Elementary Reconstruction of the Hockey Stick Curve: Discussion of Paper by Li, Nychka and Ammann

Richard L. Smith*

August 14, 2010

Abstract

The paper by Li, Nychka and Ammann (2010) has exemplified the power of Bayesian Hierarchical Models to solve fundamental problems in paleoclimatology. However, much can also be learned by more elementary statistical methods. In this discussion, we use principal components analysis, regression, and time series analysis, to reconstruct the temperature signal since 1400 based on tree rings data. Although the "hockey stick" shape is less clear cut than in the original analysis of Mann, Bradley and Hughes (1998, 1999), there is still substantial evidence that recent decades are among the warmest of the past 600 years.

The problem of paleoclimate reconstruction is a natural one for the use of Bayesian hierarchical models (BHMs). As in most BHMs, there is an unobserved "process" which is the true object of interest — in this case, the true series of temperatures. There are also various sources of "data" which are dependent on the "process" with different levels of accuracy — observational data, tree rings, boreholes, ice cores etc. The problem of paleoclimate reconstruction may be characterized as how to combine the different data series to obtain the best reconstruction of the unobserved process, with suitable measures of uncertainty. The BHM technique is especially valuable for answering non-standard uncertainty questions, for instance, "what is the probability that the 1990s were the warmest decade of the [1000–2000] millennium?"

In an earlier paper, Li, Nychka and Ammann (henceforth LNA, 2007) used an ensemble reconstruction, obtained via a combination of linear regression, bootstrapping and cross-validation, to reconstruct Northern Hemisphere average temperatures back to 1000, using 14 proxy series first discussed in Mann, Bradley and Hughes (MBH, 1999). Their results showed that there is indeed

*Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, N.C. and Department of Statistics and Operations Research, University of North Carolina, Chapel Hill. Email: rls@email.unc.edu

a high probability that the 1990s were the warmest decade of the millennium. The BHM technique has since been taken up by other authors, such as Tingley and Huybers (2010a, 2010b), Brynjarsdóttir and Berliner (2010), and promises to be the method of choice for future statistical analyses of paleoclimatic data.

In the paper under discussion, LNA (2010) have shown that it is also possible to answer "design"-type questions using BHMs. I believe that this is the logical next step in the scientific application of BHMs to paleoclimatology, and the methodology they have presented will play an important role in the selection of proxies for future paleoclimatological studies. I commend their contribution.

Although I fully support the further development of the BHM approach, it seems to me there is still some merit in looking for simpler statistical approaches, using methods that are routinely taught in first-year graduate courses in statistics, and that can (through the ready availability of the R programming language) be easily adopted by paleoclimatologists without extensive statistical training. Indeed, much of the debate over the "hockey stick curve" has focussed on the correct use of elementary statistical methods, in particular, the method of principal components (PCs). For the remainder of this note, I aim to show how routine application of PCs, regression and time series analysis can be used to resolve some issues that have caused much contention in the literature.

Brief summary of the controversy

The hockey stick curve, in the form that is currently debated, was first constructed in two papers of MBH (1998, 1999). After compiling a large number of proxies, they computed what they claimed was the first PC of the proxy series, with uncertainty bands for the reconstruction of each year's temperature mean.

Their analysis was criticized by a number of authors, but most sharply by McIntyre and McKitrick (2003, 2005a, 2005b) (henceforth, M&M). In particular, M&M (2005a) showed that the version of PC analysis used by MBH was more or less the following:

1. A data source was compiled that, for this discussion, is taken as 70 tree-ring series for 1400–1980.

2. Each series was rescaled so that the mean was 0 and the variance 1 over 1902–1980 — the "calibration period" for which real observational data was used.

3. For each series, a linear trend was fitted to 1902–1980, and the series rescaled again by dividing by the standard deviation of the residuals.

4. Based on the rescaled data matrix X, the singular value decomposition $X = UDV^T$ was formed as in conventional PC analysis (however, the X matrix has not been centered over all the series, only over 1902–1980).

5. Based on the SVD, the first PC (PC1) was calculated.

6. Finally, PC1 was rescaled so that the mean was 0 and the variance 1 over 1902–1980. This was the displayed result.

For the current discussion, I have reproduced these results using data and R code provided by Doug Nychka and Caspar Ammann (http://www.image.ucar.edu/~nychka/Temp/TreePC/). The basic data set consists of reconstructed temperatures from 70 trees for 1400–1980, in the North American International Tree Ring Data Base (ITRDB). In Figure 1, the 70 tree ring series have been plotted, after smoothing by applying a moving average filter with weights $w_i = (13 - |i|)/169$, $i = -12, -11, ..., 12$, i.e. a triangular window over the total span of 25 years with weights summing to 1. This is intended as a smooth representation of the underlying trend and will be used for trend comparisons in the subsequent analysis. As can be seen, there is little visual evidence of an overall upward or downward trend. Figure 2 shows the MBH reconstruction of the individual-year temperatures, together with a smoothed trend computed by the same triangular window. This shows the characteristic hockey stick shape, including a steady rise in temperatures since 1850.

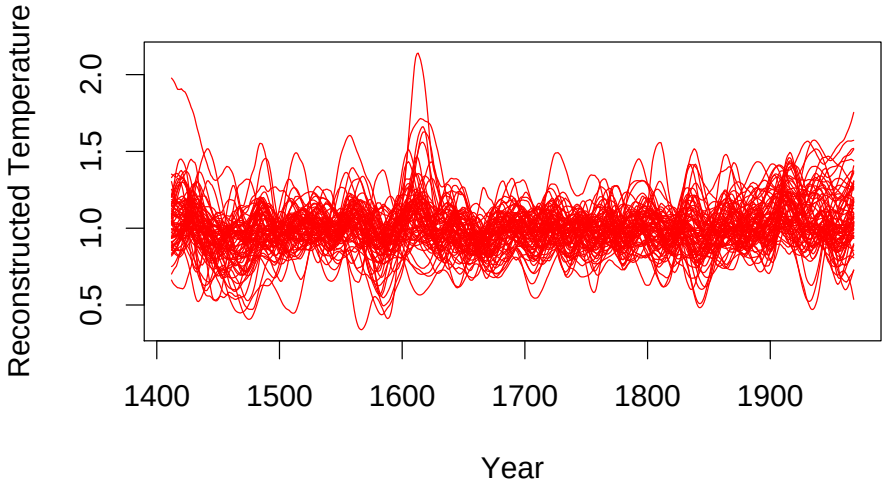

Figure 1: Reconstructed temperatures from 70 tree rings (1400-1980) in the North American ITRDB dataset. Each series has been smoothed using the 25-year triangular window described in the text.

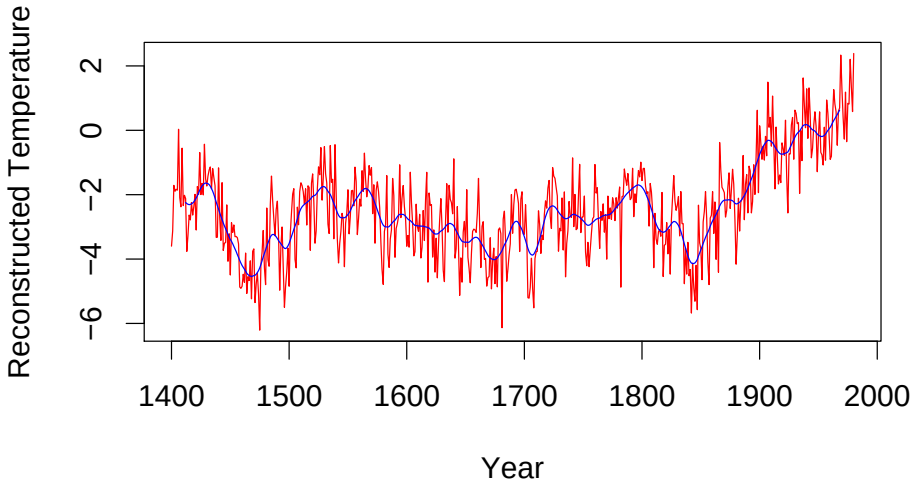

Figure 2: The original temperature reconstruction of MBH (1999), replotted for the present discussion, together with a smoothed trend computed using the same triangular window.

Steps 2 and 3 of the above algorithm are non-standard, since conventional PC analysis centers the data over the whole time period, not just part of it. In additional, conventional PC analysis sometimes standardizes the individual series so that the standard deviation is 1 — the standardized and unstandardized analyses are also commonly known as correlation-based and covariance-based PC analysis. M&M (2005a) argued that the version of PC analysis used by MBH can induce a spurious "hockey stick" shape, a point they illustrated by using simulations of red-noise time series with no trend to represent the tree-ring series. In many cases, the simulated series combined with the MBH algorithm resulted in a spurious hockey stick-like curve for the trend. Instead, M&M argued for a conventional PC analysis with centering over the whole time series. As illustration, Figure 3 shows PC1 from a correlation-based PC analysis, with the smooth trend using the same moving average smoother as previously. The output series has been rescaled to have mean 0 and variance 1 over the 1902–1980 period, to allow direct comparison with MBH. As noted by M&M, this series does not support the notion that recent decades were substantially warmer than temperatures in earlier centuries.

In a report to the House of Representatives of the U.S. Congress, Wegman, Scott and Said (2006) confirmed the M&M results and gave much additional discussion, generally supporting the conclusion of M&M that there was no hockey stick shape. At the same time, the National Research Council commissioned a report, published as North *et al.* (2006). They acknowledged the high level of uncertainty in reconstructions of medieval temperatures, but concluded that the balance of evidence, from many different studies, still supported the overall hockey stick shape of the

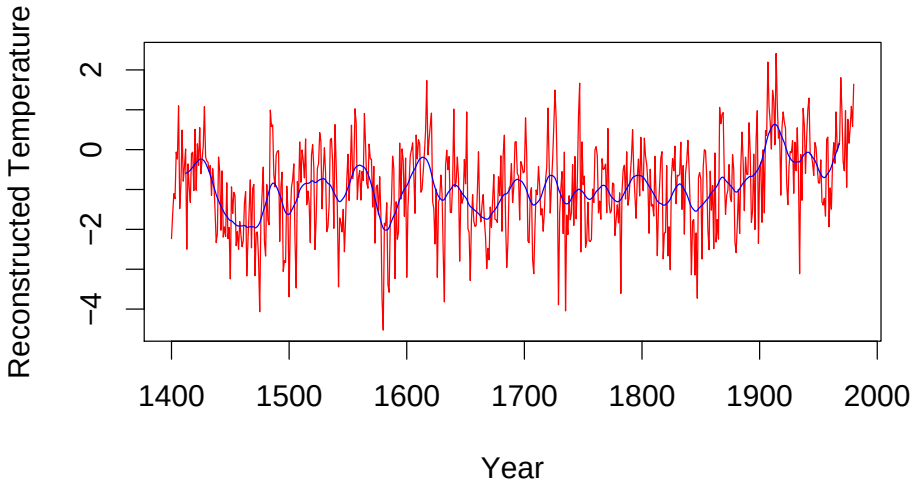

Figure 3: The first principal component, computed from the tree ring dataset using a conventional (correlation-based) PC decomposition, together with the smoothed trend.
temperature curve.

One criticism of the M&M analysis is their focus on PC1 as the main indicator of a signal in the series. M&M (2005b, pages 75–76) acknowledged this point, stating that a hockey stick shape reappears in the second PC of a standardized analysis (see Figure 4, where the 25-year smoother is also included), and the fourth PC of an unstandardized analysis, but argued against this interpretation, saying (with respect to the unstandardized analysis) that MBH's "conclusion about the uniqueness of the late 20th century climate hinges on the inclusion of a low-order PC series that only accounts for 8 percent of the variance of one proxy roster". In contrast, Ammann and Wahl (2007), Wahl and Ammann (2007) argued for the "robustness" of the MBH results, in particular the "convergence" of the reconstructed curves as the number of PCs in the reconstruction increased, and that this convergence holds regardless of whether the PCs are constructed using standardized series, unstandardized series, or the MBH technique. However, they did not discuss the standard error of the reconstructed curve, nor provide a formal criterion for selecting the number of PCs to include in the reconstruction.

Principal Components Regression

Suppose we have observed temperatures y_t for $t = 1902, ..., 1980$, and proxy series $\{x_{jt},\ j = 1, ..., q\}$ (where, here, $q = 70$) for $t = 1400, ..., 1980$. A natural way to think about the paleoclimate

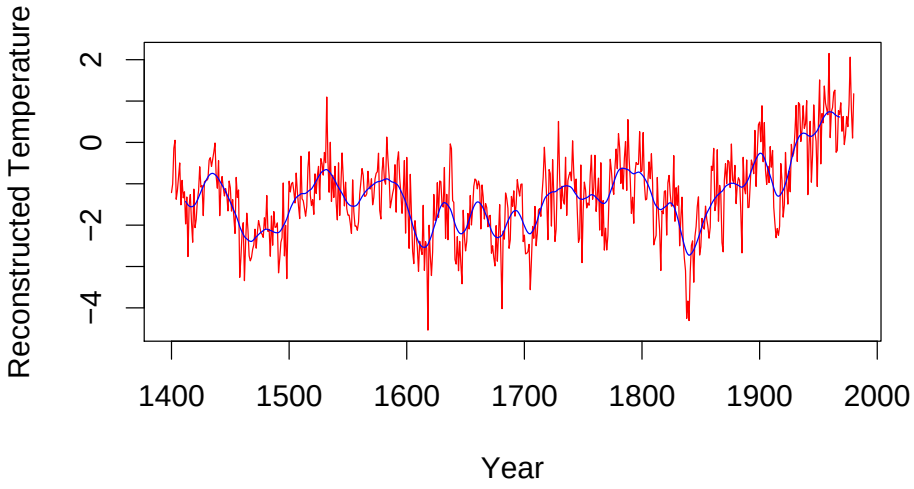

Figure 4: The second principal component, computed as in Figure 3.

reconstruction problem is first to fit the regression

$$y_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j x_{jt} + \epsilon_t \tag{1}$$

for 1902–1980, and then, having estimated coefficients $\hat{\alpha}_j,\ j = 0, ..., q$, to apply the fitted regression curve, $\hat{y}_t = \hat{\alpha}_0 + \sum_{j=1}^{q} \hat{\alpha}_j x_{jt}$, to reconstruct the temperature curve prior to 1902.

Direct application of (1), however, is open to the objection that the number of regressors (70) is very close to the number of data points used to fit the regression (79), creating potentially serious overfitting problems. A standard method for dealing with this problem is first to transform the covariates $\{x_{jt},\ j = 1, ..., 70\}$ to PCs $\{u_{kt},\ k = 1, ..., 70\}$, ordered by decreasing variance. Then, we fit the observed temperatures to a subset of the PCs,

$$y_t = \beta_0 + \sum_{k=1}^{K} \beta_k u_{kt} + \epsilon_t \tag{2}$$

where K is to be determined. The argument for this is that with moderate K, chosen to capture most of the variability in the covariates, the right hand side of (2) contains almost as much information as the right hand side of (1), but with far fewer regression coefficients to be estimated. For the moment, I assume a conventional ordinary least squares (OLS) regression analysis in which the ϵ_t are uncorrelated with mean 0 and common unknown variance σ^2.

This method was applied with the observed series y_t taken as the "HADCRUT3" global temperature mean anomalies of the Climate Research Unit of the University of East Anglia (http://www.cru.uea.ac.uk/cru/data/temperature). Based on the PC regression, I compute the

predicted temperature series $\hat{y}_t = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k u_{kt}$, and also its smoothed version $\tilde{y}_t = \sum_{i=-12}^{12} w_i \hat{y}_{t-i}$. Noting that we can also write $\tilde{y}_t = \sum_{k=0}^{K} \hat{\beta}_k \tilde{u}_{kt}$ where $\tilde{u}_{kt}$ is 1 for $k = 0$, $\sum_{i=-12}^{12} w_i u_{k,t-i}$ for $k = 1, ..., K$, the prediction intervals for $\tilde{y}_t$ can be computed by the standard formula given in regression textbooks. Therefore, we can display the smoothed reconstructed curve $\tilde{y}_t$, together with $100(1-\alpha)\%$ prediction bounds for any α, pointwise for each t. In the following examples, I have taken $\alpha = 0.1$ to give 90% prediction intervals.

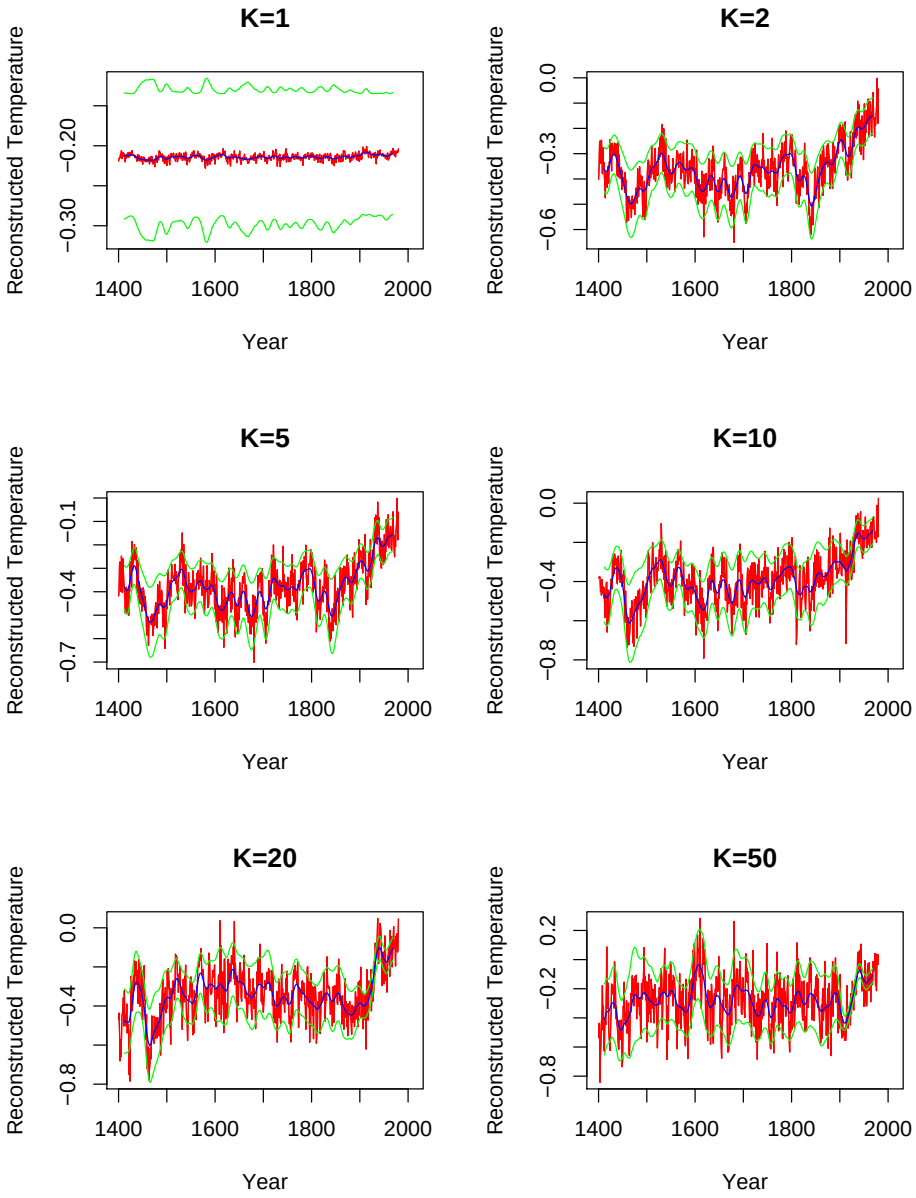

Figure 5: Six reconstructions of historical temperature anomalies, together with their smoothed trends and pointwise 90% prediction intervals on the trends.

Figure 5 shows the resulting reconstruction for $K = 1, 2, 5, 10, 20, 50$. The curve for $K = 1$ is flat, suggesting that the first PC has very little predictive power. However, each of the remaining curves has a noticeable hockey stick shape. For $K = 50$, and to a lesser extent $K = 20$, the prediction intervals are extremely narrow in the portion of the series with observational data (1902–1980), a sure sign of overfitting. For $K = 2, 5, 10$, however, the curves look very similar and there is no clear-cut choice among them.

A more systematic comparison may be made by computing various measures of fit. In comparing regression and/or time series models, three common criteria for selecting the best model are the Akaike Information Criterion (AIC; Akaike 1973), the Bayesian information criterion (BIC; Akaike 1978) and the bias-corrected Akaike Information Criterion (AICC, Hurvich and Tsai, 1989). For the PC regression model with K up to 10, but still for the moment without making any allowance for time series autocorrelations, these criteria are listed in Table 1.

K	AIC	BIC	AICC
1	–40.4	–33.3	–40.1
2	–58.2	–48.7	–57.7
3	–58.7	–46.9	–57.9
4	–57.3	–43.0	–56.1
5	–55.4	–38.8	–53.8
6	–57.3	–38.4	–55.3
7	–58.1	–36.8	–55.5
8	–63.8	–40.1	–60.5
9	–66.4	–40.4	–62.5
10	–66.1	–37.7	–61.4

Table 1: Table of AIC, BIC, AICC values for OLS regression without allowing for autocorrelation.

It can be seen that as K increases, each of AIC, BIC, AICC drops sharply (indicating improved fit) at $K = 2$ and again at $K = 8$. The minimum is at $K = 9$ for AIC and AICC and $K = 2$ for BIC: as typically happens, BIC chooses a more parsimonious model.

The omission of autocorrelation from the foregoing analyses is potentially a serious matter, since the width of the prediction interval could be substantially larger if autocorrelation is included. Therefore, I now extend the previous analysis to include a time series component.

The logical extension to the assumption that the errors ϵ_t in (1) or (2) are independent is that they form an ARMA(p, q) time series for suitable p and q (see, e.g., Brockwell and Davis (2003) for extensive discussion of ARMA modeling). The most powerful way to select a model is to treat all of $K,\ p$ and q as undetermined model parameters, to perform a generalized least squares (GLS) analysis, and to select $K,\ p$ and q to minimize one of AIC, BIC or AICC. Unfortunately, this procedure quickly produces unwieldy models and does not lead to a clear-cut conclusion. For example, fitting models by minimizing AIC up to $K = 9,\ p = 10,\ q = 5$ produced the best model at $K = 8,\ p = 2,\ q = 5$. These results were obtained by maximum likelihood fitting using the `arima` command in R (R Core Development Team, 2010), ignoring models for which the MLE algorithm did not converge or for which the resulting model fit violated the stationarity condition for the autoregressive part of the model. However, the final model is hard to interpret with so many parameters, and it seems probable that still higher-order models would be obtained if larger values of $K,\ p,\ q$ were permitted. Similar results were obtained using BIC and AICC.

As an alternative to full GLS time series regression, therefore, I used the same OLS fits for the regression components produced earlier, but selected the optimal ARMA(p, q) model fitted to the residuals, and then recalculated the width of the prediction intervals to take account of the autocorrelation. This produced more easily interpretable results. For example, with $K = 2$, the optimal ARMA model had $p = 1,\ q = 2$ when selecting by AIC and $p = 1,\ q = 0$ when using BIC or AICC. With $K = 9$, all three selection criteria resulted in AR(1) as the optimal time series model — incidentally relevant to LNA (2010), where they used AR(2) as the time series model for residuals, though Dr. Li remarked in her oral presentation that the AR(1) model appears equally suitable in practice.

For the three models just derived, the reconstructed curves, with prediction interval bounds for the 25-year moving average, are shown in Figures 6 through 8. Also shown on the plot are the actual global mean temperatures for 1902–1980, and a 25-year triangular moving average filter applied to those. The three figures look very similar to each other, though the width of the prediction intervals is about twice that in Figure 5. All three confirm that the reconstructed smoothed temperature for prior centuries was well below its value in recent decades.

Discussion and Summary

This analysis has used principal components regression combined with time series analysis of the residuals to reconstruct the global mean temperature series back to 1400. I smoothed the reconstructed series using a 25-year triangular moving average, and calculated 90% prediction

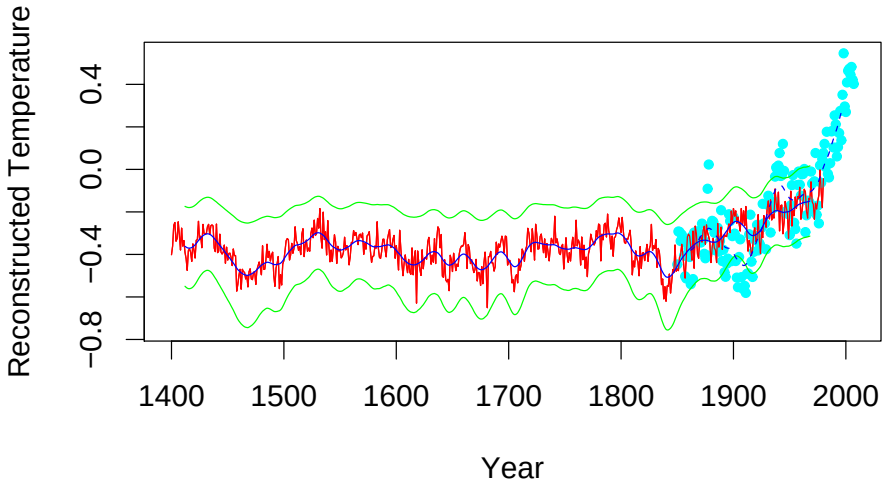

Figure 6: Reconstructed series with time series corrections: $K = 2$ PCs with AR(1) residuals. The solid curves represent the smoothed reconstructed series with pointwise 90% prediction intervals. The dashed curve at the right-hand end is the same smoother applied to the observational data points.

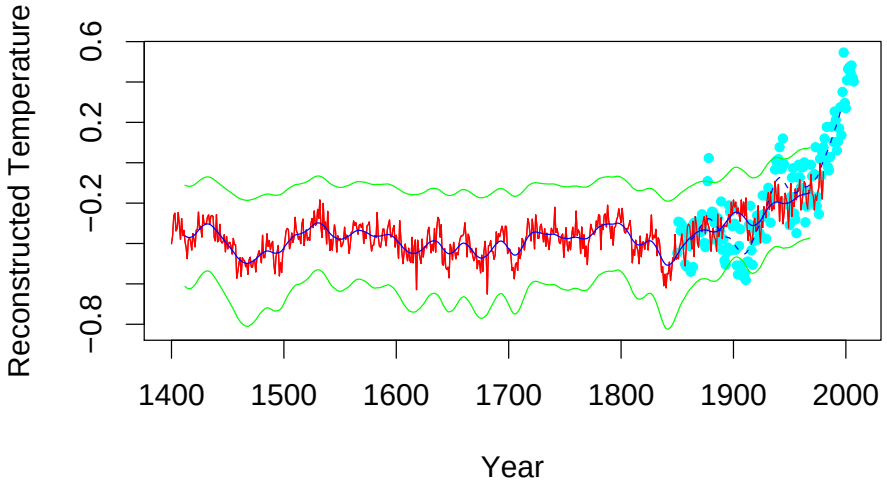

Figure 7: Reconstructed series with time series corrections: $K = 2$ PCs with ARMA(1,2) residuals

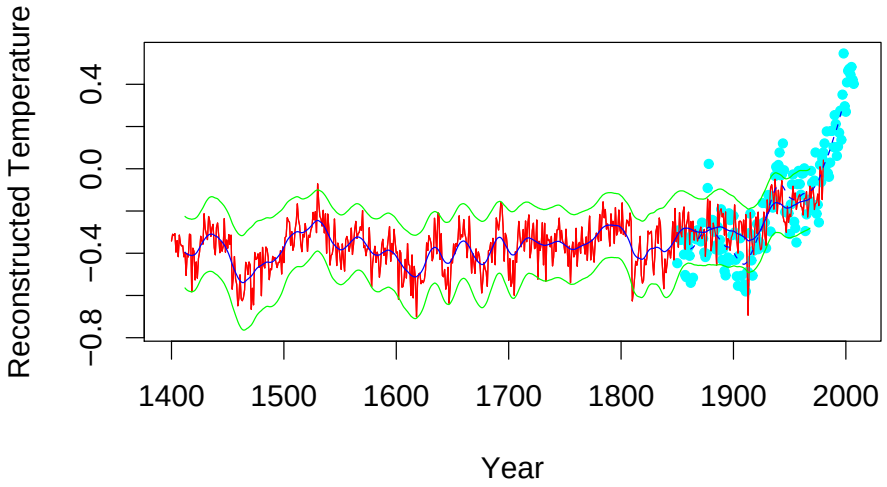

Figure 8: Reconstructed series with time series corrections: $K = 9$ PCs with AR(1) residuals intervals on the smoothed reconstruction as a measure of uncertainty. Three standard statistical model selection criteria (AIC, BIC and AICC) were used to select the model orders K (number of PCs), p and q (for the autoregressive and moving average components of the time series model fitted to the residuals). Although these criteria do not lead to clear-cut selection of the best model, the final reconstructions do not appear to depend too sensitively on the model selected. Taking into account the general desire in applied statistics for a parsimonious model, the model with $K = 2$ PCs and AR(1) residuals appears adequate.

The idea of PC regression as a technique in paleoclimate reconstruction is not new — for example, it was discussed in Chapter 9 of North *et al.* (2006) — but it does not appear to have been systematically developed.

The results support an overall conclusion that the temperatures in recent decades have been higher than at any previous time since 1400. On the other hand, none of the recent reconstructions shows as sharp a hockey stick shape as the widely reproduced Figure 3(a) of MBH 1999, so in that respect, critics of the hockey stick are also partially vindicated by these results.

I have confined this discussion to statistical aspects of the reconstruction, not touching on the question of selecting trees for the proxy series (extensively discussed by M&M, Wegman *et al.*, and Ammann/Wahl) nor the apparent recent "divergence" of the relationship between tree ring reconstructions and measured temperatures (see, e.g., North *et al.*, pp. 48–52). I regard these as part of the wider scientific debate about dendroclimatology but not strictly part of the statistical discussion, though it would be possible to apply the same methods as have been given here to

examine the sensitivity of the analysis to different constructions of the proxy series or to different specifications of the starting and ending points of the analysis.

Acknowledgments

SAMSI is supported by the National Science Foundation, grant DMS 0635449. I am grateful to Doug Nychka and Caspar Ammann for making their data and programs available.

References

Akaike, H. (1973), Information theory and an extention of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, ed. B.N. Petrov and F. Csaki, pp. 267–281. Budapest: Akademia Kiado.

Akaike, H. (1978), A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.* **30**, 9–14.

Ammann, C.M. and Wahl, E.R. (2007), The importance of the geophysical context in statistical evaluations of climate reconstruction procedures. *Climatic Change* **85**, 71–88

Brockwell, P.J. and Davis, R.A. (2003), *Introduction to Time Series and Forecasting*. Second edition: Springer.

Brynjarsdóttir, J. and Berliner, L.M. (2010), Beysian hierarchical modeling for paleoclimate reconstruction from geothermal data. Preprint, Ohio State University.

Hurvich, C.M. and Tsai, C.-L. (1989), Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Li, B., Nychka, D.W. and Ammann, C.M. (2007), The 'hockey stick' and the 1990s: a statistical perspective on reconstructing hemispheric temperatures. *Tellus* **59A**, 591–598.

Li, B., Nychka, D.W. and Ammann, C.M. (2010), The value of multi-proxy reconstruction of past climate. *Journal of the American Statistical Association*, to appear.

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. *Geophysical Research Letters* **26**, 759–762.

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

McIntyre, S. and McKitrick, R. (2005a), Hockey sticks, principal components, and spurious significance. *Geophysical Research Letters* **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.

North, G.R., Biondi, F., Bloomfield, P., Christy, J.R., Cuffey, K.M., Dickinson, R.E., Druffel, E.R.M., Nychka, D., Otto-Bliesner, B., Roberts, N., Turekian, K.K. and Wallace, J.M. (2006), *Surface Temperature Reconstructions for the Last 2,000 Years.* National Research Council ISBN: 0-309-66144-7, available online from http://www.nap.edu/catalog/11676.html

R Core Development Team (2010), *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org.

Tingley, M.P. and Huybers, P. (2010a), A Bayesian algorithm for reconstructing climate anomalies in space and time. Part I: Development and applications to paleoclimate reconstruction problems. *Journal of Climate* **23**, 2759–2781.

Tingley, M.P. and Huybers, P. (2010b), A Bayesian algorithm for reconstructing climate anomalies in space and time. Part II: Comparison with the regularized expectation-maximization algorithm. *Journal of Climate* **23**, 2782–2800.

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. *Climatic Change* **85**, 33–69.

Wegman, E.J., Scott, D.W. and Said, Y.H. (2006), *Ad Hoc Committee Report on the 'Hockey Stick' Global Climate Reconstruction.* Report presented to the Committee on Energy and Commerce and the Subcommittee on Oversight and Investigations, U.S. House of Representatives.