Tellus (2006), 58A, 227–235

TELLUS

Climate reconstruction by regression – 32 variations on a theme

By GERD BÜRGER*, IRINA FAST and ULRICH CUBASCH, *Institut für Meteorologie, FU Berlin, Carl-Heinrich-Becker-Weg 6–10, D-12165 Berlin, Germany*

(Manuscript received 18 May 2005; in final form 19 September 2005)

ABSTRACT

Regression-based methods fail to provide a sufficiently unique reconstruction of a given millennial history of Northern Hemisphere mean temperature. They instead offer a multitude of variants, depending on the specific data processing scheme. Using a simulated climate history with noise-disturbed pseudo-proxies, we systematically test a set of such configurations, each of which appears to be a priori reasonable, with existing applications elsewhere. This results in an entire spectrum between practically useless and almost perfect reconstructions. The reason lies in the fact that the training variations are not representative of the full millennium, and the regression equations have to be extrapolated. This creates an error that is proportional to both the model uncertainty and the proxy amplitudes. Estimation of that uncertainty is paramount for a useful millennial reconstruction, especially if it is of the parameter-loaded multiproxy type.

1. Introduction

Conventional wisdom has it that empirical models can simulate only what they have seen before, just slightly simpler, or smoother. Statistical regression, which is ubiquitously encountered in empirical studies of climate research, is no exception. Even dynamical models, such as general circulation models (GCMs), are not purely based on so-called first principles and contain empirical parameterizations that are not seldom derived from regression. Among numerous applications are statistical downscaling (Bardossy and Plate, 1991; von Storch et al., 1993; Bürger, 2002), seasonal ENSO forecasts (Jiang et al., 1995; Barnston et al., 1999), or short- to medium-range climate forecasts based on teleconnections (Lloyd-Hughes and Saunders, 2002; Blender et al., 2003) or lagged SST information (Rodwell and Folland, 2001; Qian and Saunders, 2003); examples for model parameterizations are Tiedtke (1989) and Fu and Liou (1993).

In all of these examples, regression can be understood as a law that describes the variation of one, usually noisy, variable – the predictand – in terms of the variation of other variables – the predictors. The law is a mathematical function of a specified form, such as linear or logarithmic, with additional parameters that have to be *fitted* on previously observed covariations. The applicability of the law, i.e. the domain of allowed predictor variation, is usually of little concern; in most cases it is tacitly assumed to be constant or at most changing with second or higher order. In the radiation example (Fu and Liou, 1993), the predictor (top of atmosphere solar radiation and distribution of clouds) variations are of a well-defined scale. In downscaling applications (e.g. von Storch et al., 1993), climatic change is supposed to occur at least one order slower than the variations that are to be downscaled.

Recently, regression-based methods have found their way into the field of climate reconstruction, using (paleo-)historic ("proxy") data such as tree rings, corals, and sediments (Mann et al., 1998, henceforth MBH98; Jones et al., 1998; Briffa et al., 2001, 2004; Esper et al., 2002; Jones and Mann, 2004). Among these studies, the most influential and controversial was certainly MBH98. Following its prominent role in the third IPCC report (IPCC, 2001), the study boosted the importance of historic climate for the current debate about anthropogenic global warming. And since it utilized regression in a rather complicated and unconventional fashion, the focus turned to the regression theme as a whole. Purely methodological studies followed where noise-disturbed temperature grid points, so-called pseudo-proxies, played the role of the proxies, either by using observations (Mann and Rutherford, 2002) or by simulations (Zorita et al., 2003; von Storch et al., 2004, henceforth S04). The MBH98 method was thoroughly tested by S04, and they conclude that for climate reconstruction purposes, any regression-based method is inherently unable to produce sufficient variability. The validity of this argument shall be discussed in Section 3.

In all these applications, the regression is trained in the period of known climate, that is, in the time of instrumental

*Corresponding author.
e-mail: gerd.buerger@met.fu-berlin.de
DOI: 10.1111/j.1600-0870.2005.00164.x

observations, and applied to periods for which proxies are sufficiently abundant. The number of potentially useful proxies is increasing steadily, and their reach back in time is overwhelming. For example, the bristlecone pines at the Methuselah Walk (California), as one of the oldest living species on earth, have tree-ring chronologies dating back to as much as 8000 yr! (Hughes and Graumlich, 1996). In view of roughly 200 yr of calibrating data, this is a bold undertaking. Specifically, the two main question to be answered are:

(1) Is the instrumental period representative for the predictor climate?
(2) If no, can one extrapolate the regression law using predictors that are not represented by that period?

Question (1) addresses the properties of the proxies and must be answered by inspecting their statistics. Question (2) can of course only be answered if the predictand, that is the past climate, is known. As this is not the case in the real-world, one must rest upon climate simulations performed by GCMs, along with using pseudo-proxies derived from it. We now describe a millennial climate simulation whose variations, including those of the predictors, deviate strongly from the present. It gives a very good playground for addressing questions (1) and (2). For the "real" proxies, however, the first question must of course be answered elsewhere (cf. Bürger and Cubasch, 2005).

2. The Erik simulation

The Erik simulation was performed by the coupled atmosphere-ocean general circulation model ECHO-G (Legutke and Voss, 1999; Zorita et al., 2003, S04). It is driven by an estimated history of the past 1000 yr of three external forcing factors: insolation, volcanic activity, and greenhouse gases, all of which have been estimated using proxy information from various sources (Blunier et al., 1995; Lean et al., 1995; Etheridge et al., 1996; Crowley, 2000). Due to internal variability, the model response to that forcing can only be sensibly verified on at least decadal timescales, and as global instrumental observations are available only since about 150 yr that possibility is rather limited. Figure 1 shows the Northern Hemispheric (NH) average of near surface temperature, as observed and interpolated from instrumental records (Jones et al., 1999), as analyzed by models (Kalnay et al., 1996), and as simulated by Erik. Note that there are significant differences between the instrumental observations and analyses (cf. Simmons et al., 2004). The discrepancies, especially those near the end of the period (not shown), are reduced if only land areas are considered for which the data coverage is larger. After about 1950, there are noticeable associations on longer timescales (5 yr and longer) between Erik and the observations/analyses. Before 1950, where analyses are unavailable, observed and simulated NH temperature diverge, with distinct simulated cooling and almost stationary observations backward in time. It is open whether or not this is attributable to discrepancies over the oceans, or to model deficiencies, or both.

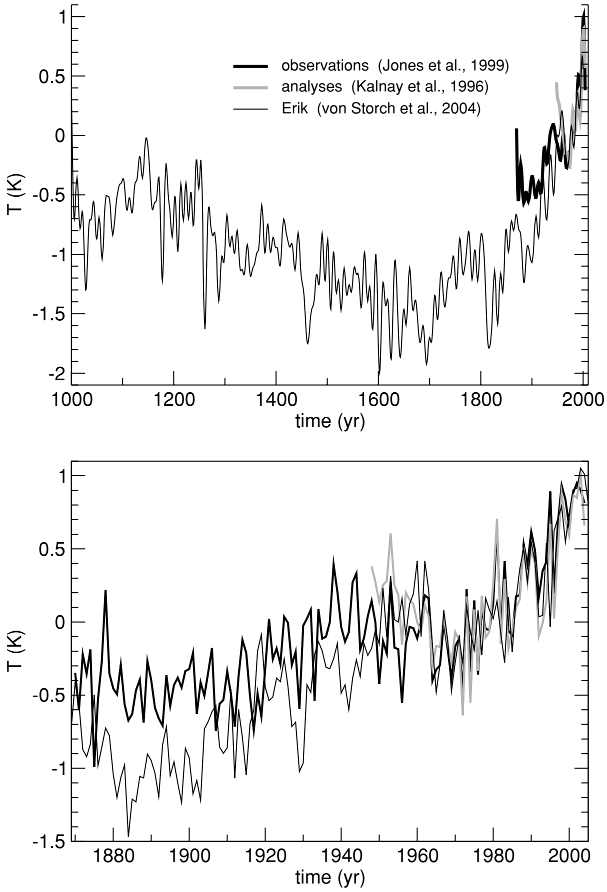

Fig. 1. Upper panel: Time series of NH temperature, from observations (heavy black), analyses (gray), and the Erik simulation (thin black). Lower panel: Closeup of upper panel for instrumental period. After about 1950, noticeable association exists between Erik and the observation-based values on longer timescales, in particular between Erik and analyses. Before 1950, Erik is generally cooler than observations, in particular in the 19th century. All curves are displayed using a 5-yr smoothing filter.

2.1. *Pseudo-proxies*

Whether or not the Erik simulation is realistic on longer timescales, it can serve as a "surrogate climate" (cf. S04) appropriate for solving methodological questions such as (1) and (2) above. Inside the world of a GCM with a known climate history, one can mimic the records of proxy data by noise-degraded values at specific temperature grid points. With two exceptions explained below, we follow exactly the experiments conducted by S04. That is, from the Erik fields we select 105 collocated grid points of the proxies used by MBH98 and superimpose, independently for each point, a white noise process of appropriate amplitude. S04 chose spatially uniform noise levels that were roughly inspired by the prevailing local proxy

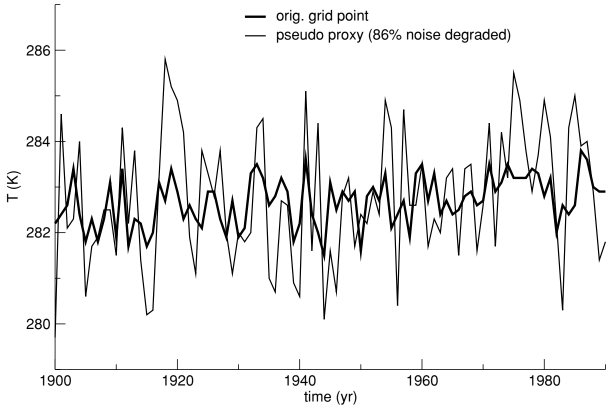

Fig. 2. Typical grid point temperature and corresponding pseudo-proxy. The disturbance is a white noise process whose variance is six times the original variance, so that 86% of the variance is noise. That level of association roughly corresponds to estimated values for real proxies (see text).

temperature correlations. But these levels do not account for the details of the multivariate reconstruction method, yielding correlations between original and reconstructed global temperature which are much higher than those of MBH98. Following MBH98, the portion of explained, spatially explicit global temperature variance is 22% for the verification phase. Using a proxy variance noise (i.e. a ratio between noise and total variance) of 86%, one does roughly match that measure, and we chose that level for the purpose of this study, which is exception one. The second exception is the following: For reasons to be discussed later, S04 applied a *detrending* in their regression calibration phase, and thus crucially divert from the MBH98 setting, which even emphasizes the importance of that trend. It turns out that this diversion has noticeable consequences for the reconstructions. Figure 2 shows a typical example of a pseudo-proxy record. Note that the degradation is effective only on the shorter timescales; the longer timescales appear more coherent. It has been pointed out (Osborn and Briffa, 2004) that using a red instead of a white noise process is possibly more realistic in the definition of pseudo-proxies.

3. The regression theme . . .

The basic idea behind regression is as follows. Once a functional form f is specified the method comes down to estimating, from an array of observed predictors x and predictands y with sufficient covariability, a set of unknown parameters, a, that satisfy as well as possible the identity

$$y = f(x, a). \tag{1}$$

Here we are dealing throughout with linear regression, so that the parameters a to be estimated are the entries of a matrix $\mathbf{L}$, as follows:

$$\hat{y} = \mathbf{L}x \tag{2}$$

(with $\hat{y}$ indicating the simulated values). Note that $\mathbf{L}$ is of dimension mn, with m and n denoting the dimensions of y and x, respectively. In our case, this gives $n = 105$ parameters to be estimated for each predictand.

Linear regression is defined as the maximum likelihood estimator of $\mathbf{L}$, which for a normally distributed error $\epsilon = \hat{y} - y$ is simply given by the least-squares solution:

$$\mathbf{L} = \mathbf{R} = \mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}, \tag{3}$$

with $\mathbf{C}_{xx}$, $\mathbf{C}_{yx}$, etc. denoting the cross-covariance between x and x, and y and x, respectively. Using the matrix

$$\mathbf{\Gamma}_{yx} = \mathbf{C}_{yy}^{-1/2}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1/2}, \tag{4}$$

eq. (2) can be written as

$$\hat{y} = \mathbf{C}_{yy}^{1/2}\mathbf{\Gamma}_{yx}\mathbf{C}_{xx}^{-1/2}x \tag{5}$$

which exactly corresponds to the one-dimensional case, the square root matrices being *SD*s and $\mathbf{\Gamma}_{yx}$ being the correlation coefficient.

3.1. Amplitude damping

Using eq. (5), the simulated local covariance can be written as

$$\mathbf{C}_{\hat{y}\hat{y}} = \mathbf{C}_{yy}^{1/2}\mathbf{\Gamma}_{yx}\mathbf{\Gamma}_{xy}\mathbf{C}_{yy}^{1/2}, \tag{6}$$

so that

$$|\mathbf{C}_{\hat{y}\hat{y}}| = |\Gamma_{yx}\Gamma_{xy}||\mathbf{C}_{yy}|. \tag{7}$$

Note that $\mathbf{\Gamma}_{yx}$ is known from canonical correlation analysis to be the matrix whose singular values, which are all ≤ 1, are the canonical correlations between y and x (e.g. Johnson and Wichern, 2002). Equation (7) expresses the fact that when regressing y on x, the simulated amplitudes are scaled by the canonical correlations between x and y. This is the exact mathematical expression of the damping effect of regression mentioned in S04. S04 argue that, with proxy temperature correlations being in the range 0.4–0.7, this effect renders regression-based reconstructions of past climate from proxies with too little variability and thus basically useless. They neglect the fact that eq. (7) is governed by *canonical correlations*, which can be much higher in a multidimensional context than their pointwise pendants. Another complicating effect is the superposition of the target quantity (NH average temperature) of other quantities (NH temperature PCs), where positive and negative effects might cancel each other.

3.2. Inverse regression

The entire damping argument is misplaced, however, if applied to the reconstructions undertaken by MBH98. MBH98 utilize a completely different scheme which one might call inverse regression. (To our knowledge, no other application makes use of this scheme.) There, one first regresses x (proxies) on y (temperature) and then inverts the result, which yields, according to eq. (5), the linear model

$$\tilde{y} = \mathbf{C}_{yy}^{1/2}\boldsymbol{\Gamma}_{xy}^{+}\mathbf{C}_{xx}^{-1/2}x, \tag{8}$$

where the "+" indicates the Moore–Penrose (pseudo) inverse. Obviously, $\boldsymbol{\Gamma}_{yx}$ and $\boldsymbol{\Gamma}_{xy}$ share the same singular values (= canonical correlations), so that eq. (7) has now the simple analogue

$$|C_{\hat{y}\hat{y}}| = \left|\Gamma_{yx}^2\right|^{-1} |C_{yy}|. \tag{9}$$

For inverse regression, the simulated amplitudes are therefore scaled by the inverse of the squared canonical correlations and thus *amplified*.

3.3. Colinearity

Equation (3) shows that to have a well-defined regression model, the covariance matrix $\mathbf{C}_{xx}$ must be invertible. If that is not the case, for example if there are more parameters than cases, this leads to an overfitted system. Likewise, the *SE* in estimating the regression matrix eq. (3) also depends on the inverse of $\mathbf{C}_{xx}$ (Johnson and Wichern, 2002). Therefore, a badly conditioned matrix $\mathbf{C}_{xx}$, for example if some predictors are correlated, leads to large estimation errors. In the current context, this so-called colinearity is relevant because the 20th century warming trend affects most of the proxy predictors and renders them highly correlated. To reduce the model uncertainty several standard measures can be applied, such as transforming to principal components (principal component regression, PCR), or simply removing the trend. The latter was applied by S04, erroneously assuming that MBH98 did the same (pers. comm. E. Zorita). Besides creating colinearity, the trend moreover reveals considerable nonstationarity in the (real) tree ring – temperature relation on centennial timescales (cf. Briffa et al., 2004) which would further degrade a regression model. In both cases, the inclusion of the trend belongs to the more general question of whether or not the calibrating variations (e.g. interannual) are representative for the millennium. While being important for real-world applications, that question must, however, clearly be separated from the purely methodological aspects such as amplitude damping.

4. . . . and 32 variations

It is customary to undertake certain pre- and postprocessings to the variates x and y before and after estimating the regression matrix. For example, one can detrend or, more generally, apply frequency filters to the data; another processing is the use of rescaling techniques (e.g. in downscaling) to correct for variance deficits. This processing is not a vital part of regression, and the literature is full of applications with all sorts of processing techniques and combinations thereof. To study the effects of such processing more closely, we have selected five criteria to define a "flavor" of regression. Since all of them are mutually independent, one has altogether as many as $2^5 = 32$ regression flavors on stock. This list is easily extendible but already enough for our purpose, which is the reconstruction of the Erik NH temperature from pseudo-proxies using all flavors.

Each flavor is determined by the validity of the criteria shown in Table 1. It can thus be identified using a binary code of length 5, indicating whether or not the criteria C1–C5 are valid. For example, 10011 refers to an inverse regression with rescaling, using trended, correlated predictors and spatially explicit predictands. This is the variant used by MBH98; the detrended version, i.e. 00011, was used by S04. The following points are important:

(ad C1): Twentieth century warming is the dominant variation in the instrumental data as well as in the Erik simulation. For example, 28% (Erik 37%) of instrumental global variance comes from the detrended, purely interannual fluctuations, contrasting to 71% (63%) stemming from the warming trend. Whether or not one builds the model on trended or detrended data should therefore strongly affect the result.

(ad C2): Applying PCR is not only a measure against the colinearity induced by the 20th century warming. By retaining only the dominant predictor PCs (in our case: 50% explained variance), most of the noise stemming from residual variations is filtered out.

(ad C3): One can use either the single predictand of spatially averaged NH temperature, or alternatively, a set of leading principal components so that spatial detail is simulated as well. Following MBH98, the predictand PCs are defined as follows: At first, the monthly fields were transformed to monthly anomalies from the 1900 to 1980 climatology and scaled by the corresponding (detrended) *SD*. Artificial grid-size-related effects were removed by weighting each grid point by the cosine of its latitude. These fields were subject to a principal component analyses, from which we selected the first nine dominant EOFs for the subsequent analysis. Principal components were obtained by projecting the original fields (from the entire period) onto the

Table 1. The five criteria used for defining a regression flavor

	C1	C2	C3	C4	C5
0	No trend	No PCR	Spatially explicit	Normal regression	No rescaling
1	Trend	PCR	Spatially averaged	Inverse regression	Rescaling

EOFs. Finally, annual averages were formed from the observed fields and principal components.

(ad C4): See above.

(ad C5): To match simulated and original variability, rescaling of the predictand is sometimes applied with scaling factors that are derived from the calibration period. This ensures adequate variability at least for that period, but introduces uncontrollable results if that domain is left. Note that if either one of C4 and C5 is satisfied, the simulated amplitude is increased.

Who has believed that the influence of the various flavors on the overall result is only moderate must rethink now. Figure 3 shows the reconstructions using the 32 flavors. We see that the entire spectrum is covered between practically no and almost perfect levels of variability. What complicates things is that except for the trend, no single factor can be isolated with a definite positive or negative influence on the variability, as shown below; they thus appear thoroughly mixed. But why should the result be any more unique? In view of Fig. 3, it is evident that the variability captured in the calibration period is only a small portion, and therefore not representative, of the full millennial variability. We shall return to this point further below.

The influence of the 20th century trend on the regression model is depicted in Fig. 4. It shows the S04 reconstruction (00011) versus the MBH98 analogue (10011). With the trend included in the training phase, much greater variability is simulated, not only for that century but for the entire historic period. For example, the Late Maunder Minimum (LMM, Luterbacher et al., 2001) – as measured by the 1681–1710 NH temperature average – is about −0.5 K versus −0.3 K in the S04 version, which is still quite imperfect relative to the true −0.9 K cold anomaly (which is, as we note again, way off the fluctuations

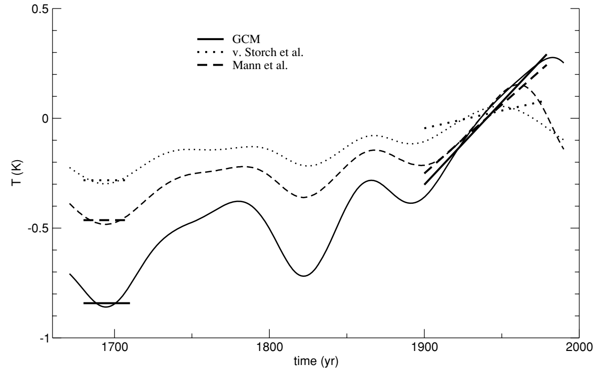

Fig. 4. Effect of using trend in calibration. Relative to the von Storch et al., 2004 method with detrended calibration data (dotted, from Fig. 3), the MBH98 analogue with the full 20th century warming trend (dashed) creates much larger variability. The LMM cooling is almost doubled, but still only 60–70% of the true cooling (solid). The straight lines indicate average LMM temperature and 20th century trend.

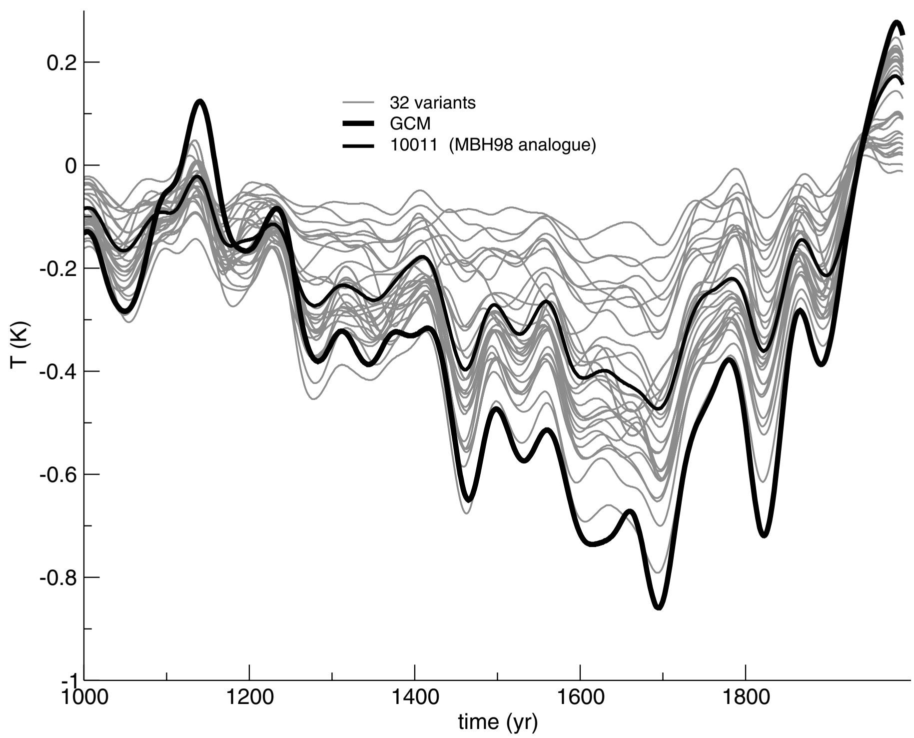

Fig. 3. Thirty-two variants of reconstructed NH temperature using regression from pseudo-proxies (gray). Relative to the true NH history (black), the entire spectrum between almost no and full variability is attained. The overall variations are well beyond those of the calibration period (1900-1980). A 30-yr smoothing has been applied to all curves.

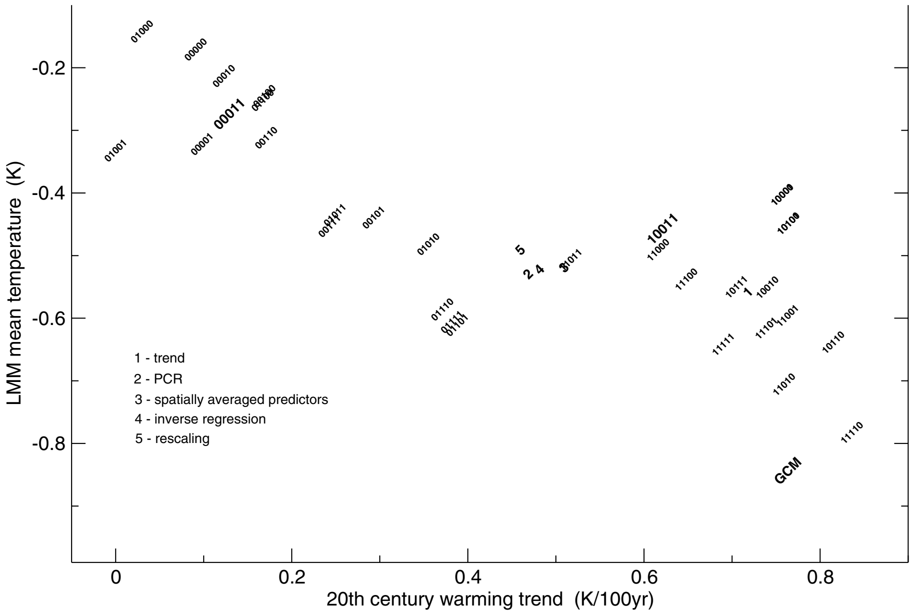

Fig. 5. Scatter of the 32 variants in terms of simulated 20th trend versus LMM climate. To be able to have a cold LMM, one evidently needs a strong warming trend. The reverse is only partly true as there are instances with a strong trend but without a cold LMM. We have emphasized (bold large) the versions of von Storch (00011) and Mann (10011) along with the true GCM values. Note that all 10 × 0x variants result from overfitting (see text). A single number *n* indicates the average overall variants with criterion *n* valid.

from the calibration period). It is trivial that if the 20th century trend cannot be reproduced, there is only little hope that fluctuations such as the LMM be simulated.

This is verified in Fig. 5, showing the scatter of the simulated 20th century trend versus the LMM climate for all 32 regression variants, along with the original Erik values. In fact, there is no variant with a cold LMM and no, or small, 20th century warming. The reverse is only partly true, as there are instances where that trend is strong but the LMM is not cold. Four of these counterexamples have the signature 10×0x. They all represent an underdetermined system (105 predictors, 81 cases) with perfect calibration statistics (so that C5 has no effect) that is prone to overfitting. As a consequence, strong 20th century trends are accompanied by reduced amplitudes for the past. It is not clear if this also applies to the variants 10×1x, i.e. those using inverse regression. Note further that the five variants closest to the GCM all have complementary choices of criteria 4 and 5 (i.e. "01" or "10" at the end) ensuring enhanced variability (see above). Four of them (11110, 11010, 11001, and 11101) are special instances of PCR with training 20th century trend. Conversely, the entire upper left corner, corresponding to bad performance, is occupied by detrended variants.

Except for the trend, no clear signs of the average behavior of each criterion can be recognized. If the trend is included (in the calibration), the simulated overall variability is enhanced, with a better representation of the LMM and especially the 20th century.

To summarize, if overfitting can be excluded it appears that the ability to reproduce the 20th century warming trend is enough to establish realistic variability for the millennium. In other words, the warming-induced 20th century covariations of pseudo-proxies and temperature are representative for the millennium – in contrast to the purely interannual (detrended) covariations.

5. Other calibration periods

With the variant 11110, that is, inverse PCR with trended data and spatially averaged predictands, the reconstruction closest to the GCM was achieved. For that variant, Fig. 6 shows the original reconstruction as in Fig. 3, along with two other versions with different calibration periods (1400–1480, 1300–1380). While it is obvious that each version shows best performance in its calibration period, they are all remarkably similar and close to the GCM in the entire millennial period. Even the period 1300–1380, with its fairly little variability, allows a 11110 calibration that is representative for the millennium. This is indicative of a large robustness of that variant and might recommend it for further use, for instance, with real proxies.

A word of caution may nevertheless be required here: In the stage of model definition, apart from things that are known a priori, only information from the calibration period must be used. To select from a large set of model variants (and 32 is large here), the one that shows the best skill for an independent period amounts to *defining* the model from that period, and therefore renders that same period no longer independent. Being without independent verification the model is, like in the case with many parameters, prone to overfitting. The only way to avoid

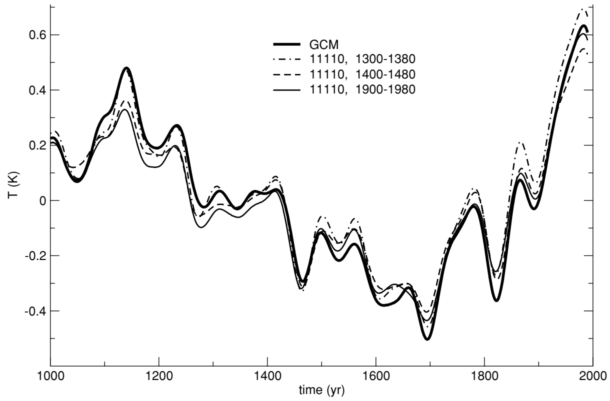

Fig. 6. 11110 reconstruction with various calibration periods. Units: anomalies from the 1000 to 1990 mean. Independent of the calibration period, performance is good for the full millennium. Note that 15th (dashed) and 20th (solid) century trends are strong, while the 14th century (dashed-dotted) has almost no trend. A 30-yr smoothing has been applied.

overfitting is to exclude it a priori, for example, by applying appropriate regularization schemes.

Whether or not one is able to select the "optimum" variant, they all differ in only a handful of details but still diverge so strongly on the millennial scale. This large divergence cannot simply be explained by overfitting. Figure 7 might throw some light on this as it concentrates, in a simple example, our view on the main issues around regression-based reconstructions. It shows two scatters between the dominant, trended proxy PC and the NH temperature. They represent the two populations corresponding to the 20th century calibration and to the entire millennium. First of all, NH temperature variations are obviously dominated by the first proxy PC. Moreover, not only are both populations evidently different [recall (1) from the first section], the calibrated linear relation is also not representative of the full millennial population. Application of the model therefore involves *extrapolation* of that relation [recall (2)], with an error that grows linearly with both the model uncertainty and the amplitude of the change. The model uncertainty, on the other hand, grows with the number of predictors, especially if they are colinear as in applications using trended multiproxies.

6. Conclusion

Inside the ideal world of a 1000-yr GCM simulation (Erik) with known climate history and noise degraded pseudo-proxies, we have tested the applicability of regression-based methods to reconstruct the NH mean temperature from the proxies. On the basis of five independent configuration switches, we defined a set of $2^5 = 32$ "flavors" of regression, each of which appearing a priori reasonable. It turned out that relative to the true NH temperature history, an entire spectrum was generated reaching from practically no variability to almost perfect reconstructions. This uncertainty is grounded in the fact that the empirical estimation applied here is not of the kind one usually encounters in the context of regression. The core assumption, that the calibrating portion is representative for the entire population, is not met

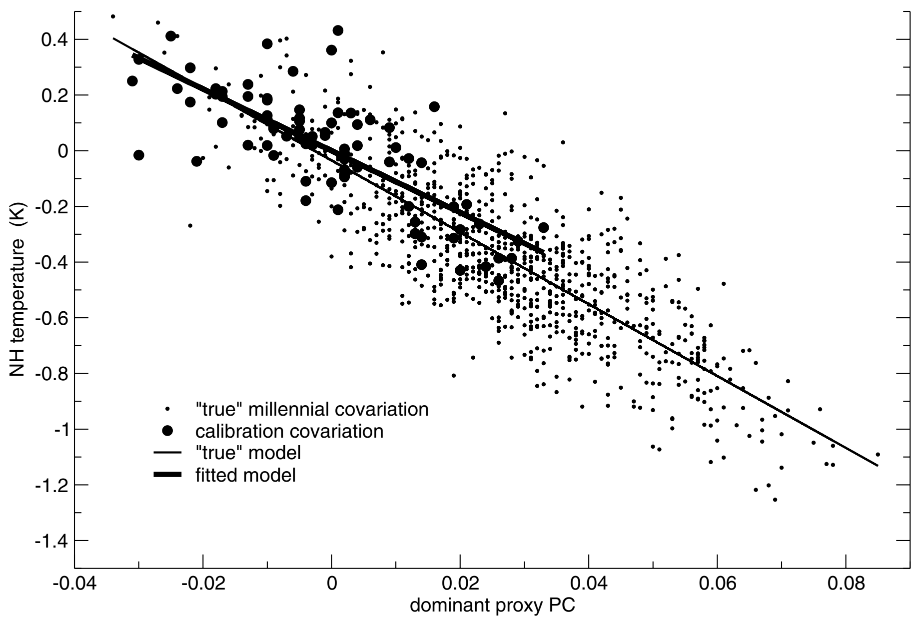

Fig. 7. Scatter plot between the dominant proxy PC and the NH temperature, as anomalies from the 1900 to 1980 mean. The calibration population (fat circles) is clearly distinct from the full millennium (dots). A strongly linear, but different relation exists for both populations. The calibration model (heavy line) predicts too small NH variations compared to the full millennium (thin line).

here as the calibrating 20th century climate is much warmer than the rest of the millennium. Hence, although a regression model is defined the context in which it is applied is an extrapolation into a domain which it has never experienced before. This is foreign to regression and explains the large spread of simulation errors.

The ability to simulate the 20th century trend is paramount for the performance of the full millennium. It turns out that the covariations defined by that trend are of a different kind than purely interannual (detrended) covariations; therefore, removing the trend is not a recommended procedure. The colinearity (estimation) problems arising from the few degrees of freedom of the trend, along with the predictor noise, can be handled using PCR.

With respect to the real-world, the question of applicability of the regression model is of foremost importance. This addresses the statistics of proxies such as tree rings, bore holes, corals, and others, and in particular the question: Are their 19th to 20th century variations representative for the entire millennium? If not, estimates of the model uncertainty are essential, and one should be aware that applying the model amounts to extrapolating the observed linearity (or whatever function is assumed) into the unknown and extrapolating the model uncertainty accordingly.

7. Acknowledgments

We thank Hans von Storch and Eduardo Zorita for helpful discussions. This work was made possible as part of the EU project SOAP.

References

Bardossy, A. and Plate, E. 1991. Modeling daily rainfall using a semi-Marcov representation of circulation pattern occurence. *J. Hydrol.* **122**, 33–47.

Barnston, A. G., Glantz, M. H. and He, Y. 1999. Predictive skill of statistical and dynamical climate models in SST forecasts during the1997–98 El Niño episode and the 1998 La Niña onset. *Bull. Am. Meteorol. Soc.* **80**, 217–243.

Blender, R., Luksch, U., Fraedrich, K. and Raible, C. 2003. Predictability study of the observed and simulated European climate using linear regression. *Q. J. R.. Meteorol. Soc.* **129**, 2229–2313.

Blunier, T., Chappellaz, J. A., Schwander, J., Stauffer, B. and Raynaud, D. 1995. Variations in atmospheric methane concentration during the Holocene epoch. *Nature* **374**, 46–49.

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

Briffa, K. R., Osborn, T. J., Schweingruber, F. H., Harris, I. C., Jones, P. D. and co-authors. 2001. Low-frequency temperature variations from a northern tree ring density network. *J. Geophys . Res.* **106**(D 3), 2929–2941.

Bürger, G. 2002. Selected precipitation scenarios across Europe. *J. Hydrol.* **262**, 99–110.

Bürger, G. and Cabasch, U. 2005. Are multiproxy climate reconstructions robust?. *Geoph. Res. Lett.* doi: 10.1029/2005GL024155 (in press).

Crowley, T. 2000. Causes of climate change over the past 1000 years. *Science* **289**, 270–277.

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

Etheridge, D., Steele, L. P., Langenfelds, R. L., Francey, R. J. Barnola, J. M. and co-authors. 1996. Natural and anthropogenic changes in atmospheric Co2 over the last 1000 years from air in Antarctic ice and firn. *J. Geophys. Res.* **101**, 4115–4128.

Fu, Q. and Liou, K. N. 1993. Parameterization of the radiative properties of cirrus clouds. *J. Atmos. Sci.* **50**, 2008–2025.

IPCC. 2001. 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, 944.

Hughes, M. K. and Graumlich, L. J. 1996. Multimillenial dendroclimatic records from Western North America. In: *Climatic Variations and Forcing Mechanisms of the Last 2000 years* (eds R. S., Bradley, P. D. Jones and J. Jouzel). Springer, Berlin, 109–124.

Jiang, N., Ghil, M. and Neelin, D. 1995. Forecasts of equatorial Pacific SST anomalies by using an autoregressive process and singular spectrum analysis. *Exp. Long-Lead Forecast Bull.* **4** (1), 35–36.

Johnson, R. A. and Wichern, D. W. 2002. *Applied Multivariate Analysis*. Prentice Hall, New Jersey, 767.

Jones, P. D., Briffa, K., Barnett, T. P. and Tett, S. F. B. 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.

Jones, P. D. and Mann, M. E. 2004. Climate over past millennia. *Rev. Geophys.* **42** (2).

Jones, P. D., New, M., Parker, D. E., Martin, S. and Rigor, I. G. 1999. Surface air temperature and its variations over the last 150 years. *Rev. Geophys.* **37**, 173–199.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D. and co-authors. 1996. The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77**, 437–471.

Lean, J., Beer, J. and Bradley, R. 1995. Reconstruction of solar irradiance since 1610: implications for climatic change. *Geophys. Res. Lett.* **22**, 3195–3198.

Legutke, S. and Voss, R. 1999. The Hamburg atmosphere-ocean coupled circulation model ECHO-G. DKRZ Technical Report 18, Hamburg, 62.

Lloyd-Hughes, B. and Saunders, M. A. 2002. Seasonal prediction of European spring precipitation from ENSO and local sea surface temperatures. *Int. J. Clim.* **22**, 1–14.

Luterbacher, J., Rickli, R., Xoplaki, E., Tinguely, C., Beck, C. and co-authors. 2001. The Late Maunder Minimum (1675–1715) – a key period for studying decadal scale climatic change in Europe. *Clim. Change* **49** (4), 441–462.

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

Mann, M. E. and Rutherford, S. 2002. Climate reconstruction using 'Pseudoproxies'. *Geophys. Res. Lett.* **29** (10), 139.

Osborn, T. and Briffa, K. R. 2004. The real color of climate change? *Science* **306**, 621–622.

Qian, B. and Saunders, M. A. 2003. Seasonal predictability of wintertime storminess over the North Atlantic. *Geophys. Res. Lett.* **30** (13), 1698.

Rodwell, M. and Folland, C. K. 2001. Atlantic air-sea interaction and seasonal predictability. *Q. J. R. Meteorol. Soc.* **128**, 1413–1443.

Simmons, A. J., Jones, P. D., da Costa Bechtold, V. Beljaars, A. C. M., Kållberg, P. W. and co-authors. 2004. Comparison of trends and low-frequency variability in CRU, ERA-40, and NCEP/NCAR analyses of surface air temperature. *J. Geophys. Res.* **109**, D24115.

Tiedtke, M. 1989. A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Mon. Weather. Rev.* **117**, 1779–1800.

von Storch, H., Zorita, E. and Cubasch, U. 1993. Downscaling of global climate change estimates to regional scales: an application to Iberian rainfall in wintertime. *J. Clim.* **6**, 1161–1171.

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

Zorita, E., González-Rouco, F. and Legutke, S. 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.