July 5, 2005

Dear Dr. Schneider,

At this point Wahl and Ammann have refused to provide the requested R2 and other cross-validation data, directly bearing upon claims of significance, and have refused to disclose their recent paper, the website reference to which shows that they used cross-validation R2 elsewhere. Climatic Change has specific policies requiring the provision of supporting data and calculations. Wahl and Ammann are in clear breach of this policy and their paper should be rejected on these grounds alone. There is an obvious reason why these authors are so obdurate in refusing to provide their supporting cross-validation statistics: they show the results are insignificant. As such I think it self-evident that the paper should proceed no further at Climatic Change.

Aside from this breach of Climatic Change policy, their presentation is verbose and poorly written (45 pages essentially discussing the results in two tables). There are errors and mischaracterizations on almost every page, which are impossible to itemize.

The following is a list of tasks that the authors would need to do to begin to have a paper which could be reviewed in detail. Carrying out these tasks would require a completely re-written paper, and would reach quite different conclusions.

1. Delete all arguments depending on the rejected GRL submission.
2. Provide an accurate rendering of MM criticisms, including direct quotations from MM05a, MM05b, as they touch on the main criticisms of MBH98, including: failure of cross-validation statistics; inaccurate benchmarking of RE significance; withholding of adverse cross-validation statistics; lack of robustness to the presence/absence of bristlecone pines; the defects of bristlecones as a temperature proxy. Correction of inaccurate renderings would also require removing all references to MM "presenting" an alternative reconstruction; removing all reference to MM or MBH "centering conventions" and using neutral language such as covariance matrix, correlation matrix, or uncentered; removing all reference to tree ring chronologies being "unstandardized" (since they are all pre-standardized); etc.
3. Provide and discuss standard cross-validation statistics for all scenarios.
4. Either provide Monte Carlo simulations to benchmark RE significance in the context of the MBH98 model or remove all attributions of skill and/or significance as they relate to RE statistics.
5. Provide an accurate account of what steps in MBH98 have been replicated and what has not (e.g. Preisendorfer calculations, selection of gridcells, calculation of confidence intervals, etc….)
6. Provide an accurate account of remaining shortfalls in replication of MBH98 results. Do not obscure the failure to replicate MBH results exactly by making irrelevant "simplifications" in the benchmark. Benchmark using MBH weights and methodology.
7. Provide an accurate account of differences between the WA emulation and the MM emulation.

More issues would undoubtedly emerge if a new and completely re-written paper were submitted.

However, I do not believe that the authors should be given this option as, in my opinion, the submission and the Response Letter exhibit **********. Even if they were innocent mistakes, the mistakes go to the heart of the paper and render it unacceptable merely on those grounds.

The GRL Submission by Ammann and Wahl

The ********* in the Response Letter, where they continue to invoke their GRL submission even though they had already been notified that GRL was **not** proceeding with a review. They said:

> The requester mentions that the RE statistic is at issue, a claim that Dr. Ammann and I have shown is made moot by the results of our indirect tests in ms #3321. Dr. Ammann and I have shown in other material referenced in mss. #3321 that the analysis of McIntrye and McKitrick in GRL (2005)--which claims RE significance levels are improperly determined by Mann, Bradley, Hughes--is itself deeply flawed.

The other material referenced is Ammann and Wahl (submitted to GRL). On June 6, 2005, I was notified by the GRL editor handling this submission that GRL had "decided not to proceed with the review of the Ammann and Wahl Comment; therefore, you need not compose a Reply to this manuscript." This was several days prior to their letter to you on June 10, 2005. In my opinion, for Wahl and Ammann to present this "other material" as justification for refusing to produce requested data, *knowing* that GRL had decided not to proceed with a review, is falsification. Various other matters are itemized below but you could save yourself the time of going through them by noting that this kind of behaviour is surely sufficient grounds for immediately closing the file on this paper.

To make matters worse, Wahl and Ammann made a second misrepresentation in this short paragraph by claiming that the "other material" shows that the RE critique made in MM05a is "deeply flawed". However, their GRL submission did not even address the RE critique in MM05a. I have also attached a copy of their GRL submission. Again, this misrepresentation rises to falsification.

MM Criticisms

The WA paper purports to be a rebuttal of "MM criticisms" as made in MM05a,b. It has sections headed "1.1 Details of MM Criticism"; "2.2 Reconstruction Scenarios Reflecting MM Criticisms"; "3.2 Evaluation of MM Criticisms" "4.2 Robustness of MBH98 Results in Relation to MM Criticisms". However, if you compare the criticisms in the Abstracts of MM05a and MM05b to section 1.1 and elsewhere in WA, you will see that WA have omitted nearly all the major MM criticisms, while substituting as the "major" MM criticism a claim that MM have explicitly denied making. The omissions and misrepresentations are so pervasive as to constitute a falsification of the record. I will discuss a number of particular instances, but the discussion here is not comprehensive merely because of limits on my time and patience.

Cross-Validation Statistics

One of the most notable discrepancies arises in connection with description of cross-validation statistics, which was also the topic of the Response Letter. The MM05a Abstract states:

> using a range of cross-validation statistics, we show that the MBH98 15th century reconstruction lacks statistical significance.

The conclusion to MM05a stated:

> An obvious guard against spurious RE significance is to examine other cross-validation statistics, such as the R^2 and CE statistics, as recommended, for example, in *Cook et al. [1994].* While there are limitations to the R^2 statistic, the analysis of statistical "skill" of *Murphy [1988]* presupposes that the R^2 statistic exceeds the skill statistic and cases where the RE statistic exceeds the R^2 statistic are of particular concern [*Cook et al., 1994].* In the case of MBH98, unfortunately, neither the R^2 and other cross-validation statistics nor the underlying construction step have ever been reported for the controversial 15th century period. Our calculations have indicated that they are statistically insignificant.

The MM criticism of the need to examine MBH98 cross-validation statistics was specifically endorsed by one of our GRL referees as follows:

> [they] also show that by not presenting other stringent verification statistics (e.g. R_2, CE, product mean test and sign test) the validity of the 1400 step is likely much weaker than is apparent from the original MBH98 study.

MM05b criticized not only the statistical insignificance of the cross-validation statistics, but also the withholding by MBH98 of adverse cross-validation statistics. Yet in section 1.1 "MM Criticisms", WA omitted both topics and obviously failed to rebut them.

In their text, WA omit the very cross-validation statistics that were at issue in the MM criticisms. This is done without any notice to the reader of the omission. Although they have withheld key cross-validation statistics themselves, they repeatedly emphasize the need for cross-validation statistics, using language such as the following:

> More generally, our results highlight the necessity of reporting skill tests for each reconstruction model, as is customary in quantitative paleoclimate reconstruction. (p. 30)

A reader would be misled by the omission of cross-validation statistics and by the many WA statements about verification as he would have no way of knowing that WA had intentionally withheld standard cross-validation statistics.

WA are acting in bad faith by refusing to provide the requested information to a reviewer. They know that the cross-validation statistics are adverse to them. The rationalizations provided in the Response Letter are simply implausible. WA argue that the RE statistic measures "low frequency" variability, while the R2 and other cross-validation statistics measure "high-frequency" variability, and that low-frequency is the topic of interest in recent paleoclimate literature. However, 99% of their article is devoted to "MM criticisms" and the salient context is MM, where the issue of these other cross-validation statistics is not merely discussed, but a highlighted issue in the Abstracts.

Even where low-frequency variability is the topic of interest WA provided no statistical reference to support their argument that the RE statistic should be considered in isolation. I have examined the original literature discussing the RE statistic [Fritts, 1976; Fritts, 1979; Gordon and Leduc, 1980; Cook et al, 1994; Wilks, 1995 and others] and, in every case, I can find no recommendation that only the RE statistic be used; in fact, the recommendations are consistently that a range of cross-validation statistics be examined, as recommended in MM05a (and WA misleadingly imply that they have done). I have also examined a number of empirical articles by authors whose primary interest is low-frequency variation, to assess their use of cross-validation statistics. Jacoby's interest is low frequency variability, but, in every case where he reports the RE statistic, he also reports the R2 statistic, including in articles published in Climatic Change. Similarly, the interests of Cook and Briffa are hardly limited to high-frequency effects, but in every instance that I have examined, where they report an RE statistic, they also report the cross-validation R2 statistic. Thus, I am unable to locate any practice justifying the withholding of the R2 statistic in cross-validation studies; indeed, the practice is exactly the opposite.

Some studies interested in low frequency variability [Briffa et al, 2001] have reported the R2 statistic between filtered versions of series in conjunction with the R2 statistic from the raw series. But again the R2 statistic is not withheld (even for the raw data).

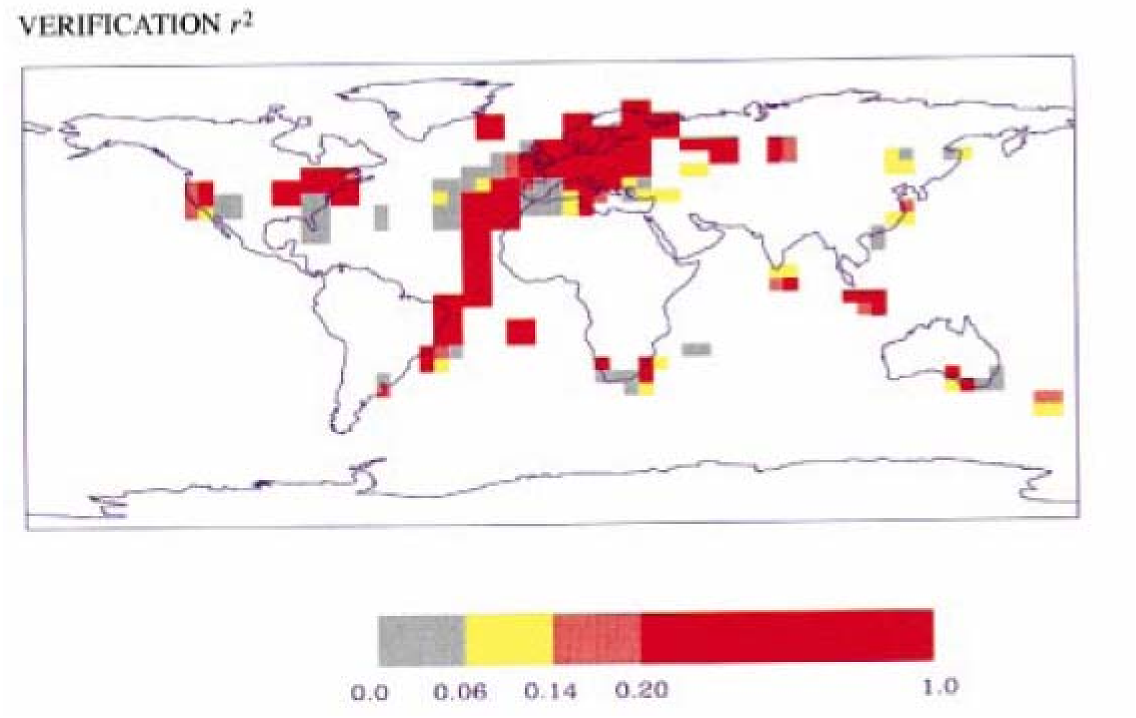
I attempted to examine Wahl and Ammann's own practices in this respect, since both of them have prior interest in low-frequency variability. I have not seen any prior article in which they applied an RE and deviation-from-calibration-mean statistic to assess low-frequency model significance, while withholding the cross-validation R2 statistic. On the contrary, a page on their website discussing a current project (http://www.assessment.ucar.edu/paleo/past_stationarity.html) shows use of the R2 statistic in model validation for low-frequency scales together with the following comment:

> This result indicates that modern-period validations of reconstructions based on relatively poor-quality proxies can give a strongly false sense of security about the likely long-term reliability of these reconstructions. "

The paper arising out of this project is the one I requested. I believe it shows that they use the R2 statistic for low-frequency cross-validation when it suits them to do so. Their refusal to provide it is further evidence of bad faith.

As to MBH98 itself there is no evidence it confined its claims of skill to the low-frequency domain, while renouncing claims of high-frequency skill. MBH98-99 is the source of the claim that 1998 was the "warmest" year of the millennium – obviously a claim in the high-frequency part of the spectrum. MBH98 purported to provide annual confidence intervals, which rely on high-frequency residuals. WA themselves describe MBH98 as "one of a growing set of **high-resolution (annually resolved)** reconstructions" and do not renounce or criticize high-frequency skill claims.

Perhaps most tellingly MBH98 itself not only stated that they calculated the R2 statistic, but made the results the subject of a prominent figure, which I have shown below together with the original caption. Even recently, Mann has not renounced claims that the MBH98 reconstruction passes the R2 test, as he repeated these claims in a letter to Natuurwetenschap in December 2004.



Original caption to MBH98 Figure 3: … bottom, verification r2 (also based on 1854–1901 data)…. For the r2 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). …Significance levels were determined for r2 from standard one-sided tables, accounting for decreased degrees of freedom owing to serial correlation. (Methods)

Hence, Wahl and Ammann are not only isolated in claiming that R2 is not used in the paleoclimate literature for low-frequency cross-validation, but their Response Letter is contradicted by their own previous and current practices elsewhere. If Wahl and Ammann now wish to withhold cross-validation statistics on the basis of the "low frequency" arguments presented in the Response Letters, **they had an obligation to explicitly disclose this reasoning in their article and to warn readers that expected cross-validation statistics were not being reported and provide a complete explanation why not, especially given the fact that this was such a big issue in MM05a,b.** Instead of doing this, they deceptively used the term "verification statistics" or "validation statistics" or "skill tests", when they were merely using the RE statistic.

RE Statistic

The handling of the RE statistic likewise shows bad faith by WA.

The RE statistic was developed in the context of linear regression models. In this limited application, the consistent premise is that it was applied after the calculation of a cross-validation R2 statistic. The source literature states clearly that there is no statistical distribution for the RE statistic and there are no tables providing benchmarks for significance. The source literature [Fritts, 1976; Gordon and Leduc, 1980, etc.] cautiously states only that an RE>0 indicates "some" skill, but this is always in the context of a linear regression model passing other cross-validation tests including the R2.

MBH98 is **not** a linear regression model, but a sui generis methodology involving inverse regression, inversion, re-scaling and eigenvector expansion. Obviously it is impossible to simply transpose a rule-of-thumb test from a completely different statistical procedure to an MBH98-type methodology. To their credit, this was recognized in MBH98, who carried out fresh benchmarking through simulations, which, by coincidence, arrived at a 99% benchmark of 0 as well. Unfortunately, their simulations were not well-constructed and arrived at an inappropriate benchmark. MM05a strongly criticized their benchmarking calculations, stating (in the Abstract):

> we show that MBH98 benchmarks for significance of the Reduction of Error (RE) statistic are substantially under-stated and,

MM05a states (in its conclusion) that:

> More generally, this example shows that changes in methodology will generally require new Monte Carlo modeling, that benchmarks carried forward from one methodology cannot necessarily be applied to a new methodology – even if the method changes may appear slight, and that great caution is required prior to concluding statistical significance based on RE statistics.

These conclusions were specifically endorsed by a GRL referee who stated:

> secondly, they question the acceptance threshold of the RE value used by MBH98 … I fully accept the analysis, interpretation and implications with their assessment of MBH98's use of RE.

In their submission, WA seriously misrepresent what MM05a said about the RE statistic. Rather than reporting the very strong cautions in the above paragraph, without providing any specific quotation from MM05a they state that the conclusions of MM05a are limited to RE in association with the use of PC algorithms and that the MM05a criticism of RE benchmarking is "moot" under circumstances not involving the use of PC algorithms (see pages 9, 13, 20, 29). This is obviously not the case and is highly misleading.

Secondly, WA state that MM05a criticized the "standard benchmark" of 0 as it applies to regression models. This is untrue. MM05a only criticized the benchmark of 0 in the MBH98 methodology using the MBH98 simulations. The use of the RE statistic in respect of a simple linear regression was not relevant and was simply not discussed.

Thirdly, even if the MM05a critique of the RE statistic were "moot" under circumstances not involving PC calculations (which is not the case), WA applied the RE statistic as evidence of skill or significance in scenarios 1,5 and 6, which do involve the use of PC algorithms. Any WA claims regarding skill or lack of skill for scenarios 1, 5 and 6 are in the face of criticisms that have not been "mooted" even by WA's incorrect characterization. More generally, WA have not "mooted" the argument that RE significance requires new simulations. Instead, they transpose a benchmark from simple linear regression which MBH98 itself did not even propose. This benchmark cannot be used.

As matters stand, WA have not refuted the MM05a arguments in respect to RE significance; they have not attempted to re-habilitate the MBH98 simulations to establish RE significance; and they are not entitled to recklessly transpose the RE significance test from a linear regression context. At present, the only published standard for RE significance in an MBH98 context is the one calculated in MM05a; by this standard, WA have misrepresented scenarios as having statistical significance when they simply do not. Literally every single claim to statistical skill or significance in their entire paper lacks a statistical foundation and cannot be made.

Misrepresentation of MM05 Thesis

While WA fail to discuss the criticisms that MM actually made, they misrepresent as the "main" thesis of MM05 a claim that MM not only do not make, but have repeatedly and explicitly denied making. WA state incorrectly that MM05 have "presented" a temperature reconstruction and spend much energy criticizing this supposed reconstruction. However, MM have consistently and explicitly stated that their work is entirely critical, that they do not endorse MBH98 methodology or purport to "correct" it and that they do not "present" any reconstruction of their own; they have explicitly said that the purpose of their reconstructions is to illustrate the non-robustness of MBH98 and the inability of MBH98 to make claims about 20th century uniqueness.

In addition to explicit caveats in the texts themselves, the FAQ section of the Supplementary Information to MM03 stated:

> *Your graph seems to show that the 15th Century was warmer than today's climate: is this what you are claiming?*
>
> No. We're saying that Mann et al., based on their methodology and corrected data, cannot claim that the 20th century is warmer than the 15th century – the nuance is a little different. To make a positive claim that the 15th century was warmer than the late 20th century would require an endorsement of both the methodology and the common interpretation of the results which we are neither qualified nor inclined to offer. http://www.uoguelph.ca/~rmckitri/research/trcqa.html

Likewise, the FAQ for MM05 stated:

> Are you saying the 15th century was warmer than the present?
>
> No, we are saying that the hockey stick graph used by IPCC provides no statistically significant information about how the current climate compares to that of the 15th century (and earlier). And notwithstanding that, to the extent readers consider the results informative, if a correct PC method and the unedited version of the Gaspé series are used, the graph used by the IPCC to measure the average temperature of the Northern Hemisphere shows values in the 15th century exceed those at the end of the 20th century. We do not think that we could be more explicit than this. http://www.climate2003.com/FAQ.htm

Further, our GRL article does not contain ***any*** reference to an alternate reconstruction. Thus, any criticisms levelled by WA against the alleged reconstructions in MM03 or MM05b (EE) (regardless of their blithe ignoring of caveats) simply have no application against MM05a (***GRL)***. Indeed, it is extremely difficult to identify any points that they make in their CC submission, which rebut or even pertain to MM05(***GRL)*** at all.

This mischaracterization of MM is an important and highly misleading error. Since it is central to the entire premise of the submission, it is impossible to see how it could be re-written to eliminate the error. WA repeat this error in their Response Letter, where they argue:

> What could possibly change is that some of the MBH "segments" (based on varying richnesses over time of the proxy data) and some of the WA scenarios we present ***might not pass*** verification significance testing at the highest-frequency domain.

This is a fundamental point. MM05a stated that MBH98 results do not pass cross-validation tests; here WA show that they are aware of this, evidencing that the withholding is intentional.

Other Issues

I am running out of time and patience and the following comments are terser than a complete criticism would require.

Robustness to Presence/Absence of Dendroclimatic Indicators

One of the fundamental representations of MBH98 (and Mann et al [2000]) was its "robustness", including robustness to presence/absence of dendroclimatic indicators. One of the central criticisms of MM05b was that (a) this MBH claim was simply untrue and (b) that MBH98 withheld adverse information about the lack of robustness to presence/absence of bristlecone pines; (c) that the bristlecone pines were flawed proxies.

WA fail to report or rebut the first MM criticism.

On the second point, WA failed to report or rebut the MM criticism that MBH98 had both withheld adverse information about the lack of robustness to presence/absence of bristlecone pines and had made misrepresentations about robustness to dendroclimatic indicators.

On the third point, despite an extensive discussion in MM05b of bristlecones (and cedars), including the potential of the proxy being contaminated by CO2 fertilization, WA fail to discuss or rebut these criticisms. Their only argument on this point is that the bristlecones increase the statistic. However, this can also be done by any series with a nonclimatic trend.

It is very obvious that WA realize that MBH98 are not "robust" to presence/absence of dendroclimatic indicators, since they realize that MBH98 results are not robust to presence/absence of bristlecones. WA argue that bristlecones should be included because they improve the RE statistics. They are entitled to argue this point (although I believe that their arguments are very weak). However, prior to doing so, they need to plainly acknowledge the base point that MBH98 results are not robust to presence/absence of bristlecones and discuss prior misrepresentations by MBH98 in a straightforward way.

Instead, they omit a discussion of these points and engage in an elaborate subterfuge of using code words like "full information" rather than a clear discussion of the validity of bristlecones as a proxy and their impact on MBH98 results.

Misrepresentation of Replication

WA provide a very misleading impression of the degree to which they have actually replicated MBH98 results. I have carefully analyzed the code and determined that they have actually only attempted to replicate a very small portion of MBH98. There is no discussion of the use of Preisendorfer's Rule N in retention of tree ring principal components; there is no discussion of the impact of weighting temperature grid cells by cosine of latitude instead of the square root of cosine of latitude, no discussion of replication of gridcell selections.

They fail to provide any statistics on their replication of the 15th century MBH reconstruction such as maximum differences, correlation etc.

The WA benchmark replication should **not** include distracting "simplifications". The benchmark replication should apply MBH weights and temperature principal component methodologies. If they wish to assert that the reconstruction is insensitive to these weights, that's a different point.

They fail to discuss any differences between their emulation and the MM emulation (or even to acknowledge or discuss the MM emulation.) In fact, MM have reported (climateaudit.org) that RPCs calculated using the MM emulation and using the WA algorithm with like weights and the MBH98 temperature principal components coincide to 10 decimal places for 15th century proxies. This needs to be acknowledged.

Misrepresentation of "Centering Conventions" and "Unstandardized"

The WA discussion of "standardization" and "centering conventions" relies on their rejected GRL submission. They state:

> The effect of using "princomp" without specifying that the calculation be performed on the correlation matrix (an alternate argument of "princomp") forces the routine to extract eigenvectors and PCs on the variance-covariance matrix of the unstandardized proxy data, which by its nature will capture information in the first one or two eigenvectors/PCs that is primarily related to the absolute magnitude of the numerically largest-scaled variables in the data matrix (Ammann, C.M. and E.R. Wahl, 'Comment on "Hockey sticks, principal components, and spurious significance" by S. McIntyre and R. McKitrick', in submission to *Geophysical Research Letters*).

Later they state that there is a "mathematical inconsistency in the published replication of the MBH North American proxy PC calculations" (p.17). There are several errors here. Tree ring site chronologies are already "standardized" to a mean of 1 so that argument set out here is inapplicable. The variances of site chronologies vary, but Fritts has argued that chronologies with less variance generally have poorer climatic relationships; equalizing the variances has to be justified by WA in dendroclimatological terms, which they have not done.

It is also incorrect to attribute the term "centering convention" to MBH. MBH98 reported that they used a conventional calculation, when they did not. It could easily have been a programming error rather than a misrepresentation in the original article. There is no basis for the use of the term “centering convention”; in addition, there are other aspects to the MBH98 procedure not encompassed by this phrase.

It is also incorrect to attribute a “centering convention” to MM. MM did not propose a “centering convention”. MM stated that they attempted to implement the stated methodology of MBH98 i.e. a “conventional” PC calculation on sites which were already standardized. The default option in PC algorithms is a covariance matrix, which was applied by MM. It is incorrect to suggest that this was a choice by MM.

All language should be in neutral terms e.g. covariance matrix, correlation matrix.

Further, some points made as reproaches by WA were already made in MM05b (e.g. the point about the “fourth PC” made on page 18, already made in MM05b.

Failure to Properly Attribute MM05b

Nearly all the scenarios presented in WA were discussed in MM05b. Even though WA purport to be discussing MM, they fail to report explicit consideration. If one examines pages 75-76 of MM05b, one sees a discussion of (as far as I can tell) every scenario of WA, except scenario 4. What statements, if any, do WA disagree with? The systematic failure to attribute or discuss the prior MM05b discussion of the very same scenarios is a serious distortion of the record.

Conclusion

On several alternate grounds, Climatic Change should reject this article. First, the errors and mischaracterizations are so numerous and affect the central conclusions so severely that dealing with the required corrections will require a completely new article and rejection of the present article is mandated. Secondly, the authors have flouted a Climatic Change policy requiring authors to provide supporting data and calculations and have provided a highly implausible rationalization for their position. Finally and most importantly, *********** in their Response Letter by citing a submission they knew had already been rejected, in support of a point it did not provide support for anyway.

Yours truly,

Stephen McIntyre