Clim. Past, 3, 397–409, 2007
www.clim-past.net/3/397/2007/

Climate of the Past

On the verification of climate reconstructions

G. Bürger

FU-Berlin, Institut für Meteorologie; Carl-Heinrich-Becker-Weg 6–10, 12165 Berlin, Germany

Received: 30 November 2006 – Published in Clim. Past Discuss.: 31 January 2007
Revised: 23 May 2007 – Accepted: 5 July 2007 – Published: 11 July 2007

Abstract. The skill of proxy-based reconstructions of Northern hemisphere temperature is reassessed. Using an almost complete set of proxy and instrumental data of the past 130 years a multi-crossvalidation is conducted of a number of statistical methods, producing a distribution of verification skill scores. Among the methods are multiple regression, multiple inverse regression, total least squares, RegEM, all considered with and without variance matching. For all of them the scores show considerable variation, but previous estimates, such as a 50% reduction of error (RE), appear as outliers and more realistic estimates vary about 25%. It is shown that the overestimation of skill is possible in the presence of strong persistence (trends). In that case, the classical "early" or "late" calibration sets are not representative for the intended (instrumental, millennial) domain. As a consequence, RE scores are generally inflated, and the proxy predictions are easily outperformed by stochastic, a priori skill-less predictions.

To obtain robust significance levels the multi-crossvalidation is repeated using stochastic predictors. Comparing the score distributions it turns out that the proxies perform significantly better for almost all methods. The scores of the stochastic predictors do not vanish, nonetheless, with an estimated 10% of spurious skill based on representative samples. I argue that this residual score is due to the limited sample size of 130 years, where the memory of the processes degrades the independence of calibration and validation sets. It is likely that proxy prediction scores are similarly inflated and have to be downgraded further, leading to a final overall skill that for the best methods lies around 20%.

The consequences of the limited verification skill for millennial reconstructions is briefly discussed.

Correspondence to: G. Bürger
(gerd.buerger@met.fu-berlin.de)

1 Introduction

Several attempts have been made to reconstruct the millennial history of global or Northern hemisphere temperature (NHT) by way of proxy information (Overpeck et al., 1997; Jones et al., 1998), (Mann et al., 1998, henceforth MBH98), (Mann et al., 1999; Crowley and Lowery, 2000; Briffa, 2000; Briffa et al., 2001; Esper et al., 2002; Moberg et al., 2005). Since past variability is essential for the understanding of, and attributing forcing factors to the present climate some of these reconstructions have played a prominent role in the third report of the IPCC (IPCC, 2001). This was followed by an intense debate about the used data and methods (McIntyre and McKitrick, 2003; von Storch et al., 2004; McIntyre and McKitrick, 2005a, henceforth MM05; Rutherford et al., 2005; Mann et al., 2005; Bürger and Cubasch, 2005; Huybers, 2005; McIntyre and McKitrick, 2005b; Bürger et al., 2006; Wahl et al., 2006; Wahl and Ammann, 2007), some of which has found its way into the fourth report of the IPCC (IPCC, 2007). While that debate mostly turned on the variability and actual shape of the reconstructions (the "hockey stick") the aspect of verification has not found a comparable assessment.

In the above models (that term used informally here to mean any empirical scheme), a limited number of proxies – usually in the order of several dozens – serve as predictors, either for the local temperature itself or for some typical global pattern of it. The models are defined/calibrated in the overlapping period of instrumental data, and predicted back to those years of the past millennium where proxies are available but temperature observations are not. Once a model is specified, e.g. as a multiple linear regression with a specified number of predictors and predictands, its parameters (the entries of the regression matrix) are calibrated from a finite sample of data (the calibration set). This is usually done by optimizing some measure of model skill, a "score", e.g. mean square error or correlation (Frank and Friedman,

Published by Copernicus Publications on behalf of the European Geosciences Union.

1993). With decreasing sample size the estimated model parameters are increasingly disturbed by "sampling noise", that is, random properties of the sample that do not reflect the envisaged relationship. This estimation error renders the model imperfect, and its "true" skill for predicting independent data is bound to *shrink* relative to the calibration skill (cf. Cooley and Lohnes, 1971).

One would assume that this shrinkage for independent data equally affects all skill scores, simply because the model quality (its "skill") is impaired. This is not true, however. Model skill, as we shall see, cannot be characterized by simple measures such as a single number. Different scores capture different aspects of a model performance. The absence of a single, perfect score is a well known phenomenon from weather forecasting for which most of the scores were originally invented (cf. Murphy, 1996). Moreover, basic conditions such as model linearity are often tacitly implied but are not per se valid, and should be checked using appropriate tests on the residuals, see below.

Instrumental temperatures are available only back until about 1850. Therefore, the period of overlap is just a small fraction of the intended millennial domain. It is evident that empirical models calibrated in that relatively short time span (or even portions of it) must be taken with great care and deserve thorough validation. This applies even more since proxy and temperature records in that period are strongly trended or *persistent*, which considerably reduces the effective size of independent samples that are available to fit and verify a model.

It is therefore essential to find robust estimates of the predictive model skill, as a basis for model selection as well as for the general assessment of the resulting temperature reconstructions. Besides analytical approaches to estimate the true predictive skill from the shrinkage of the calibration skill (Cattin, 1980; Raju et al., 1997) various forms of cross validation are utilized. Simple cross validation (MBH98; Mann et al., 1999; Cook et al., 2000) proceeds as follows: From the period of overlapping data with both proxy and temperature information a calibrating set is selected to define the model. This model is applied to the remaining independent set of proxy data (as a guard against overfitting), and modeled and observed temperature data are compared. A more thorough estimate, called double cross validation, is obtained by additionally swapping calibrating and validating sets (Briffa et al., 1988, 1990, 1992; Luterbacher et al., 2002, 2004; Rutherford et al., 2003, 2005). Multiple cross validation ("multi-crossvalidation") using random calibration sets (Geisser, 1975) is a form of bootstrapping (Efron, 1979; Efron and Gong, 1983) that has been applied only rarely for reconstructions (Fritts and Guiot, 1990; Guiot et al., 2005), but never in a hemispheric context. In this study, that approach will be applied to the NHT.

Only multi-crossvalidation fully accounts for a basic principle of statistical practice: that estimated skill scores are always affected by random properties of the sample from which they were derived. In other words: scores, be they from a calibration or validation set, are random variables, with variations that mainly depend on the sample size. And since it is unlikely that the "true" model is the one with the highest score, picking a model *after* the validation basically renders it unverified, which is therefore not a recommended procedure (cf. Bürger et al., 2006). This equally applies to any other possible variation in the model setting, as long as there is no a priori argument against its use.

Like any bootstrapping, multi-crossvalidation is blind to any predefined (temporal) structure on contiguous calibration or validation periods, such as the 20th century warming trend, and will pick its sets purely by chance. This appears to entirely conflict with a dynamical approach, since any "physical process" that one attempts to reflect (cf. Wahl et al., 2006) is destroyed that way. However, empirical models of this kind do in no way contain or reflect dynamical processes beyond properties that can be sampled in *instantaneous* covariations between the variables. The trend may be an integral part of such a model, but only as long as it represents these covariations.

To estimate whether a verification score represents a significantly skillful prediction it must be viewed relative to score levels attained by skill-less, or "nonsense", predictions. This is necessary because such predictions, in fact, may attain nonzero values for some of the scores. Inferences based on nonsense ("spurious", "illusory", "misleading") correlations turn up since the first statistical measures of association came to light (Pearson, 1897; Yule, 1926). In most cases they are a typical byproduct of small samples (Aldrich, 1995), a problem that is aggravated in the presence of nonstationarity (see below).

There is some analogy to classical weather forecasting where climatology and persistence serve as skill-less predictions whose scores are, especially in the case of persistence, not so easy to beat. While the notion of a skill-less prediction is common sense in weather forecasting, it is the subject of considerable confusion and discussion in the field of climate reconstruction. To give an example: for the reduction of error (RE, see below) in NHT reconstructions, MBH98 and MM05 report the 1%-significance level of RE to be as different as 0% and 59%, respectively. On this background, the usefulness of millennial climate reconstructions, such as MBH98 with a reported RE of 51%, depends on the very notion of a nonsense predictor. This confusion evidently requires a clarification of terms. Towards that goal, the study begins by analyzing and discussing a very basic example of a nonsense prediction with remarkable RE scores. This is followed by a more refined bootstrapping and significance analysis, with models that are currently in use for proxy reconstructions. Having obtained levels of skill and significance the consequences for millennial applications are reflected.

2 Skill calculations, and shrinkage

The study is based on proxy and temperature data that were used in the MBH98 reconstruction of the 15th century. Specifically, the multiproxy dataset, ***P***, consists of the 22 proxies as described in detail in the MBH98 supplement. To meet the bootstrapping conditions of a fixed set of model parameters, the 219 temperature grid points, ***T***, are used that are almost complete between 1854 and 1980, and which were used by MBH98 for verification (see their Fig. 1). This gives 127 years of common proxy and temperature data. Note that the proxies represent a typical portion of what is available back to AD 1400, showing a large overlap with comparable studies (cf. Briffa et al., 1992; Overpeck et al., 1997; Jones et al., 1998; Crowley and Lowery, 2000; Rutherford et al., 2005). Other studies, such as Esper et al. (2002), relied on these proxies as well but processed them differently. Note that for the relatively short time span considered here non-stationarity is hardly an issue. But it might become relevant for millennial applications since some of the proxies actually reveal rather large values of the memory parameter (cf. Robinson, 1995).

Suppose now that we have formulated a statistical model relating ***P*** and ***T***, and picked a calibration set to estimate and a validation set to verify its parameters. For the validation set we denote observed and modeled NHT as x and $\hat{x}$, respectively. Now suppose we have calculated from x and $\hat{x}$ some measure of skill, S, such as the mean square error MSE$=\langle(x-\hat{x})^2\rangle$ (brackets indicating expectation). A classic method to transform a score S into one that measures performance relative to a perfect score, S_p, and a reference score, S_r, is the "skill score", SS (cf. Wilks, 1995). It is given by

$$SS = \frac{S - S_r}{S_p - S_r}. \tag{1}$$

In numerical weather prediction (NWP) it is convenient to take climatology, μ, as a reference forecast (besides, e.g., persistence). For S =MSE, with S_p=0, this gives

$$SS_{\mathrm{MSE}} = 1 - \frac{\mathrm{MSE}}{\langle(x-\mu)^2\rangle}, \tag{2}$$

which is also known under the name "reduction of error" (sometimes also reduction of variance). But while for the stationary context of NWP "climatology" was considered a constant, its use changed in the paleoclimate community to refer to a specific period. Accordingly, the meaning of the score became somewhat ambivalent. While RE was associated with the calibration climatology, μ_c, (Fritts, 1976; Cook et al., 1994), a new score, CE, was introduced by Briffa et al. (1988) relating it to the validation climatology, μ_v:

$$RE = 1 - \frac{\mathrm{MSE}}{\langle(x-\mu_c)^2\rangle} \quad ; \quad CE = 1 - \frac{\mathrm{MSE}}{\langle(x-\mu_v)^2\rangle}, \tag{3}$$

The latter score, CE, is actually the formal analogue of the SS_{MSE} from NWP, for which the concept of a calibration period does not really exist. Note that Briffa et al. (1988) incorrectly equate CE with the "coefficient of efficiency" of Nash and Sutcliffe (1970). In that source (which moreover Cook et al. (1994) incorrectly characterize as a multiple regression study) a validation period mean is never mentioned and the coefficient is simply RE itself.

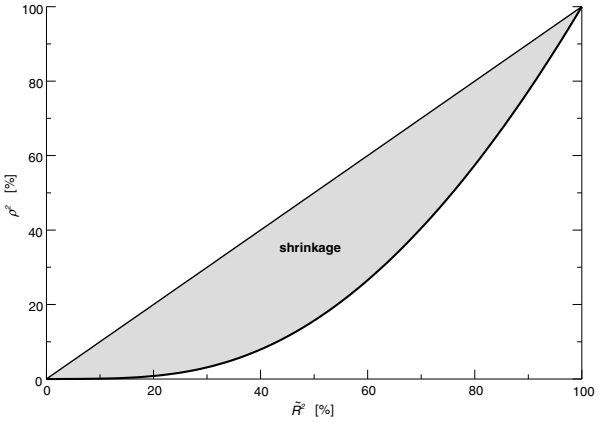

Fig. 1. Dependence of ρ^2 on $\tilde{R}^2$ and corresponding shrinkage.

Both scores are useful, but they measure different things especially when there is a climate shift from calibration to validation. Denoting this shift by $\Delta_C=\frac{\mu_v-\mu_c}{\sigma_v}$ (σ_v the validation standard deviation), both are simply related as follows:

$$RE = \frac{CE + \Delta_C^2}{1 + \Delta_C^2}. \tag{4}$$

CE, on the other hand, is related to the correlation, ρ, between x and $\hat{x}$:

$$CE = \rho^2 - (\rho - \delta_\sigma)^2 - \delta_\mu^2, \tag{5}$$

with δ_μ and δ_σ being the mean and variance bias of the modeled values (cf. Wilks, 1995, p. 256, and Appendix).

For example, applying a multiple regression for the complete population (or, equivalently, validating with the calibration set) gives $\delta_\sigma=\rho$ and δ_μ=0, and thus for the coefficient of determination $R^2=CE=RE=\rho^2$, the well known relation of the squared multiple correlation. From Eqs. (4) and (5) it follows generally $CE\leq\rho^2$ and $CE\leq RE$. That $CE\leq\rho^2$ has the important consequence that skill-less predictions, for which ρ=0, must have $CE\leq$0. Equation (4) illustrates the dependence of RE on the climate shift, Δ_C, and how large Δ_C values inflate that score. For example, if Δ_C=1, that is, one standard deviation, a score of CE=0% would yield RE=50%. This applies, e.g., to time series that exhibit persistence, such as a trend, be it deterministic or stochastic. For example, MM05 report for their MBH98 emulation RE and CE validation scores of 46% and –26%, respectively. That discrepancy is solely caused, as calculated from Eq. (4), by

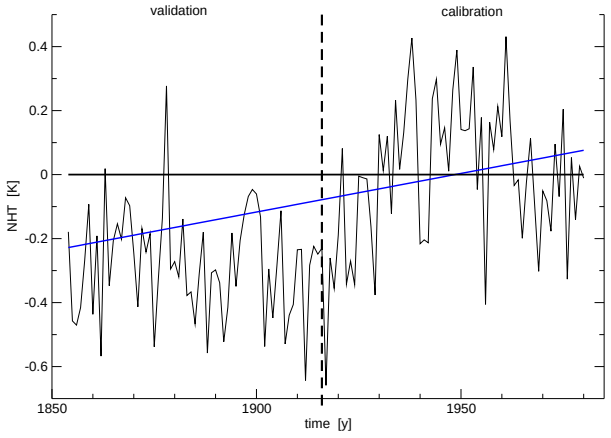

Fig. 2. NHT observed (thin black line) and predicted from the series of calendar years (blue line). The model is calibrated in the late portion (1917–1980) and validated in the early portion (1854–1916), yielding a RE score of 56%. Also depicted is the climatology forecast of the calibration period which, by definition, scores RE=0 (heavy black line).

a climate shift of Δ_C=1.2. Similar sensitivities are reported by Rutherford et al. (2005); Mann et al. (2005); Wahl and Ammann (2007).

It has been argued (Wahl and Ammann, 2007) that RE is superior to CE in measuring the low-frequency performance of models. While in fact RE better "rewards" a correct representation of climate shifts, such as Δ_C, that fact is based on merely one sample and does not warrant the definition of a proper low-frequency skill score. Due to the limited time span of little more than a century no validation skill for time scales longer than a few decades can be expected from this kind of analysis. For an impression of what skills, and in particular what shrinkage thereof, might generally be expected let us consider, as the most straightforward statistical model, a multiple regression of average NHT on p proxies, using N years of calibration. We consider the coefficient of determination, R^2, to estimate the skill. First of all, simple mathematics shows that with an increasing number of (independent) proxies (or, vice versa, with a decreasing number of years) the *calibrated* reconstruction will become perfect - and the model not even unique. R^2 does not account for this inflating effect, so its practical value is limited as it contains not much information about what can be expected from independent (past) data. But it can be adjusted for the number of predictors, as follows (cf. Seber and Lee, 2003):

$$\tilde{R}^2 \;=\; 1-(1-R^2)\frac{N-1}{N-p-1} \tag{6}$$

Eq. (6) gives an estimate of the true multiple correlation from the multiple correlation of a sample of size N. It aims, therefore, at the correlation that one can expect if the model was perfectly estimated (e.g. for $N\to\infty$). That is, however, a very rare circumstance since real models are usually imperfect. If predictions/reconstructions are made with those real models an estimate of skill is needed that takes into account this imperfection. One of the first attempts to incorporate this additional effect has been Lorenz (1956). A more refined estimate was then given by Nicholson (1960), (cf. Cattin, 1980):

$$\rho^2 \;=\; \frac{(N-1)\tilde{R}^4+\tilde{R}^2}{(N-p)\tilde{R}^2+p} \tag{7}$$

$\tilde{R}^2$ and ρ^2 must not be confused. While $\tilde{R}^2$ is of explanatory character describing the statistical population, ρ^2 explicitly represents the correlation skill of a model that is estimated from a finite sample of that population, and represents the same quantity as the corresponding term in Eq. (5) where is was estimated from cross-validation. Accordingly, ρ^2 is often referred to as "cross-validity". In the current context, Eq. (7) describes the following: suppose for our multiple linear regression model with p predictors calibrated from N years we found an adjusted calibration skill of $\tilde{R}^2$. If this model is applied to past (independent) proxies the resulting reconstructed temperature will roughly have a correlation of ρ to the true temperature.

The dependence of ρ^2 on $\tilde{R}^2$ is shown in Fig. 1 for the $\boldsymbol{P}$ and $\boldsymbol{T}$ setting with N=127 and p=22. Even with very large multiple correlations the cross-validity remains quite moderate, so that, for example, to achieve ρ^2=50% one already needs $\tilde{R}^2$=80%. Conversely, a regression of NHT on the proxies using the *full* instrumental period yields $\tilde{R}^2$=36%, which dramatically shrinks to a cross-validity of only 6%. This illustrates the order of magnitude that is to be expected from shrinking, given a ratio of predictors and sample size that is typical for millennial climate reconstructions. Estimates based on multi-crossvalidation shall be provided in §5.

It should be noted that via Eq. (3), RE provides an ad hoc assessment of the uncertainty of a reconstruction. Using a 5% significance level, that uncertainty is

$$\delta T = 2\sigma\sqrt{1-RE}, \tag{8}$$

σ denoting the standard deviation of the measured values.

3 The trivial NHT predictor

Having studied the close relation between RE and CE mathematically via the climate shift, Δ_C, let us illustrate this dependence using a very basic example. Figure 2 shows the average NHT as estimated from the set of 219 temperature grid points, $\boldsymbol{T}$. There is an obvious imbalance between the early and the late half of the period: while colder, even cooling conditions prevail in the early portion, much warmer conditions, initiated by a strong positive trend between 1920 and 1940, dominate the second half. Along with NHT, the

linear model is plotted that results from regressing the late portion (1917–1980) against a very simple predictor: the series of calendar years. I will call this the trivial model or trivial predictor. This is in effect nothing more than fitting a linear trend to that portion. And as a positive trend, the trivial model predicts colder conditions for the past earlier portion. While this does not seem to be an overwhelming performance, the model attains for that part (1854–1916) a verification RE score of 56%! Recalling that RE measures the relative improvement to the climatology forecast, μ_c, indicated by the zero line, the trivial model outperforms that forecast easily by simply predicting colder conditions.

On the other hand, the trivial prediction attains a CE of –70%. According to Eq. (4), this large discrepancy is caused by the enormous bias in the calibration mean of Δ_C=1.7 standard deviations (note that Δ_C=1.2 from the last section is based on a 1902–1980 calibration period). At this point it is important to understand what – besides the presence of the overall trend – leads to that bias. The trend is obviously only effective because of the clean temporal separation of calibration and validation sets. Large values of Δ_C, and thus high RE scores, are obtained because of *a*) a *positive* trend in the *late calibration* and *b*) *negative* anomalies in the *early validation*. In general, it needs a calibration trend of the same sign as the mean difference between late and the early portion.

To clarify the interplay between trend and the degree of temporal separation the following Monte Carlo exercise is performed. We iteratively define two series of calibration and validation sets, starting from the original, well separated partition into a 1917–1980 (1854–1916) late calibration (early validation) set. For a given calibration and validation set we randomly pick one year from each, swap them and put them back to form a new calibration and validation set. – At the end the initial separation is lost, and calibration and validation years are equally distributed and mixed. These series are now "mirrored" by swapping, at each step, the entire calibration and validation sets.

For each of the generated partitions we regress NHT on the trivial predictor using the respective calibration and validation sets, resulting in corresponding RE and CE scores. Moreover, we calculate an individual "degree of separation", Δ_T, for such a partition, using the relative difference

$$\Delta_T = \frac{\bar{T}_c - \bar{T}_v}{\bar{T}_{\text{late}} - \bar{T}_{\text{early}}} \tag{9}$$

where $\bar{T}$ indicates the mean of the respective calendar years (with subscripts c and v indicating calibration and validation, respectively, and "early" and "late" as above). This defines two series of points, (Δ_T,RE) and (Δ_T,CE), that should roughly reflect the dependence of each score on Δ_T. That dependence is noisy, however, due to the random partitions in combination with the NHT fluctuations. To remove these random effects the entire analysis is repeated 500 times, so

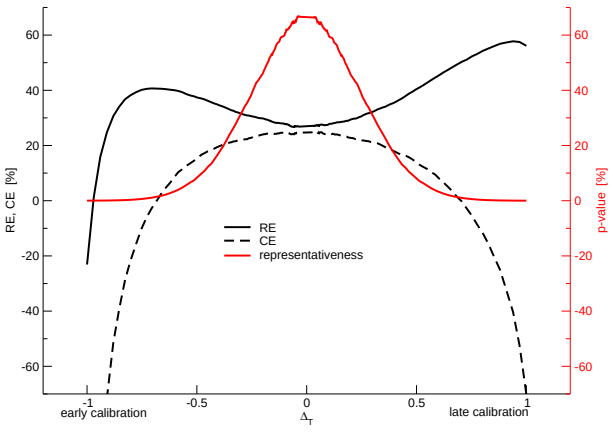

Fig. 3. The dependence of the validation scores RE and CE on the degree of temporal separation, Δ_T, for the simple NHT predictor (see text). For the full separation with a late (1917–1980) calibration and early (1854–1916) validation RE (solid black) approaches 60%, while the fully mixed case attains only about 30% RE; towards early calibration RE rises again to 40% but then sharply drops to negative values. CE (dashed black) shows somewhat opposite behavior, with strongly negative values for the full separation and values similar to RE in the mixed case. Also shown is an index (see text) of the representativeness of the corresponding calibration sets (red).

that each of the above points has now 500 realizations. Using their (vector-) average as a new point gives a graphic that is obviously a function, as shown in Fig. 3. It shows a smooth dependence of the RE and CE values on Δ_T. Both scores show opposite behavior, with RE preferring positive and CE negative values. RE values rise from about 30% for the full mixture to almost 60% for the full separation of the late calibration, while the early calibration shows much lower scores due to the missing, or negative, trend there. CE is more symmetric about the full mixture. There, CE nearly equals RE, while it strongly decreases to about –50% at both ends of the full separation. It is thus found that a trend creates enormous RE scores, but at least half of it is due to the particular selection of calibration and validation sets.

The statistics of each single calibration set are now, with varying degree, representative of the full set (population). As a simple measure of that representativeness one can, for example, test the hypothesis that the NHT values from the calibration and those of the full set are equally distributed, using the Mann-Whitney (ranksum) test, and take the resulting p-value. Averaged over the 500 realizations one finds, not surprisingly, a strong dependence of that index on Δ_T (see Fig. 3). It is symmetric about zero separation, i.e. full mixture, with a maximum attained there and calibration sets that are representative. At both ends, under full separation, the values are practically zero and the calibration sets not representative. It is at these minima where both scores, RE and CE, happen to show the most extreme values.

Note that this representativeness is closely related to the missing-at-random (MAR) criterion that is important for the imputation of missing data and algorithms such as EM and RegEM (see below; cf. Rubin, 1976; Little and Rubin, 1987). It is also relevant for the extrapolation argument given by Bürger and Cubasch (2005).

If it is not clear from the start that the trivial predictor, which was basically the trend itself, does not represent a useful model it will be so in view of the intended time span – the full millennium. The model simply extrapolates the trend backwards into the millennium and produces unrealistic cooling. Hence, the high RE scores do not convey much useful information in this simple case.

I will now turn to "real" predictors, that is, proxy information made up of tree-rings, corals, ice cores, etc., and the more sophisticated empirical models that make use of them.

4 Reconstruction flavors

Several statistical methods exist or have extra been developed to derive millennial NHT from proxy information. They are distinguished by using or not using a number of independent options in the derivation of the final temperature from the proxies. These options mainly pertain to the specific choice of the preprocessing, the statistical model, and the postprocessing.

The methods basically fall into two categories: those which employ a transfer function and those which employ direct infilling of the missing data. In the first approach, the heterogeneous proxy information is transformed to a temperature series by means of a transfer function that is estimated from the period of overlapping data. In the second approach, data are successively infilled to give a completed dataset that is most consistent (see below) with the original data. The transfer function approach uses either some a priori weighting of the proxies, based on, e.g., areal representation, or a weighting directly fitted from the data, that is, multiple regression. To reduce the number of weights in favor of significance, several filtering techniques can be applied, such as averaging or EOF truncation on both the predictor (Briffa et al., 1988, 1992) and the predictand side (MBH98; Evans et al., 2002; Luterbacher et al., 2004).

4.1 Preprocessing (PRE)

Besides using

1) NHT directly as a target, that is, calibrating the empirical model with the NH mean of the $\boldsymbol{T}$ series, so that no spatial detail is modeled at all,

intermediate targets can be defined, as follows:

2) PC truncation. Here a model is calibrated from the dominant principal components (PCs) of $\boldsymbol{T}$, and a hemispheric mean is calculated from their reconstruction. This is applied by MBH98, who have used a single PC. To be compatible with that study I also used only one PC (explaining about 20%–30% depending on the calibration set).

3) full set. The third possibility, applied by Mann and Rutherford (2002); Rutherford et al. (2003, 2005); Mann et al. (2005), does not apply any reduction at all to the target quantity, treating the entire set of temperature grid points (more than 1000 in those studies) as missing. In our emulation, the full set $\boldsymbol{T}$ of 219 temperature grid points is set to missing. From the reconstructed series the NH mean is calculated.

4.2 Statistical method (METH)

The reconstruction of temperatures from proxies can be viewed in the broader context of infilling missing values. The infilling is done by using either a transfer function between knowns and unknowns that is fitted in the calibration (1–4 below), or in a direct way using iterative techniques (5, 6):

1) Classical (forward) regression. Between the known $\boldsymbol{P}$ and unknown $\boldsymbol{T}$ quantities, a linear relation $\mathbf{R}$ is assumed, as follows:

$$\boldsymbol{T} = \mathbf{R}\boldsymbol{P} + \varepsilon, \tag{10}$$

where ε represents unresolved noise. The matrix $\mathbf{R} = \Sigma_{\boldsymbol{P}}^{-1}\Sigma_{\boldsymbol{PT}}$, with Σ_{xy} denoting the cross covariance matrix between x and y (taking $\Sigma_x = \Sigma_{xx}$), is determined by least squares (LS) regression, with $\boldsymbol{T}$ assumed to be noisy.

2) Inverse (backward) regression. This method is applied by MBH98. It also uses a linear model as in 1), but now $\boldsymbol{P}$ is assumed noisy, leading to the LS estimate $\mathbf{R}^I = \Sigma_{\boldsymbol{TP}}^{+}\Sigma_{\boldsymbol{T}}$, ("+" denoting pseudo inverse).

3) Truncated total least squares (TTLS). This form of regression, in combining 1) and 2), assumes errors in both quantities $\boldsymbol{P}$ and $\boldsymbol{T}$ (cf. Golub and Loan, 1996). In this study, the 10 major singular values were retained.

4) Ridge regression. As 1) , but with an extra offset given to the diagonal elements of the (possibly ill-conditioned) matrix $\Sigma_{\boldsymbol{P}}$ used as regularization parameters (Hoerl, 1962).

5) EM. Unlike using a fixed transfer function defined from a calibration set, there are methods that exploit all available information when infilling data, including those from a validation predictor set. A very popular method uses the Expectation-Maximization (EM) algorithm, which provides maximum-likelihood estimates of statistical parameters in the presence of missing data (Dempster et al., 1977). EM is applied using the more specialized *regularized* EM algorithm, RegEM (see below), with a vanishing regularization parameter.

6) RegEM. RegEM has been invented to utilize the EM algorithm for the estimation of mean and covariance in ill-posed problems with fewer cases than unknowns (cf. Schneider, 2001). It was intended for, and first applied to, the interpolation/completion of large climatic data sets, such as gridded temperature observations, with a limited number of missing values (3% in Schneider, 2001). The technique was then

Table 1. Table of the 3×6×2=36 reconstruction flavors.

PRE	METH	POST
219 grid points	forward regression	no rescaling
1 EOF	backward regression	rescaling
1 global average	TTLS	
	ridge regression	
	EM	
	RegEM	

extended to proxy-based climate reconstructions (with a rate of missing values easily approaching 50%) and seen as a successor of the MBH98 method (Mann and Rutherford, 2002; Rutherford et al., 2003, 2005; Mann et al., 2005). Details on RegEM are given in the Appendix.

4.3 Postprocessing (POST)

In applications (e.g. verifications) the output of the statistical model is either taken 1) as is or 2) rescaled to match the calibration variance (cf. Esper et al., 2005; Bürger et al., 2006). Note that this operation increases the expected model error.

As all of PRE, METH, and POST represent independent groups of options, they can be combined to form a possible reconstruction "flavor" (cf. Bürger et al., 2006). As a reference, each such flavor receives a code φ in the form of a triple from the set {1,2,3}×{1,2,3,4,5,6}×{1,2}, indicating which options were selected from the 3 groups above. This defines a set of 3×6×2=36 flavors. For example, the MBH98 method corresponds to flavor φ=222 and Rutherford et al. (2005) to φ=161. Table 1 illustrates the various settings.

It should be emphasized that the suite of flavors shall reflect *existing* methods, taken from the literature, that are to be validated in terms of model error. No focus is put on ways to improve them. For example, all flavors rely in some form or another on a linearity assumption that is not necessarily true, and other schemes such as nonlinear regression or neural nets might give better performance. On the other hand, corresponding Durbin-Watson statistics (not shown, cf. Seber and Lee (2003)) give no indication that the flavors are critically misspecified. Likewise, the chosen set of 22 predictors is not likely to be optimal in a pure statistical sense (using measures such as Mallows C_p, AIC, BIC, etc., cf. Seber and Lee (2003)), as that choice is also determined by external factors like spatial representativity.

5 Multi-crossvalidation of NHT reconstructions

I consider 300 random partitions π of the set $\mathbf{I}$= {1854, ..., 1980} of calendar years,

$$\mathbf{I} = \mathbf{C}_\pi \cup \mathbf{V}_\pi, \tag{11}$$

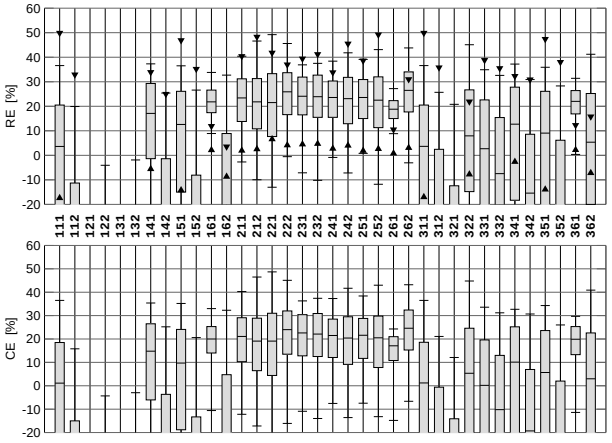

Fig. 4. Boxplot of the distribution of RE and CE for each of the 36 flavors, based on 300 resamplings of the calibration/verification period. Each box indicates the 10%, 50%, and 90% quantile, and the whiskers the minimum and maximum, of the distribution. Also shown are the scores obtained from the full separation into early (upward triangle) and late (downward triangle) calibration. For readability, some flavors/experiments are not shown (too negative).

into calibration and validation sets $\mathbf{C}_\pi$ and $\mathbf{V}_\pi$, where both sets are roughly of equal size ($|\mathbf{C}_\pi|$=64 and $|\mathbf{V}_\pi|$=63). For any of the 36 flavors, φ, it is now possible to calibrate an empirical model, with corresponding scores $RE_\varphi(\pi)$ and $CE_\varphi(\pi)$. $RE_\varphi(\pi)$ and $CE_\varphi(\pi)$ thus appear as realizations of *random variables* RE_φ and CE_φ, with corresponding distributions. Along with the 300 random partitions I also consider the two complementary partitions with full temporal separation.

The distributions of RE_φ and CE_φ are depicted in Fig. 4 as a boxplot. For most flavors the distributions show a remarkable spread, with minimum and maximum (low and high 10%-quantiles) easily departed by more than 50% (20%) of skill. Moreover, between the flavors the distributions are quite different. For example, the flavors φ=161 and φ=162 are merely distinguished by the use of rescaling. Their performance, however, is grossly different. This applies likewise to the flavors φ=141/2 and φ=151/2, so that at least in these cases skill is strongly degraded by rescaling (note, however, φ=261/2). While there is so much spread in skill within and between the flavors the distributions themselves are quite similar for both scores RE_φ and CE_φ. This indicates that, in fact, most calibration/validation partitions are temporally well mixed and RE_φ and CE_φ measure the same thing (see §3).

The skill varies, but it varies on rather low levels. The 90% quantile hardly exceeds the 30% mark, and the highest median is RE_φ=26.5% and CE_φ=24.6% for φ=262. Generally, flavors of the form 2xx, i.e. those predicting PC1 of NHT, perform much better, with almost all medians above 20%.

The other flavors are much more variable, partly caused by the degradation from rescaling mentioned above. An exception are the flavors of the form x61 which show remarkably little variance (albeit only moderate scores). This is understandable insofar as RegEM, unlike the other flavors, depends on the particular calibration set only in terms of the predictand (utilizing the full instrumental period for the predictors). This would also apply to the EM flavors (x51), but they are probably more susceptible to overfitting. Note that the flavor φ=311, which has shortly been touched in §2 to exemplify shrinking, scores very little, with RE and CE values below 5%. This is about the same order of magnitude as the estimate obtained from Eq. (7).

The mindful reader has noticed that some flavors, such as φ=111 and φ=311, have identical distributions. In fact, for direct regression, with a linear dependence of the estimated model on the predictand, cf. §4.2, they are equivalent with respect to NHT and thus redundant. (Note that the RegEM flavors φ=161 and φ=361 are similar as well.)

The triangles in the figure represent the two calibrations with full temporal separation, i.e. the periods 1917–1980 (upper triangle) and 1854–1916 (lower triangle). They are more comparable to estimates of previous studies and obviously assume the role of outliers, in a positive sense for RE and in a negative one for CE. While several RE values approach 50% the CE values are negative throughout. Models with trended and fully separated calibration sets are thus rewarded with high RE scores but penalized with low CE scores.

Based on such levels of performance it is difficult to declare one specific flavor as being the "winner" and being superior to others. Just from the numbers, the flavor φ=262 gives the best RE performance (see above). It predicts PC1 using RegEM and rescaling. But it is only marginally better than, e.g., the simpler variant 211 (simple forward regression, with median 23.4%). Note that the flavor 161 was promoted by Mann et al. (2005) and earlier to replace the original MBH98 flavor 222. From the current analysis, this cannot be justified (RE median of 21.8% compared to 25.9%). This is somewhat in agreement with Rutherford et al. (2005) who report a millennial RE of 40% (46% for the "hybrid" case), as compared to the 51% of MBH98. Moreover, for the late calibration the 161 flavor is particularly bad (RE_φ=11.9%); it improves, nonetheless, when calibrating with the "classical" calibration period 1902–1980 (28%).

6 Significance

There is an ongoing confusion regarding the notion of significance of the estimated reconstruction skill. For the same model (the one used by MBH98, here the emulated flavor 222), MBH98 (resp. Huybers, 2005) and McIntyre and McKitrick (2005b) report a 1% significance level for RE as different as 0% and 54%! Hence, with a reported RE of 51% the model is strongly significant in the first interpretation and practically useless, i.e. indistinguishable from noise, in the latter. And what might be even more intriguing: The trivial model of §3 with an RE score of 56% would be accepted as "significantly skillful" under both interpretations. Obviously, the notion of "being significant", or of being a "nonsense predictor", deserves a closer look.

A major difference in the two approaches is the allowance for nonsense regressors for the significance estimation. That is, whether stochastic time series themselves are considered, or instead the result of feeding them into a regression model. Only the latter yields higher scores. Now even in the well-mixed, representative case the trivial predictor scored about RE=20% (similarly to CE), which is still higher than the 1%-significance level of RE=0%. To avoid this, nonsense regressors must therefore be allowed. On the other hand we have seen how the temporal separation produces non-representative samples and creates RE "outliers" of up to 60%. The proposed significance level of RE=54%, which is based on these outliers, is thus equally inflated and must be replaced by something more representative.

A crucial question is: What kind of nonsense predictors should be allowed? – To derive a statistically sound significance level requires a null distribution of nonsense reconstructions. Now one can think of all sorts of funny predictors, things like calendar years, Indias GDP, the car sales in the U.S., or all together, etc., but that will not make up what mathematically is called a measurable set (to which probabilities can be assigned). Hence, a universal distribution of nonsense predictors does not exist. – A more manageable type of nonsense predictors are stochastic processes generated from white noise, such as AR, ARMA, ARFIMA, ..., (cf. Brockwell and Davis, 1998). Once we fix the number of predictors, the type of model, say ARMA(p,q), and the set of parameters, a unique null distribution of scores can be obtained from Monte Carlo experiments. From these, a significance level can be estimated and compared to the original score of the reconstruction. The only problem is then that each of the specified stochastic types creates its own significance level.

It was perhaps this dilemma that originated the debate about the benchmarking of RE, specifically, estimating the 1% level of significance, RE_{crit}. In the literature, one finds the following approaches:

1. (MBH98) simple AR(1) process with specified memory: RE_{crit}=0%;

2. (MM05) inverse regression of NHT on a red noise predictor derived from the North American tree ring network: RE_{crit}=59%;

3. (Huybers, 2005; Wahl and Ammann, 2007) as 2, with rescaling: RE_{crit}=0%. Using the matlab code accompanying Huybers (2005) I obtained RE_{crit}=36%.

4. (McIntyre and McKitrick, 2005b) as 3, but with 21 additional (uncorrelated?) white noise predictors: $RE_{\rm crit}$=54%.

One might now feel inclined to provide the "correct" or "optimum" way of representing the proxies as a stochastic process. If I now add

5. as 4, but with all noise predictors (not only the North American treering PC1) estimated from the original proxies,

the series of benchmarking attempts from 1 to 5 would in fact slowly convergence to what MBH98 and similar studies should be compared to. But so much is not required. One can and must only provide a realistic *lower bound* on the level of significance (cf. MM05), may it come from whatever stochastic process. Actually, "whatever" is not entirely true as the class of stochastic processes is not fully arbitrary, as discussed below. With regard to 5, a benchmark has not been estimated so far, and will not be estimated here. The lesson of §3 is that all benchmarks 1–4 are inflated by the temporal separation of calibration and validation sets, and more realistic values are to be expected from multi-crossvalidation.

For each of the 36 flavors I have therefore repeated the analysis of §5, with the proxies being replaced by red noise series. Specifically, for each proxy a fractionally integrated noise series is generated whose memory parameter, d, is estimated from the proxy using log-periodogram regression (Geweke and Porter-Hudak, 1983; Brockwell and Davis, 1998). To obtain more robust estimates of d I used here, like MM05, the *full* proxy record from 1400 to 1980; the corresponding estimates varied between d=−0.17 and d=0.85. Note that the log-periodogram estimation is slightly different from the method applied by MM05 (which is based on Hosking, 1984). Neither method is perfect, as both rest on various approximations (cf. Bhansali and Kokoszka, 2001) that provide little more than a rough guess of what the "true" memory parameter might be. The noise generation was redone for each of the 300 partitions (to remove sampling effects). The result is shown in Fig. 5. Like in Fig. 4, RE and CE values are similar. All scores are smaller compared to the corresponding proxy predictions, with a greater spread per flavor. They are nonetheless not negligible. Analogously to the proxies, the scores are generally better for flavors of the form 2xx, with median levels varying about 10%. For each flavor, also included are the experiments with full temporal separation. Some of the RE scores exceed 50%, like the trivial predictor (54% for φ=311). As an example, Fig. 6 shows the distribution of the 300 predictions for the flavor φ=222, in terms of validation RE and in comparison to the proxy predictions. We clearly see different distributions, the nonsense predictions being more spread and generally shifted to smaller RE values, varying roughly about 20%. Note that this is about the score of the trivial predictor for representative calibration sets, depicted in Fig. 3. There are nonetheless

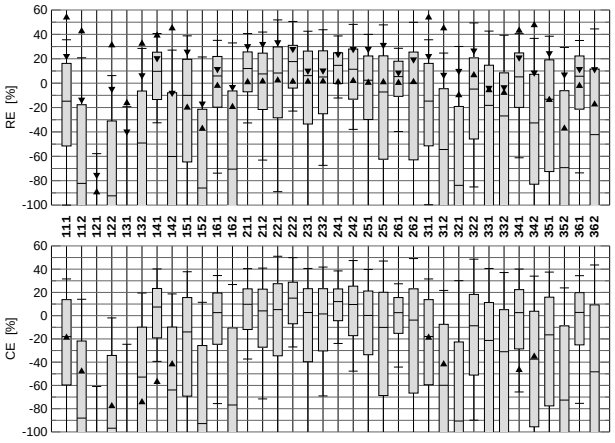

Fig. 5. As Fig. 4, using nonsense predictors.

outliers with very good scores (∼45%). These are possible, as we saw, if the predictors are sufficiently persistent, and calibration and validation sufficiently separated in the time domain.

The degree to which the proxy predictions outperform their nonsense pendants is depicted in the last Fig. 7; it shows for each flavor the respective Mann-Whitney test statistic. Except for the flavors φ=13x the values are well beyond the 1% significance level of the standard normal null distribution of the test (obtained if both samples come from the same population). The highest values are, like in Fig. 4, attained by the 2xx flavors that are based on predictand EOF filtering. The x61 flavors, i.e. those using RegEM, are also large, which is possibly due to the overall reduction in RE spread for those flavors (see above).

Now one thing is still unresolved: Why do the nonsense predictions yield non-vanishing score even for the well mixed, representative samples? – A nonsense prediction has, by definition, no skill. In an ideal world, which among other things has infinite samples and truly independent validation, it would have a cross-validity of ρ=0 and thus, using Eq. (5), $CE\leq 0$. In the real world of finite samples, this condition is violated. The 127 cases/years of instrumental data capture too few degrees of freedom to facilitate proper independent validation sets.

The inflation of scores is thus an artifact of the imperfect verification. The validity of a calibrated regression equation is partly inherited by the (no longer independent) validation set and creates skill there. This is aggravated by the presence of strong trends, such as those seen in many of the proxy and temperature series in the instrumental period. The inflating effect has two consequences: First, it affects all reconstructions that are based on the instrumental period, including those based on real proxies. The cross-validity estimates from this study, along with most others reported in the literature, have yet to be adjusted (downgraded) for this effect.

This suggests to use direct, formula based approaches such as those of Eq. (7). Second, the estimation of the significance level from the noise series requires the adequate representation of the persistence properties of the proxies. Incorrect estimation entails incorrect significance levels. For example, increasing the estimated memory parameter d of the noise series by 50% enhances their verification scores considerably, with the result that the proxy reconstructions are rendered insignificant for all flavors (not shown).

7 Conclusions

The analysis poses three questions:

1. How do we interpret the estimated *levels* of reconstruction skill?
2. How do we interpret the resulting *spread* in that skill?
3. How are possible answers to 1. and 2. affected by the *significance* analysis?

ad 1: It was found that realistic estimates of skill vary about 25%, equally for RE and CE. The results were obtained using a well confined testbed of proxy and temperature information through 127 instrumental years, with almost no gaps. The proxies represent a standard set of what is available back to AD 1400. The set of temperature grid points does not cover the entire globe, and its areal averages serve only as approximations to the full NHT average; but it is about the largest subset that is rigorously verifiable. On this background, previous estimates of NHT reconstruction skill in the range of RE=50% appear much too large. They are inflated by the use of a non-representative calibration/validation setting in the presence of trended data.

ad 2: Crossvalidation of any type (single, double, multi) is a means to estimate the distribution of unknowns (here: the reconstruction skill). As there is no a priori criterion to prefer a specific calibration set, all such sets receive equal weights before and after the analysis (this is in conflict with Rutherford et al. (2003) who seem to prefer one set *because* of its validation skill). The estimated distributions were quite similar for RE and CE, indicating that both scores actually measure the same thing. The considerable spread of most distributions simply reflects our limited ability to estimate skill any better, based on a sample size of 127 cases/years, and on an effective sample size that is even less, due to persistence.

ad 3: Reconstructions from real proxies significantly outperform stochastic (nonsense) predictions if those have comparable persistence characteristics. The scores attained by the latter do not vanish, nevertheless, with RE (CE) values varying about 10% for many flavors. This was attributed to the degraded independence of the finite validation period by memory effects, allowing portions of the calibration information to drop into the validation. As this is equally true for the proxy predictions – from this and from any comparable study – a substantial amount of the estimated verification skill is likely to be spurious and must further be downgraded. The significance level, and thus the final value of the reconstructions, depends strongly on the persistence properties. A version with memory parameters increased by 50% rendered all reconstructions insignificant. This is important insofar as well established methods for their estimation do not (yet) exist.

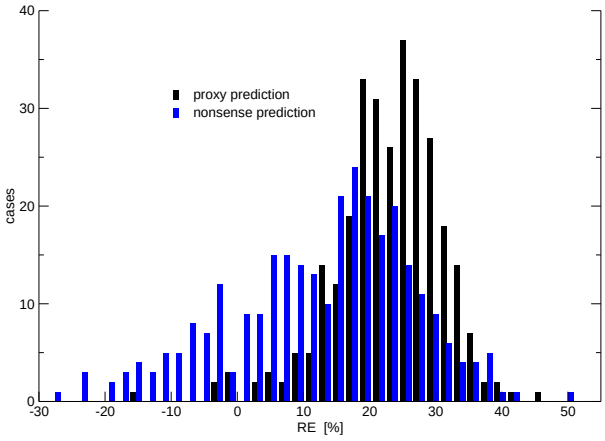

Fig. 6. Histogram of RE from proxy and nonsense prediction using flavor φ=222. Proxy predictions show less spread and generally greater skill. Note, however, that high scores are also obtained from nonsense predictions.

It is unknown how such a downgrading should be done numerically, producing a final overall verification skill that for the best flavors is likely to be around 20%, with large uncertainties. Are such levels of skill sufficient to decide the millennial NHT controversy? – Inserting a value of RE=20% into Eq. (8) gives a reconstruction uncertainty of δT=0.43 K. If one were to focus the controversy into the single question: Was there a hemispheric Medieval Warm Period and was it possibly warmer than recent decades? – that question cannot be decided based on current reconstructions alone, at least not in a verifiable sense.

Appendix A

RE, CE, and ρ

For ease of notation, we generally drop the subscript v, and write all validation values in the form mean plus anomaly, $\mu+x$. The variance relative to the calibration mean, μ_c, can thus be written (recalling that $\Delta_C=\frac{\mu-\mu_c}{\sigma}$)

$$\langle(\mu+x-\mu_c)^2\rangle = \langle(x+\sigma\,\Delta_C)^2\rangle = \sigma^2(1+\Delta_C^2) \qquad \text{(A1)}$$

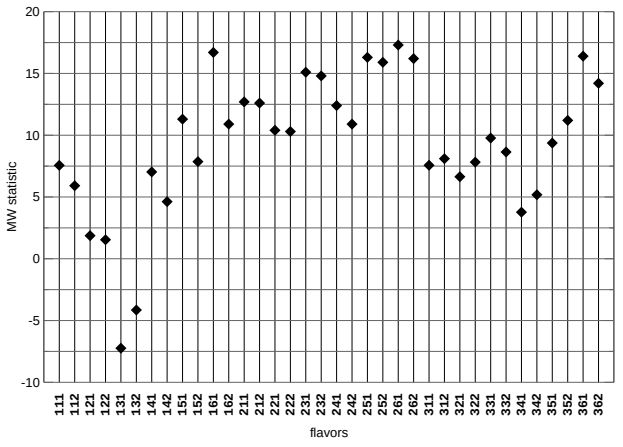

Fig. 7. Testing RE scores of proxy vs. nonsense predictions, using Mann-Whitney test, for all flavors. The null distribution is N(0,1), so that for almost all flavors the real predictions are significantly better than the nonsense predictors.

Now we can write Eq. (3) in the form

$$RE * \sigma^2(1+\Delta_C^2) = \sigma^2(1+\Delta_C^2) - MSE$$
$$CE * \sigma^2 = \sigma^2 - MSE \tag{A2}$$

Subtracting the second equation from the first immediately gives

$$RE = \frac{CE + \Delta_C^2}{1+\Delta_C^2} \tag{A3}$$

Suppose now that true and predicted values are given as $\mu+x$ and $\hat{\mu}+\hat{x}$, respectively. Let σ and $\hat{\sigma}$ denote the observed and modeled standard deviation, and $\delta_\mu = \frac{\hat{\mu}-\mu}{\sigma}$ and $\delta_\sigma = \frac{\hat{\sigma}}{\sigma}$ the bias in the mean and the variance, respectively. Then we have

$$\begin{aligned} CE &= 1 - \frac{\langle(\hat{\mu}-\mu+\hat{x}-x)^2\rangle}{\sigma^2} \\ &= 1 - \frac{(\hat{\mu}-\mu)^2 + 2(\hat{\mu}-\mu)\langle\hat{x}-x\rangle + \langle(\hat{x}-x)^2\rangle}{\sigma^2} \\ &= 1 - \frac{(\hat{\mu}-\mu)^2 + \langle\hat{x}^2\rangle - 2\langle x\hat{x}\rangle + \langle x^2\rangle}{\sigma^2} \\ &= 1 - \delta_\mu^2 - \delta_\sigma^2 + 2\frac{\langle x\hat{x}\rangle}{\sigma^2} - 1 \\ &= 2\rho\delta_\sigma - \delta_\mu^2 - \delta_\sigma^2 \\ &= \rho^2 - (\rho-\delta_\sigma)^2 - \delta_\mu^2 \end{aligned} \tag{A4}$$

Appendix B

RegEM

RegEM was originally developed for the infilling of large incomplete climate fields, for which the original EM algorithm was not applicable. One part of the EM algorithm is a regression of the unknown on the known variables of a field, representing the expectation ("E"-) step. If there are too many explanatory variables the problem is ill-posed. RegEM overcomes this by regularizing (e.g. ridge regression, principal component regression). The algorithm was quickly adopted for climate reconstructions, the role of the known part being played by the proxies. But due to the relatively small number of proxies the problem is no longer ill-posed and it is not clear why the much simpler EM algorithm had not been used from the start. Moreover, the reported millennial verification RE of RegEM is less than that of the original MBH98 (cf. Rutherford et al., 2005). The current study is the first to compare the performance of EM and RegEM.

Configuration: - To control the iteration, RegEM has a number of configuration switches that can be adjusted. The following settings gave satisfactory convergence results for most of the experiments. I used: multiple ridge regression as a regression procedure; regularization parameter determined from general cross validation (GCV); minimum relative variance of residuals: 5e-2; stagnation tolerance: 3e-5; maximum number of iterations: 50; inflation factor: 1.0; minimum fraction of retained variance: 0.95. This latter setting is borrowed from Rutherford et al. (2003) who argue that the GCV regularization estimate is too crude in the presence of too many unknowns. This was true here as well. In fact, using the GCV estimate for the flavors φ=1xx resulted in RegEM reconstructions that were hardly distinguishable from the calibration mean.

Acknowledgements. I enjoyed lively discussions with U. Cubasch, F. Niehörster and F. Kaspar. This work was partly supported by the EU project SOAP.

References

Aldrich, J.: Correlations Genuine and Spurious in Pearson and Yule, Discussion Paper Series In Economics And Econometrics 9502, Economics Division, School of Social Sciences, University of Southampton, available at: http://ideas.repec.org/p/stn/sotoec/9502.html, 1995.

Bhansali, R. J. and Kokoszka, P. S.: Estimation of the long memory parameter: a review of recent developments and an extension, in: Selected proceedings of the symposium on inference for stochastic processes. IMS Lecture notes and monograph series, edited by: Basawa., I., Heyde, C. C., and Taylor, R., 125–150, Institute of Mathematical Statistics, Ohio, USA, 2001.

Briffa, K. R.: Annual climate variability in the holocene: interpreting the message of ancient trees, Quat. Sci. Rev., 19, 87–105, 2000.

Briffa, K. R., Jones, P. D., Pilcher, J. R., and Hughes, M. K.: Reconstructing Summer Temperatures in Northern Fennoscandinavia Back to A.D.1700 Using Tree Ring Data from Scots Pine, Arctic and Alpine Research, 385–94, 1988.

Briffa, K. R., Bartholin, T. S., Eckstein, D., Jones, P. D., Karlen, W., Schweingruber, F. H., and Zetterberg, P.: A 1,400-year tree-ring

record of summer temperatures in fennoscandia, Nature, 346, 434–439, 1990.

Briffa, K. R., Jones, P. D., and Schweingruber, F. H.: Tree-ring density reconstructions of summer temperature patterns across western north america since 1600, J. Climate, 5, 735–754, 1992.

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

Brockwell, P. J. and Davis, R. A.: Time series: theory and methods, 2nd edition, Springer Series in Statistics, 1998.

Bürger, G. and Cubasch, U.: Are multiproxy climate reconstructions robust?, Geophys. Res. Lett., 32, L23711, doi:10.1029/2005GL0241550, 2005.

Bürger, G., Fast, I., and Cubasch, U.: Climate reconstruction by regression – 32 variations on a theme, Tellus A, 227–35, 2006.

Cattin, P.: Estimation of the predictive power of a regression model, J. Appl. Psychol., 65(4), 407–414, 1980.

Cook, E. R., Briffa, K. R., and Jones, P. D.: Spatial regression methods in dendroclimatology: a review and comparison of two techniques, Int. J. Clim., 14, 379–402, 1994.

Cook, E. R., Buckley, B. M., D'Arrigo, R. D., and Peterson, M. J.: Warm-season temperatures since 1600 bc reconstructed from tasmanian tree rings and their relationship to large-scale sea surface temperature anomalies., Clim. Dyn., 16, 79–91, 2000.

Cooley, W. W. and Lohnes, P. R.: Multivariate data analysis, New York: Wiley, 1971.

Crowley, T. J. and Lowery, T. S.: How warm was the medieval warm period?, Ambio, 29, 54, 2000.

Dempster, A., Laird, N., and Rubin, D.: Maximum likelihood estimation from incomplete data via the EM algorithm, J. Royal Statist. Soc., B, 39, 1–38, 1977.

Efron, B.: Bootstrap methods: another look at the jackknife, Annals of Statistics, 17, 1–26, 1979.

Efron, B. and Gong, G.: A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation, American Statistician, 36–48, 1983.

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

Esper, J., Frank, D. C., Wilson, R. J. S., and Briffa, K. R.: Effect of scaling and regression on reconstructed temperature amplitude for the past millennium, Geophys. Res. Lett., 32(7), L07711, doi:10.1029/2004GL021236, 2005.

Evans, M. N., Kaplan, A., and Cane, M. A.: Pacific sea surface temperature field reconstruction from coral delta o-18 data using reduced space objective analysis, Paleoceanography, 17, 1007, doi:10.1029/2000PA000590, 2002.

Frank, I. E., Friedman, J. H.: A Statistical View of Some Chemometrics Regression Tools, Technometrics, 35, 109, 109–148, 1993.

Fritts, H. C.: Tree rings and climate, Academic Press, 1976.

Fritts, H. C. and Guiot, J.: Methods of calibration, verification, and reconstruction, in: Methods Of Dendrochronology. Applications In The Environmental Sciences, edited by: Cook, E. R. and Kairiukstis, L. A., 163–217, Kluwer Academic Publishers, 1990.

Geweke, J. and Porter-Hudak, S.: The estimation and application of long-memory time series models, J. Time Series Analysis, 4, 221–238, 1983.

Golub, G. H. and Loan, C. F. V.: Matrix computations (3rd ed.), Johns Hopkins University Press, Baltimore MD USA, 1996.

Guiot, J., Nicault, A., Rathgeber, C., Edouard, J. L., Guibal, E., Pichard, G., and Till, C.: Last-millennium summer-temperature variations in western europe based on proxy data, Holocene, 15, 500, 2005.

Hoerl, A. E.: Application of ridge analysis to regression problems, Chem. Eng. Prog., 58, 54–59, 1962.

Hosking, J.: Modeling persistence in hydrological time series using fractional differencing, Water Resour. Res., 20(12), 1898–1908, 1984.

Huybers, P.: Comment on "Hockey sticks, principal components, and spurious significance", Geophys. Res. Lett., L20705, doi:10.1029/2005GL023395, 2005.

IPCC: Climate change 2001: the scientific basis. contribution of working group I to the third assessment report of the intergovernmental panel on climate change, Cambridge University Press, Cambridge, 2001.

Working Group I Report "The Physical Science Basis". The Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (in press).

Jones, P. D., Briffa, K. R., Barnett, T. P., and Tett, S. F. B.: High-Resolution Palaeoclimatic Records for the Last Millennium: Interpretation, Integration and Comparison with General Circulation Model Control-Run Temperatures, Holocene, 8, 455–71, 1998.

Geisser, S..: The Predictive Sample Reuse Method with Applications, Journal of The American Statistical Association, 70, 320–328, 1975.

Little, R. J. A. and Rubin, D. B.: Statistical analysis with missing data, Wiley, 1987.

Lorenz, E. N.: Empirical orthogonal functions and statistical weather prediction, Sci. Rept. No. 1, Dept. of Met., M. I. T., p. 49pp, 1956.

Luterbacher, J., Xoplaki, E., Dietrich, D., Rickli, R., Jacobeit, J., Beck, C., Gyalistras, D., Schmutz, C., and Wanner, H.: Reconstruction of sea level pressure fields over the eastern north atlantic and europe back to 1500, Climate Dynamics, 18, 545–561, 2002.

Luterbacher, J., Dietrich, D., Xoplaki, E., Grosjean, M., and Wanner, H.: European Seasonal and Annual Temperature Variability, Trends, and Extremes Since 1500, Science, 303, 1499–1503, 2004.

Mann, M. E. and Rutherford, S.: Climate reconstruction using "Pseudoproxies", Geophys. Res. Lett., p. 139, 2002.

Mann, M. E., Bradley, R. S., and Hughes, M. K.: Global-scale temperature patterns and climate forcing over the past six centuries, Nature, 779–87, 1998.

Mann, M. E., Bradley, R. S., and Hughes, M. K.: Northern hemisphere temperatures during the past millennium: inferences, uncertainties, and limitations, Geophys. Res. Lett., 759–762, 1999.

Mann, M. E., Rutherford, S., Wahl, E., and Ammann, C.: Testing the Fidelity of Methods Used in Proxy-Based Reconstructions of Past Climate, J. Climate, 4097–107, 2005.

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

McIntyre, S. and McKitrick, R.: Hockey sticks, principal components and spurious significance, Geoph. Res. Let., 32, L03710,

doi:10.1029/2004GL021750, 2005a.

McIntyre, S. and McKitrick, R.: Reply to comment by huybers on "hockey sticks, principal components, and spurious significance", Geophys. Res. Lett., 32, L20713, doi:10.1029/2005GL023586, 2005b.

Moberg, A., Sonechkin, D. M., Holmgren, K., Datsenko, N. M., and Karlen, W.: Highly variable northern hemisphere temperatures reconstructed from low- and high-resolution proxy data, Nature, 433, 617, 2005.

Murphy, A. H.: The Finley Affair: A Signal Event in the History of Forecast Verification, Weather and Forecasting, 11(1), 3, 1996.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models - Part I - A discussion of principles, J. Hydrol., 10, 282–290, 1970.

Nicholson, G. E.: Prediction in future samples, in: Contributions to Probability and Statistics, edited by: Olkin, I., 322–330, 1960.

Overpeck, J., Hughen, K., Hardy, D., Bradley, R., Case, R., Douglas, M., Finney, B., Gajewski, K., Jacoby, G., Jennings, A., Lamoureux, S., Lasca, A., MacDonald, G., Moore, J., Retelle, M., Smith, S., Wolfe, A., and Zielinski, G.: Arctic Environmental Change of the Last Four Centuries, Science, 278, 1251–1256, 1997.

Pearson, K.: Mathematical Contributions to the Theory of Evolution – On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs, Proc. R. Soc., 60, 489–498, 1897.

Raju, N. S., Bilgic, R., Edwards, J. E., and Fleer, P. F.: Methodology review: estimation of population validity and cross-validity, and the use of equal weights in prediction, J Appl. Psychol. Measurement, 21(4), 291–305, 1997.

Robinson, P. M.: Log-Periodogram Regression of Time Series with Long Range Dependence, Annals of Statistics, 23, 1048–1072.

Rubin, D. B.: Inference and missing data, Biometrika, 63, 581–592, 1976.

Rutherford, S., Mann, M. E., Delworth, T. L., and Stouffer, R. J.: Climate field reconstruction under stationary and nonstationary forcing, J. Climate, 16, 462–479, 2003.

Rutherford, S., Mann, M. E., Osborn, T. J., Bradley, R. S., Briffa, K. R., Hughes, M. K., and Jones, P. D.: Northern Hemisphere Surface Temperature Reconstructions: Sensitivity to Methodology, Predictor Network, Target Season and Target Domain, J. Climate, 2308–29, 2005.

Schneider, T.: Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values, J. Climate, 14, 853–871, 2001.

Seber, G. A. F. and Lee, A. J.: Linear regression analysis (wiley series in probability and statistics), Wiley-Interscience, 2 edn., 2003.

von Storch, H., Zorita, E., Jones, J. M., Dmitriev, Y., and Tett, S. F. B.: Reconstructing Past Climate from Noisy Data, Science, 679–82, 2004.

Wahl, E. R. and Ammann, C. M.: Robustness of the Mann, Bradley, Hughes reconstruction of Northern hemisphere surface temperatures: Examination of criticisms based on the nature and processing of proxy climate evidence, Climatic Change, in press, 2007.

Wahl, E. R., Ritson, D. M., and Ammann, C. M.: Comment on “Reconstructing past climate from noisy data”, Science, 312, p. 529b, http://www.sciencemag.org/cgi/content/abstract/312/5773/529b, doi:10.1126/science.1120866, 2006.

Wilks, D. S.: Statistical methods in the atmospheric sciences. an introduction, Academic Press, San Diego, 1995.

Yule, G. U.: Why do we sometimes get nonsense-correlations between time-series? – A study in sampling and the nature of time-series, J. Roy. Stat. Soc., 1–29, 1926.