A BRIEF RETROSPECTIVE ON THE HOCKEY STICK*

Ross McKitrick
Department of Economics
University of Guelph
May 23, 2014

* For inclusion in the compendium volume *Climate Change: The Facts 2014*, Institute for Policy Analysis, Australia.

1 THE CORE ISSUES

The best place to start when learning about the hockey stick is Andrew Montford's superb book *The Hockey Stick Illusion* (Montford 2010). Other essential sources are the original Mann et al. papers (denoted MBH98 and MBH99), the M&M papers (McIntyre and McKitrick 2003, 2005a—d), Steve McIntyre's and my presentation to the National Academy of Sciences Panel (McIntyre and McKitrick 2006), Steve's Ohio State University presentation,[1] a few survey papers and chapters of mine (McKitrick 2005, 2006), and Steve's climateaudit.org posts over the past decade on proxy quality, the Yamal substitution, the Briffa truncation, data secrecy, and some other issues.

It is sometimes said that we found Mann's algorithm would always produce a hockey stick, even from random numbers. That is not quite right: we found that it *could* do so, given the right kind of random numbers (autocorrelated, rather than independent). We also found that it mined for hockey stick shapes and overstated their dominance in the underlying data patterns, and that it understated the uncertainties of the resulting climate reconstruction (or equivalently, exaggerated the significance).

A very brief summary of the problems of the hockey stick would go like this. Mann's algorithm, applied to a large proxy data set, extracted the shape associated with one small and controversial subset of the tree rings records, namely the bristlecone pine cores from high and arid mountains in the US Southwest. The trees are extremely long-lived, but grow in highly contorted shapes as bark dies back to a single twisted strip. The scientists who published the data (Graybill and Idso 1993) had specifically warned that the ring widths should not be used for temperature reconstruction, and in particular their 20th century portion is unlike the climatic history of the region, and is probably biased by other factors.

Mann's method exaggerated the significance of the bristlecones so as to make their chronology out to be the dominant global climatic pattern rather than a minor (and likely inaccurate) regional one; Mann then understated the uncertainties of the final climate reconstruction, leading to the claim that 1998 was the warmest year of the last millennium, a claim that was not, in reality, supportable in the data. Furthermore, Mann put obstacles in place for subsequent researchers wanting to obtain his data and replicate his methodologies, most of which were only resolved by the interventions of US Congressional investigators and the editors of *Nature* magazine, both of whom demanded full release of his data and methodologies some six years after publication of his original *Nature* paper.

[1] Online at http://climateaudit.files.wordpress.com/2005/09/ohioshort.pdf.

Mann had re-done his hockey stick graph at some point during its preparation with the dubious bristlecone records excluded and saw that the result lost the hockey stick shape altogether, collapsing into a heap of trendless noise. However he never pointed this out to readers. He also stated that he had computed test scores called r^2 statistics that he said (or implied) confirmed the statistical significance of his results, yet when the r^2 scores were later revealed they showed no such thing; and by then he had taken to denying he had even calculated them.

2 OUR CRITIQUE OF THE METHOD

There are two key parts to the hockey stick-making machine. The first is the principal components (PC) step, and the second is the least squares (LS) fitting step. The PC step takes large numbers of temperature proxies and compiles them into a relatively small number of composite series. The LS step then lines up the final segment of the composites against an upward-sloping temperature graph and puts weight on them in proportion to how well they correlate. If there are many composites and only one has a hockey stick shape, the LS step will find it and put most of the weight on it. If none of the composites has a hockey stick shape, then the LS step will come up blank and the resulting graph will just look like noise.

Mann's PC step was programmed incorrectly and created two weird effects in how it handled data. First, if the underlying data set was mostly random noise, but there was one hockey stick-shaped series in the group, the flawed PC step would isolate it out, generate a hockey stick composite and call it the dominant pattern, even if it was just a minor background fluctuation. Second, if the underlying data consisted of a particular type of randomness called "red noise"—basically randomness operating on a slow, cyclical scale—then the PC step would rearrange the red noise into a hockey stick-shaped composite. Either way, the resulting composites would have a hockey stick shape for the LS step to glom onto and produce the famous final result.

The use of red noise series is necessary for testing the statistical robustness of the hockey stick method. This is a procedure called Monte Carlo analysis. For one of our 2005 papers (McIntyre and McKitrick 2005b) we generated thousands of series of trendless autocorrelated random numbers and ran them through the PC and LS steps. This generated thousands of results, each of which had an index of accuracy called the "RE" score (for Reduction of Error). Likewise the actual proxy data had an associated RE score. We set a benchmark based on the idea that, if the proxy data were actually informative about the real world they had to yield a higher RE score than most of the (uninformative) artificial data. Mann had done the same thing, but had not taken into account the effect of the erroneous PC method. The real proxy data didn't turn out to be more informative than red noise, but he set his benchmark too low, making his proxy results look statistically significant when in reality they weren't.

There was a big red flag in his calculations that should have tipped him off. Another model test is called the r^2 score. It has the nice feature that you don't need to do Monte Carlo simulations, it has standard benchmark tables available in any statistics textbooks (as MBH98 note on p. 786). While Mann reported the (favourable) r^2 scores for the later portion of his graph (MBH98 p. 781-782; see his Figure 3), he didn't mention them for the early portion (pre-1750), where they were nearly zero, indicating a lack of statistical significance. Instead he only reported the RE score, which he thought indicated significance. He showed the reader the RE test that he thought (incorrectly) was

favourable, yet he kept referring to significance *tests* in the plural[2] in support of his claims, so the reader would naturally assume the unreported r^2 scores looked good too.

They didn't, but he failed to report that in the article. And as we later showed, the r^2 and RE scores were actually saying the same thing, namely that the hockey stick was uninformative as an indicator of past temperatures.

3 STICKHANDLING

In 2005, following an article on the dispute in the Wall St Journal, Mann had been sent a list of questions by the Energy and Commerce Committee of the US Congress, one of which was whether he had computed the r^2 score. His answer was:[3]

> 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.

The answer is classic misdirection. He was not asked, "Did you rely on the r^2 score when assessing your results?" There was no need to ask that: if he *had* relied on it he would never have claimed his results were significant! He only claimed significance by *ignoring* it. The question specifically was whether he computed r^2. Tellingly, in his reply he changed the subject. But it hardly matters. Either he did not compute it, in which case he was lying in the paper by saying he had, or he did, in which case his failure to disclose it was misleading to his readers.

When we and Mann appeared before the National Academy of Sciences panel in 2006 we presented this issue in detail. We showed the panel that the Supplementary Information to MBH98 did not report the verification r^2 scores , and we urged them to ask Mann whether he had computed them. John Christy of the University of Alabama-Huntsville put the question to him. To our astonishment, Mann point-blank denied having done so, claiming it would be "silly and incorrect reasoning."[4] Mann then launched an extraordinary tirade against r^2, a well-understood statistic which is found in every statistics textbook and is the workhorse of model testing.

After the NAS panel hearings we wrote a letter to the chair, Gerald North,[5] expressing our frustration that they allowed Mann to get away with this, but we were not successful in getting them to follow up on the matter.

4 THE NAS REPORT

[2] E.g.: "For comparison, correlation (r) and squared-correlation (r^2) statistics are also determined." (MBH98 p. 785.)

[3] The text of the questions and answers were in our NAS brief, see http://www.uoguelph.ca/~rmckitri/research/NAS.M&M.pdf.

[4] See detailed notes of the hearing at http://climateaudit.org/2006/03/16/mann-at-the-nas-panel/.

[5] Posted online at http://www.rossmckitrick.com/uploads/4/8/0/8/4808045/nas-followup-mm.pdf.

More evidence that the Mann procedure exaggerated the statistical significance of his results came in the NAS Panel Report itself (North et al. 2006). While they lost the plot on the r^2 issue, they did at least look at the overall question of how to assess a statistical climate reconstruction. They came up with the most elliptical way possible to say that the Mann hockey stick was unreliable. Here is what they said:

> Reconstructions that have poor validation statistics (i.e., low CE) will have correspondingly wide uncertainty bounds, and so can be seen to be unreliable in an objective way. Moreover, a CE statistic close to zero or negative suggests that the reconstruction is no better than the mean, and so its skill for time averages shorter than the validation period will be low. Some recent results reported in Table 1S of Wahl and Ammann (in press) indicate that their reconstruction, which uses the same procedure and full set of proxies used by Mann et al. (1999), gives CE values ranging from 0.103 to –0.215, depending on how far back in time the reconstruction is carried.

(North et al. 2006 p. 91). Sir Humphrey Appleby could not have phrased it better. Unpeeling the obfuscations, here is what they said.

- Reconstructions can be assessed using a variety of tests, including RE, r^2 and the CE (Coefficient of Efficiency) scores.
- If the CE score is near zero or negative your model is junk.
- Wahl and Ammann include a Table in which they use Mann's data and code and compute the test scores that he didn't report.
- The CE scores range from near zero to negative, which tells us that Mann's results were junk.

Another exercise in obfuscation concerned the reliance on bristlecones. The NAS report said the following.

> Such trees [bristlecones] are sensitive to higher atmospheric CO2 concentrations (Graybill and Idso 1993), possibly because of greater water-use efficiency (Knapp et al. 2001, Bunn et al. 2003) or different carbon partitioning among tree parts (Tang et al. 1999). …While 'strip-bark' samples should be avoided for temperature reconstructions, attention should also be paid to the confounding effects of anthropogenic nitrogen deposition (Vitousek et al. 1997)…For periods prior to the 16th century, the Mann et al. (1999) reconstruction that uses this particular principal component analysis technique is strongly dependent on data from the Great Basin region in the western United States. Such issues of robustness need to be taken into account in estimates of statistical uncertainties.

North et al. (2006) (p. 50, 107). Stripping away the bark, here is what this means:

- Bristlecone records are sensitive to a variety of environmental conditions other than temperature and should be avoided for climate reconstructions.
- Mann's results strongly depend on the bristlecone records.
- His results are therefore not robust, an important point over and above the lack of statistical significance.

The NAS report (North et al. 2006) also made a few other points, buried in elliptical prose or scattered around the report where the press would be sure of never finding them (not that they looked). Putting them together, they upheld all the claims in our submission.

- (p. 86—87) "McIntyre and McKitrick (2003) demonstrated that under some conditions, the leading principal component can exhibit a spurious trendlike appearance, which could then lead to a spurious trend in the proxy-based reconstruction."

- (p. 106) "As part of their statistical methods, Mann et al. used a type of principal component analysis that tends to bias the shape of the reconstructions." The Report even included its own graphical replication of the artificial hockey stick effect from feeding red noise into Mann's algorithm (p. 87).

- (p. 107) The usual RE significance benchmark "is not appropriate."

- (p. 107) "Uncertainties of the published reconstructions have been underestimated."

5 THE CENSORED FOLDER

Mann also published an online review article in 2000 (Mann et al. 2000) that assured readers in categorical terms that their results were "robust" to non-climatic bias in tree ring data[6] and even to the complete removal of tree rings from their data set, though they illustrated that point only for the post-1760 interval. In the course of our analysis, Steve found some directories at Mann's FTP site (the "CENSORED" directories), which, through detective work, were found to contain assessments of the impact from dropping the bristlecones from the underlying data. In light of the claim in Mann et al. (2000), this should not have made any difference, but it did. In our NAS presentation we showed graphs of the data in Mann's "CENSORED" results, in which the hockey stick shape completely disappears. That is, even applying Mann's biased methods, after dropping the few bristlecone pine series there is no remaining hockey stick shape. The claim in Mann et al. (2000) about robustness to the exclusion of the tree ring data was obviously misleading.

In the letter from the Congressional Oversight committee to Mann he was asked:[7]

> 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

[6] "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." http://www.ncdc.noaa.gov/paleo/ei/ei_datarev.html; "MBH98 found through statistical proxy network sensitivity estimates that skillful NH reconstructions were possible without using any dendroclimatic data…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." http://www.ncdc.noaa.gov/paleo/ei/ei_nodendro.html

[7] The question and answer excerpt are in our NAS presentation (McIntyre and McKitrick 2006).

> 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).

Translation: … Yes. When we removed them the graph collapsed and the statistical scores went to zero. Oh dear, didn't we mention that? Anyway, to avoid the problem, we kept them in.

Mann's claim that MBH99 addressed "precisely these issues" was misleading. In that paper they did mention that their top-weighted PC was "essential" (p. 760) but they didn't report the results of excluding the bristlecones. Instead they applied a "correction" that they claimed (without proof) fixed the contamination pattern in the bristlecones, even though it only applied to the 19th century portion. And their 2000 paper claimed robustness both to contamination of bristlecones and removal of tree ring data. Wahl and Ammann (2007) later offered the argument that since the hockey stick fails all statistical tests without the bristlecones they ought to be retained (the logic really was that bad). In our letter to North we pointed out that we agreed with Wahl and Ammann (and Mann) that the reconstruction without the bristlecones is no good. But, we added,

> Our contention is that the reconstruction **with** bristlecones is also no good, as evidenced by the failure of verification r^2 and CE statistics.

6 CONCLUSION

The story continued on from there and much more could be said. The intensity with which so many people have followed the story, and its continuing relevance via the ongoing Mann *v.* Steyn lawsuit (as well as others), indicate to me that it is more than just an academic spat about proxy quality and r^2 scores. I suspect that the whole episode has wider social significance as an indicator of a rather defective aspect of early 21st century scientific culture.

7 REFERENCES

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.

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., 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., 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.

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

McKitrick, Ross R. (2005) What is the Hockey Stick Debate About? Presentation to the Asia Pacific Economic Cooperation Study Centre Meeting on "Managing Climate Change - Practicalities and Realities in a post-Kyoto Future", Parliament House, Canberra Australia, April 4, 2005. Online at http://www.rossmckitrick.com/paleoclimatehockey-stick.html.

McKitrick, Ross R. (2006) The Mann et al. Northern Hemisphere “Hockey Stick” Climate Index: A Tale of Due Diligence in Michaels, Patrick, ed. *Shattered Consensus: The True State of Global Warming.* Rowman and Littlefield, 2006.

Montford, Andrew (2010) *The Hockey Stick Illusion* London: Stacey International.

North, G. et al. (National Research Council, NRC) (2006). Surface Temperature Reconstructions for the Last 2,000 Years. Washington: National Academies Press.

Wahl, Eugene and Caspar Ammann (2007) Robustness of the Mann, Bradley, Hughes reconstruction of Northern Hemisphere surface temperatures: Examination of criticisms based on the nature and processing of proxy climate evidence. *Climatic Change* 85:33–69, DOI 10.1007/s10584-006-9105-7.