Presentation to the National Academy of Sciences Expert Panel, "Surface Temperature Reconstructions for the Past 1,000-2,000 Years."

Stephen McIntyre,
Toronto Ontario

Ross McKitrick, Ph.D.
Associate Professor
Department of Economics
University of Guelph

March 2, 2006
Washington DC

1 Introduction

Thank you for the invitation to speak to you today. We are flattered to be included among so many distinguished presenters and to have the opportunity to be heard by such an eminent panel. We have expressed our concern to the NAS that a couple of panel members have close associations to people who strongly oppose our views. We reiterate the need for the Panel to undertake a thorough and unbiased examination of the evidence we present, and we have prepared our submission on the assumption that this will be done.

We are not here to argue for or against the Medieval Warm Period, or to present an alternative interpretation of climate history, but to report on our efforts to evaluate the quality, robustness and statistical significance of multiproxy evidence behind the claim that 20th century climate change is unprecedented in a millennial context. In this regard we will primarily focus on the following questions put to this panel by Representative Boehlert:

> 2) (b) What are the principal scientific criticisms of their [Mann, Bradley and Hughes] work and how significant are they? (c) Has the information needed to replicate their work been available? (d) Have other scientists been able to replicate their work?

Our answers to these questions are as follows.

(b) With respect to Mann et al. [1998, 1999] (MBH98-99), our most important objections [see McIntyre and McKitrick, 2003, 2005a, 2005b, 2005c, 2005d and www.climateaudit.org] are:

- The study used "new" statistical methods that turned out to "mine" for hockey stick shaped series. These methods were misrepresented and/or inaccurately described in important particulars and their statistical properties were either unknown to the authors or unreported by them.
- The reconstruction failed an important verification test said to have used in the study. This failure was not reported and the statistical skill was misrepresented both in the original article and by the IPCC.
- Dominant weight was placed on proxies known to be inappropriate temperature proxies, along with, at best, misleading information about their impact and, at worst, actual withholding of adverse results;
- The method of confidence interval calculation leads to unrealistically narrow confidence intervals;

(c) No. Systematic obstruction was placed at every step of the way of replication attempts. The underlying data were exceedingly hard to identify and obtain. The methodology was not accurately described in the paper and the computational code was withheld until the intervention of a Congressional investigation.

(d) No. Some authors (Ammann and Wahl) claim to have replicated the MBH results. Contrary to their representations, they have not confirmed MBH claims of statistical skill and robustness or dealt with all relevant aspects of MBH. Their emulation of MBH is almost identical to ours. Differences between us pertain entirely to the characterization of the results, rather than to the calculations themselves. In fact, their code actually confirms our claims about MBH verification statistics.

We have also carefully studied the data and methods of the major multiproxy studies used in the reconstruction of surface temperatures for the past millennium [Mann et al., 1998; Mann et al, 1999; Jones et al 1998; Crowley and Lowery 2000; Esper et al, 2002; Mann and Jones, 2003; Jones and Mann, 2004; Moberg et al, 2005; Osborn and Briffa, 2006]. We will focus our discussion on the most prominent of these studies, MBH98-99, which was heavily relied upon by the IPCC, but we will also itemize issues regarding the other studies, that should be of concern to the Committee.

Our concerns with other studies frequently cited in support of MBH are related to the above. For every study, there are pointless obstacles to replication, causing long delays to any statistical researcher attempting to evaluate the results. The studies are neither independent in authorship nor in proxy selection. None of the studies describe objective protocols for proxy selection. Because they have very small populations, their results are highly sensitive to proxy choice. The repetitive use of proxies known

to be questionable as temperature proxies, but which happen to have a hockey stick shape (such as the bristlecone growth index), raises questions about potential bias in proxy selection. There is also evidence of considerable instability in well-known site chronologies depending on sample (e.g. the Polar Urals pre- and post- recent resampling), yielding remarkably divergent results even from the same site.

We can only briefly survey these questions and will leave list of major issues and questions for the Committee to consider.

With regard to Rep. Boehlert's question (3b):

> 3) How central is the work of Drs. Mann, Bradley and Hughes to the consensus on the temperature record?

MBH is the origin of the claims that 1998 was the "warmest" year and the 1990s the "warmest decade" of the millennium. It was relied upon both by the IPCC and then, subsequently by national governments, including Canada. It became a standard for every subsequent multiproxy study and is included in all representations of millennial climate. Its results and methods continue in use, directly affecting papers released as recently as last month.

2. Background

In 1990, the Intergovernmental Panel on Climate Change First Assessment Report included a graph showing the 20th century was unexceptional compared to the previous millennium (Fig 1).

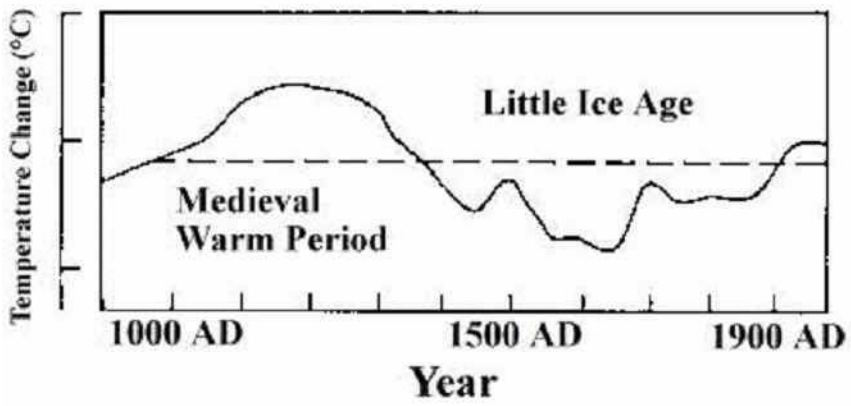

Figure 1: IPCC 1990.

In the early 1990s, a number of articles with a rotating list of core authors, like Jones, Briffa, Bradley and Hughes, began questioning the existence of the Medieval Warm Period and Little Ice Age on anything other than a local European scale. These studies included some of the first attempts at multiproxy analysis [Bradley and Jones, 1993; Hughes and Diaz, 1994] and presented new and subsequently very influential tree ring series [Briffa et al 1992 (Tornetrask); Briffa et al 1995 (Polar Urals)].

The second IPCC report in 1996 cited these studies and, while they did not present a millennial history, expressed caution about continued reliance on concepts such as the MWP and Little Ice Age. They also expressed caution about some aspects of tree ring site chronologies, including a concern over direct CO_2 fertilization, a phenomenon which had been proposed in 1984 by dendrochronologists (Lamarche, Fritts) in relation to the Sheep Mountain bristlecone site, and in 1993, for a much larger collection of 14 bristlecone sites [Graybill and Idso, 1993].

Mann et al. [1998] originated in the milieu after the IPCC Second Assessment Report. In the original article, Mann et al said they used a "new statistical approach", noting that "conventional" methods had proved "relatively ineffective". Bradley said recently that they had "originated new mathematical approaches that were crucial to identifying strong trends" [Goldscheider, 2005].

If a group of proxies is selected because they are sensitive to temperature, the simplest way to characterize their dominant pattern is to standardize the scale and calculate the mean. The simple mean of the Mann et al.[1998] data is in the top panel of Figure 2. One notes that the 20th century is unexceptional and, for what it is worth, that there is a downward trend over the 20th century. The final reconstruction, shown in the bottom panel, yields a remarkably different story, in which the historical values were low prior to the 20th century, and the data have a strong upward trend after 1900.

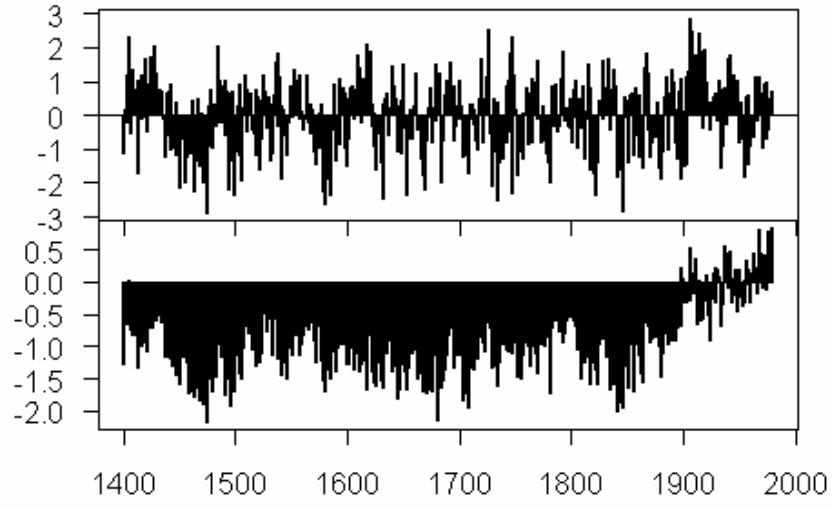

Figure 2: *Top – Average of 415 series in MBH98 "dataall" dataset archived in July 2004. Bottom – MBH98 reconstruction.*

In 1999, the results were extended 400 years back to 1000 [Mann et al., 1999]. This article, together with its national press release, were the first claims that 1998 was the "warmest year" and the 1990s "warmest decade" of the millennium – phrases that themselves became iconic.

The study immediately became influential, not just in the public but also in the specialist literature. It seemed to achieve an unprecedented level of technical and statistical sophistication. Claims (repeated or expanded in Mann et al, 2000) included the following:

- that the reconstruction had been verified in independent cross-validation tests and achieved remarkable statistical skill. The terms "skill" and "skilful" are used 20 times in MBH98 and the term "verification" 44 times. MBH98 explicitly stated that three different tests were used: the Reduction of Error (RE), correlation (r) and correlation-squared (r^2). We draw your attention to this particular fact as the actual values of two of these tests were not published.

- that, because of the multiproxy nature of the reconstruction, it was not dependent on any one class of proxies, thus alleviating potential concerns about problems with tree ring indicators. MBH98 and, Mann et al [2000] in stronger and very categorical terms, assured readers that their results were "robust" even to the total exclusion of dendroclimatic indicators (a claim recently relied upon, for example, in Bunn et al [2005].

- that confidence intervals could be estimated from their methodology and, based on these confidence intervals, probabilities could be assigned to the warmth of the 1990s on a millennial scale. No prior study had ever ventured to make such an estimate.

The IPCC described the work of Mann et al. as follows:

> Averaging the reconstructed temperature patterns over the far more data-rich Northern Hemisphere half of the global domain, they [MBH] estimated the Northern Hemisphere mean temperature back to AD 1400, a reconstruction which had significant skill in independent cross-validation tests. Self-consistent estimates were also made of the uncertainties…. Taking into account these substantial uncertainties, Mann et al. (1999) concluded that the 1990s were likely to have been the warmest decade, and 1998 the warmest year, of the past millennium for at least the Northern Hemisphere. (IPCC TAR, WG1, Page 133, 136)

These claims, and the MBH reconstruction itself, were highlighted in the Summary for Policymakers. The MBH reconstruction appeared six times in the IPCC TAR (Figure 3).

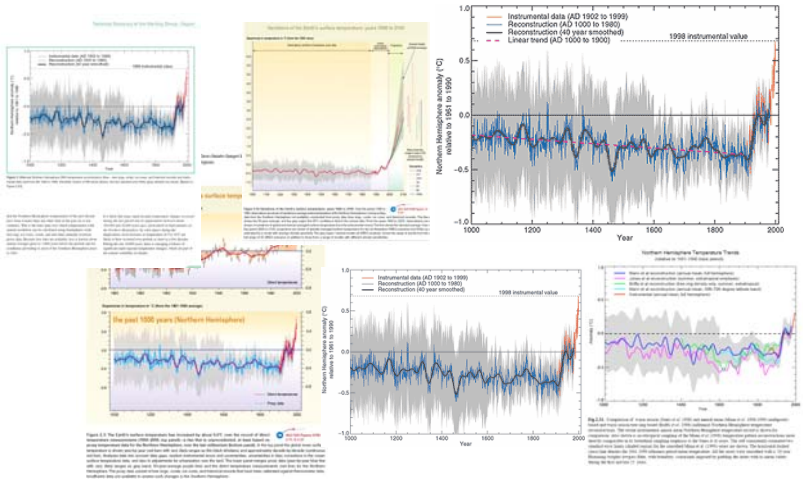

Figure 3: *Separate appearances of the hockey stick graph of Mann et al (1999) in the IPCC Third Assessment Report.*

In Canada, the hockey stick became central to the government's campaign to adopt the Kyoto Protocol. A pamphlet mailed to households across the country in 2001 opened with the statement:

> The Earth is getting warmer. The 20th century was the warmest globally in the past 1,000 years. In fact, the 1980s and 1990s were the warmest decades on record.

The graph has become iconic and the 2 underlying articles have been heavily cited in scientific and policy circles.

Since the MBH results were so remarkably different from that suggested by the simple, unweighted average of the proxies supposedly selected because of their relationship to temperature (a criterion presupposed in later tests by von Storch and Zorita, 2005), and since the results were so widely cited, one would have assumed that contemporary readers and reviewers would have been deeply interested in what, exactly, these "new" statistical methods were, and what were their properties. Yet, prior to our efforts beginning in 2003, no one else seems to have bothered to clarify the exact steps, not that they could have even if they wanted to, since the details were unavailable. It remains an astonishing commentary on this case that it required the intervention of the House Energy and Commerce Committee to force disclosure of some crucial but undisclosed computational steps.

One might have also have thought that the Intergovernmental Panel on Climate Change (IPCC), who staked so much on the validity of these results, would want to ensure an independent assessment of these methods and their properties. Yet Mann himself—the one scientist in the world who could *not* provide an independent review—was selected as a lead author for the section on climate reconstructions for the IPCC Third Assessment Report.

3 MBH98

We turn now to an exposition of: what the new statistical approach was; its statistical properties; the validity of the claims to statistical skill; how confidence intervals were calculated; its propensity to favour the most flawed proxies; and we comment on some of the obstacles to replicating and verifying the original claims.

3.1 The "New" Statistical Approach

The essential features of the "new" statistical approach were not all described in the original article or even in the 2004 Corrigendum:

1) MBH98 did provide a reasonable description of the use of principal components (with conventional centering) on a network of 1082 gridcells to obtain resulting principal component (PC) series, which they called "climate fields". No fewer than 5 such "fields" were illustrated in the original article. No analysis was provided of the stationarity of these fields; instead readers were assured that the statistical "skill" of the reconstruction itself was proof of the stationarity of the relationships,

2) A second "new statistical approach" was in how they calculated PCs on tree ring networks. First, they centered the data over the closing sub-segments and then carried out a singular value decomposition over the decentered data matrix. Their method turns out not even to be a PC method within the definition of their own cited reference [Preisendorfer 1988]. Worse, the data decentering introduces a bias in which the algorithm effectively mines for hockey stick-shaped series. Second, instead of calculating PCs over the maximum period in which all data in a network was available, they calculated them in steps – although the break points are not the same as the those in the final climate reconstruction itself. Neither of these methods was reported in the original publication, which, in this case, said that a "conventional" method was used. The decoding of this approach was significantly complicated by the provision by the authors of data sets in which PC series from different periods were spliced together and even incorrectly collated.

3) Third, in the calibration-estimation step, the proxy series (including the PC series) were used to estimate past temperature PCs. The methodology boils down to two steps: a) a piecewise multiple regression of up to 112 proxy series (including up to 31 PC series) against up to 11 temperature PC series; b) multiple regressions of proxy data (piecewise by year) against the regression coefficients from the first step. We are unaware of precise precedents for this methodology, nor was there any articulation of its statistical properties, such as whether the coefficient estimates are asymptotically consistent and unbiased, or how confidence intervals should be derived.

4) Fourth, they claimed to have used "clear a priori criteria" to select proxies and, in Mann et al 2000, attributed the robustness of the study to the use of such criteria.

Overlaid on this were some more mundane matters, which fall more in the realm of accounting than statistics. A few prominent series were truncated; one series out of 415 was used twice in different locations and, in one usage but not the other, extrapolated in its early portion; gray data was often used

even when official versions were archived; series were inaccurately cited and some series were bizarrely mislocated. Many series said to have been used were excluded along the way. In most cases, these simply present obstacles to replication, but in at least one instance there is a serious issue of whether the series was edited with a view to affecting the final result.

3.2 Bias in the Tree Ring PC Methodology

We observed above that the MBH98 methodology constructs a system of weights that results in a reconstruction that is totally different from the simple average of the proxies. This occurs both because of the tree ring PC methodology, on which we have published and will discuss first, and the multivariate methodology, on which we have not yet published, but which is relevant to the Committee's terms of reference.

One of the results of the "new statistical approach" used for calculating PCs on tree ring networks is that it effectively mines large networks for 20th century trends – either up or down. The signs of the trend are adjusted so that the maximum hockey stick result. If the proxies share a strong and consistent signal, then the signal can be recovered despite the biased method. However, if there is not a strong common signal (which is the case with MBH proxies) and/or if there are some proxies which are "bad apples" – showing a non-climatic 20th century trend, then the method will nearly always load extra weight on them and produce a leading principal component (PC) with a strong hockey stick shape even if it is obviously not a dominant feature of the data set.

We reported on this in McIntyre and McKitrick 2005a. We defined a Hockey Stick Index (HSI) as the displacement of the post-1902 mean versus the series mean, in standard deviation units (σ). Using a conventional PC method (decomposing the covariance matrix) a 1-σ hockey stick appears only 15.3% of the time and a 1.5- σ hockey stick appears only 0.1% of the time. Results using a correlation matrix are virtually identical. Using Mann's method, a 1–σ hockey stick appears 99% of the time. Figure 4 shows two histograms from MM05a. Some samples or red noise hockey sticks are shown in the panel at right, obviously with similar features to the MBH reconstruction itself, which is located top right.

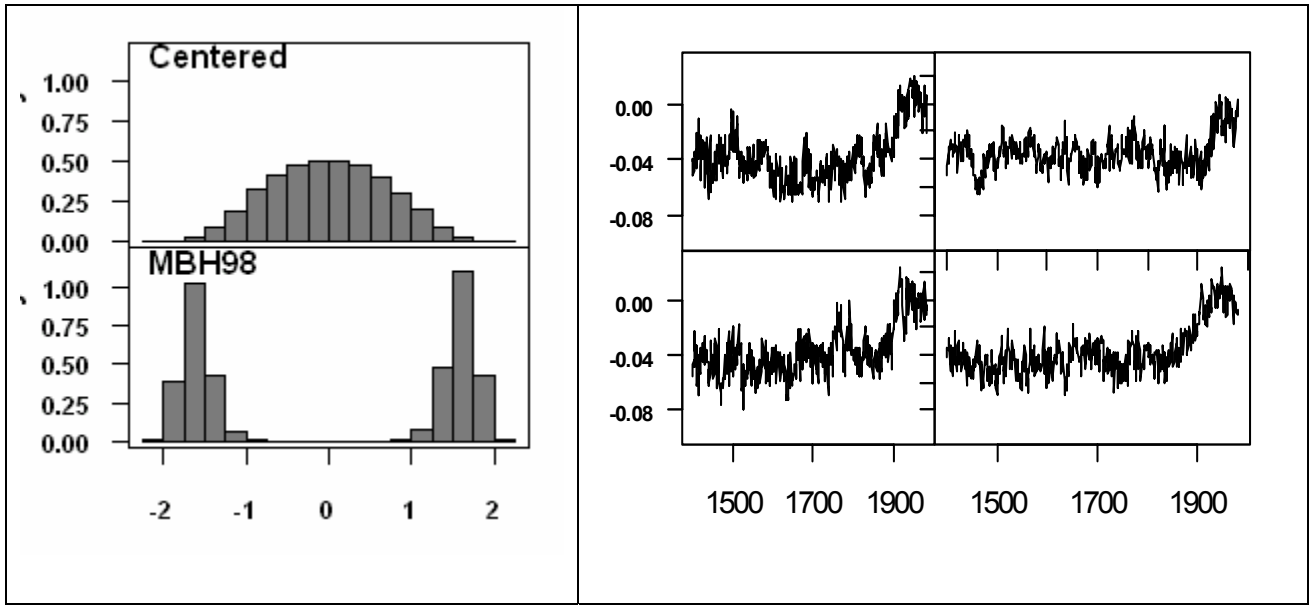

Figure 4. *Reproduction of two charts in MM05a. Left: histogram of 'Hockey Stick' Indexes of PC1s from red noise under conventional PC algorithm and under Mann PC algorithm. Right: sample of PC1s from red noise simulations with MBH98 reconstruction; MBH98 reconstruction top right, scaled to match PC.*

Further details of the bias were described in McIntyre and McKitrick [2005c]. Von Storch and Zorita [2005] argued that the bias existed only in a red noise environment where there was no actual "signal". In our Reply to their comment, we showed that, if a network has even 1-2 "bad apples" – say series with a hockey stick shape because of fertilization or some other reason not related to a temperature signal, the "bad apples" will overcome the actual signal to steer the result. We showed that even 1 hockey stick shaped series in a network of 50 can impart a hockey stick shape to a leading PC, which appears as "significant" under a Preisendorfer-type test.

One fairly practical issue on which the Committee's opinion would be useful is the existence of this bias. Von Storch and Zorita [2005] and Huybers [2005] both acknowledge the existence of the bias, although they do not agree with us on its ultimate effect on a temperature reconstruction. Mann et al. have denied the very existence of the bias. The two issues are obviously separable and we would welcome a specific opinion.

Another detrimental property of the methodology, pointed out in McIntyre and McKitrick 2005b, was the failure of PC methods to preserve orientation. PCs are weighted averages, and if it helps improve the fit the algorithm will 'flip' a series by putting a negative sign on the weight coefficient. We showed an example in MM05b of how, if an actual signal of a ring width *increase* was inserted into the North American network in the 15th century, the MBH method would flip all of these series over and *reduce* its estimate of 15th century temperature.

3.3 Bias in the Multivariate Methodology

We have referred to the fact that the MBH98 multivariate method relating temperature PCs to after-PC proxies is also sui generis, whose statistical properties were undiscussed.

In our opinion, there remains considerable misunderstanding about the properties of this multivariate method. Mann et al, Zorita et al [2003] and von Storch et al [2004] have all opined that, under an MBH98-type method, the impact of individual proxies cannot be identified (See references in MM05b). However, it is our surmise, as noted above, that this opinion is incorrect. All of the operations are linear (or easily linearized) and if the underlying linear algebra is carried through, the operations reduce to an expression in which the final NH reconstruction is a linear combination of the proxies.

In the early portion of the reconstruction where only one temperature PC is used (the AD1400 and MBH99 steps), it appears that the weighting coefficients are proportional to the correlation of each proxy to the temperature PC1. This latter procedure is used by one of our co-presenters in a very recent paper (Hegerl et al., submitted), where it is once again described as a "new" procedure.

In a situation where the proxies are close to being uncorrelated (which is how Hegerl et al describe their network and is the case with the early MBH networks), the coefficients end up being relatively close to those resulting from a multiple (inverse) regression of the temperature PC1 (or average temperature in the case of Hegerl et al) against the 15th century MBH98 network of 22 proxies (MBH99 – 14; Hegerl et al – 12) over a short calibration period of 79 years.

But the presence of substantial autocorrelation in the temperature PC1 (temperature) and many of the proxies (especially the North American tree ring PC1) is a recipe for both spurious regression and overfitting, implying that common statistical tests applied without careful consideration may provide very misleading results. In particular, as we will discuss in the next section, with the very high degrees of autocorrelation in the temperature series and some individual series, there is insufficient separation between calibration and verification period to ensure that out-of-sample properties are being tested in some key statistics.

Some of the recently-proposed recipes for salvaging MBH in the wake of our criticisms need to be closely examined on this topic. For example, Rutherford et al [2005] and Ammann and Wahl (unpublished) have argued that they can "get" a hockey stick using the multivariate methodology without using principal components. However, if this proposal is examined more closely, one observes that, in effect, an autocorrelated temperature series in a calibration period of only 79 years is being fitted against 95 (mostly) autocorrelated proxies, so the inherent weaknesses of this method in the original application are exacerbated substantially.

We believe that it is very important that the Committee consider properties of the MBH98 multivariate method when assessing whether non-PC regression models salvage the basic MBH claims.

3.4 Verification r^2 Statistic

The widespread acceptance of MBH98 was directly related to its claims of unprecedented statistical skill—as highlighted in the quoted IPCC text above. Controversy over the validity of these claims had a

direct role in the formation of this panel. On June 23, 2005, the House Energy and Commerce Committee sent a letter to Mann (with similar letters to Bradley and Hughes) asking:

> 7c. Did you calculate the R2 statistic for the temperature reconstruction, particularly for the 15th Century proxy record calculations and what were the results?
>
> 7d. What validation statistics did you calculate for the reconstruction prior to 1820, and what were the results?

This inquiry prompted considerable controversy, and, in a letter to the House Energy and Commerce Committee on July 15, 2005, Dr. Cicerone, in his capacity as President of the National Academy of Sciences, referred specifically to question 7c and offered to form an independent expert panel.

Mann's own answer to question 7c came a week later on July 23, 2005, when he advised the House Committee as follows:

> My colleagues and I did not rely on this statistic in our assessments of "skill" (i.e., the reliability of a statistical model, based on the ability of a statistical model to match data not used in constructing the model) because, in our view, and in the view of other reputable scientists in the field, it is not an adequate measure of "skill." The statistic used by Mann *et al.* 1998, the reduction of error, or "*RE*" statistic, is generally favored by scientists in the field.

We find this reply somewhat puzzling when assessed against the actual text of MBH98, which uses the term "r^2" on no fewer than 7 occasions. In the methods section (page 785), MBH98 states:

> β [or RE] is a quite rigorous measure of the similarity between two variables, measuring their correspondence not only in terms of the relative departures from mean values (as does the correlation coefficient *r*) but also in terms of the means and absolute variance of the two series. **For comparison, correlation (*r*) and squared-correlation (*r*2) statistics are also determined.** [emphasis added]

The text itself contains a prominent figure labeled as showing the "verification r^2" statistic, shown below together with the original caption. The running text described the Figure as follows:

> Figure 3 shows the spatial patterns of calibration β, and verification β **and the squared correlation statistic r^2**, demonstrating highly significant reconstructive skill over widespread regions of the reconstructed spatial domain [emphasis added]

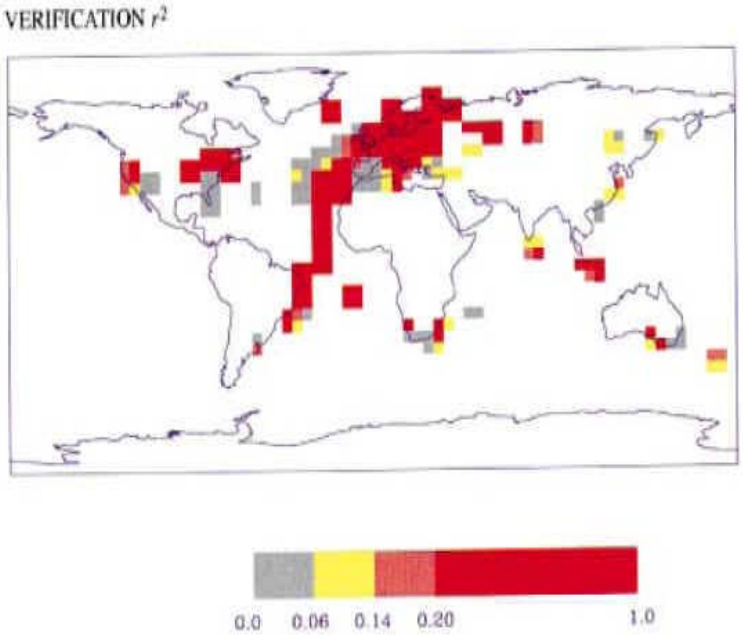

Figure 5: [MBH98 Figure 3 and caption] Spatial patterns of reconstruction statistics; bottom, verification r^2 (also based on 1854–1901 data). For the b statistic, values that are insignificant at the 99% level are shown in grey negative; but 99% significant values are shown in yellow, and significant positive values are shown in two shades of red. For the r^2 statistic, statistically insignificant values (or any gridpoints with unphysical values of correlation r , 0) are indicated in grey. The colour scale indicates values significant at the 90% (yellow), 99% (light red) and 99.9% (dark red) levels (these significance levels are slightly higher for the calibration statistics which are based on a longer period of time). A description of significance level estimation is provided in the Methods section.

While MBH98 Figure 3 only shows the verification r^2 statistic for the AD1820 step, source code provided in response to the request from the House Energy and Commerce Committee clearly shows that the verification r^2 statistic was calculated in the same step as the β [RE] statistic (see www.climateaudit.org). However, while the RE statistic was archived for all steps, the verification r^2 statistic was not archived in the original Supplementary Information at Nature (now deleted.)

Our own calculations, reported in McIntyre and McKitrick, 2005a, showed that, in the 15th century step, the period in controversy, far from the verification r^2 statistic confirming seemingly high values of the RE statistic, values were extremely low (~0), which we interpreted as contradicting claims for statistical skill for the MBH98 model. We are unable to conceive of any situation where an index with true statistical relationship to temperature has such a catastrophic failure of the r^2 test. It is essential that the committee carefully consider this situation.

In May 2005, UCAR issued a press release that our claims were "unfounded" including the following statement:

> Ammann and Wahl conclude that the highly publicized criticisms of the MBH graph are unfounded.

The UCAR press release was subsequently relied upon in evidence to Congress. Sir John Houghton cited it before the Senate Energy and Natural Resources Committee in support of his claims that our criticisms had been refuted [Houghton, 2005]; Mann cited it in his reply to the House Energy and Commerce Committee [Mann, 2005]

However, analysis of Ammann and Wahl code (which, to their credit, accompanied the press release) showed exactly the opposite. In fact, results from their code for the verification r^2 (and CE statistic) were almost identical to what we reported in MM05a. In Table 1, the top row shows the test scores reported by MBH98. The next two rows show the test scores as reported by us and by Ammann-Wahl. The unavoidable fact is that Ammann and Wahl's results confirm our findings of the insignificance of MBH98 and yet, like MBH, they did not report this information – even while issuing the opposite claim in a national UCAR media advisory.

	Verification RE	**Verification R^2**	**CE**
MBH98	0.48	n.r.	n.r.
MM05a Emulation	0.46	0.02	-0.26
Ammann&Wahl Code	0.47	0.02	-0.24

Table 1. *MBH98 (and Emulations) 15th Century Step Cross-Validation Statistics in verification interval. Note: 99% significance cut-off for R^2 is 0.34.*

As a reviewer for the Ammann and Wahl submission to Climatic Change, one of us (McIntyre) requested the values of the verification r^2 statistic for the 15th century step, which they refused to provide. Later, I repeated the request of Ammann at his presentation on this topic at the December 2005 AGU meetings, but he again refused to release the number.

3.5 "Spurious" Significance

In econometrics, there has been considerable attention paid to the phenomenon of "spurious regression", commencing with Granger and Newbold's [1974] demonstration that regressions of independent unit root series almost always yield "high" t-statistics even though the series are, by construction, independent. This has led to a rich and extensive literature on proper handling of time series data which should be considered by the Committee. Recently, Ferson et al [2003] have shown that autocorrelated (but not unit root) series are subject to a spurious regression effect if the correlation coefficients are used for data selection. High correlations in a calibration interval were shown to be routinely associated with zero predictive skill across a verification interval. The problem, as initially explained by Phillips [1986], is that in the particular circumstances of spurious regression, the usual test statistics from regression techniques have degenerate asymptotic behaviour, such that the 95% significance level converges to infinity, and lengthening the sample size does not overcome the effect. (Here we use the term "spurious" in a specific technical sense and, not, as in climate science, merely as a term of disapproval.) Using correlation

measures to select a subset of proxies out of a large library is exactly analogous to spurious data mining methods in stock market forecasting, which were vividly criticized by Ferson et al.

In McIntyre and McKitrick 2005a we explained the contradiction between MBH's failed r^2 statistic and the seemingly significant RE statistic by showing that, à la Phillips [1986], the RE critical value was underestimated in MBH98. We were struck by the remarkable similarity of many simulated PC1s, arising entirely out of operations on red noise, to the MBH reconstruction. A valid temperature reconstruction should be able to out-perform these series produced from red noise. Hence as a lower bound benchmark for skill of *actual* data, we regressed the red noise PC1s against NH temperature. This process yielded a very high 99% significance benchmark (0.56), indicating that the MBH RE statistic was not necessarily significant, thus reconciling to the information from the r^2 statistic.

This analysis was criticized by Huybers [2005], who argued that a simulation in which the simulated PC1s were re-scaled against NH temperature (a step unreported in the original article but observable in the source code released in 2005), yielded a benchmark of 0.0, which Huybers interpreted as restoring the seeming significance of the MBH98 reconstruction. In reality, this would have merely left the RE contradiction unexplained as long as the model failed the verification r^2 test. In our Reply to Huybers [McIntyre and McKitrick 2005d], we pointed out that Huybers' new simulations did not replicate all the pertinent MBH features, and we reported on new simulations in which we applied the re-scaling but also constructed proxy networks consisting of the simulated PC1s combined with 21 white noise series. These simulations yielded virtually identical results to those reported in MM05a (99% benchmark of 0.51) and supported our reconciliation of the r^2 and RE statistic.

Bürger and Cubasch [2005] pointed out a further serious problem in the handling of verification statistics by MBH. It is inappropriate to use the RE statistic for choosing between models – which is the reason advanced in Mann et al [2003] or Ammann and Wahl [unpublished] for eliminating MBH98-type reconstructions with high 15th century values. Bürger and Cubasch observe correctly that using the verification RE statistic to *choose* models amounts to incorporating the verification period into the calibration period, a procedure which is clearly a recipe for overfitting. While the r^2 test should be ordinarily consulted in any event, the need in such circumstances is overwhelming.

3.6 Bristlecones and Robustness

As we observed above, the other major selling point of MBH was its claimed robustness, even to the presence/absence of dendroclimatic indicators, about which the IPCC Second Assessment Report (1996) expressed the following reservations:

> Tree-ring records frequently represent interannual and decadal climate variability with good fidelity, as indicated by comparison with recent instrumental records. However the extent to which multidecadal, century and longer time-scale variability is expressed can vary, depending on the length of individual ring-width or ring-density series that make up the chronologies **and the way in which these series have been processed to remove non-climatic trends**. **In addition, the possible confounding**

> **effects of carbon dioxide fertilization needs to be taken into account when calibrating tree ring data against climate variations.** [emphasis added]

The concern about carbon dioxide fertilization presumably reflected the concerns previously expressed by Lamarche et al [1984] and Graybill and Idso [1993] that the 20th century growth pulse in bristlecones could not be explained by temperature, hypothesizing that the increased growth was due to carbon dioxide fertilization. In MM05b, we surveyed other possible non-climatic and non-temperature climatic effects on bristlecone growth. Graybill and Idso [1993] was endorsed by Biondi et al. [1999] (including MBH coauthor Hughes) as follows:

> The average of those sites [a network of high-elevation temperature-sensitive tree-ring sites in the Great Basin and Sierra Nevada of Hughes and Funkhouser, unpublished], plotted in Figure 5, is based on many ring-width series, each one being 500 years or longer, without individual growth surges or suppressions and from "strip-bark" five-needle upper forest border pines of great age. Such record is not a reliable temperature proxy for the last 150 years as it shows an increasing trend in about 1850 that has been attributed to atmospheric CO2 fertilization [Graybill and Idso, 1993]

Their Figure 5 is reproduced below. They interpreted it as evidence of low frequency coherence between bristlecone growth and the Idaho chronology prior to 1850, and as not showing exceptional 20th century warmth in Idaho.

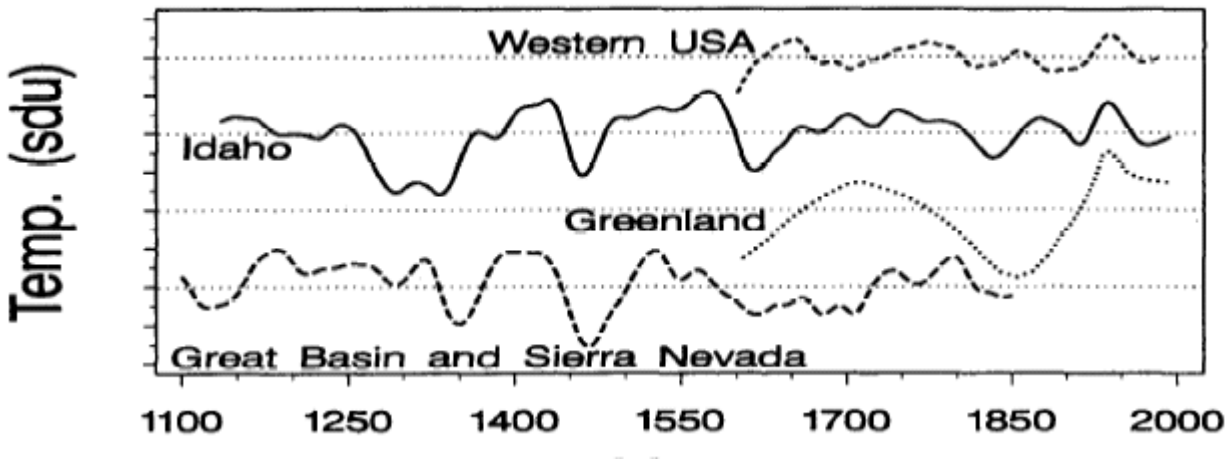

***Figure 6**. Biondi et al [1999]*

Given this prevalent concern with potential problems with carbon dioxide fertilization of bristlecones, Mann et al. took great pains to emphasize that their results were not sensitive to nonclimatic factors of this type. MBH98 stated:

> **the long-term trend in NH is relatively robust to the inclusion of dendroclimatic indicators in the network**, suggesting that potential tree growth trend biases are not influential in the multiproxy climate reconstructions. (p. 783, emphasis added.)

Mann et al., 2000 went even further:

> We have also verified that possible low-frequency bias due to non-climatic influences on dendroclimatic (tree-ring) indicators is not problematic in our temperature reconstructions…**Whether we use all data, exclude tree rings, or base a reconstruction only on tree rings, has no significant effect on the form of the reconstruction for the period in question.** … These comparisons show no evidence that the possible biases inherent to tree-ring (alone) based studies impair in any significant way the multiproxy-based temperature pattern reconstructions discussed here. (http://www.ngdc.noaa.gov/paleo/ei/ei_nodendro.html, emphasis added.)

Despite these claims, in McIntyre and McKitrick 2005a, 2005b, we showed that an important interaction exists between the proxies in question and the biased PC methodology. We observed that 15 (out of 70) series selected for top-weighting in the North American PC1 were bristlecone (or inter-related foxtail) pine series originating from a single researcher, Donald Graybill, which accounted for over 93% of the variance in the PC1. While we are not convinced that PC methods are particularly appropriate for this type of exercise, in order to assess the effect of the "new statistical approach" of MBH98, we compared the MBH98 "PCs" in the North American network to PCs conventionally calculated. There were remarkable differences. In McIntyre and McKitrick 2005a Figure 3 (shown below in Figure 7), we compared the PC1 from a conventional calculation using a covariance matrix to the corresponding MBH98 PC1.

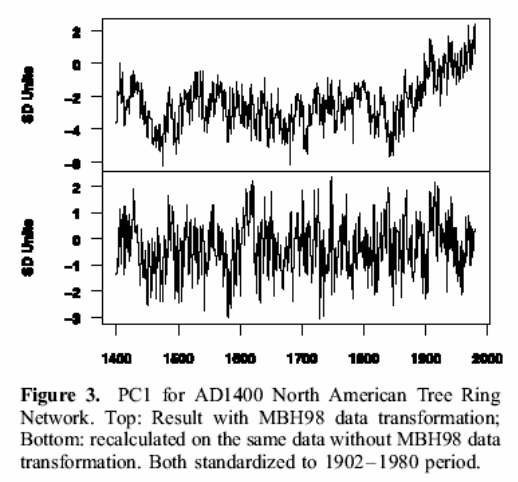

Figure 3. PC1 for AD1400 North American Tree Ring Network. Top: Result with MBH98 data transformation; Bottom: recalculated on the same data without MBH98 data transformation. Both standardized to 1902–1980 period.

Figure 7: *Comparison of Mann PC1 (top) and conventional (bottom.)*

We observed that the hockey stick shape associated with the bristlecones was demoted to the PC4 in a calculation using a covariance matrix and, in MM05b, we observed that it appeared in the PC2 using a correlation matrix. We also noted that the MBH98 method dramatically inflated the first eigenvalue (MBH98 – 38%), making the bristlecone hockey stick shape seem like the "dominant component of variance" instead of merely being a localized effect which a covariance matrix PC weights at 8%.

In McIntyre and McKitrick 2005b, we observed that, if the Graybill bristlecones (and an equally questionable cedar series) are removed from the MBH98 network, even with its questionable multivariate methodology, in an MBH98-type calculation, one obtained high 15th century values, contradicting claims to 20th century uniqueness. Equivalently, if a principal components calculation were done with a covariance matrix and the same number of PC series retained as in MBH98, then because of the occurrence of the PC representing bristlecone growth in the PC4, this also led to high 15th century values (which Mann et al. characterized as "effectively throwing out" valuable data.)

In terms of specific empirical findings on the effect of various PC procedures and bristlecone sensitivities, all calculations in the 2005 corpus by both parties substantially agree, although the parties characterize their findings in substantially different terms. What we describe as a "sensitivity analysis" or a "test of robustness", Mann et al. and Ammann and Wahl describe as "throwing out" data. For example, Mann et al [2003] showed a calculation using their methods, also with a high 15th century, in their case, obtained by excluding all 15th century North American tree rings (although the "active ingredient" in the difference was simply the bristlecones).

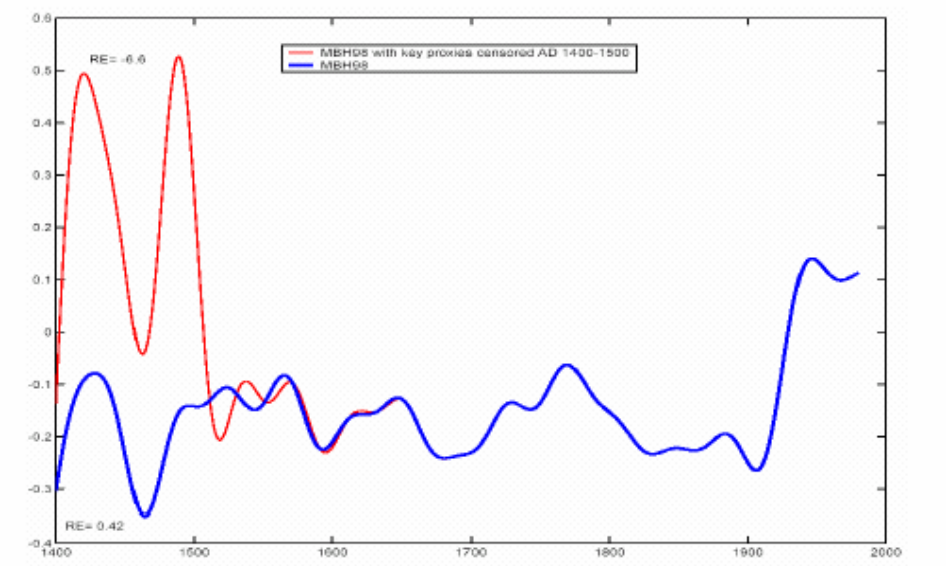

Figure 8: *Mann et al [2003] Figure 1.* Original Caption: *Comparison of MBH98 reconstruction (blue) with reconstruction resulting from the elimination of key proxy data sets (1)-(3) over the AD 1400- 1500 interval. This yields essentially the same result obtained by mm by eliminating a significant fraction of the MBH98 data available for that period (both series have been smoothed with a 40 year lowpass filter).*

Mann et al. [2003] implicitly and Wahl and Ammann [unpublished] explicitly recognized that one could get high 15th century values using MBH98-type procedures on the MBH98 proxies if the bristlecones are removed (Figure 9). However, they repudiated these variations on the grounds that they failed an RE test (without, however, discussing the fact that both variations fail an r^2 test).

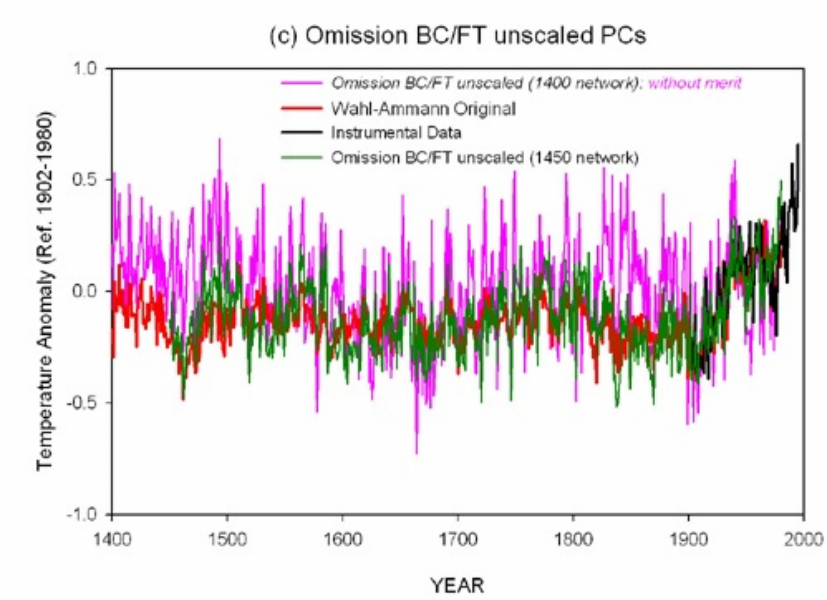

Figure 9. *Removal of the bristlecone pines from the AD1400 network changes the results from the red line to the pink line, thereby removing the characteristic "hockey stick" shape and leaving behind mere noise.* *Source: Wahl and Ammann (2005) http://www.cgd.ucar.edu/ccr/ammann/millennium/recon/WEB_examples.jpg (unpublished).*

Bürger and Cubasch [2005], a very important recent study of MBH98-type reconstruction, have placed this aspect of the controversy in a helpful new context. Building on our studies, they pointed out that very slight variations in methodological choices in an MBH98-type reconstruction – none of which seem on the surface to be preferable one to another – led to a bewildering array of results, as illustrated below.

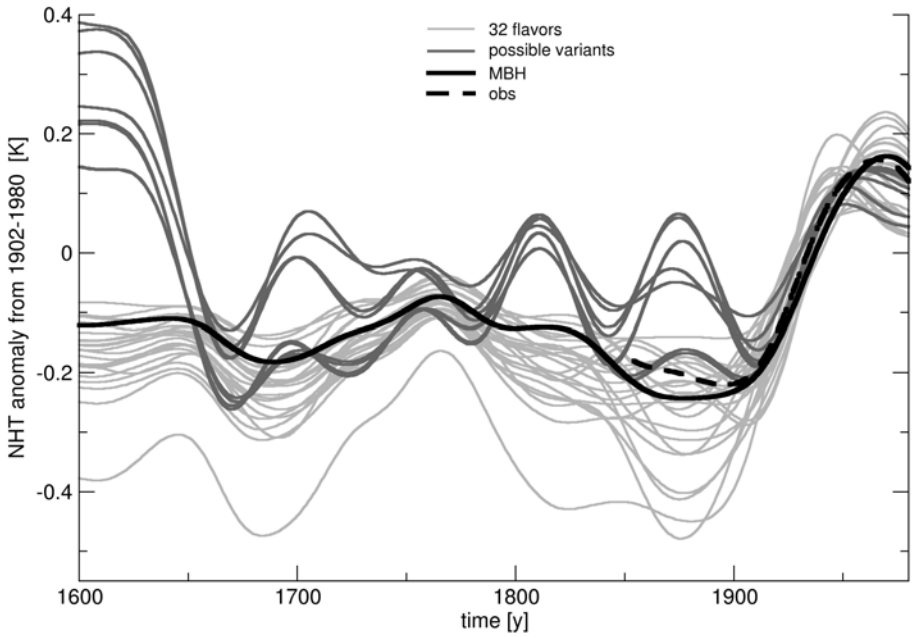

Figure 10: *Burger and Cubasch [2005] SI Figure 1.*

They pointed out that " No a priori, purely theoretical argument allows us to select one out of the 64 as being the ''true'' reconstruction" and provided the following admonition about using RE statistics to choose a model:

> On the basis of the validation RE one might be tempted to prefer the (most simple) variants 100000 or 101000, or also MBH, to the others. But that statistic must not be used to select a model; it can only serve as a check of a model, e.g. for overfitting, after it has been selected. To do otherwise amounts to extend the calibration over to the validation period.

Clearly, this completely eliminates the argument for model preferences relied upon by Mann et al. and later Ammann and Wahl.

However, let us now consider the original claims that the MBH98 model was robust to the presence/absence of **all dendroclimatic indicators**. Mann et al. argued that high 15th century values were obtained only by "throwing out" valuable data. But the relevant data being "thrown out" is only the bristlecones. Even if the bristlecones were known to be good proxies this would undermine any claims of robustness. That the proxies are not considered valid further strengthens our point. And even if the sensitivity test removes all 15th century North American tree rings, as in Figure 8, it hardly matters – it

disproves the claim that the reconstruction is robust to the presence/absence of **all dendroclimatic indicators.**

In the course of carrying out these analyses, we observed that some directories at Mann's FTP site (the "CENSORED" directories) contained assessments of the impact of excluding the questionable bristlecones from their network. Separate censored calculations were carried out for periods beginning 1400, 1300, 1200, 1100 and 1000. We show below in Figure 10 the PC series – even with the biased MBH method – on the 50-site AD1400 network excluding the bristlecones. These are graphs of data at Mann's FTP site.

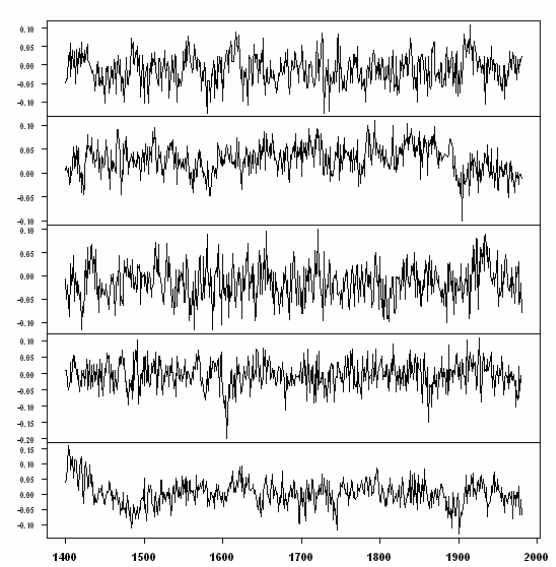

Figure 10. *The first 5 PCs in the NOAMER network after removing the bristlecone pines and applying Mann's decentered PC method. Note the contrast with the top panel of Figure 7.*
Source: ftp://holocene.evsc.virginia.edu/pub/MBH98/TREE/ITRDB/NOAMER/BACKTO_1400-CENSORED/.

It is self-evident that none of these PC series contains a hockey stick shaped series. If these PC series are carried forward in an MBH98-type calculation, even with their questionable multivariate methodology, even including all five NOAMER PCs, high early 15th century values result, as discussed in MM05b.

In the letter from Barton to Mann, the House Committee inquired:

> 7a. Did you run calculations without the bristlecone pine series referenced in the article and, if so, what was the result?

Mann's answer was lengthy, but included the following:

> For a complete scientific response, you should consult the article my co-authors and I published back in 1999 addressing precisely these issues: Mann, M.E., Bradley, R.S., and Hughes, M.K.,…Geophysical Research Letters, 26, 759-62 (1999). As my co-authors and I explained in our 1999 article cited above, given the proxy data available at that time, certain

> key tree-ring data (including the series mentioned above) were essential, if the reconstructed temperature record during early centuries were to have any climatologic "skill" (that is, any validity or meaningfulness). These conclusions were of course reached through analyses in which these key datasets were excluded, and the results tested for statistical validity. Our conclusions have been confirmed by Wahl and Ammann (see above).

The statistical "validity" discussed here is, of course, only the RE test. Whatever the merits or otherwise of the procedures described in the above answer, and regardless of whether the above procedure is clearly described in MBH99 or not, the complete change in the character of the results is inconsistent with the claim that their results are robust to the presence/absence of all dendroclimatic indicators, one of the prominent claims that was relied as part of the widespread acceptance of MBH98.

Moreover, regarding the "treatment" of the bristlecones in MBH99, it is puzzling to us that the "pre" treatment NH climate reconstruction in MBH98 is identical to the "post" treatment NH climate reconstruction where they overlap after 1400 (see Figure 11). What kind of "adjustment" leaves the result unchanged?

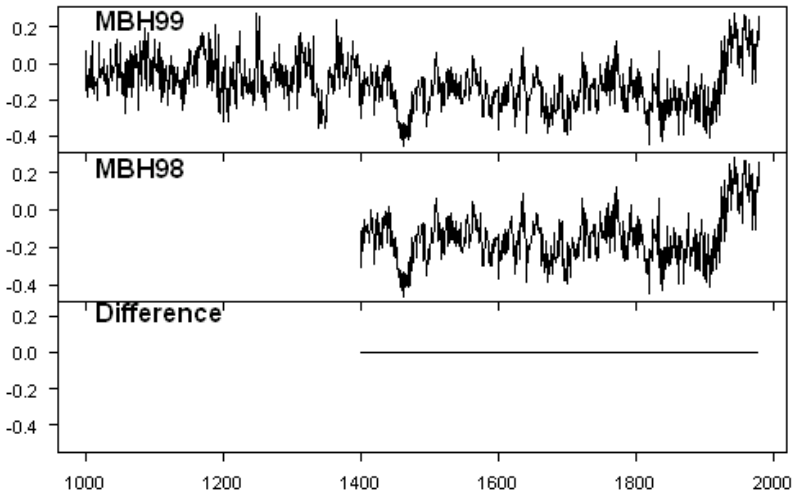

Figure 11: *Top panel: MBH99 hockey stick. Middle panel: MBH98 hockey stick. Bottom panel: differences.*

3.7 Confidence Intervals

The calculation of confidence intervals in MBH98 was new to millennial paleoclimatology and added very much to the impressiveness of the original presentation. Indeed, it was what enabled MBH98 to derive their claims about the "warmest year" and "warmest decade".

Confidence intervals in MBH98 (to which the term "self-consistent" is applied) are, as we understand it, calculated simply as twice the standard error from calibration period residuals. If there is overfitting (or spurious regression) in the calibration period, as appears almost certain, then calibration period residuals are likely to provide an extremely biased and over-confident estimate of confidence intervals. For a sui generis procedure with little knowledge of its statistical properties, at a minimum, it seems to us that confidence intervals should be calculated from the verification period residuals.

In this case, given that the verification r^2 for the early steps is ~0, this procedure would, of course, have led to very wide confidence intervals and little to no reduction from natural variability, hence a complete inability to assess the statistical significance of warmth in the 1990s.

MBH99 acknowledged that there was significant low-frequency content in the spectrum of residuals i.e. highly autocorrelated residuals. Since at least Granger and Newbold [1974], econometricians have interpreted autocorrelated residuals as evidence of a misspecification. Instead, MBH99 purported to adjust the confidence interval calculations. However, no statistical reference is provided for this calculation. Neither we nor a time series specialist who we consulted on this matter have been able to figure out how this calculation was done.

The use of calibration period residuals to estimate confidence intervals is followed in other multiproxy studies. In all cases, we see evidence of spurious relationships in the calibration period with serious out-of-sample behavior, raising in every case the spectre of over-optimistic estimation of the success of the reconstruction.

3.8 Further Issues

3.8.1 Von Storch and Zorita [2005]

Von Storch and Zorita [2005] agreed that the MBH PC methodology was severely biased, but argued that the effect was insignificant. However, they did not consider the impact on MBH itself, but on a simulation using pseudoproxies. In our Reply to VZ (MM05), we showed that their simulations failed to implement the critical data and methodological features that created the problem in MBH98.

Surprisingly it appears that they did not actually implement the MBH principal components methodology. Post-publication correspondence (see www.climateaudit.org) showed that they calculated principal components on the correlation matrix of decentered data, rather than on the decentered data matrix – a procedure which is much less biased. Secondly and equally importantly, they endowed the pseudoproxies with an average correlation to gridcell temperature of 0.3, based on information from an unrelated study (Jones and Mann, 2004) and which may not be accurate anyway. In our Reply, we showed that the proxies in the actual North American network had little actual correlation to temperature (Figure 13) and were much more correlated to precipitation. The subset of bristlecones had a significant correlation to CO2 levels. The assumptions of VZ precluded both "bad apples" and high noise - the very effects that drove the MBH problems. It is notable that in their example, while there is little difference whether or not the data are decentered, neither case yields a hockey stick to begin with.

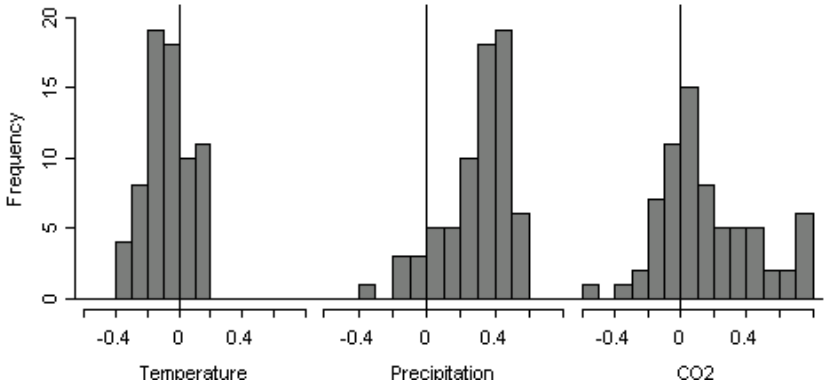

Figure 13: *(MM05c Figure 1). Correlation Histograms for AD1400 MBH98 Tree Ring Network. Left: Gridcell Temperatures; Middle – State Precipitation; Right – CO2.*

VZ failed to report their verification r^2 statistic. An essential aspect to simulating MBH is achieving both a high RE statistic and an insignificant r^2 statistic. We have seen no evidence that VZ succeeded in doing so. Perhaps the Committee can shed light on this matter.

3.8.2 Huybers and Correlation vs. Covariance PCs

Huybers [2005] pointed out that if the bias were assessed by comparing Mann's PC1 against that calculated using a correlation matrix decomposition, rather than a covariance matrix, the effect would seem less severe than shown in Figure 7 above. We gave a detailed response in our Rely to Huybers [MM05d]. We ourselves do not advocate any particular PC methodology as a means of extracting a temperature signal from the North American network; in fact, we are doubtful that any unsupervised numerical algorithm can reliably do so. Using a correlation matrix simply increases the weighting of bristlecones in the 20th century – which are the very series in dispute. Huybers (like von Storch and Zorita) acknowledged the bristlecone problem but merely suggested that the validity of bristlecones should be studied on another occasion. However, given what is already known about their deficiencies and the caveats expressly stated in IPCC [1996], this cannot be accepted. In addition, Huybers was unable to provide a statistical reference which supported the use of correlation matrices in a network already expressed in common units (in this case dimensionless units resulting from prior tree ring standardization procedures before archiving at ITRDB/WDCP). We pointed out that his own references, Rencher [1992; 2002] and Preisendorfer [Overland and Preisendorfer, 1982] had expressly used covariance matrices for networks denominated in common units – to which we would now add North [1982].

3.8.3 The Corrigendum

In July 2004, Mann et al. issued a Corrigendum to MBH98. Contrary to a general impression of Nature's policies, the Corrigendum was not subjected to external peer review and the Corrigendum SI was not even edited by Nature.

Although we had by then identified serious problems both with MBH98 PC methodology and the original disclosure of that methodology, these problems were still not disclosed in the text of the Corrigendum. In the Corrigendum SI, there was a very obscure reference to the existence of the problem and a continued denial that any results, even the PC series themselves, were affected by the problems:[1]

> All predictors (proxy and long instrumental and historical/instrumental records) and predictand (20th century instrumental record) were standardized, prior to the analysis, through removal of the calibration period (1902-1980) mean and normalization by the calibration period standard deviation. Standard deviations were calculated from the linearly detrended gridpoint series, to avoid leverage by non-stationary 20th century trends. The results are not sensitive to this step (Mann et al, in review).

4 Replication

Question 2(c) of the Boehlert Committee is:

> 2 (c) Has the information needed to replicate their work been available?

Here we can categorically answer that it has not.

In the case of MBH98, ***replication*** means more than simply obtaining a squiggly line that looks somewhat like the original MBH hockey stick. We can get that from red noise. Replication also requires verifying skill claims in the three verification statistics referenced in the original article (RE, r and r^2) and replication of the claimed robustness to the presence/absence of dendrochronological indicators.

Some information was available when we started. In the aftermath of our 2003 publication, much new information came to light, commencing with the first *public* access to MBH98 data at Mann's FTP site at the University of Virginia in November 2003. Mann then stated that 159 series needed to be used in the MBH reconstruction – a figure nowhere mentioned in MBH98 itself, but refused to identify the 159 series and their identity was impossible to determine from the site itself, which contained many more series. Only after we filed a Materials Complaint to Nature was a large new Supplementary Information archive provided at Nature in July 2004 (the Corrigendum), which for the first time identified the series used in the individual MBH98 steps (although the figure of 159 series remains unexplained to this day.) This was followed by additional, but not all, source code, archived at the University of Virginia in July 2005 in response to the Barton Committee, which does not work with the data as presently archived.

For now, we will provide a few illustrations of obstacles to replication and will send a supplement listing the information still missing that is necessary to a complete replication.

[1] http://www.nature.com/nature/journal/v430/n6995/extref/METHODS/AlgorithmDescription.txt

First, and this is quite astonishing given all the publicity, the most basic information necessary to verify calculation of the skill of each step – the actual results of each step *with no splicing* – still is not on the public record. The results for each step are needed (but particularly the first step), because MBH carried out their reconstruction in 11 steps, with substantial variations in networks for the earlier steps. The statistical skill of the AD1400 network with 22 proxies (2 of which are PC series) is not necessarily at all equivalent to the skill of the AD1820 step with 112 "proxies", including 11 instrumental temperature series. MBH reported a reconstruction obtained by splicing the results of the 11 different steps, but did not report the results of the individual steps. In 2003 we requested the residual series to allow us to compute the tests ourselves. This request was refused. We sought the assistance of the National Science Foundation and of *Nature*, neither of whom required Mann to provide the residual series. We have computed the missing test statistics ourselves based on our emulation of the unreported residuals. The information remains as unavailable today as it was at the beginning.

Secondly, the underlying proxy data required for replication was not available from the time of the publication of the paper to the summer of 2004. In 2003, one of us (McIntyre) requested an FTP location of the MBH98 proxy data set. Although this information is presently archived by *Nature* (due to our efforts), it was not then available at that site. At the time and subsequently, there was no link to the information at Mann's webpage (although other data and results were linked at that site). In response to this inquiry, Mann said that he had forgotten the location of the data and would ask an associate to locate it. The associate, Scott Rutherford, said that the data did not exist in any one location, but promised to collect it. A few weeks later, he provided a URL at Mann's FTP site. Later we noticed problems with this data, especially related to the tree ring PC series, and sought specific confirmation that this was the actual data used in MBH98. Mann said that he was too busy to respond to this or any other question. In MM03, we reported on problems in this data set, including the use of obsolete data versions, duplication of series, incorrect principal component calculations, geographic mislocations (such as use of a Paris, France precipitation series for a New England gridcell) and many other problems. In reply, Mann said that we had used the "wrong" data set, a claim repeated in Rutherford et al [2005]. Suddenly the "wrong" data set was deleted and a new version of the data appeared at Mann's FTP site, which Mann said had been there all along. Mann said that the "wrong" data had been prepared because we had requested the data in the form of an Excel spreadsheet – a claim that was untrue. The file in question had, in any case, been prepared long before our request. In any event, the FTP site in question came into existence only in July 2002 and thus the data would not have been available at that location during the preparation of IPCC TAR.

Third, while a considerable amount of source code was archived in July 2005 in response to the Barton Committee, the source code as archived is not functional with the data as archived (see www.climateaudit.org). Further, the source code only covers the calibration-estimation steps. A limited amount of source code was located at the University of Virginia FTP site, which covered the tree ring principal components step. Its value in diagnosing the biased methodology illustrates the considerable benefits of archiving source code. In MM03, as the first step in our replication exercise, we had completely re-collated over 300 tree ring chronologies using original data from WDCP and site listings from the original SI. We found inexplicable discrepancies with the series then at Mann's FTP site, illustrating the problem with an Australian series. We pointed out the discrepancy but were unable to diagnose what caused the problem. This was quickly resolved when we were able to inspect source code. We were able to identify 4 different problems. There was the unreported short segment centering. In

addition, it turned out that MBH (also unreported) had calculated principal components not over the maximum available period, but in steps, but not for every step (contrary to the information in the Corrigendum SI). In addition to these two unreported procedures, the "wrong" data set had spliced together PC series from different steps, collated some of these series to start in an incorrect year and then filled the missing 1980 values for incorrectly collated series from left to right so that in some cases up to 7 different series had identical 1980 values to 7 decimal places. Once we were able to go through the code we were able to disentangle all the problems with the MBH PC series and separate the issues of short segment centering from the stepwise calculation, splicing and collation errors. Most importantly, we were able not merely to replicate the method, but to start assessing the bias in the methodology and the impact of the bias in overweighting certain proxies.

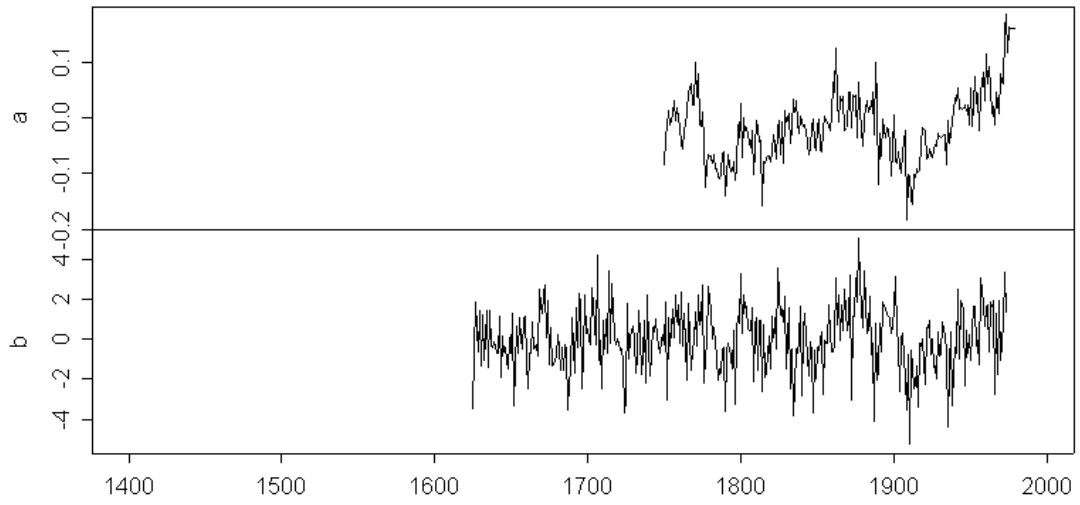

Figure 14. *(From MM03) (a) Australia PC1 in MBH98 (series #96) graphed over time. (b) PC1 for the MBH98 Australia dataset calculated using standard algorithm.*

The final replication issue that we'll mention now is how MBH selected their proxies. MBH claimed to have selected them according to a "clear a priori criteria" and, indeed, they associated this with the claimed skill of the reconstruction. But they did not state these criteria. Establishing ex ante selection protocols is an essential part of biomedical statistical procedures and we see no reason why similar protocols should not apply to multiproxy studies. We have requested a statement of these criteria from the authors, but it was refused. We also requested this information from Nature, who also refused. The only relevant information is a listing of criteria for ITRDB tree ring sites in Mann et al, 2000. ITRDB tree ring sites are an important subset of the MBH98 proxy collection, but only a subset. These criteria were appealed to in the Corrigendum (Mann et al 2004) to explain significant discrepancies in the series listings as the result of the application of quality control rules additional to those already mentioned in

Mann et al. (2000)[2], which included the criteria that: a tree ring chronology would not be used unless the mean correlation of individual tree ring records with the site chronology was at least 0.5 and it was composed of at least eight tree ring segments by 1680. In our examination of the MBH98 data we found 22 sites that did not have 8 trees by 1680, and 171 sites that had less than 0.5 mean correlation of the individual trees with the site chronology. In one of the cases the tree ring segments had so little correlation with the site chronology that one of us (McIntyre) had emailed the originating author (Roseanne D'Arrigo) to ask why. She discovered that the wrong data had been posted at WDCP. Had Mann applied these quality control rules, it is hard to see how he would not have previously noticed the discrepancy.

There remain many other issues, which we will provide in a Supplement. These sort of problems and non-disclosure greatly increase the time cost for researchers attempting to carry out replication, an effect discussed at length in McCullough and Vinod [2003].

4.2 Ammann and Wahl Replication Claims

Last year, Ammann and Wahl said that they had exactly replicated MBH, which is an exaggeration and very misleading. To their credit, Ammann and Wahl published the code for their emulation, which followed almost exactly the same approach as ours, and to which we were quickly able to reconcile our computations. It turned out that their reconstructions of "climate fields" matched ours to 9 significant digits, leaving only minor differences arising from rescaling to NH temperature target series. We reported on our reconciliation attempts in real time at climateaudit.org as an exercise. Ammann and Wahl had not achieved anything in their replication that we had not already anticipated. In particular, their differences with us did not result from a materially different emulation procedure, but from differing interpretations of more or less identical results. Further, the replication carried out by Ammann and Wahl was limited in scope and did not include the replication of confidence intervals, proxy selection, or many other MBH decisions.

As pointed out above, replication of MBH98 includes replication of their claims to statistical skill and robustness to the presence/absence of all dendroclimatic indicators. Ammann and Wahl code shows that they not only failed to replicate MBH claims of statistical skill in three verification statistics (RE, r and r^2), but again, as noted above, their code confirms our results and criticism. Secondly, Ammann and Wahl have shown that calculations without bristlecones are substantially different than results with bristlecones. Like Mann et al before them, they justify the inclusion of bristlecones merely on the basis of the RE statistic, but the very argument refutes the MBH98 claim of robustness to the presence/absence of all dendroclimatic indicators.

4.3 Model Code

Multiproxy paleoclimate data sets are very similar in form and size to some econometric data sets. There has been much serious consideration of replication issues in economics, rising most recently from McCullough and Vinod [2003], but the issue was previously raised by Dewalt et al. [1985] and in political science by King [1995]. The recently appointed chairman of the Federal Reserve System, Ben Bernanke, was previously the editor of the highly-regarded *American Economic Review*, and, in that

[2] http://www.ngdc.noaa.gov/paleo/ei/ei_nodendro.html

capacity, adopted a replication standard for empirical studies in which source code and data as used had to be submitted to the journal as a condition of review.

> Submitters should be aware that the Editors now routinely require, as a condition of publication, that authors of papers including empirical results (including simulations) provide to this office, in electronic form, data and code sufficient to permit replication. Exceptions, for proprietary data for example, must be approved by the Editor in advance of the review process. Electronic data appendices will be posted on the journal's Web site, http://www.aeaweb.org/aer/. Authors of previously published articles who wish to submit a data appendix for posting are welcome to do so. In any case, published authors are strongly encouraged to comply with the replication policy, should requests for data and programs be received. The Editors should be notified if an author of a previously published article does not make a reasonable effort to comply with the journal's replication policy. Ben S. Bernanke, Editor

We can think of no conceivable reason why similar "best practices" should not be applied by paleoclimate journals. Indeed, in the case of paleoclimate, many of the leading studies are funded by the U.S. federal government. The Global Change program specifically requires that climate data be publicly archived.

> Full and open sharing of the full suite of global data sets for all global change researchers is a fundamental objective...**For those programs in which selected principal investigators have initial periods of exclusive data use, data should be made openly available as soon as they become widely useful. In each case the funding agency should explicitly define the duration of any exclusive use period.**… In the past, some Principal Investigators have retained data for indefinite periods and this has inhibited their widespread use. This practice should be eliminated through active consideration of the tradeoffs between widespread distribution of data sets and the need to assure data quality and validity. The guiding principle is that as soon as data might be useful to other researchers they should be released, along with documentation which can be used by the other researchers to judge data quality and potential usefulness. In this way, users can determine for themselves if they want to proceed with data of questionable quality or wait for additional developments. [emphasis added]

The Earth Sciences Division of NSF (which administers relevant programs) has the following policy:

> 5. For those programs in which selected principle investigators have initial periods of exclusive data use, data should be made openly available as soon as possible, but no later than two (2) years after the data were collected.

We remain bewildered that the National Science Foundation seems unable to enforce this policy on paleoclimate authors, who routinely ignore the policy.

5 Support from "Independent" Studies

We do not have time here to survey problems with the so-called "independent" studies. This work is in progress and there are many notes on these issues at www.climateaudit.org, where there is a category tab for each of the major multiproxy studies. However a few observations:

1) The authors of the studies are not "independent" as sometimes claimed. If one looks down the masthead, Mann, Jones, Briffa and their close associates are involved in the majority of the studies. We request that the Committee consider what exactly constitutes "independence" between studies, both in terms of authorship and in terms of proxy selection.
2) All subsequent studies have been strongly influenced by MBH98-99. In virtually every case, the later studies have explicitly compared their findings to those of MBH99 and sometimes [e.g. Mann and Jones, 2003] even explicitly benchmarked their results on MBH98-99.
3) In every case, there remain obstacles to an exact (and in some cases even approximate) replication of the results either in the form of data or unclear methods. We ask that the Committee report on remaining obstacles to replication for any studies that it considers. For studies where there is no archive of the data as used or other non-compliance with federal archiving policies, we ask that the Committee explicitly note this and proscribe usage of such studies as evidence on surface temperatures in the past 1000-2000 years.
4) On an overall basis for large populations (387 sites) of temperature-sensitive tree ring series (both width and density), ring widths and densities have been declining in the latter half of the 20th century [Briffa et al, 1998; Briffa 2000]. The canonical multiproxy studies all use very small (<20, usually <10, series). Unlike the large population series, the small-sample averages tend to "support" MBH98-99 through a hockey stick shape, at least suggesting the possibility of biased selection. In particular, we note that bristlecones (or inter-related foxtails), known to be a problematic proxy, but having a distinctive hockey stick shape, are repetitively selected into these small samples either directly or though Mann's even more accentuated PC1, thus affecting, not only MBH99, but Crowley and Lowery [2000] (two series), Esper et al. [2002] (two series), Mann and Jones [2003], Jones and Mann [2004] and Osborn and Briffa [2006] (two series).

The underlying lack of robustness of a typical such study is illustrated below where the series of Crowley and Lowery [2000] are re-calculated without two problematic bristlecone series and the equally problematic Dunde δO18 series. If these proxies are problematic and reflect biased selection, then conclusions as to the relative medieval-modern levels are not at all robust. Similar results apply in varying degrees for other multiproxy studies.

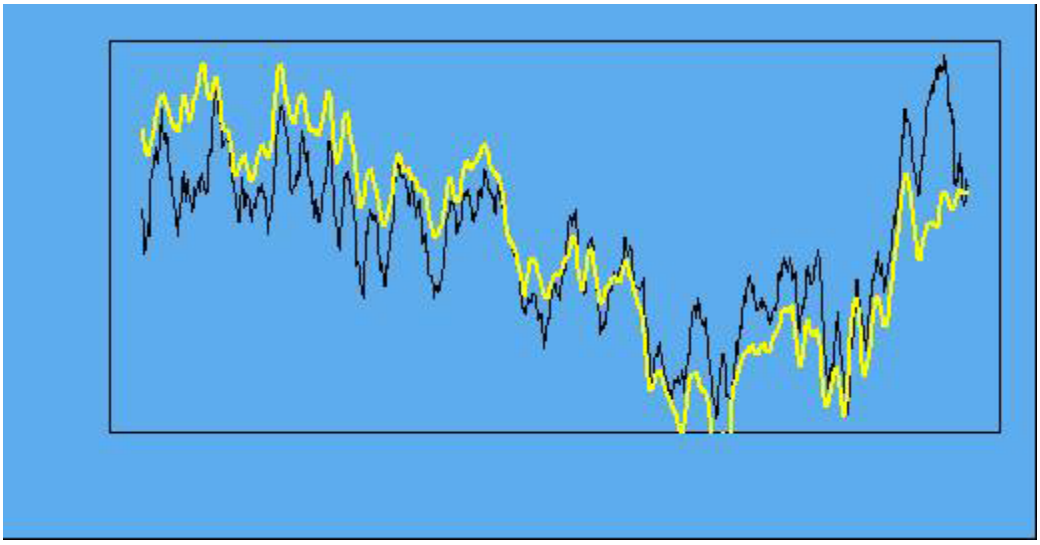

Figure 15: Crowley and Lowery (2000) original result (black) and recalculated (yellow) after removing 2 bristlecones and 1 Dunde δO18 series.

Additionally, as pointed out in McIntyre [2005b], all of the "supporting" studies manifest similar statistical problems: in the calibration period, they have failed Durbin-Watson statistics, and in a verification period, they have insignificant verification r^2 statistics, both indicating spurious relationships.

None of these studies has received thorough due diligence. While we have not published formally on the matter, in part because of the time cost of obtaining usable data and methodological disclosure, in our opinion, none of these studies provides a "safe haven" for climate history and the Committee needs to be very cautious before relying on any of them. Some issues pertaining to these "supporting studies" are listed in Appendix B and comments on all these issues are passim at www.climateaudit.org.

6 Other Presentations

We would like to make a few comments on previous presentations. The issue that gives an edge to the present topic is obviously the comparison of temperatures between the Medieval Period and the 20th century. Some of the proxies developed by earlier presenters, while interesting, are not only irrelevant to this issue, but arguably introduce inhomogeneity into the data. For example, coral data, discussed by Dr Schrag, is used in several studies, but, in no case, does the data cover periods prior to about 1600. For the purposes of the controversial medieval/modern comparison, inclusion of the data simply causes problems. Documentary evidence of the type discussed by Dr Luterbacher goes somewhat earlier, but again merely introduces inhomogeneity into the medieval-modern comparison and, in very small proxy sets such as Jones et al 1998, the inhomogeneity can disguise the actual performance of the medieval proxies.

Borehole data, of the type discussed by Dr Pollack, is used in only one multiproxy study involving the MWP - the Dahl-Jensen borehole data from Greenland shown in Figure 16 below is said to have been used in Mann and Jones [2003]. Since the final reconstruction of MJ03 bears little relationship to this shape, we presume that these proxies received very low weights.

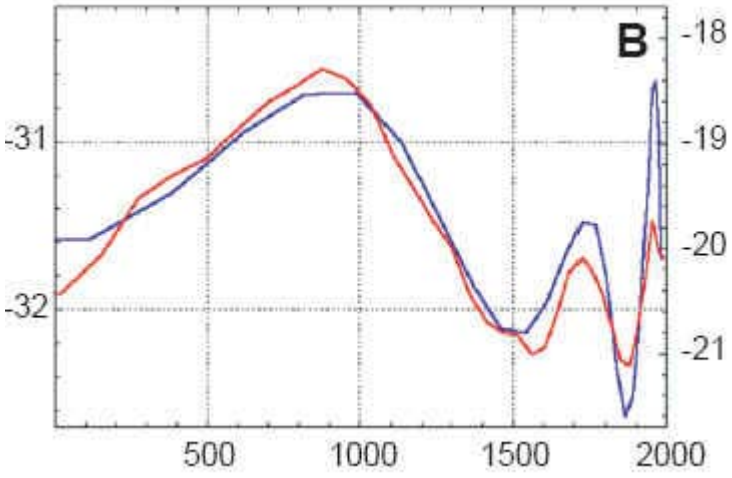

Figure 16. *Dahl-Jensen 1998 Figure 1. Original Caption. The reconstructed temperature histories for GRIP (red curves) and Dye 3 (blue curves) are shown for … the last 2 ky BP (**B**). The two histories are nearly identical, with 50% larger amplitudes at Dye 3 than found at GRIP. The reconstructed climate must represent events that occur over Greenland, probably the high-latitude North Atlantic region. Source: Dahl-Jensen et al [Science 1998] Figure 4B.*

For the purposes of assessing the canonical multiproxy studies, the panel should focus its attention primarily on indexes of tree ring growth ("site chronologies"), especially a handful of sites repetitively used in supposedly "independent" studies and secondarily on tropical δO18 cores and a couple of Moberg proxies.

In the case of two studies presented today (Hegerl et al, 2006; D'Arrigo et al, 2006), these studies were submitted to the IPCC for use in the Fourth Assessment Report. In our capacity of IPCC reviewers, we requested data used in these studies in order to provide a review. Both the authors and the IPCC refused to provide the data. The IPCC said that the function of an IPCC reviewer was only to ensure that the representation of the article by the IPCC conformed to the article and did not extend to an independent review of the data, which they said was the exclusive function of the journal. In the case of Hegerl et al, even the identity of the sites was withheld. If these studies are to be considered by the NAS Committee, we urge the Committee to ensure that data and methods archiving is completed prior to any such consideration.

Von Storch et al. [2004] considered the issue of differing variability between climate reconstructions and hypothesized that the seemingly attenuated variance of MBH99 and Jones et al [1998] was due to the their use of inverse regression. In our opinion, overly focusing on the question of variance draws attention away from the issue of "drift" in proxies, which affects the ability of achieving a medieval-modern

comparison. McIntyre [2005b] argued that the hypothesis of von Storch et al [2004] was incorrect anyway, as the multivariate methods used by MBH99 and Jones et al [1998] were not inverse regression as assumed by von Storch et al. and, in each case, actually carried out a re-scaling step, not considered by von Storch et al. [2004]. McIntyre [2005b] hypothesized that the difference in variability between proxy studies was more directly related to problems resulting from the use of standard deviations estimated over short periods under-estimating long-run standard deviations, a well-known problem with highly autocorrelated series.

7 Conclusions

Despite claims by some critics, we are not offering and have never offered an alternative climate reconstruction. In addition, some critics have purported to reduce our many criticisms of MBH98 to one claim:

- that the hockey stick is *simpliciter* a product of the flawed MBH98 principal components methodology.

Having set up this straw man, it is then argued that a hockey stick can be obtained in some other way, thus supposedly disproving all of the many criticisms. It should be abundantly clear that we do not claim that the hockey stick is *simpliciter* a product of the flawed principal components methodology. The MBH PC methodology is simply one form of making hockey stick shaped series – an effective and interesting methodology, but nonetheless, only one method. Hockey stick shaped series can be made in a low-tech way simply by cherrypicking series and averaging them. If there are flawed proxies in a network, the MBH98 multivariate method will tend to concentrate weight on them as well. We have never argued that the MBH PC methodology is the *only* way to create a hockey stick shaped index from flawed proxies. In the case of the "supporting" studies, the statistical issue is much lower tech; the populations are very small and the primary issues are proxy validity and selection bias.

We sometimes hear that climate science has "moved on" from the MBH hockey stick and associated issues and that consideration of it should therefore end. We disagree for several obvious reasons.

First, the study continues to be in active use. The MBH99 reconstruction, as well as the Mann and Jones [2003], reconstruction are both nearly always used in spaghetti graphs showing collections of climate reconstructions, as shown in the example below [Wikipedia; also see Kerr, 2005].

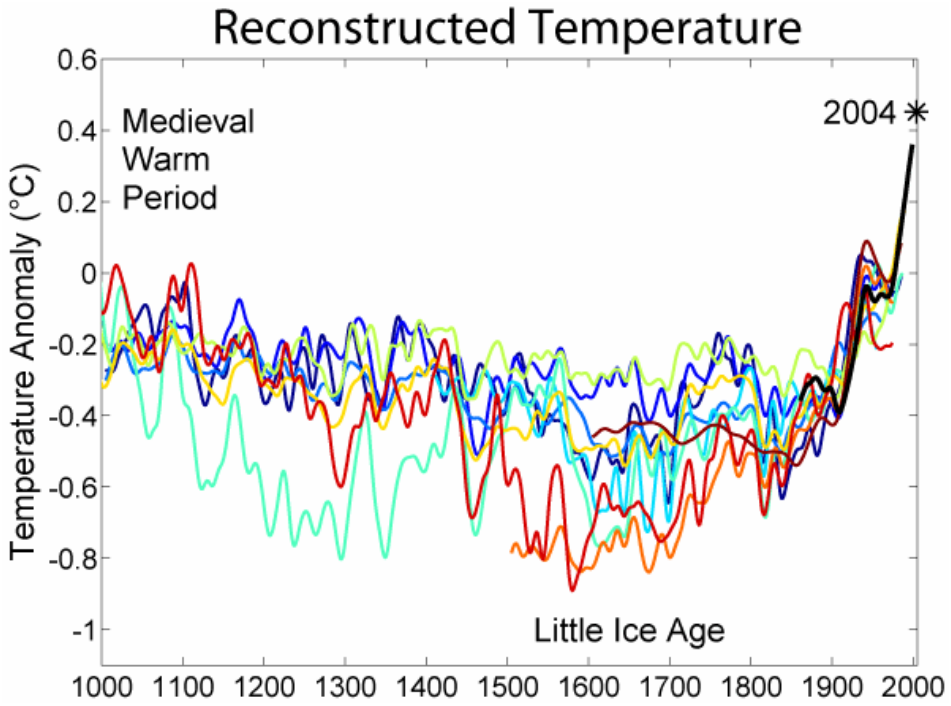

Figure 17. *Spaghetti graph of millennial temperature reconstructions. Source Wikipedia, Feb. 2006*

Second, MBH has been directly used to benchmark other studies. For example, Mann and Jones [2003], while purporting to be a different method, benchmarked against MBH98-99. Virtually all subsequent multiproxy studies benchmark themselves against MBH, which thereby has almost certainly influenced proxy selection in these later studies. This may even extend to any detection and attribution studies which have been influenced by MBH98-99.

Even the principal components methodology, which has been vociferously criticized, has continued in use in prominent studies: Mann and Jones [2003], Jones and Mann [2004]. One of the networks in Rutherford et al. [2005] (coauthored by the MBH authors) was identical to the MBH98 network, including the identical PC series. Only a few weeks ago, Osborn and Briffa [2006] used the North American PC1 as one of only 14 proxies in a *Science* article.

Given the particular attention of our articles to the problems with bristlecones as a temperature proxy, "moving on" from MBH would clearly require renunciation of bristlecones – something that has obviously not happened.

For us, in Canada, there is a significance quite different than any in the United States, as there was heavy reliance on this study for policy purposes. Thus, if there are serious defects in the study, and, especially if relevant information was not reported, that is of particular and ongoing concern.

We will leave behind a list of various unresolved issues as Appendix 1, which the panel is in an ideal position to resolve.

APPENDIX A

PRIMARY QUESTIONS

1. Is the MBH98 PC method biased towards yielding a hockey stick shaped PC1?
2. MBH have only reported a spliced reconstruction; the unspliced versions need to be disclosed.
3. What are the verification r^2 statistics as calculated by MBH for each step? Did MBH98 omit to report these results? What do they imply about the statistical skill of the early portion of the hockey stick graph? Did the omission of these results lead to inaccurate representation of the research record?
4. Is the claim that their results are robust to the "removal of all dendroclimatic indicators" true? Or do the removal of bristlecones lead to substantially different results in the 15th century? Do the calculations contained in the directory *ftp://holocene.evsc.virginia.edu/pub/MBH98/TREE/ITRDB/NOAMER/BACKTO_1400-CENSORED/* show that Mann et al were aware that removal of bristlecones led to substantially different results? If so, did the claim that their reconstruction was robust to the "removal of all dendroclimatic series" lead to an inaccurate representation of the research record?
5. Are the Graybill-Idso bristlecone pine series valid climate proxies through to 1980? If they are not, should they be used in paleoclimate studies?
6. Is there any plausible physical mechanism for the MBH99 CO2 adjustment?
7. Was any adjustment actually carried out on the MBH98 network? If not, why not?
8. In the on-line SI to the *Nature* Corrigendum of July 2004, Mann et al., referring to their use of a decentered PC methodology, state: "The results are not sensitive to this step." In the printed portion of their Corrigendum they also deny the errors affected their previously published results. Was this claim peer reviewed? Is it true?
9. Mann et al. 2000 claimed to have applied a list of quality control rules when selecting proxy records. What were the rules applied, and is there any documentation to show that all the series eventually selected passed the tests?
10. Is the use of calibration period residuals an appropriate method of calculating confidence intervals in a "new" multivariate method with serious risk of overfitting? What is the impact on MBH98-99 confidence intervals of using verification period residuals?
11. Is it prudent to analyze a multivariate statistical model without considering the verification r^2 statistic? What is the large-population behavior of temperature-sensitive tree ring sites in the last half of the 20th century? Is the behavior of proxies selected into canonical multiproxy studies consistent with random selection from the larger population?
12. Was the truncation of the Briffa MXD series by IPCC TAR misleading? Have Briffa et al. provided a valid justification for truncating this series after 1960?
13. Are there objective proxy selection criteria in any of the multiproxy studies?
14. What are the statistical properties of the "new" multivariate method implemented in MBH98? Are there precedents in the statistical literature? When new estimators are proposed, what kinds of asymptotic properties and limiting distributions for parameter estimates need to be published to establish the properties and desirability of the estimators?
15. Many multiproxy studies (e.g. Mann et al. 1998-1999, Hegerl et al. 2006) apply methods that equate to linear regression on time series data, yet routinely fail to report basic time series regression diagnostics, like coefficient t-statistics, unit root tests, Durbin-Watson statistics, tests for ARCH

residuals and so forth. What is a reasonable minimum list of time series diagnostics that should be reported in any application of regression methods on time series data?

16. Do journals have adequate policies for archiving data and methods? Do they have adequate systems in place to enforce existing policies?
17. Does the National Science Foundation have adequate policies to ensure archiving data and methods in compliance with USGRCP policies? Does it have adequate policies to enforce its existing policies? Do other agencies funding paleoclimate e.g. DOE, NOAA? Does IPCC?
18. Can all or any of the leading multiproxy studies be replicated based on existing information on data and methods?

FURTHER QUESTIONS

19. Were the conclusions of Briffa et al 1995 about a cold early 11th century sustained by additional information from the Polar Urals obtained in 1998? If not, were the new results reported by Briffa et al?
20. Is the adjustment to the widely used Tornetrask temperature reconstruction of Briffa et al [1992] valid? Should the existence of this adjustment, together with potential sensitivity, have been disclosed in Jones et al [1998]?
21. Did the 11th century portion of the Polar Urals series of Briffa et al 1995 meet quality control standards for replication? If not, should that have been disclosed? Are there cross-dating problems in this portion that might invalidate this portion of the reconstruction?
22. Are the results for the Polar Urals site of Esper et al [2002] consistent with the results of Briffa et al [1995]? For example, Briffa et al 1995 concluded that 1032 was the "coldest" summer of the millennium. Did Esper et al 2002 arrive at different conclusions? Have these differences ever been analysed and reconciled?
23. Are the results between the two different Polar Urals samples widely divergent? If so, how can one assign any confidence to site chronologies? Can the procedures for confidence interval estimation of Wigley et al 1984 still be endorsed?
24. Were the results from the Yamal site, as re-processed by Briffa, different again from the Polar Urals site of Esper et al 2002? What accounts for the differences?
25. Briffa discontinued use of the Polar Urals site in favor of the Yamal site. What is the justification for this substitution?
26. Is it possible that the reason for declining ring widths and densities in large-population samples of temperature-sensitive sites is due to a nonlinear and non-monotonic relationship between temperature and ring widths (densities)? Is there specialist literature to support this? If so, can information from the past be unambiguously interpreted?
27. In order to carry out RCS procedures to create long tree ring chronologies, is it necessary that the populations be homogeneous? Is this the case for the long chronologies used in multiproxy studies?
28. Are there significant altitude fluctuations at any of the sites? Were medieval treelines higher at any sites used in multiproxy studies? Do dendrochronologists record the altitude of samples? What procedures do dendrochronologists carry out to adjust for altitude fluctuations? Is it possible that "drift" or bias exists as a result of lack of homogeneity?
29. Are there significant changes in age composition in any of the long chronologies? Is the age composition of modern samples different from the age composition of (say) medieval samples? Is it

possible that bias between medieval and modern site chronology values exists as a result of inappropriate age adjustment?

30. What is "Modern Sample Bias"? Does it indicate a form of inhomogeneity that might affect comparisons between the medieval and modern periods?
31. What is "pith offset bias"? Does it indicate a form of inhomogeneity that might affect comparisons between medieval and modern periods?
32. What were the conclusions of Naurzbaev et al [2004] and Shiyatov [1995] about changing tree lines in Siberia? Are the conclusions of these studies consistent with the temperature histories of the site chronologies for Polar Urals [Briffa et al, 1995] and Yamal [Briffa 2000]? If not, have any attempts been made to reconcile the contrasting views?
33. Have there been changes in treeline at bristlecone sites? What are these changes and what do they signify for climate reconstructions?
34. What is the "amount effect" in respect to δO18 in tropical glacier ice cores? Is it possible or even probable that changing δO18 levels in tropical ice cores is due to changing amounts of summer precipitation as opposed to changing annual temperature? What would be the impact on the Dunde and Guliya ice cores? What would be the impact on the China composite of Yang et al 2002?
35. Are there different and inconsistent grey versions of the Dunde δO18 ice core in circulation in multiproxy studies? What accounts for the differences between these versions? Should the original author (Thompson) archive his sample isotope and chemical data for these series so that the inconsistencies can be reconciled?
36. Is the statistical argument of Thompson et al [1993] supposedly validating use of Dunde δO18 as a temperature proxy a valid argument?
37. in addition to the bristlecones, are any other series used repetitively in multiproxy studies? If so, which ones? Is this consistent with unbiased selection protocols?
38. Do multiproxy studies (and, in particular Moberg et al, 2005) adequately control for non-normality?

APPENDIX B – REPLICATION ISSUES

Study	Replication Issues
MBH98-99	Actual reconstructions through calibration period for steps prior to 1820 Cross-validation statistics: R2 Procedure for calculating confidence intervals
Jones et al 1998	A reasonable facsimile of the data has been assembled either from other sources or (in two cases) was supplied by Prof Jones. Exact replication has been unsuccessful – reasons are unknown. Source code and data as used are not archived and authors refused to supply. Replication is close enough to permit sensitivity analyses. This data is closely related to MBH data as 14 of 17 series are also used in MBH. The overlaps extends even to the use of some identical grey versions.
Crowley and Lowery 2000	The original data set has been lost by Dr Crowley, who, after months of email, finally located a smoothed and transformed version of the data, which is useful but not satisfactory. Some digital versions do not match paper citations and Dr Crowley was unable to recall the source.
Briffa [2000] (RW))	The measurement data for 4 of 7 series is unarchived.
Briffa et al [2001] (MXD)	The identity of the 387 sites is not provided. The non-identification of sites affects other papers by Briffa in many different journals.
Esper et al [2002]	Site chronologies became available only in Feb. 2006 after extended correspondence with Nature. In some cases, site chronologies do not match archived measurement data. In some cases, no measurement data is archived. The methodology is presently unintelligible as no operational definition of a distinction between "linear" and "nonlinear" trees is provided.
Mann and Jones [2003]	No archive of data, although many series overlap with Jones and Mann [2004] which does have an archive. No information or methodology for weights is provided and Prof Jones was unable to provide the weights, as he did not have the information. The North American PC1 is not properly documented, as the AD200 PC1 using the MBH method is spliced with the already adjusted AD1000 PC1.
Jones and Mann [2004]	Archive of all data except Law Dome.
Moberg et al [2005]	Two series were not available. Only after a Materials Complaint to Nature were these two series provided in Feb. 2006. The methodology is not standard and no source code is provided. We have achieved substantial but not exact replication without the missing series.
Osborn and Briffa [2006]	Only smoothed versions of sites provided. Some series could be matched to original sources.

REFERENCES

Ammann, C. and E. Wahl (2006), Comment on ‘‘Hockey sticks, principal components, and spurious significance’’ by S. McIntyre and R. McKitrick, *under review*.

Anderson, R. W. H. Greene, B. D. McCullough and H. D. Vinod, 2004. The Role of Data/Code Archives In the Future of Economic Research, Paper prepared for presentation at the American Economic Association meeting, Philadelphia PA, January 9, 2005. www.aeaweb.org/annual_mtg_papers/2005/0109_1300_0303.pdf

Arctic Climate Impact Assessment, 2004. *Impacts of a Warming Arctic: Arctic Climate Impact Assessment*, Cambridge University Press.

Bernanke, Ben S. (2004). “Editorial Statement,” American Economic Review, 94(1),404.

Biondi, F., D.L. Perkins, D.R. Cayan and M.K. Hughes, 1999. July temperature during the second millennium reconstructed from Idaho tree rings. *GRL* 26, 1445-1448.

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

Briffa, K.R., 2000. Annual climate variability in the Holocene: interpreting the message of ancient trees. *Quat . Sci. Rev.* 19, 87-105.

Briffa, K.R. and T.J. Osborn, 1999. Climate Warming: Seeing the Wood from the Trees, *Science* 284, 926.

Briffa, K.R., Jones, P.D. and Schweingruber, F.H., 1992a. Tree-ring density reconstructions of summer temperature patterns across western North America since A.D.1600, *Journal of Climate* 5, 735-754.

Briffa, K.R., Jones, P.D., Bartholin, T.S., Eckstein, D., Schweingruber, F.H, Karlen, W., Zetterberg, P. and Eronen, M., 1992b. Fennoscandian summers from A.D.500: temperature changes on short and long timescales. *Climate Dynamics* 7, 111-119.

Briffa, K.R., Jones, P.D., Schweingruber, F.H., Shiyatov, S.G. and Cook, E.R., 1995. Unusual twentieth-century summer warmth in a 1,000-year temperature record from Siberia. *Nature* 376, 156-159.

Briffa, K.R., Schweingruber, F.H., Jones, P.D., Osborn, T.J., Shiyatov, S.G. and Vaganov, E.A. 1998a: Reduced sensitivity of recent tree-growth to temperature at high northern latitudes. *Nature* 391, 678–82.

Briffa, K.R., Jones, P.D., Schweingruber, F.H. and Osborn, T.J., 1998b. Influence of volcanic eruptions on Northern Hemisphere summer temperature over the past 600 years. *Nature* 393, 450-455.

Briffa, K.R., F. H. Schweingruber, P. D. Jones, T. J. Osborn, I. C. Harris, S. G. Shiyatov, E. A. Vaganov and H. Grudd, 1998c. *Phil.Trans. R. Soc. Lond. B* 353, 65-73

Briffa, K.R., Osborn, T.J., Schweingruber, F.H., Harris, I.C., Jones, P.D., Shiyatov, S.G., Vaganov, E.A., 2001. Low-frequency temperature variations from a northern tree ring density network. *Journal of Geophysical Research* 106, 2929– 2941.

Briffa, K.R., Osborn, T.J., Schweingruber, F.H., Jones, P.D., Shiyatov, S.G., Vaganov, E.A., 2002. Tree-ring width and density around the Northern Hemisphere: Part 1. Local and regional climate signals. *Holocene* 12, 737–757.

Briffa KR, Osborn TJ and Schweingruber FH (2004) Large-scale temperature inferences from tree rings: a review. *Global and Planetary Change* 40, 11-26 (doi:10.1016/S0921-8181(03)00095-X).

Bunn, A.G., R.L. Lawrence, G.J. Bellante, L.A. Waggoner, and L.J. Graumlich, Spatial variation in distribution and growth patterns of old growth strip-bark pines. *Arctic, Antarctic, and Alpine Research*, 2003, 35:323-330.

Bürger, G., and U. Cubasch (2005), Are multiproxy climate reconstructions robust?, *GRL*, 32, L23711, doi:10.1029/2005GL024155.

Ciccerone, R., 2005. Letter to House Energy and Commerce Committee, July 15, 2005. http://www.realclimate.org/Cicerone_to_Barton.pdf

Cook, E.R. and Peters, K. 1997. Calculating unbiased tree-ring indices for the study of climatic and environmental change. *The Holocene* 7(3):359-368.

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

Crowley, T.J. and Lowery, T.S., 2000. How warm was the Medieval warm period? *Ambio* 29, 51-54.

Dahl-Jensen, D., K. Mosegaard, N. Gundestrup, G. D. Clow, S. J. Johnsen, A. W. Hansen, N. Balling, 1998. Past Temperatures Directly from the Greenland Ice Sheet, *Science* 282, 268-271.

D'Arrigo, R. D., Robert K. Kaufmann, Nicole Davi, Gordon C. Jacoby, Cheryl Laskowski, Ranga B. Myneni and Paolo Cherubini, 2004. Thresholds for warming-induced growth decline at elevational tree line in the Yukon Territory, Canada. *Global Biogeochemical Cycles*, 18, GB3021, doi:10.1029/2004GB002249.

D'Arrigo, R.D. and Jacoby, G.C., 1992 in *Climate Since A.D. 1500,* (eds. Bradley, R.S. & Jones, P.D., 246-268, Routledge.

Dewald, W.G., J.G. Thursby, and R.G. Anderson, 1986. Replication in empirical economics: The journal of money, credit and banking project. *American Economic Review*, 76(4):587-603.

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.

Ferson, W., S. Sarkissian and T Simin, 2003. Spurious regressions in financial economics, *Journal of Finance,* 58(4), 1393-1413;

Goldscheider, E., 2005, Never Mind the Weather?, *UMassMag Online*, Retrieved from www.umassmag.com/Fall_2005/Never_Mind_the_Weather__934.html

Graumlich, L.J., 1991. Subalpine tree growth, climate, and increasing CO2: an assessment of recent growth trends. *Ecology* 72: 1-11;

Granger, C.W. and P. Newbold, 1974. Spurious Regressions In Econometrics, *Journal of Econometrics* 2, 111-120.

Graybill, D.A., and S.B. Idso. 1993. Detecting the aerial fertilization effect of atmospheric CO2 enrichment in tree-ring chronologies. *Global Biogeochemical Cycles* 7:81-95.

Hegerl, Gabrielle, Thomas J. Crowley, Myles R. Allen , William T. Hyde, Henry N. Pollack, Jason E. Smerdon & Eduardo Zorita, 2006 (submitted). A new 1500 yr climate reconstruction: enhanced low-frequency variability and the fingerprint of anthropogenic warming.

House Energy and Commerce Committee, 2005. Letter to Michael Mann, June 23, 2005. http://energycommerce.house.gov/108/Letters/062305_Mann.pdf

Hughes, M. K., and H. F. Diaz, 1994. Was there a "Medieval Warm Period", and if so, where and when? *Climatic Change*, 26, 109-142.

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

Huybers, P. (2005), Comment on "Hockey sticks, principal components and spurious significance" by McIntyre and McKitrick, *GRL*, L20705, doi: 10.1029/2005 GL023395.

Intergovernmental Panel on Climate Change, 1990. Climate Change. The IPCC Scientific Assessment. J.T. Houghton, G.J. Jenkins, and J.J. Ephrauns (editors). Cambridge University Press, Cambridge, United Kingdom

Intergovernmental Panel on Climate Change, 1996. Climate Change 1995: The Science of Climate Change, edited by Houghton, J.T., L.G. Meira Filho, B.A. Callander, N. Harris, A. Kattenberg, and K. Maskell. Cambridge University Press, Cambridge, UK. 572 pp

Intergovernmental Panel on Climate Change, 2001. Climate Change 2001: The Scientific Basis. Retrieved from http://www.grida.no/climate/ipcc_tar/.

Jacoby, G.C. and R.D. D'Arrigo, 1989. Reconstructed northern hemisphere annual temperature since 1671 based on high-latitude tree-ring data from North America, *Climatic Change* 14, 39-59.

Jacoby, G. C. and R.D. D'Arrigo, 1997. Tree rings, carbon dioxide, and climatic change, *Proc. Natl. Acad. Sci. USA* 94, 8350-8353.

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

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., T. J. Osborn, and K. B. Briffa, 2001. The Evolution of Climate Over the Last Millennium, *Science,* 292, 662– 667.

Kerr, Richard A., 2005: Millennium's Hottest Decade Retains Its Title, for Now. *Science* 307 (5711), 828-829.

King, Gary, 1995. Replication, Replication, *PS: Political Science and Politics*, with comments from nineteen authors and a response, A Revised Proposal, Proposal, 28(3), 443-499.

LaMarche, V.C., D.A. Graybill, H.C. Fritts, and M.R. Rose., 1984. Increasing atmospheric carbon dioxide: tree ring evidence for growth enhancement in natural vegetation. *Science* 225:1019-1021.

Lloyd, AH and LJ Graumlich. 1997. Holocene dynamics of treeline forests in the Sierra Nevada. *Ecology* 78, 1199-1210.

Mann, Michael E. (2003). Senator Inhofe's Follow-up Questions for Dr. Michael Mann. Retrieved from http://cfa-www.harvard.edu/~wsoon/July29-2003-EPWtestimony-d/MannInhofeQuestions-answers.pdf

Mann, M.E. 2005, Letter to Barton, http://www.realclimate.org/Mann_response_to_Barton.pdf

Mann, M.E. and Jones, P.D., Global surface temperature over the past two millennia, Geophysical *Research Letters*, 30 (15), 1820, doi: 10.1029/2003GL017814, 2003.

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., Northern Hemisphere Temperatures During the Past Millennium: Inferences, Uncertainties, and Limitations, *Geophysical Research Letters*, 26, 759-762, 1999.

Mann, M.E., E. Gille, R.S. Bradley, M.K. Hughes, J.T. Overpeck, F.T. Keimig, and W. Gross. 2000. Global temperature patterns in past centuries: An interactive presentation. *Earth Interactions* 4-4:1-29. Retrieved from NOAA website at http://www.ngdc.noaa.gov/paleo/ei, which includes additional note http://www.ngdc.noaa.gov/paleo/ei/ei_nodendro.html.

Mann, M.E., Bradley, R.S. and Hughes, M.K., 2003. Note on Paper by McIntyre and McKitrick in "Energy And Environment". Retrieved from <ftp://holocene.evsc.virginia.edu/pub/mann/EandEPaperProblem.pdf>

Mann, M.E., Bradley, R.S. and Hughes, M.K., 2004b. Corrigendum: Global-scale temperature patterns and climate forcing over the past six centuries, *Nature* **430**, 105(2004).

Mann, M.E., Bradley, R.S. and Hughes, M.K., 2004a. False Claims by McIntyre and McKitrick regarding the Mann et al. (1998) reconstruction. Retrieved from website of realclimate.org at < http://www.realclimate.org/index.php?p=8>

McCullough, B.D. and H. D. Vinod, 2003. Verifying the Solution from a Nonlinear Solver: A Case Study, *American Economic Review* 93, 873-892 with comments and replies (2004) at *American Economic Review* 94, 391-403.

McIntyre, S. 2005a. "More on Hockey Sticks: the Case of Jones et al [1998]" Poster presentation to U.S. Climate Change Workshop, Nov 14, 2005, Arlington Virginia. www.climateaudit.org/pdf/mcintyre.workshop05.pdf

McIntyre, S. 2005b. Low-Frequency Ranges in Multiproxy Climate Reconstructions, Presentation to the AGU Fall Meeting, December 2005. www.climateaudit.org/pdf/agu05.ppt

McIntyre, S. and R. McKitrick, 2003. "Corrections to the Mann et. al. (1998) Proxy Data Base and Northern Hemispheric Average Temperature Series" *Energy and Environment* 14, 751-771.

McIntyre, S. and R. McKitrick, 2004a. Global-scale temperature patterns and climate forcings over the past six centuries: a comment. . Retrieved from http://www.uoguelph.ca/~rmckitri/research/fallupdate04/submission.1.final.pdf

McIntyre, S. and R. McKitrick, 2004b. Global-scale temperature patterns and climate forcings over the past six centuries: a comment. . Retrieved from http://www.uoguelph.ca/~rmckitri/research/fallupdate04/MM.resub.pdf

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

McIntyre, S. and R. McKitrick (2005b), The M&M Critique of the MBH98 Northern Hemisphere Climate Index: Update and Implications, *Energy and Environment*, 16, 69-99.

McIntyre, S., and R. McKitrick (2005c), Reply to Comment by Von Storch and Zorita, *GRL*, 32, L20713, doi:10.1029/2005GL023089.

McIntyre, S., and R. McKitrick (2005d), Reply to Comment by Huybers, *GRL*, 32, L20713, doi:10.1029/2005GL023586.

McIntyre, S. and R. McKitrick, (2006), Reply to Comment by Ammann and Wahl (under review at *GRL*). Available at www.climateaudit.org.

Moberg, Anders, Dmitry M. Sonechkin, Karin Holmgren, Nina M. Datsenko and Wibjörn Karlén, 2005. Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data, *Nature*, 433 (7026), 613-617.

Naurzbaev, Mukhtar M., Malcolm K. Hughes and Eugene A. Vaganov, 2004. Tree-ring growth curves as sources of climatic information, *Quaternary Research* 62, 126– 133.

North, G.R., F. J. Moeng, T. J. Bell and R. F. Cahalan, 1982: Sampling Errors in the Estimation of Empirical Orthogonal Functions. *Mon. Wea. Rev.,* 110, 699-706

Osborn, Timothy J. and Keith R. Briffa, 2006, The Spatial Extent of 20th-Century Warmth in the Context of the Past 1200 Years, *Science* 311, 831-834.

Osborn., T.J., K. R. Briffa et al. , 2005, submitted to GPC http://www.cru.uea.ac.uk/~timo/papepages/osborn_summertemppatt_submit2gpc.pdf

Osborn TJ, Briffa KR, Schweingruber FH and Jones PD (2006) Annually resolved patterns of summer temperature over the Northern Hemisphere since AD 1400 from a tree-ring-density network. Submitted to *Global and Planetary Change.*

Overland, J.E. and R.W. Preisendorfer, (1982), A significance test for principal components applied to a cyclone climatology, *Mon. Weather Rev.,* 110, 1-4.

Phillips, P., 1986. Understanding Spurious Regressions in Econometrics. *Journal of Econometrics*, 33, 1-30. http://cowles.econ.yale.edu/P/cp/p06b/p0667.pdf

Preisendorfer, R.W. (1988), Principal Component Analysis in Meteorology and Oceanography, Elsevier.

Rencher, A.C. (1992), Interpretation of canonical discriminant functions, canonical variates and principal components, The American Statistician 46, 217-225.

Rencher, A. C., (2002). Methods of multivariate analysis, 2d edition. John Wiley and Sons.

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, Jou*rnal of Climate*, 18, 2308-2329, 2005.

Shiyatov, S.G., 1995. Reconstruction of climate and the upper treeline dynamics, *Publications of the Academy of Finland* 6/95, 144-147.

UCAR, 2005. Media Advisory: The Hockey Stick Controversy: New Analysis Reproduces Graph of Late 20th Century Temperature Rise, May 11, 2005 http://www.ucar.edu/news/releases/2005/ammann.shtml

von Storch, H., and E. Zorita (2005), Comment on ‘‘Hockey sticks, principal components, and spurious significance’’. by S. McIntyre and R. McKitrick, *Geophys. Res. Lett*., 32, L20701, doi:10.1029/2005GL022753.

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

Wahl, Eugene R. and Caspar M. Ammann, 2006 (under review). Robustness of the Mann, Bradley, Hughes Reconstruction of Surface Temperatures: Examination of Criticisms Based on the Nature and Processing of Proxy Climate Evidence.

Wigley, T.M.L., Briffa, K.R. and Jones, P.D., 1984. On the average value of correlated time series with applications in dendroclimatology and hydrometeorology. *Journal of Climate and Applied Meteorology* 23, 201-213

Wigley, T.M.L., P.D. Jones and K.R. Briffa, 1987. Cross-dating Methods in Dendrochronology. *J. Arch. Sci*. 14, 51-64.

Yang Bao, Achim Braeuning, Kathleen R. Johnson and Yafeng Shi, 2002, General characteristics of temperature variation in China during the last two millennia. *GRL* 10.1029/2001GL014485.

Zorita, E., F. González-Rouco and S. Legutke, 2003. Testing the Mann et al. (1998) Approach to Paleoclimate Reconstructions in the Context of a 1000-Yr Control Simulation with the ECHO-G Coupled Climate Model . *Journal of Climate* 16, 1378-1390.