Are multiproxy climate reconstructions robust?

Gerd Bürger and Ulrich Cubasch
Institut für Meteorologie, Freie Universität Berlin, Berlin, Germany

Received 30 July 2005; revised 7 October 2005; accepted 1 November 2005; published 14 December 2005.

[1] 64 climate reconstructions, based on regression of temperature fields on multi-proxies and mutually distinguished by at least one of six standard criteria, cover an entire spread of millennial histories. No single criterion is accountable for the spread, which appears to depend on a complicated interplay of the criteria. The uncertainty is traced back to the fact that regression is applied here in an extrapolative manner, with millennial proxy variations exceeding the standard calibration scale by a factor of 5 and more. Even if linearity still holds for that larger domain the model error propagates in a way that is proportional to both the estimation error and the proxy variations, and is thus extrapolated accordingly. This is particularly critical for the parameter-loaded multiproxy methods. Without a model error estimate and without techniques to keep it small, it is not clear how these methods can be salvaged to become robust. **Citation:** Bürger, G., and U. Cubasch (2005), Are multiproxy climate reconstructions robust?, *Geophys. Res. Lett.*, *32*, L23711, doi:10.1029/2005GL024155.

1. Introduction

[2] Among several proxy-based approaches of reconstructing real or synthetic millennial climate [*Overpeck et al.*, 1997; *Jones et al.*, 1998; *Briffa*, 2000; *Crowley and Lowery*, 2000; *Briffa et al.*, 2001, 2004; *Esper et al.*, 2002; *Zorita et al.*, 2003; *Jones and Mann*, 2004; *von Storch et al.*, 2004] the most prominent and most disputed of all is certainly the one of *Mann et al.* [1998], henceforth MBH98. Besides its prominent role in the third IPCC report [*Intergovernmental Panel on Climate Change*, 2001], the study partly owes its popularity a number of methodological issues that are left unsettled in the original version, and which after several critical remarks [cf. *McIntyre and McKitrick*, 2003] led to the publication of a corrigendum. The discussion, nevertheless, continued [*von Storch et al.*, 2004, *McIntyre and McKitrick*, 2005a, 2005b; *Rutherford et al.*, 2005; *Bürger et al.*, 2005], indicating that several issues are still unsettled, all related to the problem of reproducibility and robustness. For instance, assertions made by MBH98 and later about certain steps (such as rescaling) being "insensitive" to the method were hard to quantify and thus of little help. *Bürger et al.* [2005] showed that the method is, on the contrary, highly sensitive to the variation of 5 independent standard criteria (as we call the steps here), resulting in an entire spectrum of possible climate histories. Those experiments were conducted in the synthetic world of a climate model, with noise-disturbed temperature grid points serving as pseudo-proxies, and it turned out that the amplitude of the reconstructions ranged between about 20% and 100% of the true (simulated) millennial history. Whether or not these results extend to the real-world case, i.e. whether or not the MBH98 and relative approaches are robust, including the predictor selection issues as argued by *McIntyre and McKitrick* [2005a], is the subject of the current study.

0094-8276/05/2005GL024155$05.00

2. $2^6 = 64$ Flavors of Regression

[3] The climate reconstruction employed by MBH98 applies an inverse regression (see below) between a set of multiproxies on the one hand and the dominant temperature principal components (PCs) on the other. The decreasing availability of proxy data back in time is accounted for by estimating the regression for seven successive time periods. For reasons of simplicity we skip this latter step in our study, and approximate the MBH98 setting as follows (following MBH98 SI): We use the proxies that are available in the time period 1400–1450 (18 single dendro and ice core proxies identical to MBH98 plus, depending on the reference period, up to 6 leading principal components representing two denser dendro sub-networks used by MBH98). The north-hemispheric temperature (NHT) field is represented by its first PC, exactly as in the work by MBH98, with a spatial coverage of 1052 (219) grid points for the 1902–1980 calibration (1854–1901 validation) phase. From these data an empirical model is fitted, and then applied to the full proxy record to reconstruct the climate history from 1400 to 1980.

[4] This is the statistical nucleus of MBH98, and if it is robust certain refinements such as rescaling should *not* affect the essence of the final result. The method should, moreover, be robust against the successive addition of further proxy predictors, as in the work by MBH98 from 1450 onwards; for more on this see below.

[5] The following 6 criteria were considered, all belonging to the standard toolbox of empirical climatology:

2.1. TRD

[6] 20th century warming is the dominant variation in the instrumental data. It covers about half of the full variance, while the other half stems from purely interannual variations. Whether or not one builds the model on trended or detrended data should therefore affect the result. Note that *von Storch et al.* [2004] detrended the data (E. Zorita, personal communication, 2005) while MBH98 did not. From other studies [cf. *Briffa et al.*, 1998] it is known that inconsistencies exist between proxy and instrumental trends in the 20th century.

2.2. PCR

[7] Before estimating the regression model, the proxy predictors undergo a PC transformation (PC regression

Table 1. The Six Criteria Used to Define a Regression Flavor

	TRD	PCR	GLB	INV	RSC	CNT
0	no trend	no PCR	spatially explicit	direct	no rescaling	PCA decentered
1	trend	PCR	global	inverse	rescaling	PCA centered

[e.g., *Briffa et al.*, 2001]). This is a useful measure against *colinearity*, a complication that inflates the model error [*Johnson and Wichern*, 2002]. In the present context, colinearity is induced through the common positive 20th-century trend in many proxies. PCR moreover serves as a noise filter by retaining only the dominant predictor PCs (in our case: 50% explained variance).

2.3. GLB

[8] One can use either the single predictand NHT or, alternatively, a set of leading principal components so that spatial detail is simulated as well. But note that like MBH98 we use just one PC.

2.4. INV

[9] Direct regression is the kind of regression that is normally applied, here, as a regression of the instrumental temperature fields (predictand) on the proxies (predictor). Inverse regression goes vice versa, first, by regressing the proxies on temperature and, second, by finding for a given proxy the temperature field with the closest (in a least squares sense) image to the proxy under the regression map. This is the same as inverting the regression map using the pseudo inverse. It is noteworthy that the simulated amplitudes of a multiple direct regression are scaled by the *canonical correlations* between predictor and predictand field, while the inverse form is scaled by the inverse of those correlations [see *Bürger et al.*, 2005]. No error estimate of the model coefficients was given by MBH98; *Sundberg* [1999] has some material on this.

2.5. RSC

[10] To match simulated and original variability, rescaling of the predictand is sometimes applied with scaling factors taken from the calibration period. This ensures adequate variability at least for that period, but introduces uncontrollable results if that domain is left. RSC is frequently encountered in statistical downscaling under the name inflation [cf. *Karl et al.*, 1990]. Note that if either one of INV and RSC is applied the simulated amplitude is increased relative to observations; this is in conflict with the damping arguments given in [*von Storch et al.*, 2004]. We have not found any reference regarding the effect of rescaling on model uncertainty.

2.6. CNT

[11] The MBH98 choice of calculating the PCs of some proxy clusters from anomalies of the 20th century climate has been criticized for reducing off-calibration amplitudes and favoring hockey stick shaped results [cf. *McIntyre and McKitrick*, 2005a, 2005b]. Under the CNT criterion those PCs are determined from the full period to temper the impact of a strong positive 20th-century trend. We applied Preisendorfers rule N for selecting the PCs.

[12] Note that each single criterion is a priori sound, with numerous applications elsewhere, and can hardly be dismissed purely on theoretical grounds. Note further that all of the above criteria are independent, mutually consistent and can thus arbitrarily be mixed, so that any combination thereof defines one of $2^6 = 64$ reasonable "flavors" of the regression model. Following Table 1 we identify a flavor using a binary code of length 6, indicating whether any of the 6 criteria is valid or not. For example, 100110 refers to an inverse regression with rescaling, trend, and spatially explicit predictands, and without using PCR; this is the variant used by MBH98, and we denote it by MBH.

3. NHT Reconstructions

[13] Figure 1 shows the 64 variants of reconstructed millennial NHT as simulated by the regression flavors. Their spread about MBH is immense, especially around the years 1450, 1650, and 1850. No a priori, purely theoretical argument allows us to select one out of the 64 as being the "true" reconstruction. One would therefore check the calibration performance, e.g. in terms of the reduction of error (RE) statistic. But even when confined to variants better than MBH a remarkable spread remains; the best variant, with an RE of 79% (101001; see supplementary material[1]), is, strangely, the variant that most strongly deviates from MBH.

[14] It may be important to stress the following: On the basis of the validation RE one might be tempted to prefer the (most simple) variants 100000 or 101000, or also MBH, to the others. But that statistic must not be used to *select* a model; it can only serve as a check of a model, e.g. for overfitting, *after* it has been selected. To do otherwise amounts to extend the calibration over to the validation period. In that case, i.e. using the calibration 1854–1980, the simulations look remarkably different (not shown).

[15] We have analyzed the influence of each of the criteria on the overall behavior of the simulation. Here it appears that only TRD (=1) induces slightly higher amplitudes, all other criteria are thoroughly mixed. They have nevertheless a significant, however non-unique, influence on the simulations. For any criterion, the range of values with that criterion held fixed is considerable and decreases only for TRD and RSC. We note that MBH is not very different from the original MBH98 version (not shown) where proxies are added successively, indicating that our selection of proxies (those reaching back to AD 1400) is already representative for the purpose of this study. We have nevertheless conducted the same experiments under the

[1] Auxiliary material is available at ftp://ftp.agu.org/apend/gl/2005GL024155.

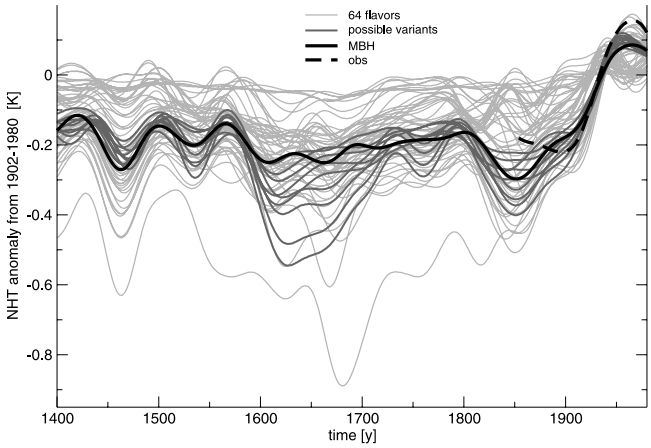

Figure 1. $2^6 = 64$ variants of millennial NH temperature, distinguished by smaller (light grey) and larger (dark grey) calibration RE than the MBH98 analogue (MBH, black). Instrumental observations are dashed. All curves are smoothed using a 30y filter.

setting of the AD 1600 step where more proxies (57) are available. The variations are comparable to those seen in Figure 1. The spread is particularly large in the earliest part of the simulations, especially among those with a calibration RE higher than MBH (cf. SM). But they have a negative validation RE, which indicates overfitting.

4. Uniformitarianism and Extrapolation

[16] Fundamental to all dendrochronological inferences on climate is the following principle of uniformitarianism, as stated by *Fritts* [1976, p. 15]: "Therefore, one can establish the relationship between variations of tree growth and variations in present-day climate and infer from past rings the nature of past climate." The principle obviously generalizes to the broader context of multiproxies, but evidently our results do not give such a relationship, at least not one that is sufficiently robust. But as *Fritts* [1976, p. 15] continues: "In order to make this kind of inference, however, it is important that the entire range of variability in climate that occurred in the past is included in the present-day sampling of environment." This is, in fact, the basic condition of statistical regression - but only one half of it. The other half applies to the tree ring variations: They also must lie in a range that is dictated by the calibrating sample. This, however, is not the case here. For almost all of the 24 proxies, the range of the millennial variation is considerably larger than the sampled one, with numerous cases of proxies exceeding 7 and more calibration standard deviations (cf. SM). As a consequence, the regression model is *extrapolated* beyond the domain for which it was defined and where the error is limited.

[17] This is illustrated by the example of Figure 2. From the simplest variant 100000 the part of the model related to the proxy predictor #20 (P20) is shown. While the model is calibrated using a P20 standard deviation of 1.0, for the year 1644 it is applied to the case P20 = 4.1. For that scale, it is unknown whether the linearity assumption on which the regression model is built still holds. But even if does, for a given linear model $y = \mathbf{B}\ \boldsymbol{x}$ the error (indicated by δ) propagates as

$$\delta y = \mathbf{B}\ \delta \boldsymbol{x} + \delta \mathbf{B}\ \boldsymbol{x}. \quad (1)$$

[18] The larger $\boldsymbol{x}$, the more dominant becomes the second term, especially if $\delta\mathbf{B}$, the model estimation error, is significantly nonzero. Following *Johnson and Wichern* [2002], we estimated $\delta\mathbf{B}$ to be in the range of 20% for P20 and the model 100000.

[19] It is evident that estimates of $\delta\mathbf{B}$ are indispensable to adequately assess the model behavior under extrapolation. Unfortunately, we were not able to find or derive such estimates for models with criteria INV and RSC. But due to phenomena such as colinearity (see above) and overfitting ($\delta\mathbf{B}$ generally increases with the number of model parameters. Models of the kind considered here are susceptible to both, and this would at least partly explain the large spread of the reconstructions.

5. Conclusions

[20] By combining 6 standard criteria to define variants of the basic regression method used in MBH98 we have found an enormous spread in the resulting millennial NHT reconstructions from AD 1400 onwards, with none of the criteria being solely accountable for the spread. This uncertainty persists even among the best performing variants, and we believe that we were able to trace it back to a scale mismatch between the full millennial and the calibrating proxy variations. Under such circumstances, the regression model leaves its generic domain of validity and is applied in an extrapolative manner. Even if linearity still

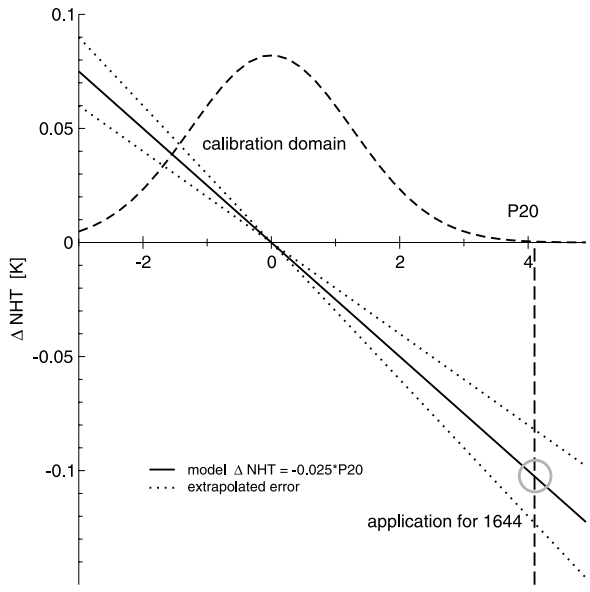

Figure 2. Extrapolation of regression model 100000. The dashed curve indicates the distribution of the calibration domain of Proxy #20 (P20), with a standard deviation of 1.0. For the year 1644, the model is extrapolated to more than 4 times of that scale (grey circle). Ordinate is the relative contribution of P20 to the simulated NHT. Error propagation indicated by two dotted lines (see text).

holds for the larger scales, the error is prone to be linearly inflated by those scales.

[21] Any robust, regression-based method of deriving past climatic variations from proxies is therefore inherently trapped by variations seen at the training stage, that is, in the instrumental period. The more one leaves that scale and the farther the estimated regression laws are extrapolated the less robust the method is. The described error growth is particularly critical for parameter-intensive, multi-proxy climate field reconstructions of the MBH98 type. Here, for example, colinearity and overfitting induce considerable error already in the estimation phase. To salvage such methods, two things are required: First, a sound mathematical derivation of the model error and, second, perhaps more sophisticated regularization schemes that can keep this error small. This might help to select the best among the 64, and certainly many more possible variants. In view of the relatively short verifiable period not much room is left.

[22] **Acknowledgments.** We thank Irina Fast for fruitful discussions. This work was funded by the EU project SOAP.

References

Briffa, K. R. (2000), Annual climate variability in the Holocene: Interpreting the message of ancient trees, *Quat. Sci. Rev.*, *19*, 87–105.

Briffa, K. R., F. H. Schweingruber, P. D. Jones, T. J. Osborne, S. G. Shiyatov, and E. A. Vaganov (1998), Reduced sensitivity of recent tree-growth to temperature at high northern latitudes, *Nature*, *391*, 678–682.

Briffa, K. R., T. J. Osborn, F. H. Schweingruber, I. C. Harris, P. D. Jones, S. G. Shiyatov, and E. A. Vaganov (2001), Low-frequency temperature variations from a northern tree ring density network, *J. Geophys. Res.*, *106*(D3), 2929–2941.

Briffa, K. R., T. J. Osborn, and F. H. Schweingruber (2004), Large-scale temperature inferences from tree rings: A review, *Global Planet. Change*, *40*, 11–26.

Bürger, G., I. Fast, and U. Cubasch (2005), Climate reconstruction by regression—32 variations on a theme, *Tellus, Ser. A*, in press.

Crowley, T. J., and T. Lowery (2000), How warm was the medieval warm period?, *Ambio*, *29*, 51–54.

Esper, J., E. R. Cook, and F. H. Schweingruber (2002), Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability, *Science*, *295*, 2250–2253.

Fritts, H. C. (1976), *Tree Rings and Climate*, 567 pp., Elsevier, New York.

Intergovernmental Panel on Climate Change (2001), *Climate Change 2001: The Scientific Basis: Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, edited by J. T. Houghton et al., 944 pp., Cambridge Univ. Press, New York.

Johnson, R. A., and D. W. Wichern (2002), *Applied Multivariate Analysis*, 767 pp., Prentice-Hall, Upper Saddle River, N. J.

Jones, P. D., and M. E. Mann (2004), Climate over past millennia, *Rev. Geophys.*, *42*, RG2002, doi:10.1029/2003RG000143.

Jones, P. D., K. Briffa, T. P. Barnett, and S. F. B. Tett (1998), High-resolution palaeoclimatic records for the last millennium: Interpretation, integration and comparison with general circulation model control-run temperatures, *Holocene*, *8*(4), 455–471.

Karl, T. R., W. C. Wang, M. E. Schlesinger, R. D. Knight, and D. Portmann (1990), A method of relating general circulation model simulated climate to observed local climate. part I: Seasonal statistics, *J. Clim.*, *3*, 1053–1079.

Mann, M., R. Bradley, and M. Hughes (1998), Global-scale temperature patterns and climate forcing over the past six centuries, *Nature*, *392*, 779–787.

McIntyre, S., and R. McKitrick (2003), Corrections to the Mann et al. (1998) proxy data base and Northern Hemispheric average temperature series, *Energy Environ.*, *14*(6), 751–771.

McIntyre, S., and R. McKitrick (2005a), Hockey sticks, principal components, and spurious significance, *Geophys. Res. Lett.*, *32*, L03710, doi:10.1029/2004GL021750.

McIntyre, S., and R. McKitrick (2005b), The M&M critique of the MBH98 Northern Hemisphere Climate Index: Update and implications, *Energy Environ.*, *16*(1), 69–100.

Overpeck, J., et al. (1997), Arctic environmental change of the last four centuries, *Science*, *278*, 1251–1256.

Rutherford, S., M. E. Mann, T. J. Osborn, R. S. Bradley, K. R. Briffa, M. K. Hughes, and P. D. Jones (2005), Northern Hemisphere surface temperature reconstructions: Sensitivity to methodology, predictor network, target season and target domain, *J. Clim.*, *18*, 2308–2329.

Sundberg, R. (1999), Multivariate calibration—Direct and indirect regression methodology, *Scand. J. Stat.*, *26*(2), 161–207.

von Storch, H., E. Zorita, J. M. Jones, Y. Dmitriev, and S. F. B. Tett (2004), Reconstructing past climate from noisy data, *Science*, *306*, 679–682.

Zorita, E., F. González-Rouco, and S. Legutke (2003), Testing the Mann et al. (1998) Approach to paleoclimate reconstructions in the context of a 1000-yr control simulation with the ECHO-G coupled climate model, *J. Clim.*, *16*, 1368–1390.

U. Cubasch and G. Bürger, Institut für Meteorologie, Freie Universität Berlin, Carl-Heinrich-Becker-Weg 6-10, D-12165 Berlin, Germany. (gerd.buerger@met.fu-berlin.de)