigma}$. This model takes uncertainties from both the instrumental data and the proxies into account. In case of very large noise on the proxy data, that is, if the noise dominates, then model (1) will identify that this particular proxy has little skill in predicting NH temperatures and so its contribution will be automatically down weighted.

We assume that there is a linear relationship between the temperature and proxies, and that this relationship does not change throughout the entire time period. This is central to all climate reconstructions, and in addition assuming that the conditional distribution of temperature given proxies is normally distributed, we can reconstruct the past temperature based on the multivariate linear regression. A more subtle assumption made in the linear model is that this linear relationship holds at all timescales. Given the relatively short time period for calibration this assumption

is difficult to check and any reconstruction based on such a stationary assumption must be executed with the necessary caution. However, as Ammann and Wahl (2007) point out, the series is likely long enough to recognize the key geophysical mechanism of long-term radiative forcing changes, and thus dangers of over interpretation based on the calibration process are probably relatively small (see also Mann et al., 2007). There remain, however, sizable uncertainties about the proxies themselves and how they represent the climatic conditions over multiple timescales.

We are now interested in studying the uncertainties underlying this linear regression reconstruction, and, more specifically, discovering the distribution of the maximal temperatures after taking the uncertainties into consideration. Accordingly we address three types of uncertainties:

(i) Effects of possible autocorrelation in the errors of the linear model, owing to the assumption of independent errors for the ordinary linear model.

(ii) Potential prediction errors due to the overfitting from the calibration period. </p>

(iii) Uncertainty induced by the calibration procedure, such as by the variance of the estimated model parameters.

We choose the full time period from year 1850 to 1980 as the calibration period for model (1), and then apply the estimated model to the proxies from year 1000 to 1849 to reconstruct the temperature for that time period. To explore whether $e_t$ should be modeled as correlated or independent for our data, we fit an ordinary least squares (OLS) model to the temperature and proxies assuming the errors are independent. One can see in Fig. 1 that the regression residuals exhibit a clear temporal correlation and thus suggest that an autoregressive (AR) model for residuals is more appropriate. In order to decide the order of the AR model, we study the temporal correlation structure in the regression residuals by examining the autocorrelation function (ACF) and partial autocorrelation function (PACF) shown in Fig. 2 and by computing the Durbin–Watson statistic (e.g. Fox, 2002). All these techniques reveal the existence of correlation for the first few temporal lags. Noting that ACF appears to tail off while PACF roughly cuts off after lag 2, we propose an AR(2) model to the residuals. Then we investigate ACF of the innovations (or residuals) based on the fitted AR(2) model and compute $Q$-statistic to assess the goodness of fit (GOF) of the AR(2) model (Shumway and Stoffer, 2006). Both these diagnostics suggest that an AR(2) model is sufficient to capture the temporal correlation in the residuals. We also assess the GOF of an AR(1) model since we seek the most parsimonious but still sufficient model. However, we find many significant $p$-values corresponding to the $Q$-statistic which indicates that there are autocorrelations remaining in the innovations from an AR(1) model. These disappear under AR(2), thus we settle on an AR(2) model.

An AR(2) process is defined as:

$$e_t = \phi_1 e_{t-1} + \phi_2 e_{t-2} + \epsilon_t,\ \epsilon_t \sim \text{iid Normal}(0, \sigma^2), \tag{2}$$

where $\phi_1$ and $\phi_2$ are coefficients governing the correlation of time lag 1 and time lag 2. 'iid' is the abbreviation of 'independent identically distributed'. In Section 3.1 below, we exploit a nice property of an AR(2) process that cov $(e_t, e_{t+k})$ can be calculated iteratively as in (p. 100 Shumway and Stoffer, 2006). Note that the AR(2) process of errors leads to a non-diagonal form of $\mathbf{\Sigma}$.

Various regression methods have been used in past temperature reconstruction. For example, MBH98 and MBH99 applied essentially an OLS to fit the linear model, and Hegerl et al. (2006) employed the total least squares to calibrate proxy reconstructions. Bürger et al. (2006) studied a range of regression techniques and found a large variability in the reconstruction using different regression flavours. Motivated by the temporal correlation in the errors, we fit the model (1) using generalized least-squares (GLS), which allows the errors to be correlated (see Carroll and Ruppert, 1988). This differs from the previous work where the errors have been generally treated as indepedent. We set the correlation structure of errors as an AR(2). The parameters involved in model (1) are $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2, \phi_1, \phi_2)'$. We choose

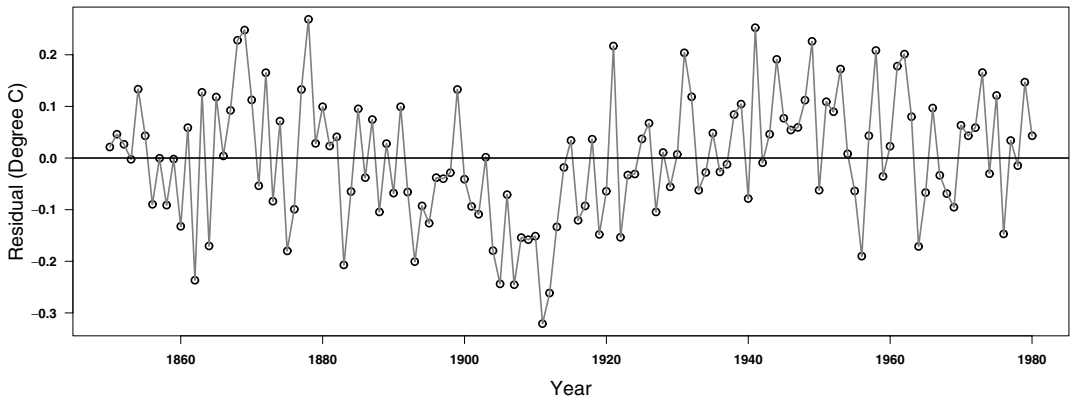


*Fig. 1.* The time series of ordinary least squares (OLS) regression residuals (circles) of instrumental temperature regressed on the 14 proxies during the calibration period from 1850 to 1980 as in Section 2.2.

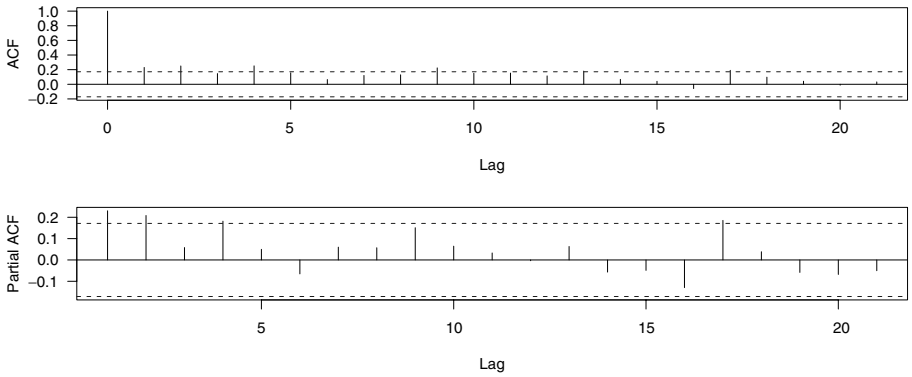


*Fig. 2.* Autocorrelation function (ACF) and partial autocorrelation function (PACF) of residuals in Fig. 1.

the maximum likelihood method in the GLS fitting and obtain the parameter estimates $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}', \widehat{\sigma}^2, \widehat{\phi}_1, \widehat{\phi}_2)'$ simultaneously.

Being concerned with the possible overfitting problem during the calibration period, we use 10-fold cross-validation (e.g. Hastie et al., 2001, p. 214) to quantify how much overfitting there is if any. We sequentially partition the 131 yr (1850–1980) into 10 groups of approximately equal size. We withhold one of the 10 groups in turn, use the other nine groups of data to calibrate our linear model. We then use this estimated model to predict the withheld group. The difference between the prediction and the actual observations is an unbiased estimate of the statistical prediction error. If there is no overfitting, the variance of the observed prediction error is expected to be equal to the prediction variability derived from our linear model. However, a larger variance of the observed prediction error is a sign of overfitting and calls for an inflation adjustment to account for the overfitting. By applying 10-fold cross-validation, the sample mean of the inflation coefficients obtained from the 10 sets of prediction errors estimate the inflation needed for our linear model. For simplicity, we fix $\widehat{\phi}_1$ and $\widehat{\phi}_2$ but leave other parameters varying in our cross-validation. This approach suggests an inflation factor of 1.30.

To account for the uncertainty from parameter estimates, $\widehat{\boldsymbol{\theta}}$, we employ a parametric bootstrap (Davison and Hinkley, 1997, p. 15) to determine the sample distribution of $\widehat{\boldsymbol{\theta}}$. This begins with generating ensembles of temperature based on model (1). Particular to each ensemble, we first generate an AR(2) error, $\mathbf{e}_0$, as defined in (2) taking $\widehat{\sigma}^2, \widehat{\phi}_1$ and $\widehat{\phi}_2$ as true parameters. Then let $\widetilde{\mathbf{T}} = (\widetilde{T}_{1850}, \ldots, \widetilde{T}_{1980})'$ denote an ensemble of temperature and $\mathbf{P}$ denote the known proxy matrix containing rows $\mathbf{p}_{1850}, \ldots, \mathbf{p}_{1980}$, $\widetilde{\mathbf{T}} = \mathbf{P}\widehat{\boldsymbol{\beta}} + \mathbf{e}_0$ produces one valid ensemble. These temperature ensembles have the same mean function $\mathbf{P}\widehat{\boldsymbol{\beta}}$, however, each ensemble has its own noise, which makes any individual ensemble feature differently from the others. For each ensemble, we repeat the GLS model fitting procedure to get the parameter estimates, denoted by $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\beta}}', \widetilde{\sigma}^2, \widetilde{\phi}_1, \widetilde{\phi}_2)'$, which are considered to have the same distribution as $\widehat{\boldsymbol{\theta}}$ estimated from the real data. We generate 1000 temperature ensembles and thus obtain 1000 $\widetilde{\boldsymbol{\theta}}$. The sampling distribution of these $\widetilde{\boldsymbol{\theta}}$ is a valid estimate of the distribution of $\widehat{\boldsymbol{\theta}}$. For example, the sample standard deviation of $\widetilde{\sigma}^2$ estimates the standard error of $\widehat{\sigma}^2$. Our approach for accounting for uncertainties is conceptually different from the significance thresholds applied in other studies (e.g. MBH98; MBH99).

## 3. Temperature reconstruction and results

### 3.1. Ensemble reconstruction

Again resorting to ensembles, we integrate the various uncertainty sources discussed above in our reconstruction. The principal advantage of ensembles lies in the novelty of simplifying complex problems. For example, predicting the uncertainty of decadal maximum via ensembles avoids the difficulty that there is no closed form for the distribution of the estimated decadal maximum. To illustrate how to generate ensembles for our purpose, we revisit model (1) in a new context. Let $\widetilde{\mathbf{T}} = (\widetilde{T}_{1000}, \ldots, \widetilde{T}_{1849})'$ and $\mathbf{P}$ be the proxy matrix containing rows $\mathbf{p}_{1000}, \ldots, \mathbf{p}_{1849}$ in this section. An ensemble is given by $\widetilde{\mathbf{T}} = \mathbf{P}\widetilde{\boldsymbol{\beta}} + (e_{1000}, \ldots, e_{1849})'|(e_{1850}, \ldots, e_{1980})'$, where $|$ means 'conditioned on'. The conditional error term ensures the ensemble to be temporally correlated with the instrumental temperature. The concise form of the conditional distribution of Gaussian random vectors and the explicit form of the AR(2) covariance matrix allows one to compute the conditional distribution of $(e_{1000}, \ldots, e_{1849})'|(e_{1850}, \ldots, e_{1980})'$ easily (e.g. Stein, 1999, p. 229). We pick each $\widetilde{\boldsymbol{\theta}}$ in turn from the 1000 members of $\widetilde{\boldsymbol{\theta}}$ generated in Section 2.2, inflate its $\widetilde{\sigma}^2$ by the inflation coefficient to get $\widetilde{\widetilde{\sigma}}^2 = 1.30\widetilde{\sigma}^2$, and take the selected $\widetilde{\boldsymbol{\theta}}$ with $\widetilde{\widetilde{\sigma}}^2$ in place of $\widetilde{\sigma}^2$ as true parameters to generate ensembles. This brings all the components of uncertainty into the ensembles. In this way, we generate 1000 temperature ensembles corresponding to the 1000 individual $\widetilde{\boldsymbol{\theta}}$.

### 3.2. Decadal maxima

Figure 3a shows one example of a reconstructed ensemble member with its annual values (small grey circles) and their moving

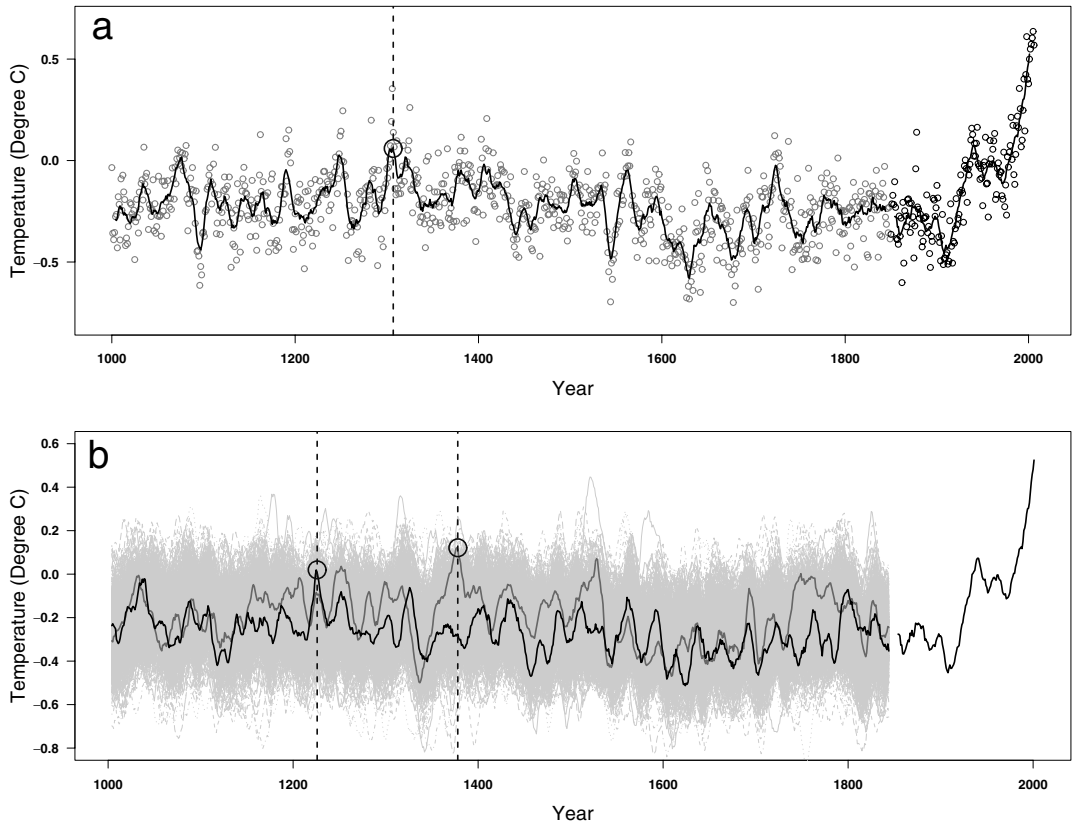


*Fig. 3.* (a) An ensemble of reconstructed temperature (small grey circles) and its decadal average (the curve embedded in grey circles), together with the instrumental temperature (small black circles) and its decadal average (the curve embedded in the black circles). The maximum of the decadal average (large black circle) and its corresponding year (dashed line) are identified. (b) 1000 decadal averaged ensembles (fine grey lines), and two examples in solid black and solid grey. Large circles and the dashed lines indicate the maximum decadal averages and their corresponding years for each of the two specified ensembles.The black line after 1850 represents the decadal average of the instrumental temperature.

10-yr average value, accompanied by the annual instrumental data (small black circles). Because little skill is to be expected at the interannual timescale (Ammann and Wahl, 2007), we are focusing on the decadal average temperature. Further, as we desire to compare the reconstruction period with the instrumental record, the decadal maximum that we now investigate is of particular interest. Therefore, we calculate the running decadal average (the solid curve) using 10 yr as the moving boxcar window. We then identify the maximum decadal average and record its corresponding year (marked by the large circle and the vertical dashed line). We do this to each of the 1000 temperature ensembles that we generate in this section, thus obtain 1000 decadal maxima with their corresponding years. The result enables us to compute the probability of each year corresponding to the maximum decadal average across all the ensembles. All decadal averaged ensemble members (fine grey lines) are given in Fig. 3b. For illustration purpose, we highlight two ensembles in solid black and solid grey, respectively. For each of these two ensembles, we identify the maximal temperature as well as the central year of the decade when the maximum occurs.

All the temperature ensembles shown in Fig. 3 are possible realizations with equal chance of occurrence given the same setting of the proxy and instrumental data, though each one owns the unique pattern that reflects the uncertainties of the reconstruction. Hence the statistical inference obtained from each ensemble is an equally valid estimate of that inference, and moreover, the variability of inferences from different ensembles shapes the distribution of that inference. It is widely recognized that the ensemble members can efficiently provide solutions in studying systems with a large number of coupled degrees of freedom. This essentially is the idea of Monte Carlo and has been employed extensively. For instance, ensembles are used in Allen and Smith (1996) to study how to distinguish the signals from an arbitrary noise process; Gel et al. (2004) successfully quantify the uncertainty of mesoscale weather field forecasting also via ensembles. Particular to our interest, the decadal maxima from ensembles constitute a sample of the estimated decadal maximum, thus this sample distribution is a reliable estimate of the distribution of the estimated maximum decadal average. We summarize the 1000 ensembles together with the running decadal average of instrumental temperature from year 1850 to 2006 in Fig. 4.

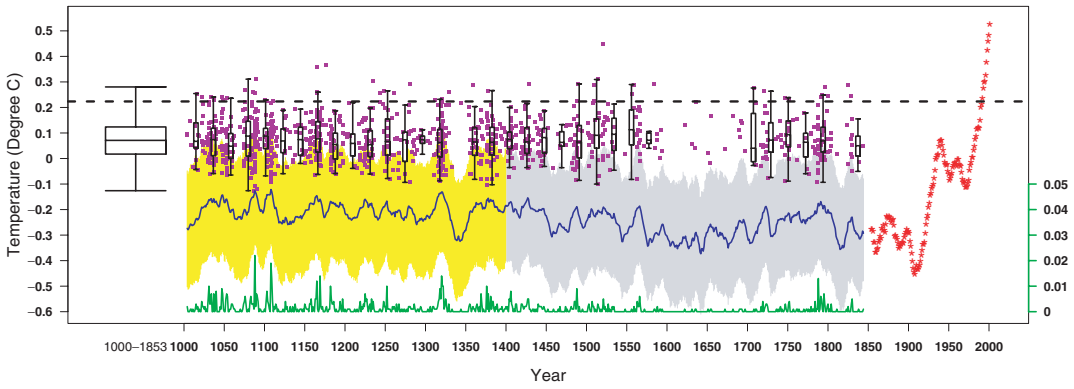


*Fig. 4.* Summary of 1000 temperature ensembles: the decadal average of mean temperature over the 1000 ensembles (blue curve) with its 95% confidence region (yellow and grey band); decadal maxima from 1000 individual ensembles (purple dots) and the chance of each year corresponding to the decadal maximum (green curve scaled by the green label); the upper bound of the 95% confidence interval of the decadal maxima (dashed line) and decadal instrumental temperatures (red asterisks). The small box plots overlapped with purple dots show the distribution of decadal maxima in small groups with each group containing about 20 yr, and the leftmost big box plot shows the distribution of all the decadal maxima. The decadal average of mean temperature before 1400 (the blue curve embedded in the yellow band) and its 95% confidence region (yellow band) can be compared to the corresponding section of Fig. 3(a) in MBH99.

*Table 1.* Distribution of decadal averages of four decades with highest probabilities of decadal maxima in the reconstruction period (upper section) and the decadal average of the four warmest decades in the instrumental period (lower section). Decade: 10 yr over which the temperature is averaged; Prob. of Max: the probability of the decade corresponding to the decadal maximum. Mean (95th quantile; 99th quantile; Max): the mean (95th quantile; 99th quantile; Max) of the decadal averages from 1000 ensembles.

| Decade | Prob. of Max | Mean (°C) | 95th quantile (°C) | 99th quantile (°C) | Max (°C) |
|---|---|---|---|---|---|
| 1084–1093 | 0.022 | −0.126 | 0.062 | 0.146 | 0.295 |
| 1104–1113 | 0.019 | −0.121 | 0.070 | 0.127 | 0.217 |
| 1165–1174 | 0.014 | −0.159 | 0.034 | 0.148 | 0.304 |
| 1316–1325 | 0.014 | −0.131 | 0.060 | 0.175 | 0.288 |
| 1970–1979 | – | −0.097 | – | – | – |
| 1980–1989 | – | 0.086 | – | – | – |
| 1990–1999 | – | 0.301 | – | – | – |
| 1997–2006 | – | 0.525 | – | – | – |

The blue curve before 1400 and the yellow shaded band in Fig. 4 can be compared to the corresponding section of Fig. 3(a) in MBH99. The period 1400–1849 shown here is based on the same limited 14-proxy network, while in MBH99 the network is continuously updated to the highest possible density (i.e. MBH98). Therefore, the details of our reconstructed temperatures after 1400 are different than those in the original papers, yet the general conclusions are directly applicable. Our ensemble reconstruction allows us to obtain the uncertainties for each individual year as well as for decadal average temperature. Such an inference about the decadal averages cannot be obtained in MBH98 and MBH99, because their errors are independent in time. For each ensemble, we calculate the running decadal average and identify the maximum (purple dots in Fig 4) of these decadal averages. Using the maxima from ensemble members we can describe the distribution of past decadal maximal temperature as illustrated by the black box plots in Fig. 4. The decadal maxima are not uniformly distributed along the time, instead, they cluster around specific time periods that represent most likely candidates for 'warmest decades' during the reconstruction period. The maxima mainly concentrate around 1100, the later 12th and the early 14$^{th}$ century as well as to a lesser degree in the late 18th and the early 11th century, leaving a large gap centred at the 17th century. The green curve shows the probability of each year to be the centre of the decade that has the maximum decadal average. All the maximum decadal NH temperatures before 1850 remain below 0.22 °C with 95% confidence, and none surpass 0.45 °C. We then selected four decades with highest probabilities in terms of decadal maxima over the reconstruction period. The upper section of Table 1 summarizes the distribution of the simulated decadal averages within these four decades.

To make a comparison between the past maxima and recent temperatures, Fig. 4 shows in red the running decadal average of the instrumental temperatures after 1850. After varying within the range of the reconstructed temperature, the instrumental

*Table 2.* Distribution of decadal averages of four decades with highest probabilities of decadal minima in the reconstruction period (upper section) and the decadal averages of the coldest decade in the instrumental period (lower section). Decade: 10 yr over which the temperature is averaged; Prob. of Min: the probability of the decade corresponding to the decadal minimum; mean (5th quantile; 1st quantile; Min): the mean (5th quantile; 1st quantile; Min) of the decadal averages from 1000 ensembles.

| Decade | Prob. of Min | Mean (°C) | 5th quantile (°C) | 1st quantile (°C) | Min (°C) |
|---|---|---|---|---|---|
| 1638–1647 | 0.043 | −0.373 | −0.576 | −0.655 | −0.775 |
| 1672–1681 | 0.029 | −0.350 | −0.532 | −0.618 | −0.674 |
| 1811–1820 | 0.027 | −0.356 | −0.555 | −0.639 | −0.813 |
| 1600–1609 | 0.021 | −0.343 | −0.532 | −0.622 | −0.840 |
| 1904–1913 | – | −0.453 | – | – | – |

temperature shows a sharp increase since 1980 and reaches 0.52 °C. Based on the proxy data set of MBH99, the reconstructed decadal maxima remain substantially lower than the observed decadal mean temperature in the late–20th-century. The lower section of Table 1 offers the direct comparison. Up to the 1980s, instrumental temperatures are not significantly higher than the reconstructed maxima of before 1850 (below 0.22, the upper bound of the 95% confidence interval of the reconstructed decadal maxima). With the 1990s, however, observed temperatures begin to rise clearly above the maxima from Monte Carlo ensembles. For example, the 1990s is 0.08 warmer than the upper bound of the 95% confidence interval, and the most recent decade (1997–2006) is yet another 0.22 warmer.

Although here we are mostly interested in decadal maxima, our ensembles equally allow us to obtain inferences for decadal minima. Table 2 is equivalent to Table 1, but applied to minima. The last row of Table 2 shows the decadal average of the coolest decade among the instrumental temperatures. All the four most likely coolest decades in our reconstruction concentrate in the little ice age. Those pre-industry cold periods do not seem necessarily colder than the coolest instrumental decadal average.

## 4. Discussion and conclusions

In this paper, we explicitly quantify the uncertainty of the reconstructed temperature while taking the errors in instrumental records and proxies into account through a rigorous statistical approach. Specifically, we quantify errors from various components and synthesize them at different levels to ensure that all those errors are reflected in ensembles. A benefit of introducing the notion of ensemble reconstruction is that ensembles make it easy to draw more sophisticated inferences about past temperature evolution such as the decadal mean, and in particular about the decadal maximum. There are several important differences in this analysis from previous work. Adjustment was made for temporal correlation in errors, cross-validation was used to adjust for overfitting and bootstrapping was used to determine uncertainty in the estimated parameters. Our approach essentially parallels the advances of a Bayesian model, which allows the introduction of various uncertainties from different stages.

Our AR(2) error model compares conservatively with previously applied error estimates, such as the temporally independent errors (e.g. MBH98; MBH99; Jones et al., 2001), or even with an AR(1) error discussed in Mann et al. (2005, 2007) when generally a temporal correlation coefficient of 0.3 has been found sufficient.

Although we focus on presenting a methodology for the uncertainty analysis, it is worth to mention that the reconstruction is only robust under the given assumptions. However, there is a possibility of violations of those assumptions. For example, the increase of $CO_2$ may accelerate the growth of trees (e.g. MBH99), so that makes the recent relationship between tree rings and temperature differ from the past. If this is true, it will break the important statistical assumption of a stationary relationship between temperature and proxies. In addition, because the proxy records end in 1980, the warm decades since then cannot be reconstructed. Hence, based on the stationarity assumption, we use the instrumental data for this period.

It is also worth noting that the spectral density of the temperature at low frequencies, say, longer than 100 yr of their corresponding periods, cannot be captured by the calibrated temperature. Therefore, should the relationship between proxies and climatic parameters at frequencies that can be represented by the instrumental record be different from the higher frequencies, then no statistical model could resolve this feature.

Finally, this work is only based on the knowledge of the 14 proxies in MBH99. Today, almost 10 yr after the original publication of MBH99, many more records have become available and it will be interesting to see how robust the conclusions are, even with our way of estimating uncertainty being incorporated. Our code and example data can be obtained from the website (http://www.image.ucar.edu/∼boli/research.html).

## 5. Acknowledgments

The authors thank Dr. Michael Mann for providing the proxy data. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

## References


Allen, M. R. and Smith, L. 1996. Monte Carlo SSA: detecting irregular oscillations in the presence of colored noise. *J. Climate* **9**, 3373–3404.

Ammann, C. M. and Wahl, E. R. 2007. Importance of the geophysical context for statistical evaluation of climate reconstruction procedures. *Clim. Change*, doi: 10.1007/S10584-007-9276-X

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

Briffa, K. R., Osborn, T. J., Schweingruber, F. H., Harris, I. C., Jones, P. D. and co-authors. 2001. Low-frequency temperature variations from a northern tree ring density network. *J. Geophys. Res.* **106**, 2929–2941.

Brohan, P., Kennedy, J. J., Haris, I., Tett, S. F. B. and Jones, P. D. 2006. Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. *J. Geophys. Res.* **111**, D12106, doi:10.1029/2005JD006548.

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

Bürger, G., Fast, I. and Cubasch, U. 2006. Climate reconstruction by regression 2 variations on a theme. *Tellus* **58A**, 227–235.

Carroll, R. J. and Ruppert, D. 1988. *Transformation and Weighting in Regression*, Chapman and Hall, New York.

Cook, E. R., Esper, J. and D'Arrigo, R. D. 2004. Extra-tropical northern hemisphere land temperature variability over the past 1000 years. *Quater. Sci. Rev.* **23**, 2063–2074.

Crowley, T. J. and Lowery, T. 2000. How warm was the medieval warm period?. *Ambio* **29**, 51–54.

D'Arrigo, R., Wilson, R. and Jacoby, G. 2006. On the long-term context for late twentieth century warming. *J. Geophys. Res.* **111**, D03103, doi:10.1029/2005JD006352.

Davison, A. C. and Hinkley, D. V. 1997. *Bootstrap Methods and Their Application*, Cambridge University Press, United Kingdom.

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.

Fox, J. 2002. Time-series regression and generalized least squares. Appendix to: *An R and S-PLUS Companion to Applied Regression*, http://socserv.mcmaster.ca/jfox/Books/Companion/appendix.html.

Gel, Y., Raftery, A. E. and Gneiting, T. 2004. Calibrated probabilistic mesoscale weather field forecasting: the geostatistical output perturbation method (with discussion). *J. Amer. Stat. Assoc.* **99**, 575–587.

Hastie, T., Tibshirani, R. and Friedman, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

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

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

Jones, P. D., Osborn, T. J. and Briffa, K. R. 2001. The evolution of climate over the last millennium. *Science* **292**, 662–667.

Luterbacker, J., Dietrich, D., Xoplaki, E., Grosjean, M. and Wanner, H. 2004. European seasonal and annual temperature variability, trend, and extremes since 1500. *Science* **303**, 1499–1503.

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

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., Rutherford, S., Wahl, E. and Ammann, C. 2005. Letters testing the fidelity of methods used in proxy-based reconstruction of past climate. *J. Climate* **18**, 4097–4107.

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

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

North, G. R., Biondi, F., Bloomfield, P., Christy, J. R., Cuffey, K. M. and co-authors 2006. *Surface Temperature Reconstructions for the last 2000 years*. Nat. Acad. Press, 144 pp.

Rutherford, S., Mann, M. E., Osborn, T. J., Bradley, R. S., Briffa, K. R. and co-authors. 2005. Proxy-based Northern Hemisphere surface temperature reconstructions: sensitivity to method, predictor network, target season, and target domain. *J. Climate* **18**, 2308–2329.

Shumway, R. H. and Stoffer, D. S. 2006. *Time Series Analysis and Its Applications: With R Examples*, Springer, New York.

Stein, M. L. 1999. *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York.

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

Wahl, E. R. and Ammann, C. M. 2007. 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. *Clim. Change*, doi: 10.1007/S10584-006-9105-7

Xoplaki, E., Luterbacker, J., Paeth, H., Dietrich, D., Steiner, N. and co-authors. 2005. European spring and autumn temperature variability and change of extremes over the last half millennium. *Geophys. Res. Lett.* **32**, L15713, doi:10.1029/2005GL023424.