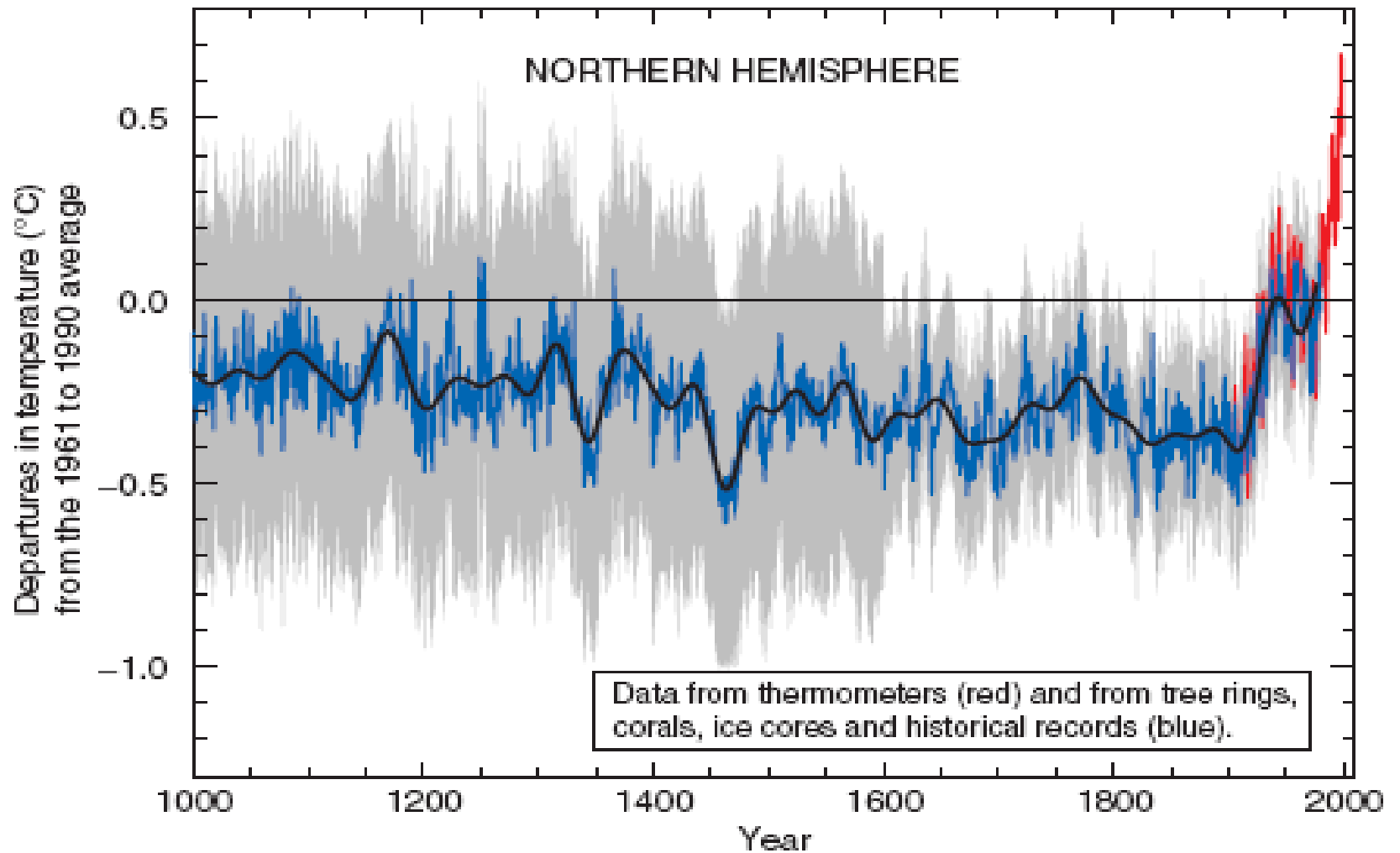


# **THE “HOCKEY STICK” CONTROVERSY AND PCA**

In 1998/1999, Mann, Bradley and Hughes (MBH) wrote two papers reconstructing temperatures over the last 1000 years from proxy data (tree rings, corals, ice cores etc.). This paper led to the “hockey stick” graph which featured prominently in the 2001 IPCC Report.

Their methodology was based on Principal Components Analysis (PCA).



From IPCC (2001), *Summary for Policymakers*  
Based on Mann-Bradley-Hughes (1998, 1999)



In 2006, statistician Edward Wegman wrote a report for the House Energy Committee, then chaired by Rep. Joe Barton (R - Texas).

Based on earlier work by McIntyre and McKittrick (2003, 2005), Wegman argued that the hockey stick shape was an artifact of the statistical analysis of MBH. The key point of the argument was the failure of MBH to center the data correctly prior to calculating PCs.

This attracted wide attention. The *Wall Street Journal* editorialized “[Wegman’s] conclusions make ‘consensus’ look more like group-think”. Wegman was profiled in a book called *The Deniers*. And a session that Wegman and I co-organized at the 2006 JSM attracted over 300 participants.



## EMPLOYMENT

- Assistant Professor, University of North Carolina, Department of Statistics, 1968-1973
- Associate Professor, University of North Carolina, Department of Statistics, 1973-1978
- Visiting Professor, University of Manchester (England), Department of Mathematics (on leave from the University of North Carolina), 1976-1977
- Director, Statistics and Probability Program, Office of Naval Research, 1978-1983
- Head, Mathematical Sciences Division, Office of Naval Research, 1982-1986
- Professor and Director, Center for Computational Statistics, George Mason University, 1986- 2006
- Founding Chairman, Department of Applied and Engineering Statistics, George Mason University, 1992-1999
- Chairman, Data Sciences Program, School of Computational Sciences, George Mason University, 2004-2006

## Outline of MBH Method

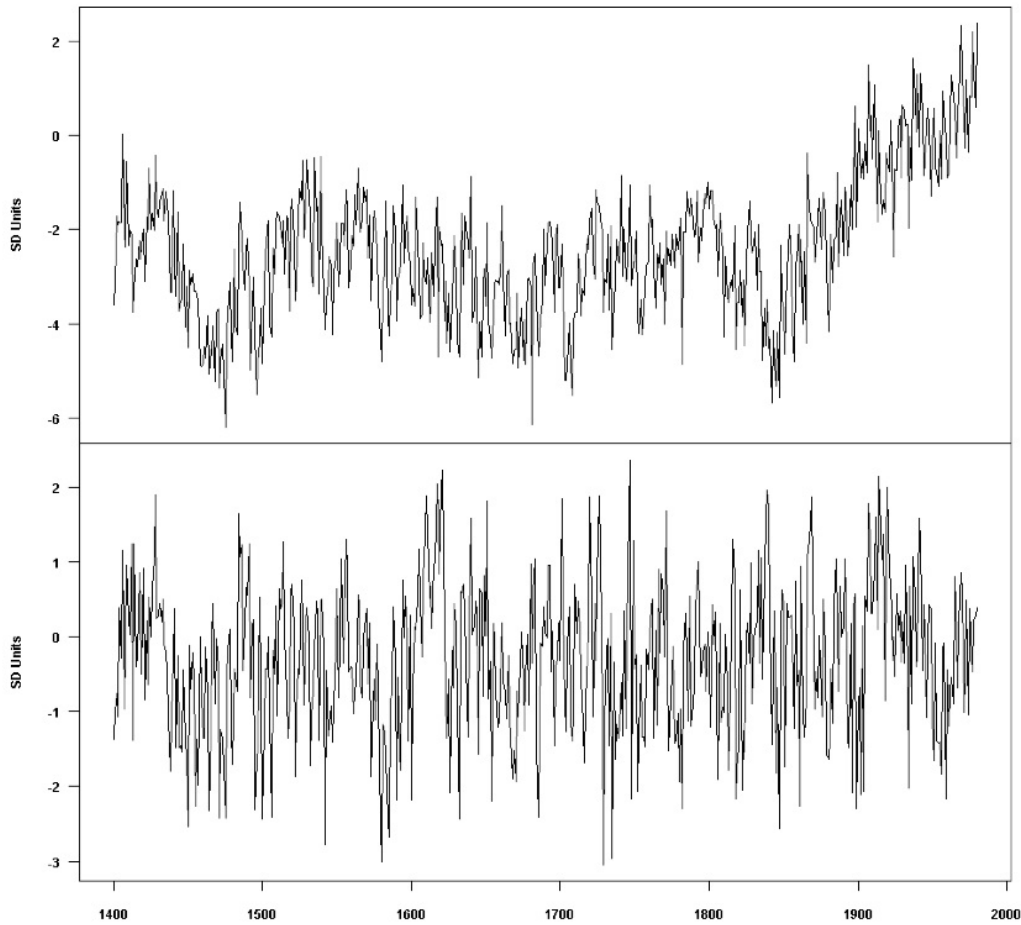
1. Data source (for this presentation): 70 proxy temperature series based on tree rings, for 1400–1980.
2. Rescale each series so that mean=0 and variance=1 over 1902–1980
3. Fit linear trend 1902–1980 — rescale again by dividing by SD of residuals
4. Form SVD  $X = UDV^T$  — *however, the  $X$  matrix hasn't been centered over all the series, only over 1902–1980*
5. Based on SVD, calculate first PC.
6. Rescale again so that mean=0 and variance=1 over 1902–1980. Plot the result.

## Main Point of the Criticism

McIntyre and McKittrick (results confirmed by Wegman) argued that failure to center the data over the whole period was a fatal defect of the method.

In essence, those proxies that already have a hockey-stick-like shape will have a mean over 1400–1980 that is far from the mean over 1902–1980. They will have a larger variance than they would have if correctly centered. Therefore, the PCA will give greater weight to these series than it should.

Instead, they argued for an ordinary centered PC (here I've used PCs based on the correlation matrix but it shouldn't make much difference if they are based on the covariance matrix)



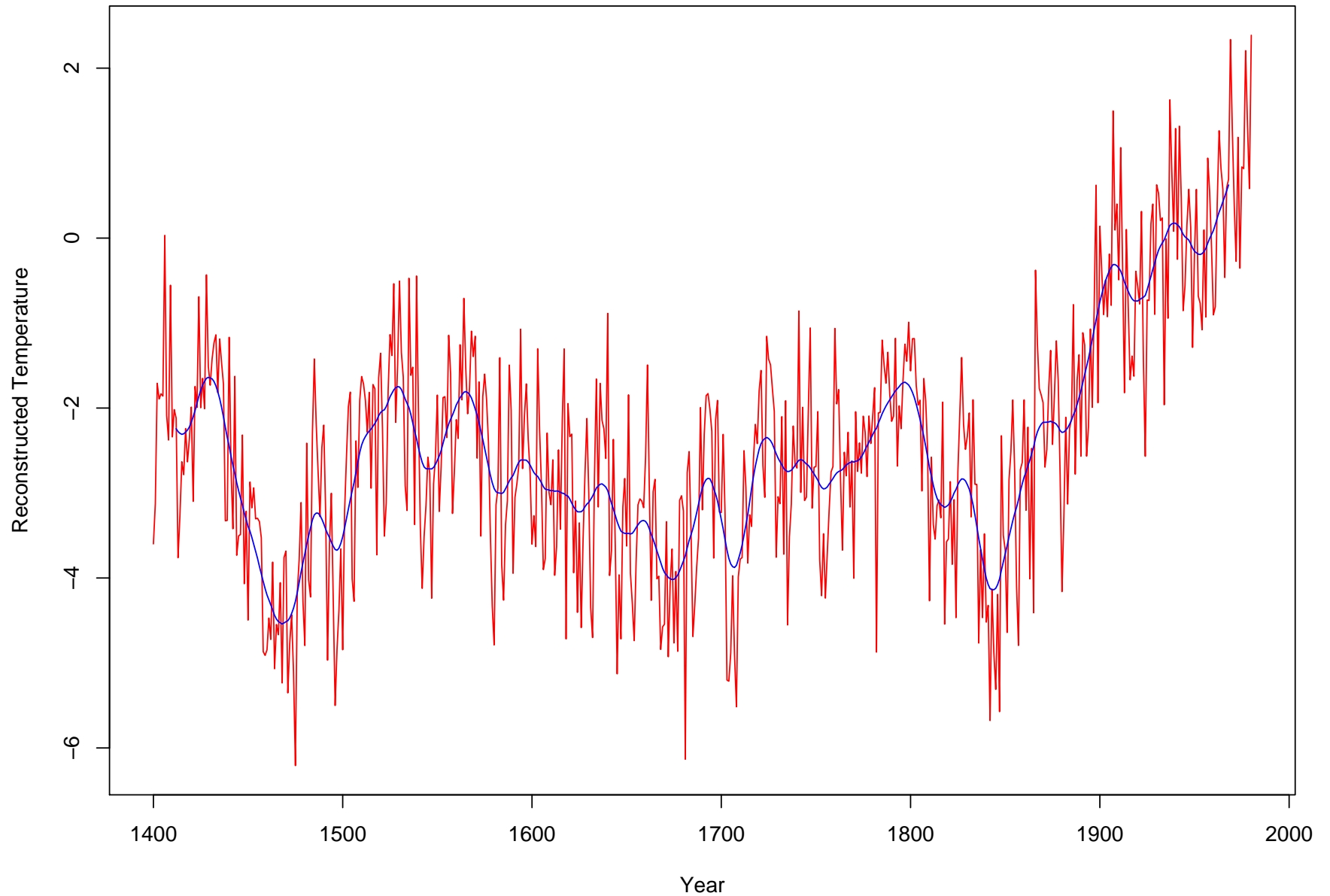
Top: MBH curve, reproduced by McIntyre-McKittrick (2005)  
Bottom: The first (centered) PC, calculated by M&M



The data have been reanalyzed by Bo Li, Doug Nychka and Caspar Ammann, at NCAR.

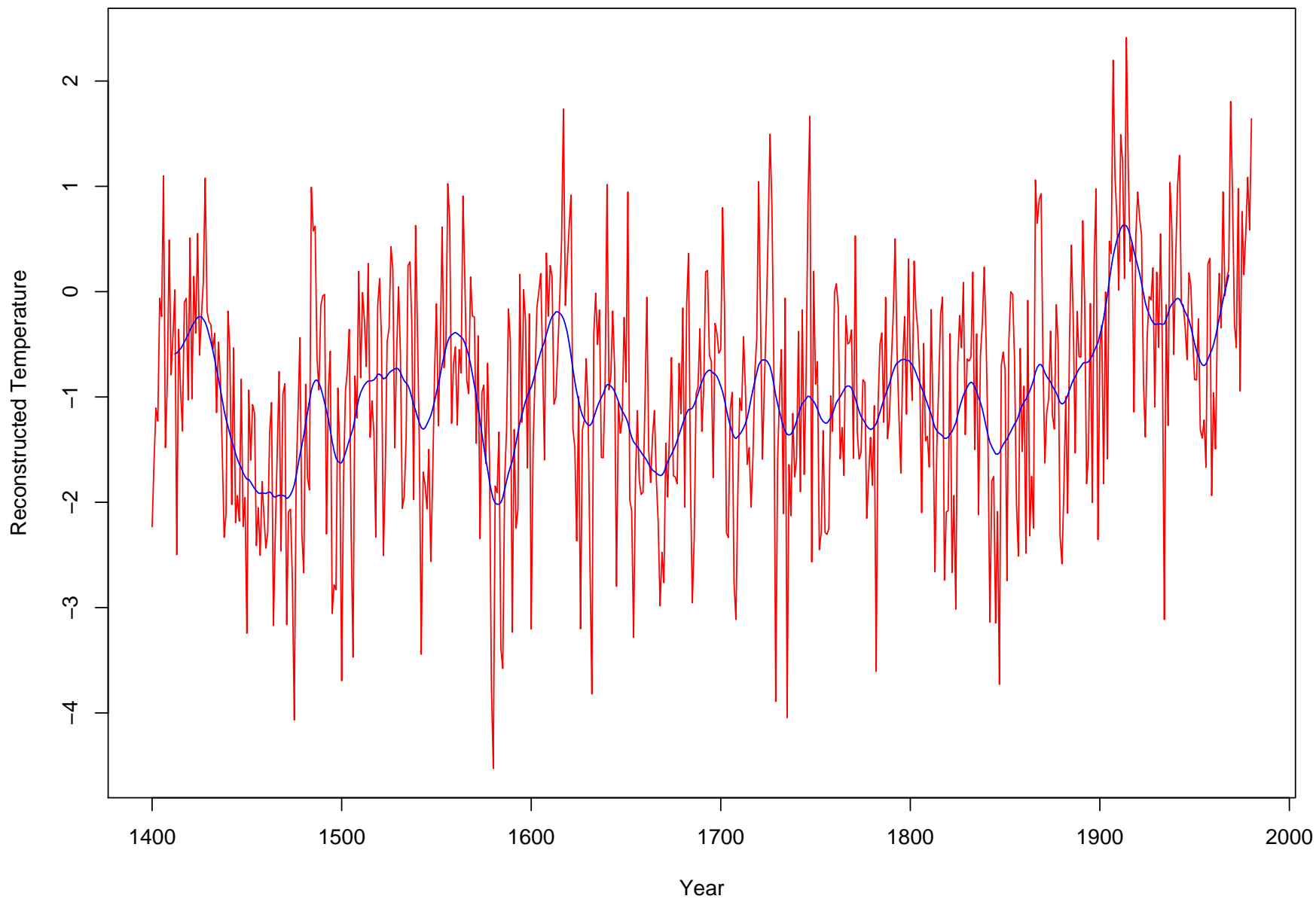
<http://www.image.ucar.edu/nychka/Temp/TreePC/>

## Original MBH Method



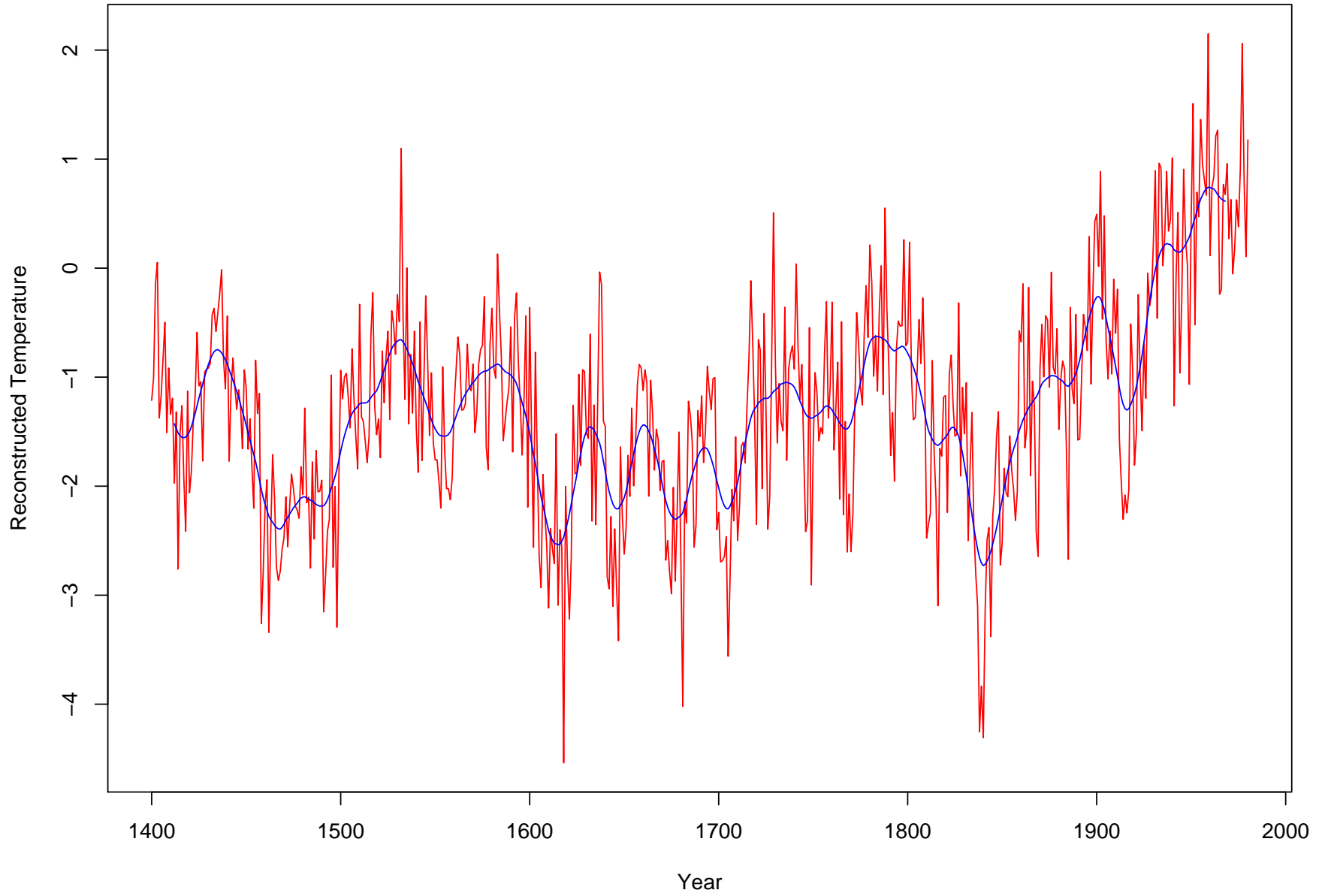
Original MBH plot, reconstructed for this presentation  
(blue curve: smoothed by tapered 25-year MA)

### First PC, Centered to 1902–1980



First PC recalculated, scaled and centered to 1902–1980

### Second PC, Centered to 1902–1980



Second PC recalculated, scaled and centered to 1902–1980

Here's an alternative approach. The idea is to use centered PCs, but not just the first PC.

## Proposed New Method

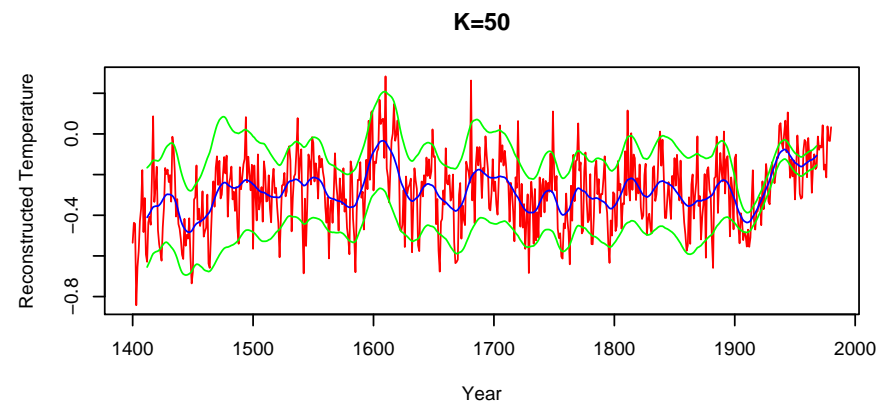
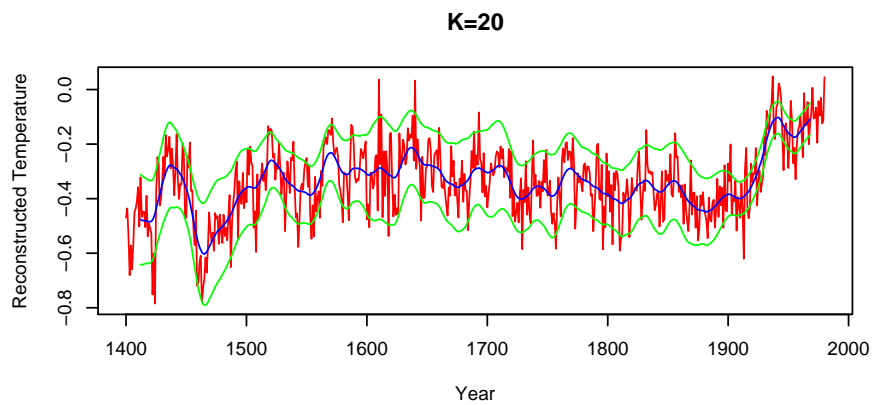
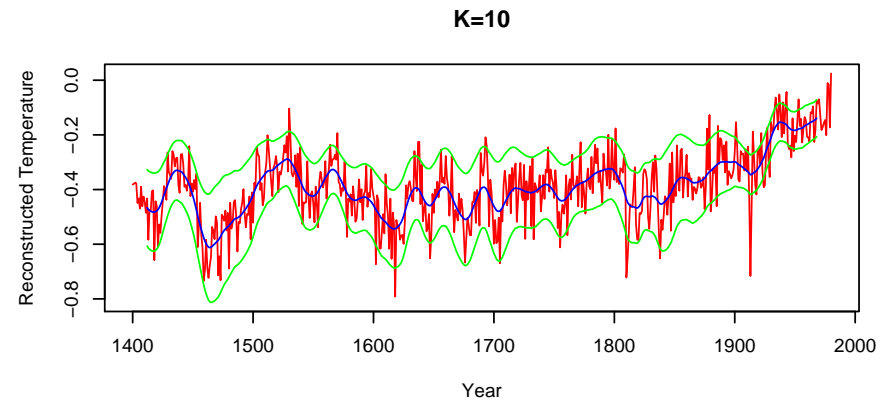
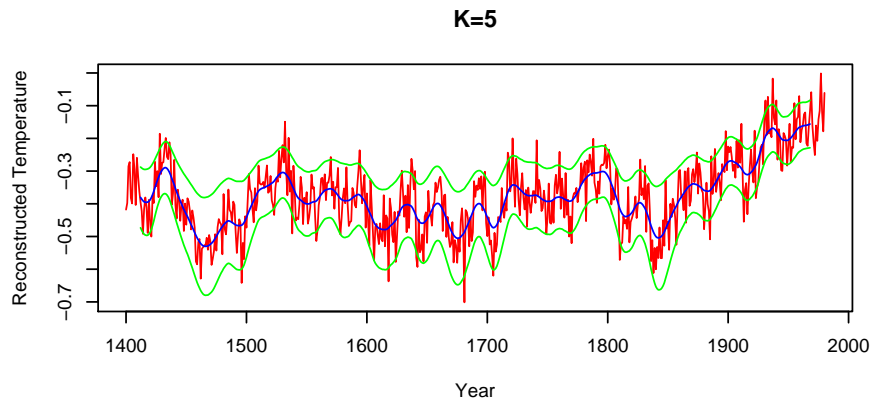
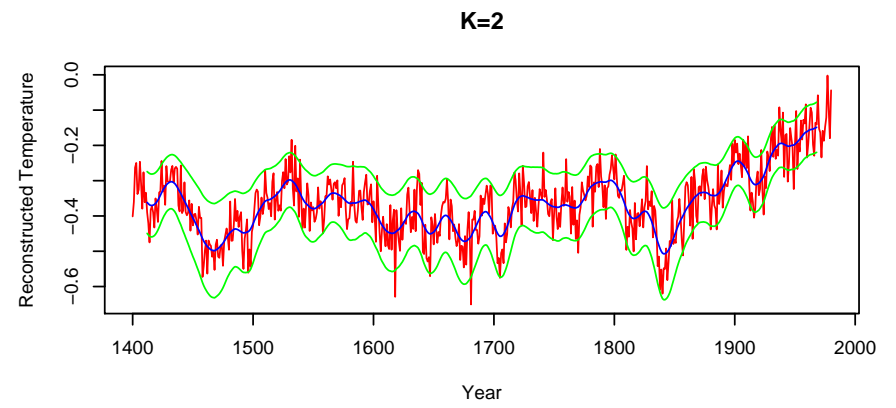
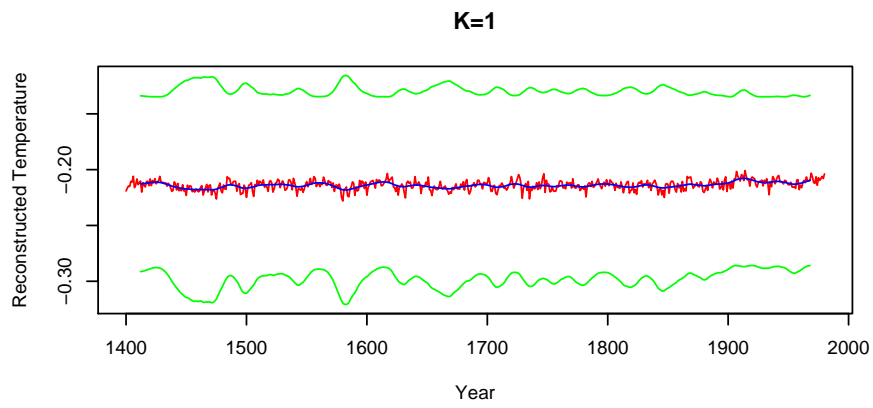
1. Calculate all the PCs by usual (correlation-based) method
2. For given  $K \geq 1$ , regress true global temperature anomaly  $y_t$  on first  $K$  PCs for 1902–1980:

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

3. Form reconstructed  $\hat{y}_t$  for 1400–1980:

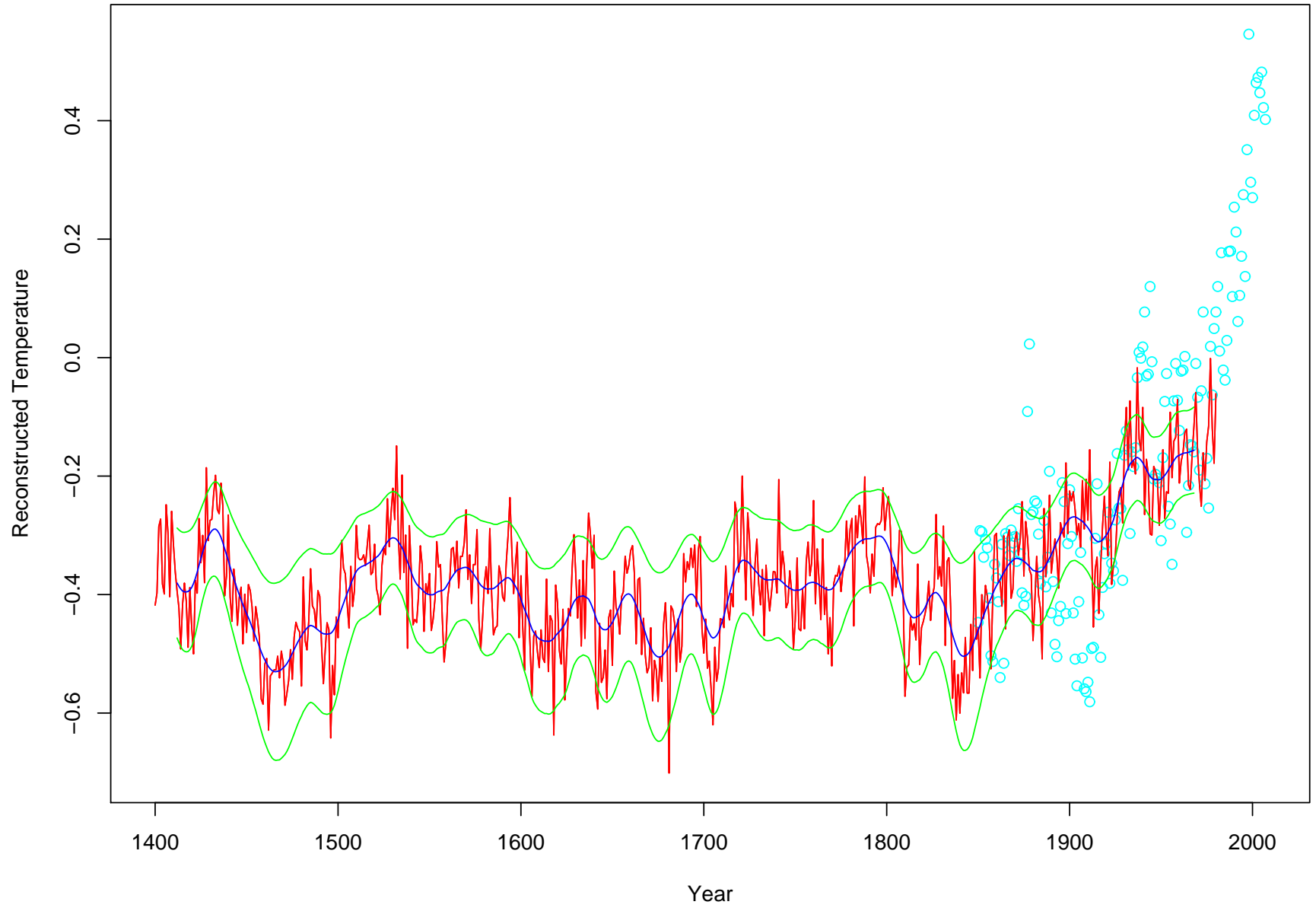
$$\hat{y}_t = \hat{\beta}_0 + \sum_{k=1}^K \hat{\beta}_k u_{kt}$$

4. I've also computed weighted 25-year tapered MAs from  $\hat{y}_t$ , and 90% prediction intervals by the standard method used to compute prediction intervals from linear regression
5. Repeated for  $K = 1, 2, 5, 10, 20, 70$ .



Six reconstructions of historical temperature anomalies with pointwise 90% prediction intervals

K=5



Reconstruction K=5 with directly measured temperatures

## Conclusions and Comments

1. Taking just the first PC ( $K = 1$ ) seems useless — reconstruction has very small variance (not apparent from M&M-Wegman because they rescaled PCs to variance 1)
2. However, even  $K = 2$  is much better and shows a clear “hockey stick” shape
3. Results similar for  $K = 2, 5, 10$ . For  $K = 20, 50$  there is evidence that the regression is overfitting in 1902–1980
4. Analysis doesn't (yet) account for autocorrelation
5. For a more comprehensive analysis of a different dataset, see Li, Nychka and Ammann (*Tellus* **59A**, 591–598, 2007)