# Comment on "Agnotology: Learning from Mistakes" by R.E. Benestad, H.O. Hygen, R. van Dorland, J. Cook and D. Nuccitelli

**Submitted to *Earth Systems Dynamics Discussion***

Reviewer:
Ross McKitrick
Department of Economics and Finance
University of Guelph

I should disclose that I was a reviewer of this paper when it was submitted previously to another journal. Since the authors declined to correct most of the problems that led to its earlier rejection, my comments here simply repeat much of what I have already said.

The BHDCN paper is a bait-and-switch, in which the authors propose a scholarly essay on the methodology of science, then proceed to deliver something quite different: a scattershot of shallow commentary on a list of climatology papers with which they disagree. It is perfectly valid to publish critiques of papers, but they should be written as such, not offered parenthetically in an essay supposedly on another topic. We are asked to take it as "proven" that the authors of this paper have so decisively rebutted all the papers in their Appendix that we can now turn to a philosophical debriefing on the question of why they ever got published in the first place. Yet as proof, all the Appendix offers is a rehash of old, and mostly unpublished, blog posts. The whole paper is thus a waste of readers' time.

The usage of the concept of "agnotology" is confused and contradictory. They introduce the term in paragraph 1 on page 452 as the counterpart of epistemology, that is, as a branch of philosophy. Then they use it in paragraph 2 as a *method* ("An agnotological study of the climate sciences can shed light…"). Then it is used synonymously with replication analysis and as a rhetorical technique ("the communication of misleading claims is a case of agnotology." para 2 p. 463). Its use in the heading of Section A4 implies it is a form of misrepresentation. Finally the title itself implies agnotology means learning from mistakes, but the paper does not allege "mistakes", instead it alleges widespread research malfeasance, such as ignoring data that doesn't fit a hypothesis or ignoring known physical theories.

Presumably, "agnotology" means an absence of information and a lack of basis for knowing. Yet all the authors' examples allege the opposite situation, namely cases in which (BHDCN assert) there is so much information, and matters are so decisively settled, that we can now assume the debates are all over and BHDCN won them all. There is no special philosophical issue behind their analysis, it is just garden-variety argumentation, most of it at a very trivial level. The agnotology angle appears to be a contrivance to try and make a weak paper sound erudite.

The authors repeatedly discuss replication as an essential part of science, insinuating that the papers they critique are at fault regarding disclosure of data and methods. Yet they provide no evidence that non-disclosure was an issue for the papers they study. The authors of the papers they critique appear to have made their methods and data freely available, and BHDCN do not claim that their work was thwarted by non-disclosure. While they point out that replication work is rare, they don't present any case studies in which replication was actively impeded by failure to release data and/or code. Examples of such studies would be Dewald et al. (1986) and Anderson et al. (1994).

Worse, BHDCN insinuate that their analysis of MM04 and MM07 was hampered by secrecy, by ending their discussion on page 490 with the statement: "Another problem was the lack of openness and transparency, which prevented finding out why the conclusions in some of these cases differed to attempts to replicate (Le Page, 2009)." This is completely misleading. The Le Page article refers to an unrelated incident involving different authors, whereas the data and code for MM04 and MM07 have always been available, and nobody has ever claimed to be unable to replicate the findings. I have pointed this out to BHDCN in response to their previous drafts and it is very objectionable to see them repeat their falsehood here once again.

Case 7 (A2.5) refers to McKitrick 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 McKitrick and Michaels reply to Benestad, which BHDCN do not mention. The McKitrick 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.

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 McKitrick 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.

Schmidt's results did *not* show, contrary to BHDCN's claim, that the MM07 coefficient estimates were inside the model-generated distribution. As is clearly shown in McKitrick and Nierenberg Section 2.2 (emphasis added), they were unambiguously outside the distribution:

> With regard to the [claim by Schmidt], the distributions of the coefficients estimated on GCM data do not encompass the coefficients from either the MM07 data set or any other observational grouping in Table 2. In the next section this will be shown after

> reestimating the model using a correction for spatial autocorrelation. Anticipating the findings, **for none of the socioeconomic coefficients does the 95% Confidence Interval estimated on model-generated data encompass the coefficients estimated on observed data. Consequently the null hypothesis as stated by Schmidt (that there is no contamination) *is* rejected.**

I am at a loss to think of wording that could be any clearer, yet BHDCN repeatedly make the opposite statement about the findings in question.

Finally, Schmidt's remarks about "Japan, Western Europe and the USA" appear only in regards to his Figure 3a which is not part of his discussion of MM07, so its repetition in Case 7 is misleading.

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.

Their discussion of the Douglass et al. paper (Case 6) focuses on the idea that the confidence interval around the mean is not the appropriate measure of the distribution of model results, instead people should examine the range of the data. But the purpose of the literature is to say something about the distribution of GCM outputs on the assumption that they are taken to be varied implementations of the same underlying physics, i.e. that they represent samples of a single data generating process. To characterize the central tendency of a data generating process one uses the first and second moments. That is why Douglass et al. and Santer et al. (and others) have argued about the correct definition of the standard deviation around the mean trend. The range, by contrast, can be made arbitrarily wide simply by running the models often enough, and it is not the appropriate measure for the question being posed. If BHDCN want to argue that the debate is totally misplaced they need to develop their argument in proper depth and address the literature in a scholarly way, not through brief, peremptory commentary.

Foster and Rahmstorf (2011) is cited as support for Santer et al. (2008), yet it does not test the model-data mismatch so the usage is misleading. Also they ignore McKitrick et al. (2010) who used longer data sets than Douglass et al. or Santer et al. and applied panel and HAC estimators robust to non-zero covariance and higher-order AR processes. The McKitrick et al. findings were closer to those of Douglas et al. than Santer et al. regarding the significance of the model-observational mismatch, especially in the Correction (2011) that fixed an error in the GISS data. BHDCN make no mention of this, despite the obvious importance they attach to the question of a model-observational mismatch.

Their discussion of Long Term Persistence (Case 10) is lacking in technical depth, yet the authors dismiss the work of Cohn and Lins (and others) without even attempting to present a statistical rebuttal. It is difficult to see the purpose of this section. The IPCC and others often use an AR1 error model on which to base claims of trend significance. The Cohn and Lins findings are consistent with a wide range of papers on the subject (e.g. Rybski et al. 2006, Lennartz and Bunde 2009, Mills 2010, McKitrick et al. 2010, plus the many others discussed at http://www.climatedialogue.org/long-term-persistence-and-trend-significance/) that find more complex long-memory and higher-order AR processes in climatic time series, thus showing quite categorically that AR1 models exaggerate trend significance. BHDCN seem to disagree with all this, but do not present their own statistical model, much less defend it. Their statement "All processes involving a trend also exhibit some LTP"

is vague and nonsensical. After all, the IPCC uses a trend+AR1 model for all its standard error calculations on the assumption that the data do not exhibit LTP.

The climatedialogue.org exchange features Benestad and van Dorland trying to argue that forcing trends can induce LTP, hence its detection might be interpreted as evidence for forcing rather than random natural variability. Koutsoyiannis commented: "I agree that (changing) forcing can introduce LTP and that it is omnipresent. But LTP can also emerge from the internal dynamics alone as the above examples show. Actually, I believe it is the internal dynamics that determine whether or not LTP would emerge." And I particularly refer Benestad to Bunde's comment:

> "When testing to what extent GHG is responsible for LTP, we found it is not, please have a look at our 2004 GRL, where we also specified the methods. It is very unfortunate that Rasmus does not seem to be able to read this and our other articles on LTP."

These are sound views from knowledgeable experts, and Bunde in particularly directs the BHDCN lead author to papers explaining the methods available to them to make their arguments. Yet rather than doing so, BHDCN resort to handwaving and insinuations that the many LTP papers in the literature are forms of misinformation.

Case 12 refers to McKitrick 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.). In fact, these issues have received so much airing elsewhere that it is hard to see the point of Case 12 at all, especially when the authors resort to a Wikipedia entry as one of their main sources. The authors do not seem to have taken the trouble to properly research the issue, and as such their brief commentary lacks credibility. Nor does it illustrate their elusive concept of "agnotology", it just seems yet another axe to grind at the end of a long, tendentious paper.

## References


Anderson, R.G., and W.G. Dewald (1994). Replication and scientific standards in applied economics a decade after *the Journal of Money, Credit and Banking* project. *Federal Reserve Bank of St. Louis Review* (Nov): 79-83.

Benestad, R.E. (2004) Are temperature trends affected by economic activity? Comment on McKitrick & Michaels Climate Research CR 27:171-173

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson (1986). Replication in empirical economics: the *Journal of Money, Credit and Banking* project. *American Economic Review* 76(4): 587-603.

Douglass, D. H., J. R. Christy, B. D. Pearson, and S. F. Singer (2008), A comparison of tropical temperature trends with model predictions, Int. J. Climatol., 28, 1693–1701, doi:10.1002/joc.1651.

Foster, G. and S. Rahmstorf (2011) Global temperature evolution 1979–2010, *Environ. Res. Lett.* **6** 044022 doi:10.1088/1748-9326/6/4/044022

Lennartz, S., and A. Bunde (2009), Trend evaluation in records with long-term memory: Application to global warming, *Geophys. Res. Lett.*, 36, L16706, doi:10.1029/2009GL039516.

Mann, M.E., Bradley, R.S. and Hughes, M.K., (1999). Northern Hemisphere Temperatures During the Past Millennium: Inferences, Uncertainties, and Limitations, *Geophysical Research Letters*, 26, 759-762.

McKitrick, 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.

McKitrick, Ross R., Stephen McIntyre and Chad Herman (2010) Panel and Multivariate Methods for Tests of Trend Equivalence in Climate Data Sets. *Atmospheric Science Letters* DOI: 10.1002/asl.290.; Correction to "Panel and Multivariate Methods for Tests of Trend Equivalence in Climate Data Series" *Atmospheric Science Letters* October 7 2011, DOI: 10.1002/asl.360.

Mills T. C. (2010) Skinning a cat: alternative models of representing temperature trends. *Climatic Change* 101: 415-426, DOI 10.1007/s10584-010-9801-1.

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

Rybski, D., A. Bunde, S. Havlin, and H. von Storch (2006), Long-term persistence in climate and the detection problem, *Geophys. Res. Lett.*, 33, L06718, doi:10.1029/2005GL025591.

Santer, B. D., et al. (2008), Consistency of modelled and observed temperature trends in the tropical troposphere, *Int. J. Climatol*., 28, 1703–1722, doi:10.1002/joc.1756.

Schmidt, GA (2009) Spurious correlation between recent warming and indices of local economic activity. *International Journal of Climatology* 10.1002/joc.1831