Supplement to Presentation to the National Academy of Sciences Expert Panel, "Surface Temperature Reconstructions for the Past 1,000-2,000 Years.": Recommendations

Stephen McIntyre,
Toronto Ontario

Ross McKitrick, Ph.D.
Associate Professor
Department of Economics
University of Guelph
April 4, 2006

Panelist Otto-Bliesner asked us whether we had any opinions on how reconstructions should be carried out. At the presentation, we did not offer to answer the question. With a little reflection, we are prepared to provide some short comments aimed at improving methodologies.

Journal Policies

1. Paleoclimate journals should adopt replication policies based on those currently implemented at economics and econometrics journals, e.g. the *American Economic Review*, which requires all authors to archive data and code. Authors should provide precise data citations for digital data versions used in public archives as part of a detailed SI. In principle, AGU policies require something close to this, but the policies are not implemented. Journals should pass as much of the responsibility as possible to authors (not reviewers) and should require authors to assert (as an online form) that they have complied with journal policies on code and data. The code and data should actually run.
2. Paleoclimate journals should require authors to state exactly what population of proxies was sampled and exactly what *ex ante* data selection criteria was used to select the subset entered into the analysis. This applies to site sampling as well as selecting data series for inclusion in paleoclimate reconstructions. We are very sceptical of field methods (associated with, e.g., Jacoby and Briffa) in which proxies of a similar type are selected or rejected ***ex post*** depending on their correlation to temperature. We are also sceptical of methods in which reconstructions just manage to maintain a slight ranking of the MWP below the present, based on a small number of proxies selected out of the large number available, with no explanation why these were chosen and others rejected. Journals should require authors to demonstrate that post hoc cherry-picking was avoided. While a relation to temperature is obviously essential, an essential safeguard against spurious correlations requires proper population sampling and specific *ex ante* selection criteria. It would be inconceivable in a drug trial only to report "favourable" cases. Yet climate scientists routinely select the "most temperature-sensitive sites" and then may not even archive or report the other sites.

Funding Agency Policies

3. Funding agencies have responsibilities and authority that are distinct from the journals and they should not rely on journals to discharge data oversight. Funding agencies should establish procedures to verify that authors archive data resulting from their studies on a timely basis and ensure that the archives are complete (and not merely selective.)
4. Funding agencies should make a concerted effort to require those authors who are presently noncompliant with U.S. federal policies to bring past archiving into compliance as a precondition to receiving further funding.
5. Data archiving should be comprehensive and detailed. For example, a comprehensive ice core archive would present data on a sample-by-sample basis, not on 10-year

averages. If 10-year averages are used as an intermediate working paper, those should be archived as well.

6. Data should be archived either at the time of any publication using the data or, in the case of federally funded data, within 2 years of collection (as prescribed under federal policy.)
7. The validity of "classic" proxies to 1990s warmth urgently needs to be demonstrated. It is scandalous that multiproxy studies use obsolete proxy series ending in the 1980s or even 1970s and link these to out-of-sample instrumental records without verifying proxy validity. Sometimes proxy collectors blame funding agencies for this [Esper et al, QSR 2005]. The panel should recommend to NSF that some money be set aside for updating "classic" sites.
8. Dr. Alley pointed out that updating may not be of interest to tenure-track PhD programs. However, NSF is not limited to that form of labour supply and should investigate other methods of updating e,g. simply hiring someone to update the tree ring data for immediate measurement archiving.
9. The ecological and locational information on tree ring samples urgently needs to be upgraded. For example, it is trite that tree lines go up and down with temperature and that temperature changes with altitude. However, altitude information on cores is not recorded in the ITRDB data bank. With GPS instruments, it is trivial to record altitude information on a sample by sample basis in case it proves to be relevant.

Statistical Issues

10. Hughes made a distinction between two "approaches" towards millennial paleoclimate data: what he called the "Schweingruber approach", in which a large collection of **temperature-sensitive** proxies is collected; and what he called the "Fritts approach" in which proxies, especially tree-ring proxies, are collected without distinguishing between temperature and precipitation proxies, relying on statistical methods to extract "climate fields". We are not sure that Fritts would necessarily wish to be the eponym for the latter approach, which more closely resembles MBH98 and might be labelled the "Mann approach". Critical attention needs to be paid to the risk that both methods are prone to data mining. With regards to the latter, given the extreme noise levels of proxies, it is unwise to run an unsupervised algorithm without rigorously evaluating significance against red noise benchmarks.
11. The use of calibration period residuals for calculating confidence intervals clearly exaggerates the quality of a reconstruction. While it is doubtful that verification periods in these calculations are truly independent in a statistical sense, there is much evidence of statistical overfitting in the calibration period in typical multiproxy studies and the use of calibration period residuals is unacceptable. This obvious point seems unknown to the paleoclimate community, either to authors or to journal reviewers. The NAS Panel should send a strong message to the community.
12. We loathe the repeated use of stereotype proxies in studies purporting to be "independent". The existence of any overlap raises the spectre that it is the overlapping series that are creating the "signal" rather than the non-overlapping

signals. A very few HS shaped series (bristlecones, Yamal, Jacoby's Mongolia, Dunde, Briffa's Polar Urals) appear repeatedly in virtually all the well-known studies and are pivotal to the ranking of the present era against the MWP. Hence the repeated use of these series destroys any claim they might make to provide "independent" evidence.

13. For small subsets, such as typical multiproxy studies, we think that robust statistics (e.g. median) should be preferred to non-robust statistics (e.g. mean). Likewise for estimates of variance.

14. Far more attention needs to be paid to autocorrelation. While many paleoclimate authors are aware that it diminishes the effective degrees of freedom, a small adjustment to the denominator of a variance formula is an inadequate treatment of the problem. Variance calculation requires a model of the residual that yields a truly uncorrelated error term. This may entail, for example, an ARMA model, and testing for ARCH and other features of correlated time series. Also, the presence of autocorrelation indicates potential misspecification of the calibration model, including possible failure to treat nonlinearities. Finally, autocorrelation in temperature and proxy data creates the conditions for spurious correlation, as was explored in Ferson et al [2003]. If proxy selection is based on spurious correlation scores, the outcome will be an apparently strong in-sample fit coupled with poor out-of-sample prediction properties. This may account for the "divergence" problem – a problem all too familiar with stock market forecasting systems.

15. The practice of cherry-picking which verification statistics to report should be strongly discouraged. If authors have adverse results in some verification statistics, all codes of conduct require them to be reported; the authors may, of course, discuss why these adverse results should be disregarded. The panel should not be perceived as acquiescing in any relaxation of this standard.

Data Issues

16. A much more concerted effort needs to be made assessing impact of non-homogeneity in tree ring collections. In **all** collections, there is significant non-homogeneity in age distribution; in important collections, there is significant non-homogeneity in altitude. Non-homogeneity needs to be routinely assessed and disclosed.

17. We are not convinced that attempts to extract low-frequency information from "site chronologies" as currently calculated sufficiently avoids important biases, and some other approaches need to be evaluated. We think that information on changing treelines and changing altitudes at individual sites is significantly under-utilized in the present generation of millennial climate studies. Such information was widely considered in the older generation of millennial paleoclimate studies [e.g. Bray 1971]. Two recent studies utilizing this class of information in novel and sophisticated ways, which have strongly impressed us are Naurzbaev et al [2004] (including MBH coauthor Hughes) and Millar et al [2006]. (See companion note.) We urge the NAS

Panel to recommend to NSF that it fund additional studies of these types in other regions.

18. We think that authors of multiproxy studies must make a more serious effort to reconcile seemingly discrepant results from the same or nearby sites. For example, results from Yamal and the updated Polar Urals are strongly discrepant. Yamal is also discrepant with Naurzbaev et al [2004]. These series are widely used in climatic reconstructions. The fact that resampling from the same site yields such discrepant results is a fundamental challenge to the meaning of proxy data. If these trees are recording a climate signal, then re-sampling from the same site ought to yield similar chronologies. If they are starkly different, as is the case for ring width chronologies from Yamal/Polar Urals, a simple conclusion is that the tree ring width chronologies are inherently poor carriers of climatic information and their use in temperature reconstructions is opportunistic. Instead of dealing with these challenges head-on, the two most recent multiproxy authors [Osborn and Briffa, 2006; D'Arrigo et al 2006] failed to disclose the data discrepancies and used the most HS shaped alternative available to them.