



COLLEGE OF BUSINESS AND ECONOMICS  
Department of Economics and Finance

From: Ross McKitrick, Professor of Economics, University of Guelph  
[ross.mckitrick@uoguelph.ca](mailto:ross.mckitrick@uoguelph.ca)

To: Robert Doe, Publishing Editor, *Theoretical and Applied Climatology*  
[robert.doe@springer.com](mailto:robert.doe@springer.com)

CC: Hartmut Grassl, Managing Editor, *Theoretical and Applied Climatology*  
[hartmut.grassl@zmaw.de](mailto:hartmut.grassl@zmaw.de)

September 1, 2015

Re: **Learning From Mistakes in Climate Research by Benestad, Nuccitelli, Lewandowski et al.**  
DOI 10.1007/s00704-015-1597-5

I am writing to request retraction of false statements your journal has published by the writing team of Benestad et al. in the above paper, which I denote as BNL. The statements were shown to be untrue when the authors submitted earlier versions of the same paper to *Climatic Change* in 2012 and in *Earth System Dynamics* in 2013. On both these occasions the paper was rejected. In the case of *Climatic Change* the editors invited a resubmission from the authors, but they failed to correct the problems and the paper was rejected. The same paper has also been rejected by other reviewers and editors at several more journals, according to a recent [blog post](#) by the lead author (Benestad).

### **1. MM04 and MM07 Out of Sample Testing**

On pages 36—38 of their supplement, BNL discuss McKitrick and Michaels 2004 (MM04), of which I was a coauthor. They state or imply that the model of MM04 was not tested for out-of-sample predictive power, that Benestad (2004) pointed this out, that MM04 claimed to be ignorant of such testing and that they subsequently ignored the issue in McKitrick and Michaels 2007 (MM07).

These claims are false, as a simple reading of the papers in question shows. Moreover, BNL doctored the meaning of a key supporting quotation by including a false parenthesis.

Both MM04 and MM07 (attached) conducted and presented out-of-sample model tests. Section 5 of MM04 is even called “Out-of-Sample Properties” and presents an out-of-sample test in which the data for North and South America were withheld and predicted based on data from the remaining continents. Section 4.5 of MM07 is likewise called “Out of Sample Prediction” and presents the results of 500 tests in which 30% of the sample was randomly removed each time and predicted based on the withheld portion. BNL did not mention this and implied that no such testing was done.

BNL also misrepresent our response to the Benestad 2004 comment by selectively quoting it and falsifying the context. On page 37 of the BNL supplement they quote us as follows (emphasis added):

While McKitrick and Michaels (2004b) responded to this, they defended their original positions by dismissing the criticism stating they were “*unaware of any paper in the refereed applied climatology literature that has performed the test [splitting the sample, using one for model calibration and the other for validation] suggested by Dr. Benestad... if he has ever seen such a test applied anywhere in a published atmospheric science paper he should have provided an example, which he did not*”.

Our 2004 response is attached to this email. The full quotation makes clear that we were fully aware of such testing and had no objection to it (rather obviously, since we performed the tests ourselves); we only objected to the extreme and implausible version implemented by Benestad:

After a close replication of our results Dr. Benestad proposes a rather odd test of robustness: half the data are discarded, and more than half of the predictor variables are discarded, and on this basis he tries to predict the behaviour of the dependent variable in the other half of the data set. Moreover, he threw out the northern hemisphere data, which is arguably the better quality data. So he is trying to use the worst half of the data set to predict the better half while using only a subset of explanatory variables. We are unaware of any paper in the refereed applied climatology literature that has performed the test suggested by Dr. Benestad; indeed, if he has ever seen such a test applied anywhere in a published atmospheric science paper he should have provided an example, which he did not.

Thus it is clear BNL have misrepresented our position by inserting a parenthesis that falsified the meaning of what we said. To emphasize, we were not only aware of “splitting the sample, using one for model calibration and the other for validation” but *we did this testing ourselves*. What is worse, BNL are aware of their misrepresentation. As a reviewer of this paper for two previous journals I already pointed out this out. In my 2012 referee report for *Climatic Change* (attached) I wrote (note that “BHDCN” refers to the author names of the earlier draft):

Example 7 refers to McKitrick and Michaels (2004) and insinuates that their results were not tested using a withholding/prediction test. But Section 5 of MM04 presents just such a test, and Section 4.1 additionally tests the results against the influence of atypical outliers, and in neither case are the conclusions affected. The Benestad (2004) comment only shows that it is possible to devise an extreme version of the withholding test, namely trying to predict the Northern Hemisphere data from the (smaller) Southern Hemisphere subset, but the failure to pass this test has no general implications, as explained in McKitrick and Michaels reply to Benestad, which BHDCN do not mention. The McKitrick and Michaels (2007) paper (Section 4.5, Figure 2) presents 500 withholding/prediction tests in which 30% of the data were randomly withheld each time and predicted by a model fit to the remaining 70%. Section 4.2 tests against the influence of outliers. The skill of the model is amply demonstrated by the reported findings, which BHDCN do not mention, even while claiming the MM07 paper was flawed for not doing such tests.

In my review of their resubmission (attached) I wrote:

Regarding MM04, they quote my comment by saying “the description ‘an extreme version of the withholding test’ is very strange, whatever that means.” My meaning was clear, since I defined it in context: “The Benestad (2004) comment only shows that it is possible to devise an extreme version of the withholding test, *namely trying to predict the Northern Hemisphere data from the (smaller) Southern Hemisphere subset*, but the failure to pass this test has no general implications, as explained in McKittrick and Michaels reply to Benestad, which BHDCN do not mention.”

In my 2013 review for *Earth System Dynamics* (attached) I wrote

Case 7 (A2.5) refers to McKittrick and Michaels (2004) and insinuates that the results were not tested using a withholding/prediction test, an accusation made even more explicitly on page 490. But Section 5 of MM04 presented just such a test, and Section 4.1 additionally tested the results against the influence of atypical outliers, and in neither case were the conclusions affected. The Benestad (2004) comment only showed that it is possible to devise an extreme version of the withholding test, namely trying to predict the Northern Hemisphere data from the (smaller) Southern Hemisphere subset, but the failure to pass this test had no general implications, as explained in McKittrick and Michaels reply to Benestad, which BHDCN do not mention. The McKittrick and Michaels (2007) paper (Section 4.5, Figure 2) presented 500 split sample withholding/prediction tests in which 30% of the data were randomly withheld each time and predicted by a model fit to the remaining 70%. MM07 Section 4.2 tested against the influence of outliers. The skill of the model is amply demonstrated by the reported findings, which BHDCN do not mention, even while falsely claiming the MM07 paper was flawed for not doing such tests.

Consequently, pages 36-38 present a deliberate misrepresentation of a published exchange in the literature, the nature of the misrepresentation was known to the authors when they submitted it to *Theoretical and Applied Climatology*, and its effect is derogatory and harmful to my professional reputation as well as that of my coauthors.

## **2. MM04 and MM07: Spatial Autocorrelation**

In the same section of the Supplement, BNL claim that the results in these papers were invalidated by spatial autocorrelation, and that this was pointed out by the 2004 Benestad comment. The Benestad comment did not provide any evidence that the model residuals were spatially autocorrelated, as we pointed out in our response. All it did was raise the possibility that the dependent variable exhibited spatial autocorrelation. But this does not mean it affects the *residuals*, which is what matters for the computation of valid hypothesis tests.

BNL failed to mention that in a follow-up paper (attached):

- McKittrick, Ross R. and Nicolas Nierenberg (2010) “Socioeconomic Patterns in Climate Data.” *Journal of Economic and Social Measurement*, Vol 35 No. 3-4 pp. 149-175.

my coauthor and I conducted a thorough analysis of the spatial autocorrelation issue in the context of the MM07 data set and model. We showed that while the dependent variable was spatially

autocorrelated, as conjectured by Benestad (and later, Schmidt), the regression *residuals* were not, hence the inferences for MM07 were not biased in the way that Benestad had argued.

In my 2012 reviews for *Climatic Change* I pointed this out:

Benestad (2004) conjectured that spatial autocorrelation (SAC) would reduce the effective degrees of freedom in MM04 sufficiently to undermine the significance of the conclusions, but provided no test statistics on the matter. Schmidt (2009), cited by BHDCN, repeated this claim but once again did not test it, and he confused SAC in the dependent variable with that in the residuals. BHDCN make no mention of the extensive treatment of the SAC issue in McKittrick and Nierenberg (2010), who presented a suite of robust LM tests on both dependent variables and residuals, and showed that MM07-type model residuals were not affected by this issue, and even if the models are re-estimated with a correction for SAC the conclusions are upheld. They showed, moreover, that Schmidt's regression on GCM-generated data was affected by SAC which he neither tested nor corrected for, and had he done so his results would be insignificant. ... Altogether, Example 7 is a highly misleading bit of editorializing, and anybody who has read the relevant papers would immediately see it as such.

In my review of their resubmission I wrote:

The authors' failure to revise their Case 7 on spatial dependence in MM04 and MM07 papers is a surprise since I pointed out the extensive treatment of the issue in the McKittrick and Nierenberg paper, which they still do not cite in that context.

In my review of their *ESD* submission I wrote:

Their discussion of spatial autocorrelation (SAC) in the MM07 results omits all the relevant aspects of that debate. Benestad (2004) conjectured, without providing any evidence, that SAC would reduce the effective degrees of freedom in MM04 sufficiently to undermine the significance of the conclusions. Schmidt (2009), cited by BHDCN, repeated this claim but once again did not test it, and he confused SAC in the dependent variable with that in the residuals. BHDCN make no mention of the extensive treatment of the SAC issue in McKittrick and Nierenberg (2010), who presented a suite of robust LM tests for SAC on both dependent variables and residuals, and showed that MM07-type model residuals were not affected by this issue, and even if the models were re-estimated with a correction for SAC the conclusions were upheld. They showed, moreover, that Schmidt's regression on GCM-generated data was affected by SAC for which he neither tested nor corrected, and had he done so his own results would be insignificant.

... It is unacceptable that BHDCN repeat their untrue statements on these matters since they have been presented with the above information in response to both of their two previous drafts. To the extent they want to claim that "agnotology" arises from authors deliberately ignoring contrary information, they are themselves serving as striking examples.

Consequently, pages 36-38 present a deliberate misrepresentation of the publication record on this topic, the relevant omission was known to the authors when they submitted it to *Theoretical and Applied Climatology*, and what they have written is false and derogatory towards me and my coauthors.

### 3. MM04 and MM07: Clustered standard errors

On page 37 of their supplement, BNL state:

Proper testing needs to account for the fact that the economic co-variables would contain the same data within the border of each country,... The criticism presented in Benestad (2004) was not heeded; McKitrick and Michaels, (2007) repeated the claim of made [sic] in MM04 without acknowledging the criticism presented in Benestad (2004).

BNL fail to acknowledge that we dealt with this issue in MM07, paragraph [30] of which states:

In our database some of the socioeconomic variables are constant within the 81 countries in our sample, resulting in possible nonindependence (clustering) of errors within country groups. Denote the country groups as  $C(1), \dots, C(81)$ . To allow within-cluster nonindependence the estimator (5) is rewritten as [there follows the mathematical representation of clustering-robust standard errors.]

Once again our work was misrepresented by BNL.

### 4. MM05 Principal Component Analysis

Please note carefully that this refers to a different coauthor (McIntyre rather than Michaels).

On pages 38—41 BNL make the following accusation against McIntyre and I:

MM05 neglected the calibration involved in the process of reconstructing the past temperatures, and failed to address the important question of how many PCs were included in the calibration and how much of the variance they could describe.

We published two related papers in 2005. Contrary to BNL's untrue statement, MM05 (our *GRL* paper, attached) addressed, among other things, the calibration issue, PC ordering and explained variance. For example, paragraphs [8-13] state, *inter alia*:

Under the MBH98 data transformation, the distinctive contribution of the bristlecone pines is in the PC1, which has a spuriously high explained variance coefficient of 38% (without the transformation – 18%). Without the data transformation, the distinctive contribution of the bristlecones only appears in the PC4, which accounts for less than 8% of the total explained variance... McIntyre and McKitrick [2005] discuss, *inter alia*, problems relating to the interpretation of bristlecone/foxtail pine growth as a temperature proxy, and we show the impact of using conventional (centered) PC methods on the MBH98 northern hemisphere temperature index, which has a significant effect on the relative values in the 15<sup>th</sup> and 20<sup>th</sup> centuries.

The companion paper referred to therein as McIntyre and McKitrick [2005] was listed in the references of the *GRL* paper:

- McKitrick, Ross R. and Stephen McIntyre (2005) "The M&M Critique of the MBH98 Northern Hemisphere Climate Index: Update and Implications." *Energy and Environment* 16(1) pp. 69-100.

This one contained an extensive discussion of how changing the number of PCs in the calibration affects the overall results. While surely aware of this paper, BNL not only fail to mention it, they falsely accuse us of ignoring the topics we discussed therein at great length. BNL then go on to say:

This failure suggest that they did not understand the process, as the shape of each individual PC, which they stressed, is less relevant as regression analyses weight the different PCs according to how well they match the calibration data.

Our 2005 *Energy and Environment* paper discussed these issues at length. Section 4 of that paper in particular presented a sequential analysis of the effect of the shape of the PCs as determined by the centering methodology on the weights that come through the regression analysis, with contrasting results shown in Figures 3 and 4. It is outrageous for BNL to ignore our detailed publications on these topics and then accuse us of neither understanding nor discussing them. BNL then say:

The arguments presented in MM05 were irrelevant for the question they wanted to address, i.e. whether the PCA used in Mann et al. (1999, 1998) would lead to spurious results.

This again is utterly untrue, as even a cursory glance at the paper in question would prove. The title of the *GRL* paper was “Hockey Sticks, Principal Components and Spurious Significance.” The Editors would not have permitted that title if the paper had nothing to do with PCA and spurious results. The paper carefully explains the flaw in Mann’s PCA step and the way in which it gave rise to spurious test scores in the calibration step. The 2006 National Academy of Science report (which BNL refer to) summarized this point by saying we “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. 86). The claim by BNL that our paper “does not address the actual question” is unambiguously false and derogatory.

In my review of the *CC* resubmission I pointed out the untruthful nature of BNL’s discussion:

The MM05 paper focused on the bias arising from using decentered data in a PCA algorithm that is only valid when the data are centered. The question of the shape of the first PC is very much relevant in PCA, and was at the heart of the MM2005 papers, as it was central to the argument in Mann et al. 1999. Additionally, the MM2005 article was heavily focused on biased goodness-of-fit measures that understated the uncertainty of the reconstruction. These points were all discussed at length in, among other places, the 2006 NRC report (North et al.). The authors do not seem to have taken the trouble to properly research the issue, and as such their brief commentary lacks credibility.

In my review for *ESD* I wrote:

Case 12 refers to McKittrick and McIntyre 2005, which focused on the bias arising from using decentered data in a PCA algorithm that is only valid when the data are centered. BHDCN dismiss the bias as irrelevant, ignoring the fact that Mann et al 1999 placed explicit emphasis on the shape of the PC1 in their analysis, and that many subsequent authors used the biased PC1 in their own reconstructions, and that the PC1 error biased the computation of critical values, a topic which was central to the MM2005 article as well as the later exchange with Huybers. Many salient details of these points were discussed at length in, among other places, the 2006 NRC report (North et al.).

Thus, once again, the issues I raise herein have already been brought to the attention of the authors. To claim we neither understood nor addressed them, while ignoring or misrepresenting our published discussions, is derogatory and un scholarly.

I therefore request immediate retraction of pages 36—41 of the Supplement to BNL, and publication of a notice to the effect that this material has been removed from *Theoretical and Applied Climatology*.

Yours truly

A handwritten signature in black ink, appearing to read 'R. McKittrick', written in a cursive style.

Ross McKittrick