Clim. Past Discuss., 2, S173–S179, 2006
www.clim-past-discuss.net/2/S173/2006/

Climate
of the Past
Discussions

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

Anonymous Referee #3

Received and published: 8 July 2006

General comments

Bürger and Cubasch (BC) set out to reassess the skill of "proxy-based reconstructions of Northern [H]emisphere temperature" by applying several variants of reconstruction methods to samples of proxy data and instrumental temperature data generated by a resampling technique. There are several open questions about statistical methods for climate reconstructions—for example, questions about the magnitude of biases of estimated variances and covariances. Resampling techniques such as that used by BC (bootstrap) are well suited to assess such questions.

However, BC's paper lacks focus and diligence in assessing such questions. There are numerous errors and inaccuracies in the use of statistical concepts and methods (for example, in the use of the regularized EM algorithm, see the specific comments below). Concepts are used loosely in a way that leaves their specific reference in the context unclear (for example, ’degrees of freedom’, ’sample’, ’population’, again see specific comments). And the writing is often rambling, without clear focus (for example, it is unclear why it is relevant for an assessment of methods for climate reconstructions that BC “have not been able to find [in the referenced literature] any reference to a validation mean, nor in any of the articles [they] checked from the hydrologic literature . . . ” [361, 14–19]).

To answer the Editor’s evaluation questions, the paper attempts to address relevant scientific questions within the scope of CP, but in doing so does not meet the required standards. The paper does not present novel concepts, ideas, tools, or data. In light of apparent conceptual and methodological errors, it is difficult to assess whether the conclusions reached are substantial. The methods and assumptions are not sufficiently outlined, and the overall presentation is unclear and lacks focus. The title likewise lacks specificity. See the specific comments for details.

Some specific comments

1. The term ’skill’ is used in the abstract quantitatively but without definition, making it impossible to read the abstract as a self-contained entity. Similarly, ’degrees of freedom’ is used without clear referent, and it is unclear how the number of degrees of freedom BC are referring to is defined.

2. The Introduction makes clear that the scope of this paper is narrow, being primarily concerned with a subset of the data analyzed by Mann et al. (1998). A paper making broad methodological claims (as in the abstract) and drawing negative conclusions (“we doubt that [the question whether there was a Medieval Warm Period] can be decided based on current reconstructions alone” [366, 20–22]) should have a wider scope, to put particularly the negative conclusions on a

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

firmer footing.

3. BC state that “these few degrees of freedom also initiated the debate on using trended or detrended calibration” [359, 11–12]. There is a conceptual misunderstanding here (as, apparently, in the referenced paper by von Storch et al. (2004)). The climate reconstruction methods they consider are based on multivariate linear regressions (or inverse regressions) of variables with missing values (temperatures) on available values (proxies), potentially with iterations as in the EM algorithm and variants. Estimating the matrix of regression coefficients in the regression model from detrended data in a calibration period amounts to including a predictor t indexing time (e.g., years) in the regression model. If the regression model *without the predictor* t is then used to infer missing values given available values, a different regression model is used for the inference than was estimated—a procedure that is difficult to justify. (Alternatively, the additional predictor can be viewed as $tH(t - t_0)$, where H is the Heaviside step function and t_0 is the time of the beginning of the calibration period, so that the same model is used for inferring missing values as was estimated. This would mean an arbitrary change point t_0 is specified a priori in the model—likewise a procedure that is difficult to justify.)

4. “The error grows proportional to both the model uncertainty and the proxy scale” [360, 2–3]. This is an instance of vague statements that abound in this paper. Which error? An error variance? What exactly is it proportional to? What is the measure of model uncertainty? Such vagueness is unacceptable.

5. “A regression/verification exercise is generally nonsense if calibration and validation samples are not drawn from one and the same population” [360/361, 27/1]. The authors seem to imply that in some climate reconstructions, the ’calibration sample’ and ’verification sample’ may not be drawn from the same ’population.’ However, the samples are trivially from the same population, for example, if the

Full Screen / Esc
Printer-friendly Version
Interactive Discussion
Discussion Paper

population is 'the climate of the Holocene.' This is not the issue in question in climate reconstructions, neither is it, directly, the stationarity of the timeseries involved. The key assumption necessary in climate reconstruction methods that ignore the reasons why temperature data are missing (all methods considered by BC) is the assumption that temperature values, in an incomplete dataset consisting of temperatures and proxies, are missing at random (Little and Rubin 2002). That is, the probability that a temperature value is missing is independent of the missing temperature value. BC's use of the number of available temperature measurements as a predictor of mean temperature shows that there are, as is well known, correlations between temperatures and missingness, so temperature values are not missing at random. The consequences of the violation of this assumption are what needs to be assessed carefully. (Some aspects of this have been assessed by Rutherford et al., but there are open questions, for example, about biases in estimated variances and about the circumstances in which the biases in reconstructed temperatures become significant.)

6. "Once a calibration subset of these data is defined . . . " [362, 3]. In the regularized EM algorithm, it is not necessary to separate a 'calibration subset.' How is this dealt with in the BC's use of the algorithm?

7. "Note that the informational flow goes strictly from GLB through RSC" [362, 15]. This is not correct for iterative methods such as the EM algorithm and its variants.

8. Numbered list following l. 25, p. 362: The R matrices are evidently matrices of regression coefficients, but they are not defined, and neither is the regression model to which they belong.

9. "The result is rescaled to match the calibration variance" [363, 6]. This step rests on a misconception. If reconstructed temperatures are estimated as conditional expectation values given the available proxies and regression coefficients

relating proxies and temperatures, the sample variance of the reconstructed temperatures will be smaller than the actual temperature variance because possible variations of the missing temperature values around the conditional expectation values are ignored (Little and Rubin 2002). Rescaling reconstructed temperatures may be an ad hoc way of obtaining approximately unbiased estimates of variances as sample variances of the reconstructed time series, but the reconstructed temperature values acquire a bias. A consistent way to estimate variances from a completed dataset is to take the variations about the reconstructed temperature values (i.e., the imputation error variance) into account. See, e.g., Little and Rubin (2002) or any other textbook on the estimation of statistics from incomplete data.

10. Supplement 1: “With the number of proxies being rather limited . . . , it is no longer an ill-posed problem, so why should one apply RegEM, instead of EM?” First, there appears to be a confusion of concepts. The regression problem may not be *rank-deficient*, but if its effective rank given the uncertainties in the stochastic model underlying the reconstruction is small, the regression problem may still be *ill-posed*, such that biased (regularized) regression methods may yield better estimates of missing values (and of the statistics of the data). Second, the regularized EM algorithm of Schneider (2001) reduces to the conventional EM algorithm in the limit of no regularization. This limit should be attained to the extent that the regularization parameter estimation (by generalized cross-validation in the version of the algorithm available online) is adequate. Do the authors in fact find zero regularization to be optimal? This question is not addressed in the paper.

11. Supplement 1: “Does the RegEM algorithm actually converge? . . . No convergence was achieved in several cases.” If the limit of no regularization is adequate and the regularized EM algorithm reduces to the EM algorithm for Gaussian data, convergence is assured because the likelihood function is monotonically increas-

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

ing from iteration to iteration (Dempster et al. 1977). The regularized EM algorithm may not inherit this convergence property if regularization parameters are chosen adaptively, but it would be surprising (and interesting) if it does not converge fairly reliably. (If regularization parameters are fixed, the regularized EM algorithm with ridge regression becomes a constrained maximum likelihood algorithm for Gaussian data (or, from a Bayesian perspective, a maximum a posteriori algorithm with an inverse Wishart prior), for which similar convergence properties as for the EM algorithm hold (cf. Schafer 1997, Little and Rubin 2002).)

Could it be that the authors did not choose an adequate stopping criterion for the iterations? The stopping criterion they use is not documented in the paper. If they used the default stopping criterion of the implementation of the algorithm currently available on the referenced website, the alleged lack of convergence may simply be due to an inappropriately small stagnation tolerance (a relative change in imputed values of less than 5×10^{-3}, which is unreasonably small given the accuracies that can be expected in paleo-reconstructions). That is, the algorithm may have converged and may simply oscillate around the solution within the errors to be expected.

Claims such as the alleged lack of convergence of the algorithm require more detailed scrutiny.

12. Supplement 2: The documentation of the methods used is inadequate and makes an assessment of the results impossible. For example, with regard to the preceding point, it is inadequate to refer to “defaults” of Schneider (2001) for parameters entering the regularized EM algorithm. No “defaults” of stopping criteria and other parameters are given in Schneider’s paper. There are default settings in the algorithm available at the referenced website, but this is not an archival source and may change over time (and may have changed since the authors downloaded the code).

    What motivates the choice of a truncation parameter of 5 for truncated total least

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

squares regression? In Supplement 1, BC questioned the need for any regularization. Why choose such a (seemingly drastic) regularization here? The authors' broad negative conclusions about the merits of different methods are lacking evidence if questions such as the extent to which the truncation affects the results are not examined. (Adressing the question of the relative merit of ridge regression vs. truncated total least squares systematically would be interesting.)

This paper is not ready for publication. There are worthwhile conceptual and methodological questions to be addressed about methods for climate reconstructions, for example: To what extent does the fact that temperature values may not be missing at random lead to biases of reconstructed temperatures? To what extent are error estimates of reconstructed temperatures biased as a result of regularization procedures? Which regularization procedures and procedures for the choice of regularization parameters perform best? This paper does not address such questions sufficiently systematically, clearly, and accurately.

References

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Roy. Stat. Soc. B*, **39**, 1–38.

Little, R. J. A., and D. B. Rubin, 2002: *Statistical Analysis with Missing Data*. 2nd ed., Series in Probability and Mathematical Statistics, Wiley, 381 pp.

Schafer, J. L., 1997: *Analysis of Incomplete Multivariate Data*. Monographs on Statistics and Applied Probability, Vol. 72, Chapman and Hall, 430 pp.