Clim. Past Discuss., 2, S163–S170, 2006
www.clim-past-discuss.net/2/S163/2006/

Climate
of the Past
Discussions

Interactive comment on "On the verification of climate reconstructions" *by* G. Bürger and U. Cubasch

Anonymous Referee #2

Received and published: 7 July 2006

The Reply by Burger and Cubasch (henceforth "BC") provides further substantiation of the lack of merit in their arguments, and the unpublishability of their submission in the peer-reviewed literature. BC have provided no substantial responses to the large number of critical errors and problems noted in my original review, and instead introduce a number of new red herrings which serve only to obfucscate, rather than clarify the key issues:

I'm pleased that BC raise the subject of the recent NRC support. The reported supported the key findings of Mann et al (1998,1999) and emphasized that they are now bolstered by a large number of independent studies coming to the same conclusions. The report also specifically criticized the Burger and Cubasch (2005;2006) method of putative significance estimation as incorrect.

BC introduce considerable additional spurious claims and arguments in their latest comments. It would take more time than it is worth for a thorough point-by-point rebuttal. It should be sufficient to refute their main points, as is done below:

1. BC cite some supposed inconsistency between verification RE scores for reconstructions based on two different methods (RegEM and MBH98 approach) using the same proxy data set. The comment is baffling. For Mann et al (1998), the minimum Northern Hemisphere mean verification RE score is 0.49 (for the network available back to AD 1500). For Rutherford et al (2005; henceforth 'R05') from the table 2 cited by the author, the minimum value is 0.46 (for the network available back to AD 1400). Given that calibration intervals (1902-1980 in MBH98 and 1901-1971 in R05) and validation intervals (1854-1901 in MBH98 and 1856-1900 in R05) were different, and the methods are different, such a close similarity is quite a convincing ıconsistency, hardly the inconsistency BC would like readers to believe exists. The number 0.40 cited by BC is disingenuous cherry-pick. Any reader of R05 knows that the authors favoured the hybrid method which separately calibrates interannual and decadal patterns of variance. The authors apparently chose to ignore the hybrid RegEM analysis which gives consistently better results, and focus on the non-Hybrid method because it happens to return a slighly lower (RE=0.40 vs R=0.46), but nonetheless quite statistically significant result for the earlier period. This the sort of cherry-picking one is used to seeing on contrarian websites claiming to "debunk" global warming by finding one or two locations on the globe that have cooled during the 20th century. One doesn't expect this sort of behavior, however, from serious scientists in the field.

2. Remarkably, BC still appear to be attempting to defend the ridiculous detrending step. Their argument seems to be that since this step increase the "sensitivity" of the method (i.e.,it produces spurious reconstructions which of course fatten the distribution of estimates returned), it should be used. Huh?

Mann et al (in press) have shown indisputedly that this step is simply erroneous. They a very general set of examples from a well-behaved (NCAR) simulation of the past 1000 years, and demonstrating that the correctly implemented RegEM procedure gives an excellent reconstruction even at every low proxy signal-to-noise ratios and even at very high levels of redness ($\rho = 0.71$) in the proxy noise, and passes statisticaly validation. They then show that implementation of the erroneous Zorita/von Storch/Cubasch/Berger step of detrending the data prior to calibration produces a meaningless reconstruction, and it fails validation. There is no room for further meaningful discussion on this scientific point. The detrending step is erroneous, and it distressing that those who have used it have backed themselves into a corner where they are unwilling to admit this simply fact which is now widely obvious to the rest of the community.

The reader is referred to the following online articles for further discussion:

http://www.realclimate.org/index.php/archives/2006/04/a-correction-with-repercussions/ http://www.realclimate.org/index.php/archives/2006/05/how-red-are-my-proxies/

3. RegEM as implemented in Schneider (2001) and subsequently used by Rutherford, Mann, and collaborators in paleoclimate reconstruction, provides both error estimates in the regression coefficients (elements of B) and a residual error covariance. The variance estimates for the estimated values (i.e., the reconstruction) is what is of primary interest, and can be obtained from the latter.

The authors erreously argue that for a small number of proxies (e.g. 22), no regularization is required. This completely misunderstands the concept of regularization. The issue isn't the nominal number of predictors (e.g. proxies), but rather that effective number of degrees of freedom in the predictor set vs. the predictand. In practice, there are potentially only a few dozen degrees of freedom, at most, in the instrumental surface temperature record. So even with modest (say, 22) number of proxies, a strict regression is likely to be ill-posed. It would be foolhardy to, as BC suggest, assume well-posedness from the start given the likelihood for ill-posedness. It is far more sensible to used a regularized procedure such as RegEM under these circumstances, and alllow the themselves to determine (through generalized cross-validation) the degree of regularization that is required. In the limit where no regularization appears to be required, RegEM will reduce to the EM algorithm. Surely BC must understand this?

4. BC are their most disappointing (and least convincing) here. It is clear that they insist on distorting the actual procedure used by Rutherford, Mann and collaborators, i.e. they apparently seek to introduce a different "playing field" because they can't compete on the only legitimate one that exists: If the authors truly believe their claims have any validity at all, they should be able to take the ıactual unaltered RegEM procedure as used by Mann and collaborators, the actual target of their reconstruction approach (the surface temperature field), and find fault with the clear demonstration by Mann et al (2005;2006) that this method produces skillful reconstructions that agree closely with the actual data (i.e. are within the estimated uncertainties) in rigorous tests using a forced long-term model simulation with synthetic "pseudoproxies" that have signal-to-noise (SNR) ratios lower than those estimated by Burger and Cubasch themselves (SNR=0.4), a short (1900-1980) calibration period, and even when the proxies have a much redder noise spectrum ($\rho = 0.71$) then is estimated for actual proxies used by MBH98 and others. The authors would also have to find fault with the clear demonstrino in Mann et al (2006) that detrending data prior to calibration completey undermines climate field reconstructino performance. The model surface temperature field, synthetic proxies, source codes for implementing the RegEM and hybrid RegEM method and uncertainty and significance estimation are all publically available. Until the authors can somehow demonstrate that the conclusions provided by Mann et al (2005;2006) are incorrect, it is simply impossible to take any of the nonsense they offer up here, at all seriously.

7. BC increasingly lose credibility in this discussion. Surely, they are aware that Rutherford et al (2005) performed split calibration/verification with a completed version of the instrumental record back to AD 1856, and get essentially the same results as MBH98

who used only a sparse (219 gridbox) subset for the early (1854-1901) period. This addresses the issue of stationarity over the calibration period too, as latter and early halves of the data are alternatively used by Rutherford et al (2005) for calibration and validation. The validation statistics indicated originally by MBH98 are clearly demonstrated to be robust, regardless of the increasingly implausible claims by BC to the contrary.

Moreover, the focus on "22" proxies is somewhat specious. BC appear to be arguing now that the behavior for a sparse subset of 22 proxies such as used by MBH98 back to AD 1400, is fundamentally different from the behavior using larger later proxy networks. This has been easily falsified. Mann et al (2006) perform experiments similar to Mann et al (2005), but using the sparse network available back to AD 1400, and even the sparser network back to AD 1000 (12 proxies). Even with the sparsest network, and SNR values lower than estimated by Burger and Cubasch (2006), they obtain skillfull reconstructions that agree with the true model series within estimated uncertainties.

8. More pure nonsense. Rutherford et al (2005) clearly stated that the hybrid approach performs better in climate model simulation tests than the non-hybrid approach. This is quite sensible for the reasons described by Rutherford et al (2005) (i.e. the patterns that dominate interannual variability such as ENSO are quite different from those that dominate decadal and longer-term variaiblity). This has now been demonstrated rigorously in experiments using "pseudoproxies" based on the NCAR CSM 1.4 model of the past millennium (Mann et al, 2005;2006). It has also been shown (Rutherford et al, 2005; Mann et al 2005;2006) that which procedure is used (hybrid or non-hybrid) hardly makes much of a difference at all in the end, so this is basically another red herring put out by BC. The hybrid approach is objectively defensible based on both ıa priori considerations about the timescale dependence of patterns of climate variability, and its performance in independent tests with model simulation data. It is not, despite the disingenuous efforts by BC to convince readers otherwise, in any way an ıa postiori choice

9. I would refer the readers back to basic work that go back decades (Lorenz, Brier, etc.; see the introductory textbook by D.S. Wilks "Statistical Methods in the Atmospheric Sciences" as well) on forecast verification and significance. The literature is quite clear on what constitutes significance, and this applies equally well to paleoclimate reconstructions (which are a "backward forecast" based a "modern training period") as they do weather or climate forecasting.

To say that a forecast or reconstruction has ‘skill’ is simply to say that its performance is significantly greater than would be expected from some ‘null’ forecast at some appropriately high threshold for random occurrence. One possible null forecast is ‘climatology’ (i.e., that the forecast or reconstruction just has the long-term mean of the ‘calibration’ set). In this case, a statistically significant reconstruction would be one that performs greater (as measured by some appropriate skill diagnostic) than Gaussian data with that mean some appropriate (e.g. 95% corresponding to significance at the $p = 0.05$ level) fraction of the time. Generally, however, a more challenging null forecast is required, which recognizes the existence of serial correlation in all climatic times series, and instead, invokes an null hypothesis of ‘red noise’ (in the forecasting literature known as ‘damped persistence’). Such a null hypothesis is invoked through the use of an AR(1) autocorrelated process to represent a realization of the null hypothesis, where the autocorrelation coefficients, means, and variances are taken from the calibration set. Again, statistical significance of a forecast (or in this context, a reconstruction) is equated with the ability to reject the null hypothesis at an appropriate (e.g. $p = 0.0$ level) threshold. This is precisely the approach used by Mann et al (2005;2006)and BC are simply referred there.

10. More plain nonsense. Rutherford et al (2005) "choose" the sub-periods periods 1856-1925 and 1926-1995 because they precisely split the full available interval for calibration (1856-1995) in half. I suppose BC think that Rutherford et al (2005) tried every possible partitioning into fraction x and $1 - x$ and some converged on x=0.5 because it gave especially favorable results? Give us a break!

Moreover, the performance of skill diagnostics and uncertainty estimates diagnosed over short available validation intervals (which is all that is available in the real world) is explicitly compared with diagnostics available from a millenial-long validation interval (which is possible in model-based pseudoproxy experiments) by Mann et al (2006), who demonstrate thathe inferences available from these precise long-term validation intervals are consistent with the sampling distributions of skill and uncertainty estimates diagnosed from short validation intervals. In other words, the uncertainty estimation and statistical validation approach used by Rutherford et al (2005) has been extensively tested using extensive independent experiments which employ "outside sample" validation of "within sample" estimates. The authors would do well to read Mann et al (2005;2006) in greater detail.

11. This is especially disappointing. Based on the above, the authors appear to have gotten little at all out of their reading of Mann et al (2005). Morevoer, Mann et al (2006) have already dispelled the specious claim that the findings for the AD 1400 sparse network are any different for those for the full network. They are not. Moreover, what can the authors possibly mean by "non-sense" predictors if not predictors that are composed entirely or almost entirely of noise. At SNR=0.25, for which Mann et al (2005) show a skillful reconstruction is still produced, the pseudoproxies are composed of 94% noise by variance. In more recent work Mann et al (2006) have shown this is true even if the noise is substantially more 'red' than is supported for actual proxy records. Mann et al (2006) show that the performance for a fixed SNR=0.4 (86% noise by variance) are very similiar that for a multiproxy data set with the same average SNR (0.4), but for which the SNR for individual pseudoproxies ranges from SNR=0.1 to SNR=0.7. Would BC try to seriously argue now that pseudoproxies with SNR=0.1 (essentially, entirely composed of noise) are not 'nonsense predictors' by their definition. Lets think a bit about what the real "nonsense" is here.

In summary, it is clear that BC cannot engage in any productive discussion with regard to the critical points raise. Instead, they have somewhat defensibly introduced a number of other specious claims and red herrings that are easily dispelled, as above.

A continued back-and-forth with BC and their supporters (Zorita and Von Storch) is therefore unlikely to shed any further light on the discussion. My original review and response to their comments thereon should provide more than an adequate basis for the re-rejection by the editors of the current Burger and Cubasch submission.

The fledgling "Climate of the Past" journal simply deserves far better, especially in this critical early stage for the journal.

Interactive comment on Clim. Past Discuss., 2, 357, 2006.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper