

Science and Environmental Policy-Making: Bias-Proofing the Assessment Process

Ross McKittrick*

Department of Economics, University of Guelph, Guelph, ON, Canada N1G 2W1
(e-mail: rmckitri@uoguelph.ca).

Scientific assessment panels are playing increasingly influential roles in national and international policy formation. Although they typically appeal to the standard of journal peer review as their quality control criterion, there seems to be confusion about what peer review actually does. It is, at best, a necessary condition of reliability, but not a sufficient condition. There is also the problem that assessment panels may be biased in favor of one side or another when evaluating areas in which the science is unclear. In this paper I argue that additional checks and balances are needed on the information going into scientific assessment reports when it will be used to justify major policy investments. I propose two new mechanisms to bias-proof the outcome: an Audit Panel and a Counterweight Panel. The need for such mechanisms is discussed with reference to the "hockey stick" debate in climate change.

Les comités scientifiques d'évaluation jouent des rôles de plus en plus influents dans la formulation des politiques nationales et internationales. Bien que pour les revues scientifiques, la révision par les pairs représente un critère pour assurer le contrôle de la qualité, la confusion semble régner quant au rôle de cette révision par les pairs. Elle est, au mieux, une condition nécessaire pour assurer la fiabilité, sans être toutefois une condition suffisante. Il y a aussi le fait que les comités d'évaluation peuvent avoir un parti pris pour un aspect ou un autre lorsqu'ils évaluent des domaines où la science n'est pas claire. Dans le présent article, nous avons maintenu que l'information publiée dans les rapports d'évaluation scientifique devrait faire l'objet de vérifications supplémentaires lorsqu'elle est utilisée pour justifier d'importants investissements en matière de politique. Nous avons proposé deux mécanismes pour vérifier et contre-vérifier les résultats: créer un comité de vérification et un comité de contre-vérification. Nous avons discuté de la nécessité d'instaurer ce genre de mécanismes en faisant référence au débat sur les changements climatiques.

INTRODUCTION

My ancestors homesteaded in Southern Manitoba in 1879, and against long odds the farm is still going. The promoters who enticed settlers to western Canada in the 19th century used the slogan "rain follows the plow" to allay fears about the dryness of the prairies. Surprisingly, the paleoclimate record drawn from prairie lake beds (Leavitt et al 2002) suggests that the 20th century was unusually moist compared to the previous millennium. Therefore, perhaps the widespread changes in land use did change the prairie climate. But the concern many people have today is not whether the plow changed the climate but whether the tractor did: in other words whether climate change can be attributed to emissions from fossil fuel use. There are well-known arguments for believing so, though for reasons spelled out at length in my book *Taken By Storm* (2002, co-authored with Chris

*Note: Invited keynote address at the International Policy Forum on Greenhouse Gas Management, University of Victoria, Victoria, BC, Canada, April 28, 2005.

Essex), and in other writings, I think there are substantial problems with the idea that a global warming hazard exists and that it can be attributed to carbon dioxide emissions. Having argued this case over the past few years, while watching the global warming idea nevertheless gain steam and drive major policy undertakings, it comes as a surprise to find voices on the other side of the debate who attribute great influence to people on my side. Here is what environmentalist Bill McKibben had to say in *Mother Jones* magazine recently:

For now and for the foreseeable future the climate skeptics have carried the day . . . In short, the deniers have done their job, and done it better than the environmentalists have done theirs. They've delayed action for fifteen years now, and their power seems to grow with each year. How, even as the science grew ever firmer and the evidence mounted ever higher, did the climate deniers manage to muddy the issue? It's one of the mightiest political feats of our time, accomplished by a small group of clever and committed people.¹

However, tempting as it is to accept this flattery, I must demur. Mr. McKibben is confusing two different conspiracies, one which is rather hard to pull off and one which is quite easy. Arguing against the "consensus" position on global warming science is hard, and skeptics cannot (yet) claim to have "carried the day." Delaying the implementation of rapid cuts in greenhouse gas emissions, however, is trivially easy, for reasons McKibben himself spelled out in the same article:

Carbon dioxide, a.k.a. CO₂, or just "carbon" for short, is not a conventional pollutant . . . there's no easy way to get rid of it, no catalytic converter you can stick on your tailpipe, no scrubber you can fit to your smokestack. To reduce the amount of CO₂ pouring into the atmosphere means dramatically reducing the amount of fossil fuel being consumed. Which means changing the underpinning of the planet's entire economy and altering our most ingrained personal habits. Even under the best scenarios, this will involve something more like a revolution than a technical fix.

Notwithstanding the problem of equating "carbon" and "carbon dioxide" (graphite and diamonds are not the concern here, only carbon attached to two oxygen atoms), I find this a good summary of the policy challenge. If we really must reduce global CO₂ emissions in any climatically meaningful amount in the next decade or two, it will not be like reducing SO_x or particulate emissions: we are looking at an enormous and costly restructuring of the global economy, involving substantial permanent reductions in income for developed and developing countries alike.

However, I am not going to talk about the economics of climate policy, instead I want to pick up on the question of how we would know if this revolution is really necessary. It is a scientific question, and points to another of Mr. McKibben's frustrations, that the so-called "climate deniers" have managed to muddy an issue he believes to be quite clear. The expectation of clarity is actually a sign of muddied thinking. The problem is not that clever and committed people like me managed to make a clear issue go artificially cloudy, the problem is that the issues are inherently muddy, but policy makers are pursuing a strategy that requires them to claim the science is clear.

When the Government of Canada introduced Action Plan 2000 five years ago,² they claimed: "Our scientific understanding of climate change is sound and leaves no doubt

that it is essential to take action now to reduce emissions” (p. 15). When they introduced the Climate Change Plan for Canada two years ago they said “Countries around the world have recognized the urgent need to take action to reduce GHGs in order to address the climate change challenge.”³ When they announced Project Green last month they made statements like (emphasis in original):

Our planet’s **temperature is rising** and this is **cause for deep concern**. Over 2000 **leading scientists** contributed to the United Nations’ Intergovernmental Panel on Climate Change. Their prediction: by 2100, Earth’s average temperature will climb between 1.5 and 6 degrees Celsius . . . Altering climate patterns will cause more frequent and severe **extreme weather** events, imperiling the northern Aboriginal way of life and **threatening the health and safety of Canadians and people around the globe** . . . That is why Canada is a strong supporter of the Kyoto Protocol.⁴

When they sent a pamphlet out to every household in the country back in 2002 to promote their climate change strategy it began:

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.⁵

They invoke science this way because they want people to believe the issue is clear and their hands are tied. The agenda is urgent, there is no alternative. Resistance is futile.

I can understand why this line of argumentation is appealing for policy-makers. It is easier to convince people to support a costly and ambitious plan if they believe it is a necessary response to an urgent and certain threat. That last quotation is especially compelling: the 1990s are the warmest decade on record, and by implication, the warmest in a millennium. It is a great line to open a pamphlet with.

Therefore, it must be a problem for politicians when a skeptic comes along and claims that the picture is muddier than that, and they have a way of making it a problem for their scientists. Not long ago there was an article in *New Scientist* about this particular dispute, explaining it as follows.

It is at this point, however, that uncertainty starts to creep in. Take the grand claim made by some climate researchers that the 1990s were the warmest decade in the warmest century of the past millennium. This claim is embodied in the famous “hockey stick” curve, produced by Michael Mann of the University of Virginia in 1998, based on “proxy” records of past temperature, such as air bubbles in ice cores and growth rings in tree and coral (see “Hotly contested”). Sceptics have attacked the findings over poor methodology used, and their criticism has been confirmed by climate modellers, who have recently recognised that such proxy studies systematically underestimate past variability. As one Met Office scientist put it: “We cannot make claims as to the 1990s being the warmest decade.”⁶

Skeptics at the UK Met Office—who knew? It is noteworthy that this scientist is quoted anonymously. Politicians have gone out on a policy limb based on claims of scientific certainty, and there are repercussions for saying the science is uncertain.

It is easy to be judgmental, but it is only human to form preferences over truth, even though we know in principle that we probably should not. It is perilous in scientific scholarship to *want* certain results. The discipline of academic research involves learning

to set aside such desires in order to let the data and the theory speak. The old joke about taking the data down to the basement and beating the truth out of it is funny because the temptation does exist. That is why we have the peer review system. For all its imperfections it serves a necessary purpose. It puts studies in front of two or three other readers for the purpose of advising a journal editor whether a paper ought to be published. Sometimes errors are caught and corrected, sometimes bad arguments are kept out of print, and sometimes trivial tautologies are spotted so the expert audience is spared a waste of time. However, sometimes good papers get an unfair runaround, sometimes faulty results slip into print, and so forth.⁷ Peer review serves a purpose internal to the scientific community, as a first line of defense against faulty research. No warranties are expressed or implied. Referees are only saying that a paper should be published. They are not guaranteeing the results are true, only that they deserve the attention of the audience of that particular journal.

However, in recent years, a problem has emerged as a result of misinterpreting the nature of the peer review system. We are seeing more and more use of expert panels to produce scientific assessment reports, which are accorded a very high level of public trust. Such reports are often the basis for major public policy decisions, including investments worth billions of dollars. The assessment process makes deliberate appeal to the concept of peer review as a filter for the information it presents. Although it makes sense to invoke peer review as a *necessary* condition for use of scientific information in assessment reports, the problem is that it seems to be regarded as a *sufficient* condition. If a major public policy decision is going to be based on the conclusions of an assessment report, and if the report draws its conclusions based on a journal article, and if the journal article is considered reliable because the journal has a peer review process, but meanwhile the journal referees only provided a quick scan of the article and a supportive email to the editor, then obviously we run the risk of making a poor public investment decision.

I will discuss this issue primarily with reference to climate change, where the issues are vivid and the stakes are high. But let me emphasize that there is a much more general problem at hand: how should science be integrated into the policy-making process? Consider the following sorts of questions, all of which pertain to agricultural and land policies:

- Should municipalities ban cosmetic herbicide use on private lawns?
- Are Canadian cattle safe enough to be allowed back into U.S. meat packing plants?
- Should the Canadian and U.S. governments reduce the role of dairy products in their Food Guides?
- Is bovine growth hormone safe for use in our livestock?
- Does gas flaring from Alberta oil fields cause stillbirths in nearby livestock?

The reader can no doubt add to the list. In every case a difficult science question must be answered, and the answer may drive policy decisions with large implications for peoples' lives. The process that is used to answer the science question typically involves appointing some kind of expert assessment group, whose job is to produce a Yes or No answer. But there are two problems lurking in the background. The first is that the politicians who commission the panel want a clear, clean answer, even if the science is not there. They do not want muddiness. They want certainty and—if at all possible—urgency.

They want to be able to say that “Our scientific understanding of herbicide toxicology is sound and leaves no doubt that it is essential to take action now to eliminate lawn chemicals.” If you are going to put a dozen lawn-care companies out of business, you need some clear scientific cover.

The second problem is that the panelists themselves may have strong preferences. Some may just inwardly dislike the concept of bovine growth hormones, and when they study the literature they will be especially alert for evidence that confirms this view. It is hard to be balanced. When a group of authors share the same bias, it is hard to prevent a report from tilting to that side, by favoring one line of evidence over another, or by framing the questions in ways that leave important counterarguments at an artificial disadvantage. The phenomenon of a self-selecting group reinforcing each others’ prior biases rather than moderating them has been extensively explored in the social psychology literature (see the review in Sunstein 1999).

We cannot do away with expert assessment panels. If policy is to have any foundation in good, relevant science, as we hope it will, then our aim should be to take a hard-headed look at bias-proofing it. In this essay I will propose two mechanisms that I think can do this. I will illustrate them with reference to the climate change debate, but the application is more general.

The first mechanism is a Science Audit and the second is a Counterweight Panel.⁸ A Science Audit panel, or permanent agency, would act independently of an assessment panel, and would identify the key studies on which the assessment panel based its conclusions. They would then audit those studies with a view to verifying that, at a minimum:

- the data are publicly available;
- the statistical methods were fully described, correctly implemented and computer code that can reproduce the results is published; and
- if the findings given maximum prominence in the assessment report are at odds with other published evidence, clear reasons are provided in the text as to the choice of studies to emphasize.

The Audit Panel would *not* be offering any judgment on the scientific conclusions of the paper, they would simply be verifying that basic disclosure conditions are met. They are, after all, standards of transparency that people already assume to be operational in science. All I am proposing is that the assumption be verified.

A Counterweight Panel would be convened to prepare the strongest possible counterargument to the conclusions of an assessment panel. If a question is complex enough to merit an expert assessment panel, it is complex enough for there to be at least two credible views. Knowing that a Counterweight Panel will be convened to provide a strong incentive for an assessment panel to guard against groupthink or other biasing mechanisms, so as not to leave any openings for obvious criticism. It will also deal with the problem of how to respond when competent experts outside the assessment process complain after the fact of bias or distortion by the assessment panel members, since these experts will have a natural process in which to make their arguments. It will also make it harder for politicians to cherry-pick phrases from a report to assert a spurious level of certainty when trying to justify a policy decision.

I believe that if these mechanisms had been in place back in 2001, the Third Assessment Report (TAR) of the Intergovernmental Panel on Climate Change (IPCC)

would have run into serious trouble. I will show that the central icon of that report, the hockey stick graph, would have failed a science audit. I have argued elsewhere⁹ that there is sufficient evidence to show that if a Counterweight Panel had been convened by the IPCC, then rational skepticism over anthropogenic global warming would be quite widely shared. The IPCC would have found itself presenting credible evidence adverse to the view of carbon dioxide as a major climate driver, and it would not have been able to downplay fundamental uncertainties in the underlying science.

AUDITING THE HOCKEY STICK GRAPH

In 1998, *Nature* published the first hockey stick paper, authored by Michael Mann, Raymond Bradley and Malcolm Hughes and commonly called “MBH98.” Mann et al followed up in 1999 with a paper in *Geophysical Research Letters* (“MBH99”) extending their results from AD 1400 back to AD 1000.¹⁰ In early 2000, the IPCC released the first draft of the TAR. The hockey stick was the only paleoclimate reconstruction shown in the summary, and was the only one in the whole report to be singled out for repeated presentation. In the final version of the TAR, the graph appears as Figure 1b in the Working Group 1 *Summary for Policymakers*, Figure 5 in the *Technical Summary*, twice in Chapter 2 (Figures 2–20 and 2–21) of the main report, and Figures 2, 3, and 9-1B in the *Synthesis Report*. Each time the graph is used it is in color, and often occupies half the page or more. Referring to this figure, the IPCC *Summary for Policymakers* (p. 3) claimed it is likely “that the 1990s has been the warmest decade and 1998 the warmest year of the millennium” for the Northern Hemisphere. It was, without question, a central piece of evidence for the argument and conclusion of the TAR.

Mann et al used a “multiproxy” technique, combining a variety of proxies for past temperatures. The most numerous, and influential, proxies in their dataset are tree ring chronologies. The method required mapping a large sample of proxies to a large sample of temperatures, so the dimensions of the data matrices had to be reduced. This was done through principal component (PC) analysis. Principal component analysis involves replacing columns of a matrix with a weighted average of the columns, where the weights are chosen so that the new vector (called the first principal component or PC1) explains as much of the variance of the full matrix as possible. This leaves a matrix of unexplained residuals, but this matrix can be reduced to a PC as well, which is called the second PC, or PC2. And there will be residuals from it too, yielding PC3, PC4, etc. The higher the number of the PC, the less important is the pattern it explains in the original matrix. PC1 is the dominant pattern, PC2 is the secondary pattern, etc. In many cases a large number of data series can be summarized with relatively few PCs.

MBH98 applied PC analysis to simplify both temperature and proxy data. For temperatures, they represent 1,082 series with 16 PCs. They used 112 proxies, of which 71 were individual records and 31 were PCs from six regional networks containing over 300 underlying series in total. The networks are from geographical regions with labels like “NOAMER” (North America) and “SWM” (Southwest Mexico).

In the spring of 2003, Stephen McIntyre sent an email to Michael Mann requesting the MBH98 dataset. Steve is not a scientist or an economist, he was just a business man who had seen the hockey stick over and over and was curious how the graph was made. Since the people promoting the graph did not seem to know he decided to find out for

himself. His experience in mineral financing had taught him the importance of looking at raw data. He wanted to see if the raw data looked like hockey sticks too. After some delay Mann arranged provision of a file which was represented as the one used for MBH98. One of the first things Stephen discovered was that the PCs used in MBH98 could not be replicated. In the process of looking up all the data sources and re-building Mann's dataset from scratch, Steve discovered a quite a few errors concerning location labels, use of obsolete editions, unexplained truncations of available series, etc. Most of these had small effects on the final results, but re-doing the PCs had a big effect.

In the late summer of 2003, Steve contacted me to explain what he had found out. I agreed to help him write up his work and we published a paper (McIntyre and McKittrick 2003) in October 2003, explaining the errors we found in Mann's data. We showed that when these errors were corrected the famous hockey stick disappeared.

In his initial response to our paper, Mann argued that we had studied the wrong dataset—in other words that the one he provided had mistakes in it and we ought instead to have used one in a newly identified FTP archive at his university. Over the next month we examined his FTP archive and discovered that, in fact, it corresponded almost exactly to the file we had originally been working with. However, it differed in important ways from the description of the dataset in the original *Nature* paper. We supplied a list of these discrepancies to *Nature* and after their own investigation they ordered a Corrigendum from Mann et al, which appeared in the summer of 2004.¹¹

Mann also objected that we did not exactly replicate his computational steps or sequence of proxy rosters. No one had ever replicated his results, and we have since learned that others tried but were also unsuccessful. To date we are the closest anyone has been able to come in print. We were not bothered by Mann's response on this point, but it did seem pointless to differ over trivial issues. Therefore, we requested his computational code to eliminate these easily resolved differences. To our surprise he refused to supply his computer code, a stance he maintains to today.

As for the proxy sequence, in building his PCs it turns out he had spliced together a number of different series in order to handle segments with missing data in the earliest part of the analysis. This was not explained in his *Nature* paper, therefore, Steve had not implemented it in the emulation program. We requested identification of the splicing sequence, which Mann refused to provide, therefore, Steve worked out an emulation as best he could. In the end nothing turned on it, though Mann continues to point to it as a knock against our efforts. It is still not possible to identify the final form of the data used in MBH98 as it requires forming sequences of spliced proxy PC segments, and Mann has given conflicting counts of the number of underlying vectors involved. Still, Steve's emulation program is very close to reproducing the original hockey stick, and is as close as anyone is able to get in the absence of cooperation from Mann and his colleagues.

Bent Principal Components

In our analysis of Mann's FTP archive, we found some remnant computer code files that turned out to be the Fortran routines he used to compute his principal components. This is the only portion of his computer code we have been able to inspect. In these files we discovered why his PCs could not be replicated. In a conventional PC analysis, if the data are in differing units it is common to standardize them to a mean of 0 and a variance of

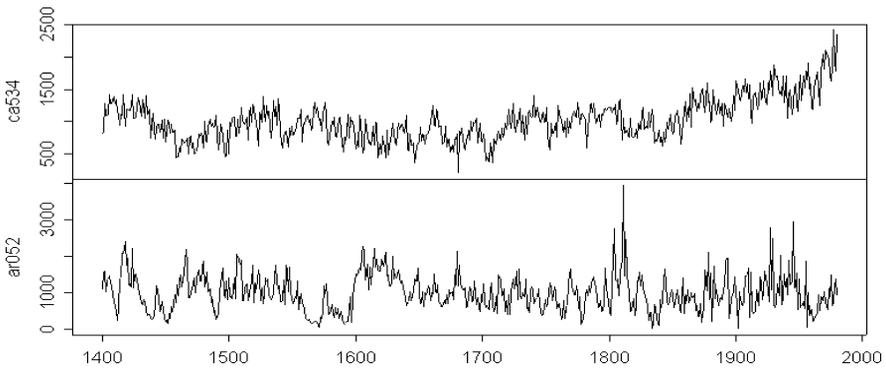


Figure 1. Two tree ring chronologies from the MBH98 dataset. *Top*: Sheep Mountain, CA, USA. *Bottom*: Mayberry Slough, AR, USA. Both series are the same length, but due to the 20th century trend in the top panel, Mann's algorithm gives it 390 times the weight of the bottom series in the PC1

1. Tree ring data are converted into unit-free index numbers before archiving, therefore, no re-scaling is needed for doing a PC computation.

In Mann's program, he applied a further scaling, but with an unusual twist. Rather than subtract the mean of the entire series length, he subtracted the mean of the 20th century portion, then divided by the standard error of the 20th century portion.¹² Most of his proxy series do not look like hockey sticks, they look like flat static, and since they do not trend up or down in the 20th century this procedure did not make much difference. The mean of the last section is roughly the same as the mean of the whole series (as is the standard error), therefore, either way of standardizing yields more or less the same result. But some of the series trend upwards in the 20th century. For these, the Mann method has a huge effect. Because the mean of the 20th century portion is higher than the mean of the whole series, standardizing on the closing subsegment de-centers the series and inflates the variance above unity.

Principal component algorithms choose weights to maximize the explained variance of a group of vectors. If one series in the group has a relatively high variance, its weight in the PC1 gets inflated. The Mann algorithm did just this. It would, in effect, look through a dataset and identify series with a 20th century trend, then load all the weight on them. In other words it 'data-mines' for hockey sticks.

Figure 1 gives an example of the effect. It shows two of the 90 full-length series in Mann's database. Both are part of the North America ("NOAMER") proxy roster, whose PC1 is the most influential series on the hockey stick's final shape. The top panel is a tree ring chronology from a stand of bristlecone pines at Sheep Mountain, California. The bottom panel is a tree ring chronology from Mayberry Slough, Arkansas. In the bottom panel, the mean over the last 80 years is roughly equal to the mean for the previous 500 years, but in the top panel the post-1900 mean is above that for the pre-1900 portion. Mann's algorithm gives 390 times as much weight to the top series as to the bottom series in the PC1.

Figure 2 shows the contrasting effects on the NOAMER PC1. The top panel is the MBH98 PC1 for North America, which they deem the "dominant pattern" in the

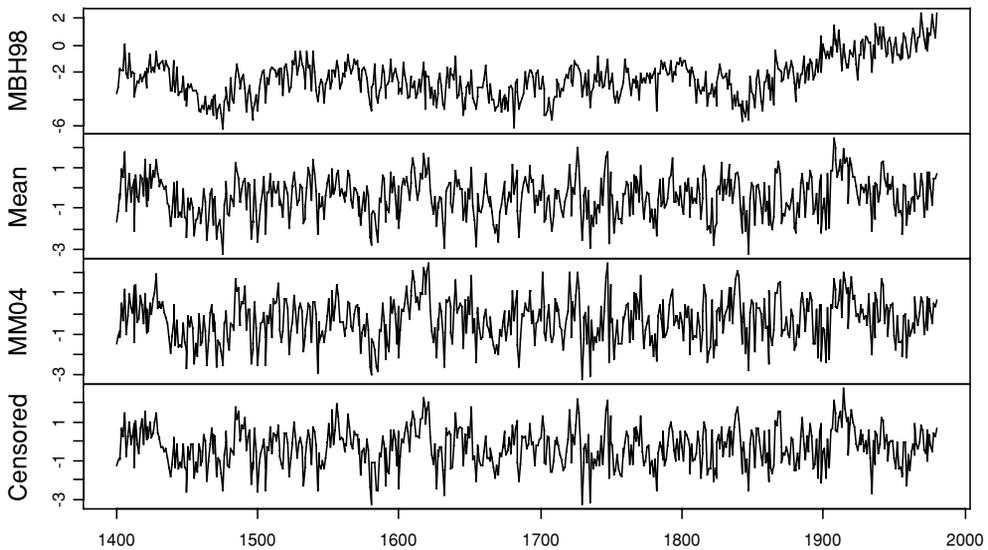


Figure 2. *Top panel:* PC1 of the post-1400 NOAMER tree ring network, calculated by MBH98 using short-segment standardization. *Second panel:* Simple mean of proxies. *Third panel:* PC1 using standard software without short-segment standardization. *Bottom panel:* Unreported PC1 calculated by MBH after censoring Graybill–Idso high-altitude series. All normalized to 1902–1980

data, and which has a distinct hockey stick shape. The second panel shows the simple average of the NOAMER proxies. Note that most proxies look more like Mayberry Slough—only a handful have the 20th century growth spurt. The third panel shows the PC1 computed using a common statistical package in which the data are standardized in the usual way. It looks like the simple mean, indicating that the dominant pattern in the data does not have a hockey stick shape. I will explain the bottom panel (“Censored”) later.

To test the power of Mann’s data-mining algorithm, we ran a Monte Carlo experiment in which we developed sequences of random numbers tuned to have the same autocorrelation pattern as the NOAMER tree ring data (“red noise”). In an autocorrelated process, a random shock takes a few periods to drift back to the mean. Initially, we used a simple first-order autocorrelation model, but later we implemented an ARFIMA¹³ routine that more accurately fits the entire autocorrelation function associated with tree ring data. The ARFIMA data is trendless random noise, simulating the data you would get from trees in a climate that is only subject to random fluctuations with no warming trend. In 10,000 repetitions on groups of red noise, we found that a conventional PC algorithm almost never yielded a hockey stick-shaped PC1, but the Mann algorithm yielded a pronounced hockey stick-shaped PC1 over 99% of the time. The reason is that in some of the red noise series there is a ‘pseudo-trend’ at the end, where a random shock causes the data to drift upwards, and before it can decay back to the mean the series comes to an end. The Mann algorithm efficiently looks for those kinds of series and flags them for maximum weighting. It concludes that a hockey stick is the dominant pattern even in pure noise.

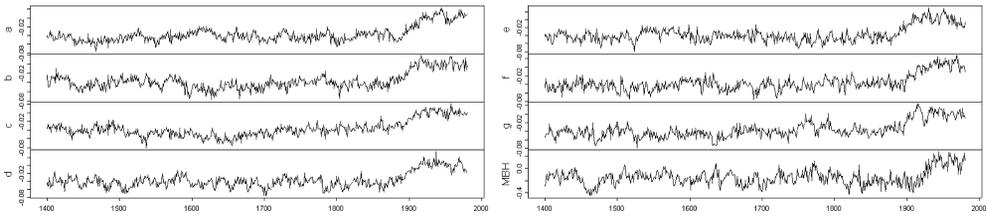


Figure 3. Seven panels are PC1's from red noise data fed into MBH98 algorithm. One panel is the MBH98 hockey stick itself

In Figure 3, seven of the panels show the PC1 from feeding red noise series into Mann's program. One of the panels is the MBH98 hockey stick graph (pre-1980 proxy portion). See if you can tell which is which.

We submitted a letter to *Nature* about this flaw in the MBH98 procedure, who declined to publish it principally (they said), because we could not condense our argument to fit the 500-word limit for the relevant section of the journal. Instead, Mann et al were permitted to make a coy disclosure in their July Corrigendum. In an online Supplement (but not in the printed text itself), they revealed the nonstandard method, and added the unsupported claim that it did not affect the results.

We extended our study in two ways. First, we showed that the data mining procedure did not just pull out a random group of proxies, instead it pulled out an eccentric group of bristlecone pine chronologies published by Graybill and Idso (1993). These trees (the Sheep Mountain series in Figure 1 is an example) were studied because of their pattern of cambial dieback. They all exhibit a 20th century growth spurt that has not been fully explained, but is likely to be at least in part due to CO₂ fertilization and is known not to be a temperature signal as it does not match nearby temperature records. The original authors (and others) have stressed that they are not proper climate proxies. Therefore, we felt it was important to examine what would happen to the MBH98 results if the Graybill–Idso proxies were excluded from the NOAMER group.

The result is in the bottom panel of Figure 2 (“Censored”). It shows what happens when Mann's PC algorithm is applied to the NOAMER data after removing 20 bristlecone pine series. Without these hockey stick shapes to mine for, the Mann method generates a result just like that from a conventional PC algorithm, and shows that the dominant pattern is not hockey stick-shaped at all. Without the bristlecone pines, the overall MBH98 results would not have a hockey stick shape: the late 20th century would look unexceptional compared to the variability of previous centuries.

As an aside, the data for the bottom panel of Figure 2 is from a folder on Mann's FTP site. He did this very experiment himself and discovered that the PCs lose their hockey stick shape when the Graybill–Idso series are removed. In so doing he discovered that the hockey stick is not a global pattern, it is driven by a flawed group of U.S. proxies that experts do not consider valid as climate indicators. But he did not disclose this fatal weakness of his results, and it only came to light because of Stephen McIntyre's laborious efforts.¹⁴

Another extension to our analysis concerned the claims of statistical significance in Mann's papers. Using the Monte Carlo simulation, we found that meaningless red

noise could yield hockey stick-like proxy PCs. This allowed us to generate a significance benchmark for the RE and other statistics.¹⁵ The idea is that if you fit a model using random numbers, you can see how well they do at “explaining” the data. Then the “real world” data, if they are actually informative about the climate, have to outperform the random numbers. We calculated significance benchmarks for the hockey stick algorithm and showed that the hockey stick did not achieve statistical significance, at least in the pre-1450 segment where all the controversy is. In other words, MBH98 and MBH99 present results that are no more informative about the millennial climate history than random numbers.

Gaspé Cedar

Another oddity in MBH98 is that some series are duplicated within the database. One of these, the Gaspé “northern treeline” series¹⁶ is included as a separate proxy (treeline #11), but it is also in the NOAMER PC collation as cana036. The data begin in 1404, but the chronology is based on only one tree up to 1421 and only two trees up to 1447. Because the earliest segment is so poorly replicated, the authors who originally sampled the Gaspé data do not use the data before AD 1600. When used as treeline #11, MBH98 listed the start date as 1400 and filled the empty first four cells by extrapolation. The misrepresented start date enabled them to avoid disclosure of the unique extrapolation; the extrapolation enabled them to include this series in the calculations going back to AD 1400, rather than withholding it until the AD 1450 step.

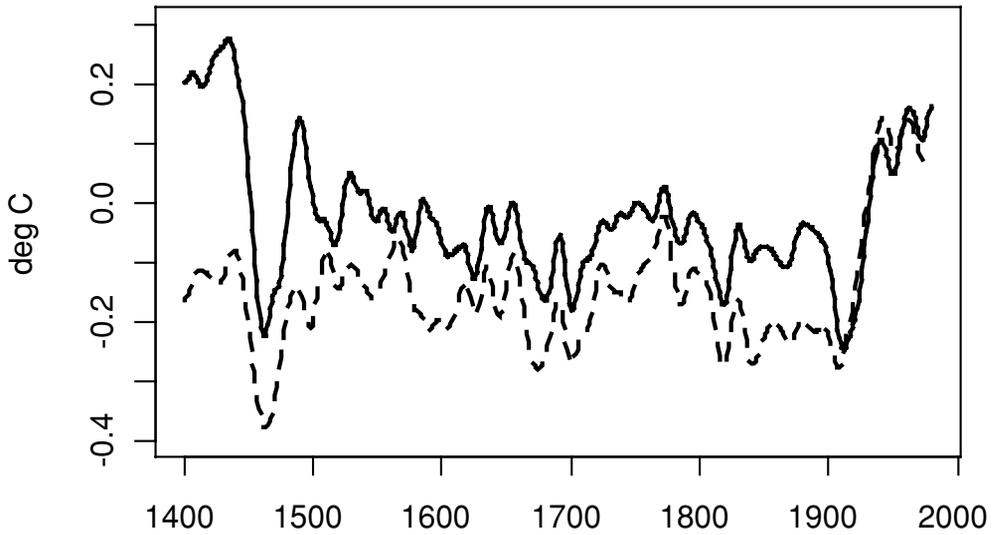
We wanted to see what would happen if the Gaspé data were not introduced until AD 1450. By rights we could have withheld it until 1600, and only used it once in the database, but that much alteration to the MBH98 procedure turned out to be unnecessary. Simply removing the pre-1450 portion had a large effect on the final graph, as will be shown in the next section. We wrote up the red noise experiment and significance benchmarking material into a paper, which was submitted to *Geophysical Research Letters*. We wrote up the information on the Gaspé cedar and the bristlecone pines and submitted it to *Energy and Environment*. Both papers were accepted and came out in February 2005.¹⁷

Hockey Score

Figure 4 shows two versions of the hockey stick chart. The dashed line is the MBH98 version. The solid line applies the corrections to methodology and data as discussed in this paper and the published references, especially McIntyre and McKittrick (2005a). The Mann multiproxy data, when correctly handled, does not show the 20th century climate to be exceptional compared to earlier centuries.

Our critics, including Mann himself, have mounted several counterarguments, chiefly that if the PC algorithm is corrected, but instead of only using two PCs from the NOAMER group we use five PCs, a hockey stick shape can be partly recovered. This is because the bristlecone pine imprint shows up on PC4. However, there are four flaws with this argument.

1. MBH98 identified the hockey stick shape as the dominant pattern (PC1) in the proxy data by using a flawed PC method. Under a corrected method, the hockey stick shape is demoted to the fourth PC where it accounts for less than 8% of the total explained variance, making it at best a small background signal. If the inclusion of a single



Source: Stephen McIntyre (pers. comm.).

Figure 4. *Dashed line*: MBH98 proxy-based Northern Hemisphere temperature index reconstruction. *Solid line*: Series resulting from using corrected PCs (retaining five PCs in the North America network), removing Gaspé extrapolation and applying CO₂ fertilization adjustment to full length of bristlecone pine series.

higher-order PC accounting for less than 8% of the variance in a single region changes the overall conclusions this does not prove that the PC4 is actually the “dominant climate pattern¹,” instead it shows that the model lacks robustness and the conclusions are unstable.

2. If the flawed bristlecone pine series are removed, the hockey stick disappears regardless of how the PCs are calculated and regardless of how many are included. The hockey stick shape is not global, it is a local phenomenon associated with eccentric proxies. Mann discovered this long ago and never reported it.
3. The MBH98 model fails to attain statistical significance regardless of the number of PCs used and regardless of whether the bristlecone pines are included or not. It is no more informative about the early millennial climate than random numbers.
4. MBH99 acknowledged that the bristlecone series are flawed and need an adjustment to remove the CO₂ fertilization effect. But they only applied the correction to the pre-1400 portion of the series. When we apply the correction to the full series length, the hockey stick shape disappears regardless of how many PCs are retained.

In February 2005, the German television channel *Das Erste* interviewed climatologist Ulrich Cubasch, who revealed that after hearing about Steve’s and my work, he too looked into the issue and discovered that the hockey stick was wrong:

He [Climatologist Ulrich Cubasch] discussed with his coworkers—and many of his professional colleagues—the objections, and sought to work them through . . . Bit by bit, it became clear also to his colleagues: the two Canadians were right. . . . Between 1400 and

1600, the temperature shift was considerably higher than, for example, in the previous century. With that, the core conclusion, and that also of the IPCC 2001 Report, was completely undermined.¹⁸

BIAS-PROOFING THE ASSESSMENT PROCESS

Recently, someone in Australia wrote to his Environment Minister (Ian Campbell) to discuss his concerns about the weakness of the global warming argument. He got a letter back that read, in part (emphasis added):

... Greenhouse science is complicated and we are constantly learning more about the Earth's climate and the impact of human activities and natural processes on it.

The Australian Government, together with about 100 other nations, has accepted the findings of the Intergovernmental Panel on Climate Change Third Assessment Report. This report was prepared by several hundred scientists from all over the world, from various scientific disciplines and with differing opinions on global warming. **The material that went into the report was from scientific research papers that go through a rigorous process of peer review in order to be published. The report itself also goes through a rigorous process of preparation, review, and debate.**

Here is the problem. Mr. Campbell assumes that the assessment panel involves a several hundred scientists conducting a rigorous review of papers that already went through a rigorous review before publication in the science journals. On this basis he and governments around the world have accepted the IPCC conclusions and actively resist taking in information that does not back it up. The IPCC made the hockey stick its central piece of evidence. But despite Mr. Campbell's perceptions, it never verified the data were published or that the results could be reproduced, nor did the publishing journal, in this case *Nature*. The results turned out to be wrong, something that would have been obvious long ago had these things been checked.

Audit Panel

In section "Introduction" I proposed an Audit Panel or agency that would identify the key studies supporting a conclusion and would verify the following:

- The data are publicly available.
- The statistical methods were fully described, correctly implemented and computer code that can reproduce the results is published.
- If the findings given maximum prominence in the assessment report are at odds with other published evidence, clear reasons are provided in the text as to the choice of studies to emphasize.

Because the IPCC report leaned so heavily on the hockey stick graph, it would have failed such an audit. This alone proves the need for an Audit Panel.

The process I have in mind would involve the Audit Panel actually reproducing the results of key papers, or reporting that the results cannot be reproduced, if that is the case. It would bear some resemblance to the "JMCB Project," in which a team of economists attempted to replicate the results published over several years in the *Journal of Money, Credit and Banking*, only to find widespread problems obtaining usable data

and code from the original authors and further problems reproducing results (Dewald et al 1986). The experiment was recently repeated and only 15 of 150 papers could be replicated (McCullough et al forthcoming). Similar difficulties have been admitted at the *American Economic Review* (Editorial Statement 2004). Most likely an Audit Panel would hire consultants or expert staff to do the actual computations, but it would have the task of replicating every result in every key study.

Counterweight Panel

As for the Counterweight Panel, in the climate context I envision an IPCC Working Group 4 assembled from among the expert community to publish, under the imprimatur of the IPCC, the case against the conclusions presented by the other IPCC Working Groups, primarily Working Group I.¹⁹ Such a panel would have to deal with both economic and scientific aspects of the IPCC's work since the emission scenarios are intrinsic to the Working Group I conclusions. It would be ideal to have the other IPCC Working Groups prepare a response, to which Working Group 4 would then prepare a reply.

This exercise should be sponsored by client governments and published by the IPCC. It is no good waiting for the expert community to self-organize into such a panel, or to expect private firms or foundations to sponsor such a panel, since people cannot seem to get past their suspicions of ulterior motives when private sponsors have undertaken such work in the past. It is in government's interest to test the IPCC's output carefully, so they ought to take the lead in organizing the work.

In making this proposal I have encountered a few common objections. One is that there is no need for it as the IPCC is already balanced. If that were so, the IPCC would have nothing to fear from such a process, and indeed it could silence a lot of critics with the outcome. However, I think the hockey stick episode establishes adequate grounds to doubt this optimistic claim.

Another is the concern that it suggests a lack of trust, or an impugning of motives. In a business or legal setting, however, there are checks and balances in place, including the entire system of independent auditing, not because we assume people in business are dishonest but because we want a system that still works even if some people are not always honest and unbiased. Checks and balances are a fact of life.

A third objection is that it will create confusion by giving "equal" time to the other side. People would not know whom to believe. This point is rather revealing. Are people worried that if the contra-IPCC position were carefully put before the public, it might appear surprisingly compelling? But isn't the IPCC so confident in its position that they dismiss the very existence of credible counterevidence? If their critics are truly unqualified and incompetent, it will only strengthen the hand of the IPCC by putting each side's best arguments side-by-side, thereby laying to rest the idea that the IPCC systematically ignores good arguments from its critics.

However, perhaps the opposite will happen. Perhaps, the science really is actually rather muddy, and allowing the public to see the opposing arguments would show this to be so. If that is the case, then that should be the message of the assessment process, even if it makes life more difficult for politicians making major policy decisions.

CONCLUSIONS

My observation of the climate change debates over the past few years has convinced me that there is a mismatch between what journal peer review actually does and what users of scientific research think it does. Politicians and policy-makers appeal to the concept of ‘peer-reviewed’ research as a foundation for decision-making. Yet peer review does not typically guarantee that data and methods are open to scrutiny or that results are reproducible. If decision-makers want a guarantee of those things, then a further process must be established to provide it. Likewise, when science assessment reports are provided, decision-makers might like to believe that the panelists who wrote the report were not one-sided in their prior views and that they undertook rigorous independent verification of the results that support their conclusions. It would be nice if this were always true, but if we really want to be sure then we have to find a mechanism to ensure it. Considering the far-reaching implications of policy decisions that are being based on expert assessments, audits and counterweight panels ought to be integrated into the process by which science is used to guide decision-making, so as to ensure the best possible foundations for public policy.

NOTES

¹http://www.motherjones.com/news/feature/2005/05/mckibben_introduction.html, accessed April 27, 2005.

²http://climatechange.gc.ca/english/publications/ap2000/Action_Plan_2000_en.pdf, accessed April 26, 2005.

³http://climatechange.gc.ca/english/publications/plan_for_canada/climate/kyoto.html, accessed April 26, 2005.

⁴http://www.climatechange.gc.ca/kyoto_commitments/brochure/, accessed April 26, 2005.

⁵<http://climatechange.gc.ca/english/publications/supplement/thescience.html>, accessed April 26, 2005.

⁶Pearce, F. “Climate change: Menace or myth?” *New Scientist*, February 12, 2005, <http://www.newscientist.com/article.ns?id=mg18524861.400>. Accessed April 26, 2005.

⁷Pannell (2002) presents examples that highlight some of the striking limitations of the peer review process.

⁸I have developed each of these ideas at greater length in earlier writings, sometimes using other terminology (and I am still not entirely satisfied with that used herein). Both are discussed in McKittrick (2004, 2005). On the concept of a Counterweight Panel, see also Essex and McKittrick (2002), chapter 10.

⁹For example, McKittrick (2003, 2004), also Essex and McKittrick (2002).

¹⁰Mann et al (1998, 1999).

¹¹Mann et al (2004).

¹²He also divided again by the detrended standard deviation, though this step is of little consequence.

¹³Autoregressive fractionally integrated moving average.

¹⁴And lest the reader wonder: neither McIntyre nor me are paid for this research. My day job is as an economics professor. Steve left his job in 2003 and has been working on this project full time for two years now, completely unpaid.

¹⁵The Reduction of Error (RE) statistic measures the extent to which a statistical model reduces mean squared forecast errors compared to extrapolating the simple average of the observed variable.

¹⁶This series was included in the North American “northern treeline” network even though the Gaspé peninsula is nowhere near the northern treeline.

¹⁷McIntyre and McKittrick (2005a,b). For copies please see www.climateaudit.org

¹⁸See http://www.daserste.de/wwiewissen/thema_dyn~id.pmhkzlh24crqytp5~cm.asp (accessed April 27, 2005).

¹⁹As presenting this paper, I learned from an engineer that the concept of a Counterweight Panel is routine in the development of large aerospace projects, where it is called a “Red Team.” The Red Team is independent of the team that prepares a project proposal, and its function is to tear the proposal to pieces looking for weaknesses.

ACKNOWLEDGMENTS

I am grateful to G. C. van Kooten and the conference organizing committee for the invitation to speak on this topic and for their hospitality. My thanks also to two anonymous reviewers.

REFERENCES

- 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.
- Editorial Statement. 2004.** *American Economic Review* 94 (1): 404.
- Essex, C. and R. McKittrick. 2002.** *Taken by Storm: The Troubled Science, Policy and Politics of Global Warming*. Toronto: Key Porter Books.
- 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.
- Leavitt, P., G. Chen, J. Rusak and B. Cumming. 2002.** The past, present and future of prairie droughts: How bad is bad? Report of the Prairie Drought Project, University of Regina, mimeo.
- Mann, M. E., R. S. Bradley and M. K. Hughes. 1998.** Global-scale temperature patterns and climate forcing over the past six centuries. *Nature* 392: 779–87.
- Mann, M. E., R. S. Bradley and M. K. Hughes. 1999.** Northern hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. *Geophysical Research Letters* 26: 759–62.
- Mann, M. E., R. S. Bradley and M. K. Hughes. 2004.** “Corrigendum,” *Nature* 430: 105.
- McCullough, B. D., K. A. McGeary and T. D. Harrison. 2005.** Lessons from the JMCB Archive. *Journal of Money, Credit and Banking* (forthcoming).
- McKittrick, R. 2003.** An economist’s perspective on climate change and the Kyoto Protocol. Invited presentation to the University of Manitoba Economics Department Fall Workshop. Available at: <http://www.uoguelph.ca/~rmckitri/research/papers.html>
- McKittrick, R. 2004.** Bringing balance, disclosure and due diligence into science-based policymaking. Presented to “Public Science in Liberal Democracy: The Challenge to Science and Democracy,” University of Saskatoon, October 2004. Available at: <http://www.uoguelph.ca/~rmckitri/research/papers.html>
- McKittrick, R. 2005.** What is the hockey stick debate about? Presentation to the Asia Pacific Economic Cooperation Study Centre, Parliament House, Canberra Australia, April 4, 2005. Available at: <http://www.uoguelph.ca/~rmckitri/research/papers.html>
- McIntyre, S. and R. McKittrick. 2003.** Corrections to the Mann et al (1998) proxy data base and Northern Hemisphere average temperature series. *Environment and Energy* 14 (6): 751–71.
- McIntyre, S. and R. McKittrick. 2005a.** The M&M critique of the MBH98 Northern Hemisphere climate index: Update and implications. *Energy and Environment* 16 (1): 69–100.
- McIntyre, S. and R. McKittrick. 2005b.** Hockey sticks, principal components and spurious significance. *Geophysical Research Letters* 32 (3): L03710 10.1029/2004GL021750 12 February 2005.
- Pannell, D. J. 2002.** Prose, psychopaths and persistence: Personal perspectives on publishing. *Canadian Journal of Agricultural Economics* 50 (2): 101–16.
- Sunstein, C. R. 1999.** The law of group polarization. University of Chicago John M. Olin Law and Economics Working Paper No. 91, Series 2. Available at: http://papers.ssrn.com/paper.taf?abstract_id=199668