Clim Dyn (2009) 32:381–395
DOI 10.1007/s00382-008-0395-5

Comparison of climate field reconstruction techniques: application to Europe

Nadja Riedwyl · Marcel Küttel · Jürg Luterbacher · Heinz Wanner

Received: 16 September 2007 / Accepted: 11 March 2008 / Published online: 25 April 2008

Abstract This paper presents a comparison of principal component (PC) regression and regularized expectation maximization (RegEM) to reconstruct European summer and winter surface air temperature over the past millennium. Reconstruction is performed within a surrogate climate using the National Center for Atmospheric Research (NCAR) Climate System Model (CSM) 1.4 and the climate model ECHO-G 4, assuming different white and red noise scenarios to define the distortion of pseudoproxy series. We show how sensitivity tests lead to valuable "a priori" information that provides a basis for improving real world proxy reconstructions. Our results emphasize the need to carefully test and evaluate reconstruction techniques with respect to the temporal resolution and the spatial scale they are applied to. Furthermore, we demonstrate that uncertainties inherent to the predictand and predictor data have to be more rigorously taken into account. The comparison of the two statistical techniques, in the specific experimental setting presented here, indicates that more skilful results are achieved with RegEM as low frequency variability is better preserved. We further detect seasonal differences in reconstruction skill for the continental scale, as e.g. the target temperature average is more adequately reconstructed for summer than for winter. For the specific predictor network given in this paper, both techniques underestimate the target temperature variations to an increasing extent as more noise is added to the signal, albeit RegEM less than with PC regression. We conclude that climate field reconstruction techniques can be improved and need to be further optimized in future applications.

Keywords Paleoclimate · NCAR CSM 1.4 · ECHO-G 4 · Pseudoproxy data · RegEM · Principal component regression · European temperature reconstruction

1 Introduction

Knowledge of temperature amplitudes is of utmost importance in gaining a better understanding of past temperature evolution and change. Reconstruction of past temperature variability based on paleoclimatic data can provide insights into the interpretation of the role of climatic forcings. Many existing reconstructions place the twentieth century warming at continental to global scale into a broader context (Mann et al. 1998, 1999, 2005; Esper et al. 2002; Luterbacher et al. 2004, 2007; Mann and Rutherford 2002; Xoplaki et al. 2005; Rutherford et al. 2005; Casty et al. 2005a, 2007; Guiot et al. 2005; Moberg et al. 2005; Jansen et al. 2007). However these reconstructions have various limitations, primarily related to the availability of proxy data and their quality. It is a methodological challenge to filter out the climatic signal from a range of different proxy archives, given the short instrumental period for calibration and the increasing lack of predictors back in time.

Electronic supplementary material The online version of this article (doi:10.1007/s00382-008-0395-5) contains supplementary material, which is available to authorized users.

N. Riedwyl (✉) · M. Küttel · J. Luterbacher · H. Wanner
Oeschger Centre for Climate Change Research (OCCR) and National Centre of Competence in Research on Climate (NCCR), University of Bern, Erlachstrasse 9, 3012 Bern, Switzerland
e-mail: riedwyl@giub.unibe.ch

N. Riedwyl · M. Küttel · J. Luterbacher · H. Wanner
Institute of Geography, Climatology and Meteorology, University of Bern, Hallerstrasse 12, 3012 Bern, Switzerland

Reconstruction is generally approached in two ways. One possibility is to reconstruct the average, i.e. a single time series over a specific time period, e.g. the Northern Hemisphere average over the past millennium. The average series is reconstructed by making a composite of multiple proxy series, centered and scaled according to the target, i.e. composite-plus-scaling (CPS) (see Jones and Mann 2004; Esper et al. 2005). The other possibility is to focus on the whole climatic field of interest. In this case climate field reconstruction (CFR) techniques provide temporal and spatial information (Jones and Mann 2004, and references therein). The CFR approach provides a distinct advantage over averaged climate reconstructions, for instance, when information on the spatial response to external forcing (e.g. volcanic, solar) is sought (e.g. Shindell et al. 2001, 2003, 2004; Waple et al. 2002; Fischer et al. 2007). The results of both approaches, CFR and CPS, have led to some controversy over temperature amplitudes, raising questions about associated uncertainties, and the robustness and skill of the various reconstructions, as well as the influence of trends, and the length and climatology of the calibration period (von Storch et al. 2004; Buerger and Cubasch 2005; Thejll and Schmith 2005; von Storch et al. 2008; Wahl and Ammann 2007; Moberg et al. 2008). Recent studies provide some answers to these questions (Wahl and Ammann 2007) and introduce improved methodologies for reconstructions, e.g. the application of different parameter estimation techniques (Schneider 2001; Hegerl et al. 2006), the use of wavelet analysis (Moberg et al. 2005) or state space models (Lee et al. 2007). In this contribution we concentrate on CFR techniques.

Principal component (PC) regression is the classical method used to reconstruct past European climate field information and has been widely applied (Briffa et al. 1987; Cook et al. 1994; Luterbacher et al. 2004; Casty et al. 2005b, 2007; Xoplaki et al. 2005; Pauling et al. 2006). With PC regression, CFR is commonly performed under the assumption that no errors are inherent to the predictor data, and regression coefficient estimates are achieved using ordinary least squares (OLS). However, if noise is inherent to the predictor data, these estimates are negatively biased towards an underestimation that results in loss of variance (Lee et al. 2007). Several authors (Hegerl et al. 2006; Mann et al. 2005, 2007; Rutherford et al. 2005; Brohan et al. 2006; Esper et al. 2007; Lee et al. 2007; Li et al. 2007) have recently discussed the necessity of taking into account not only the uncertainties of the statistical model, i.e. the residuals, but also the errors inherent to the predictand and predictor data:

$$\mathbf{Y} + \mathbf{e}_{\text{instr}} = \mathbf{B}(\mathbf{X} + \mathbf{e}_{\text{proxy}}) + \mathbf{e} \quad (1)$$

where $\mathbf{e}$ are the residuals, $\mathbf{e}_{\text{instr}}$ the errors associated with the instrumental measurements, i.e. the predictand and $\mathbf{e}_{\text{proxy}}$ the errors associated with the predictors. Thus the methodological problems can be partly solved by better incorporating the different uncertainties in the statistical reconstruction models. Studies by Schneider (2001), Mann et al. (2005, 2007) and Rutherford et al. (2005) have proved the capability of the Regularized Expectation Maximization (RegEM) algorithm to more accurately reconstruct past temperature variations. One reason for this is that RegEM integrates $\mathbf{e}_{\text{proxy}}$ in the reconstruction technique, as ill-posed problems are regularized. Mann et al (2007) found truncated total least squares (TTLS) to be a particularly successful option for undertaking the regularization. RegEM with TTLS is used here as proposed by Mann et al. (2007) and following the instructions therein.

Some of the studies mentioned above found differences between the results obtained by using PC regression, on the one hand, and those achieved by means of the more sophisticated RegEM approach, on the other. However, these studies are limited to the hemispheric to global scale and, mainly, to annual resolution (Rutherford et al. 2005; Mann et al. 2007; Lee et al. 2007). One might expect to obtain different results when applying these techniques at a smaller spatial scale, such as Europe, and considering seasonal, rather than annual, data. In this study we therefore examine the sensitivity of the reconstruction skill at the continental scale, with seasonally resolved synthetic proxy data, i.e. proxies derived from climate model data. We use data from two simulations—one generated by the National Center for Atmospheric Research (NCAR) Climate System Model (CSM) 1.4 (Ammann et al. 2007), and the other generated by ECHO-G 4, which consists of the atmosphere and ocean general circulation models (GCM) ECHAM4 and HOPE-G (González-Rouco et al. 2006). Both simulations are likely to provide realistic opportunities for testing CFR approaches (Mann et al. 2005, 2007; von Storch et al. 2004; González-Rouco et al. 2006; Lee et al. 2007). Utilizing climate model data in a systematic experiment setup to attempt to reconstruct simulated past temperatures helps to understand the two techniques better. This would be less easily undertaken with real world multiproxy data as input, due to their heterogeneous nature and limited availability. The evaluation of CFR techniques is an important step in the process of identifying methodological deficits and limitations, providing "a priori" knowledge about the performance of the methodologies. Testing the techniques is therefore a good preparation for the next step: the improvement of reconstruction using real world proxy data. Apart from the choice of the reconstruction technique, there are several other factors limiting the skill of reconstructions of past climate variability, e.g. the varying number and spatial distribution of proxies over time (Pauling et al. 2003; Kuettel et al. 2007; Mann et al. 2007). However, here we focus on three things: on the

dependence of reconstruction skill on a specific predictor network, comparable in size and spatial distribution to a millennial European real world network, on the two techniques applied, and on the quality of the predictor data. We evaluate RegEM (Schneider 2001; Rutherford et al. 2005; Mann et al. 2007) for European summer and winter temperatures over the past millennium. In this study, RegEM is for the first time applied to spatial scales smaller than the hemispheric. Furthermore, we compare RegEM to PC regression, the basic multivariate regression model applied at the European scale, e.g. in Luterbacher et al. (2004, 2007), Casty et al. (2005b, 2007) and Xoplaki et al. (2005). In Sect. 2 we describe the NCAR CSM 1.4 and ECHO-G 4 climate model data and the experimental setting. Then we introduce the two CFR techniques and the criteria for comparison. In Sect. 3 we present the results. We begin by looking at the European average temperatures and diagnosing the skill. Then, we evaluate the spatial skill. The results are compared and discussed in Sect. 4, followed by a summary of our principal conclusions and a glance at future research in Sect. 5.

2 Data and methods

We test the performance of PC regression and RegEM in the surrogate climate of the two global coupled models NCAR CSM 1.4 and the ECHO-G 4. The use of climate model data permits an evaluation of the skill of the European reconstructions over a time period of 1,000 years and not only during the twentieth century verification period, as would be the case in reality. The brevity of the real world instrumental period for calibration makes it very difficult to compare techniques and assess reliability of their performance (e.g. Lee et al. 2007). Moreover, different virtual scenarios can be created by altering the input data of the statistical models, in order to better understand their performance and their sensitivities.

2.1 Simulated European surface air temperature data

NCAR CSM 1.4 (Ammann et al. 2007) and ECHO-G 4 (González-Rouco et al. 2006) are both global coupled models. NCAR CSM 1.4 has a grid resolution of 3.75° × 3.75° and is forced over the period 850–1999 AD. ECHO-G 4 has a grid resolution of 3.75° × 3.75° for the atmospheric component and 2.8° × 2.8° at low latitudes for the ocean, and is forced over 1000–1990 AD. NCAR CSM 1.4 forcings included are observation-based time histories of solar irradiance, aerosol loadings from explosive volcanism, greenhouse gases and anthropogenic sulfate aerosols (Ammann et al. 2007). Orbital parameters and land use changes are not included as forcings in NCAR CSM 1.4. Any potential long-term drift is removed by subtracting a millennial-scale spline fit for individual months of the annual cycle, obtained from the control integration, at each gridpoint (Ammann et al. 2007). ECHO-G 4 forcing includes natural (solar irradiance, radiative effects of stratospheric volcanic aerosols) and anthropogenic (greenhouse gas concentrations) estimates (González-Rouco et al. 2006) of past millennial external forcings. A flux adjustment constant in time and zero spatial average are used to inhibit climate drift (González-Rouco et al. 2006). The NCAR CSM 1.4 simulation used here is the one with 'medium' solar irradiance scaling (0.25% Maunder Minimum reduction) in the terminology of Ammann et al (2007). The ECHO-G 4 simulation (using 0.3% Maunder Minimum reduction) is the one sometimes known as 'Erik 2' (González-Rouco et al. 2006), which has cooler initial conditions than the older 'Erik 1' simulation used in several previous pseudoproxy studies.

The predictand in the reconstruction experiments is the simulated gridded surface air temperature field, generated by the NCAR CSM 1.4 and the ECHO-G 4 simulations respectively. To represent Europe we selected the area 52.5° W–71.25° E and 28.125° N–76.875° N of the global model run, which gives a rather coarse picture of the European area, namely 476 gridboxes (land and sea). Gridded model surface temperature information with a higher spatial resolution is not available for the past millennium. Nevertheless, testing and comparing CFR techniques in this experimental setting is reasonable. The original NCAR CSM 1.4 and ECHO-G 4 simulation temperature data are monthly resolved. We have calculated seasonal mean temperatures for summer (JJA) and winter (DJF) starting in December 1000 AD and ending in August 1990 AD. Some analyses are made on a gridpoint basis, while others are made for the (latitude weighted) European average temperature.

2.2 The Pseudoproxy data

The predictor data used for this study corresponds to NCAR CSM 1.4 and ECHO-G 4 model gridpoints closest to real world proxy locations in Europe. As the proxies are derived from the model data, we call them synthetic proxies or "pseudoproxies" (Mann and Rutherford 2002; Rutherford et al. 2005; von Storch et al. 2004, 2006). The pseudoproxy locations are chosen according to published data (Mann et al. 1999; Briffa et al. 2001; Klimenko et al. 2001; Proctor et al. 2002; Shabalova and van Engelen 2003; Luterbacher et al. 2004, 2007; Casty et al. 2005b; Rutherford et al. 2005; Guiot et al. 2005; Mangini et al. 2005) and some other data that will be potentially available from current research projects (NCCR Climate and MILLENNIUM). The real world proxy data referred to

consists of 1,000 year long series and some series covering several centuries. Additionally, a few gridpoints refer to shorter real world series, which are primarily used to optimize the spatial distribution of the network towards Eastern and Southern Europe. We argue that if the techniques already fail using input data covering the full length of 1,000 years, they certainly can be expected to do so, if the number and spatial distribution are reduced and change through time. Thus the pseudoproxies derived and used in this paper are idealized, as we assume them all to be constantly available over the full time period of 1,000 years (e.g. as been done in Mann et al. (2007)). Keeping the spatial distribution and the number of proxies constant over time allows us to focus on the actual variable of interest, the performance of the two different CFR techniques. We limit ourselves to considering mainly one network, equal for both seasons and without changing availability over time. The predictor network (Fig. 1) consists of 30 gridpoints and is seen as a reasonable selection of a predictor network for a 1,000 year European temperature reconstruction. Additional testing has been made with a smaller pseudoproxy network, which consists of 12 gridpoints (not shown). These pseudoproxies refer to real world proxy series available to reconstruct the late Maunder Minimum (Kuettel et al. 2007). The conclusions drawn are conditional upon the specific network configuration considered. Accordingly, this study can not apply in complete generality. Moreover, we restrict our analysis based on the assumption that our pseudoproxies have seasonal resolution and do not combine temporally low and high resolved climate proxies such as those for instance in Moberg et al. (2005). Generally, the quantity and, even more, the spatial distribution of the proxy information plays a crucial role in determining the reconstruction skill. Even a single point, if optimally situated, has an impact on the reconstruction result, and thus improves the skill (Kuettel et al. 2007). However, the focus of this study lies more on the performance of the two reconstruction techniques as such.

We use different scenarios for errors in the local pseudoproxy series, i.e. the predictors are characterized by the addition of red or white noise with varying signal to noise ratios (SNR) to the simulated temperature signal. Noise is added as indicated e.g. in Mann and Rutherford (2002) and von Storch et al. (2004), with the difference that the pseudoproxies here are constructed based on seasonal means, correlations etc., i.e. separately for summer and winter, taking into account different responses of real-world proxy data to warm and cold season conditions.

The predictand is regarded as "perfect", i.e. no noise is added. The noise is intended to mimic errors inherent to the predictor data (Eq. 1). White noise is added to be consistent with the premises given by the regression model used, i.e. the residuals are independent and identically distributed (i.i.d.). We have selected the five SNR 0.25, 0.4, 0.5, 1 and ∞ (no added noise) according to Mann et al (2007). With $r = SNR/(1 + SNR^2)^{1/2}$ the SNR is related to the associated root–mean–square correlation between the predictor data and their associated local climate signal (Mann et al. 2007). We obtain $r = 0.24$, 0.37, 0.45, 0.71 and 1.0 for the five SNR values under consideration, respectively (Mann et al. 2007). As it is plausible that errors in proxy series are serially autocorrelated, we use red noise to make the uncertainties more realistic. The red noise is modeled as a first- order autoregressive AR(1) process (Mann et al. 2007) and represented by $X_t = \phi X_{t-1} + Z_t$, where $Z_t \sim WN(o,\sigma^2)$ and $\phi \neq 0$. For AR(1) processes the autoregressive parameter ϕ is equal to the sample lag-1 autocorrelation

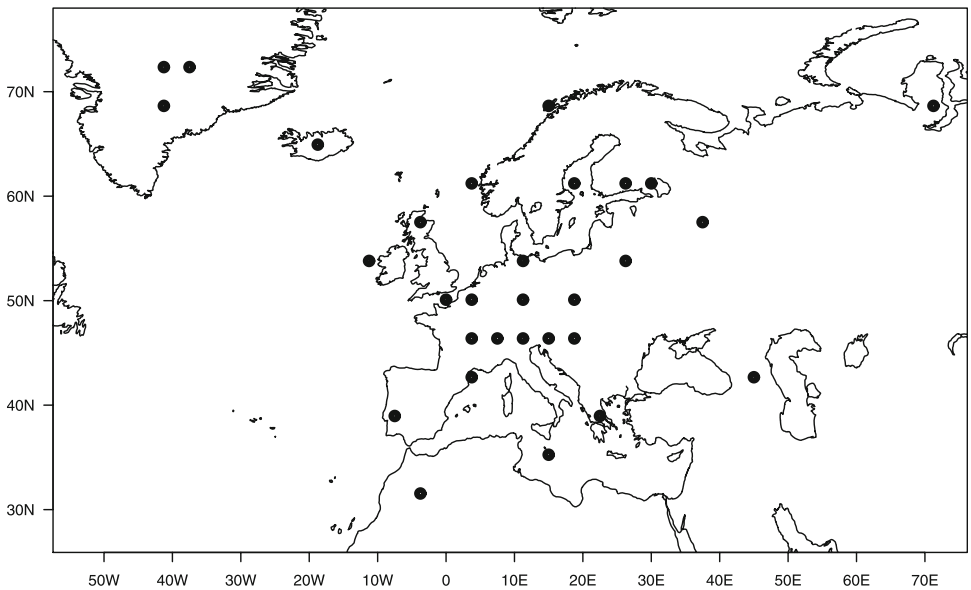

Fig. 1 The distribution of the 30 pseudoproxies used in this study. Each dot corresponds to the north-western corner of one 3.75° × 3.75° gridbox of the NCAR CSM 1.4 and ECHO-G 4 model

coefficient ρ, here $\rho = 0.32, 0.71$. The sample lag-1 autocorrelation coefficients for red noise as well as the five SNR for white noise are the same as those evaluated in Mann et al. (2007). This allows for direct comparison, making it possible to determine whether RegEM performs better than PC regression at the continental scale as well, and how the increase in temperature variability due to the downscaling affects the reconstruction results.

2.3 PC regression versus RegEM

RegEM was first described by Schneider (2001). It has only recently been further developed and implemented by Rutherford et al. (2005) and Mann et al. (2007), and is compared to PC regression in the present paper. The two reconstruction techniques each take a different approach to the reconstruction "problem" (Fig. 2). With PC regression, past temperature values are "retrodicted", i.e. predicted into the past, whereas with RegEM missing values are imputed, i.e. missing values are replaced by plausible ones. While for RegEM the input is the whole data matrix including the missing and available values, as indicated in red (Fig. 2), for PC regression only the available predictand and predictor values are part of the input, as shown in green (Fig. 2).

2.3.1 Multivariate principal component regression

Multivariate PC regression seeks to reconstruct the past temperature field using the PC of both the predictand and the predictors:

$$\mathbf{y}_{\mathrm{pc}} = \mathbf{x}_{\mathrm{pc}}\mathbf{B} + \mathbf{e} \tag{2}$$

where $\mathbf{B}$ are the regression coefficients relating the explanatory variables $\mathbf{x}_{\mathrm{pc}}$, i.e. the predictor information, and the target $\mathbf{y}_{\mathrm{pc}}$, i.e. the predictand. The relationship is assumed to be a linear function of parameters stationary over time. The regression coefficients of the calibration period, here $\mathbf{B}$, are estimated by OLS and then used to "retrodict" past temperature values. Predictand and predictors are transformed to their PC to obtain orthogonal series and make it possible to reduce the dimensionality of the data while still retaining most of the variability contained in the full dataset (Wilks 1995). This allows for climatic interpretation of temperature fields, as first few PC typically capture large-scale modes. Here the calculation of the PC is based on the correlation matrix as for instance in Luterbacher et al. (2004). Furthermore, they are truncated as in that study, i.e. most of the variance is captured by considering only the most important directions of the joint variations, thus avoiding redundancy (Wilks 1995).

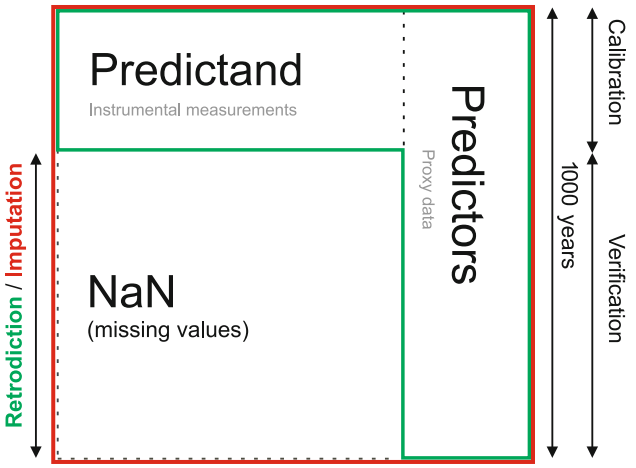

Fig. 2 Scheme of the analogousness/differences between PC regression (*green*) and RegEM (*red*). PC regression corresponds to "retrodiction" and RegEM to the imputation of past temperature values. The input matrix for both techniques is indicated in *colors*

2.3.2 Regularized expectation maximization

RegEM is a covariance-based iterative CFR technique based on the idea of gradual linear modeling of the relationship between missing values and available values, also taking into account ill-posed or under-determined settings (Mann et al. 2007). The input data matrix combines both predictand and predictor data over the full reconstruction period:

$$\mathbf{x}_m = \mu_m + (\mathbf{x}_a - \mu_a)\mathbf{B} + \mathbf{e} \tag{3}$$

where $\mathbf{B}$ refers to the regression coefficients relating available values $\mathbf{x}_a$ and missing values $\mathbf{x}_m$ within the multivariate data set. $\mathbf{e}$ is the random vector representing the error with mean zero and the according covariance matrix $\mathbf{C}$ to be determined (Schneider 2001; Mann et al. 2007). The conventional iterative Expectation Maximization algorithm (EM) estimates the mean and the covariance matrix of an incomplete data matrix and imputes values for the missing ones (Schneider 2001). The EM algorithm is used under the assumption that the predictand and predictor data are Gaussian. With each iteration step, estimates of the mean μ and the covariance-variance matrix Σ of the input matrix are calculated, followed by the computation of estimates of the coefficient matrix $\mathbf{B}$ and the residual covariance matrix $\mathbf{C}$. The iteration is repeated step by step until the convergence criterion is fulfilled (Schneider 2001; Mann et al. 2007).

In cases where the number of variables exceeds sample size the EM algorithm has to be regularized, as ill-posed problems lead to singularity of the covariance-variance matrix Σ (Schneider 2001). Instead of estimating the coefficients $\mathbf{B}$ by the conditional maximum likelihood method given the estimates of μ and Σ, the parameters are estimated

by TTLS. Thus, in order to regularize the covariance matrix Σ its PC are truncated, i.e. only a specific number of PC is considered, according to the truncation parameter. For further information and a more detailed description of RegEM see Schneider (2001), Rutherford et al. (2005) and Mann et al. (2007). In our study the non-hybrid, revised version of RegEM is used (Mann et al. 2007). We standardized the available values with regard to the calibration period, 1900–1990 AD, to ensure that the testing of climate reconstruction methods relies on the appropriate application of real world constraints (e.g. Smerdon and Kaplan 2007). The truncation parameters for TTLS are chosen in two ways. The first is as explained in Mann et al. (2007). Mann et al. (2007) identify optimal truncation parameters based on the estimate of the noise continuum to the log-eigenvalue spectrum (Wilks 1995). This procedure serves to determine leading eigenvalues that lie above the estimated noise continuum. The second way is by evaluating a range of possible other truncation parameters and then selecting the parameters leading to reconstruction results with smallest differences in mean and standard deviation to the target over the verification period. As stated in Mann et al. (2007), the choice of the truncation parameters is not unique. This is illustrated here: validation scores of reconstruction results obtained with the log-eigenvalue spectrum criteria are shown together with those of reconstruction results (see supplementary online material) using alternative truncation parameters.

Furthermore, the reconstructions were performed both with and without the PC of the predictand. However, analyses indicated that results using or not using PC analysis do not differ much (not shown), and therefore, in this paper, we restrict our results to the case of not using the PC of the predictand. In this way another ambiguous choice is avoided and the whole range of variability is retained for reconstruction.

2.4 The comparison criteria

PC regression and RegEM are compared to each other in the same experimental setting. As mentioned above, the reconstructions are performed within the surrogate climate of the NCAR CSM 1.4 and ECHO-G 4 climate models using 30 pseudoproxies with different SNR, all constant over time. We investigate how and to what extent the quality of the predictor data affects the reconstruction skill. Furthermore, we evaluate the results of the two techniques. On the one hand, the skill of the reconstructions is analyzed focusing on the European average only. For this reason, figures display the target, the European average temperature from 1001–1990 AD, in comparison to the reconstruction results, accompanied by a quantitative summary of the skill. The commonly used reduction of error (RE) and coefficient of efficiency (CE) skill scores are calculated. Tables 1 and 2 indicate the RE and CE skill scores over the verification period 1001–1899 AD, both for NCAR CSM 1.4 and ECHO-G 4. On the other hand, we concentrate on the climate field information, i.e. the spatial patterns. Our focus here lies on the averaged reconstruction bias, and RE calculated for the 30-year filtered reconstruction results at each gridpoint, both over the verification period 1001–1899 AD. In first comparing the target with the reconstruction results for each technique separately, and subsequently comparing the results of PC regression with those of RegEM, we determine how well the techniques perform, depending on the influence of the errors inherent to the predictor information.

3 Results

3.1 Impact of the quality of the predictor data

The subsequent figures all refer to results obtained using NCAR CSM 1.4, whereas the results produced with ECHO-G 4 are provided in the supplementary online material, with the exception of the skill scores tables (Tables 1, 2), which are shown for both climate models.

Figures 3 and 4 show the methodological comparison for averaged European summer (Fig. 3, suppl. Fig. 3) and winter (Fig. 4, suppl. Fig. 4) temperature reconstructions from 1001 to 1990 AD (land and sea). The figures display temperature anomalies with regard to the calibration period 1900–1990 AD. The target, i.e. the average of the simulated European surface air temperature over the past millennium, is shown in black, while the average of the reconstructed summer and winter temperature fields are given in color. All curves are smoothed with a 30-year running mean. The results differ according to the five white noise scenarios used in the reconstructions. The NCAR CSM 1.4 target exhibits variability with quite large quasi-periodic amplitude variations over the past millennium, both in summer and in winter. The variability for summer and winter average temperatures is similar to that exhibited by the ECHO-G 4 run (von Storch et al. 2004; González-Rouco et al. 2006).

The reconstructions realized with PC regression (Fig. 3, suppl. Fig. 3, top) and the perfect pseudoproxy set, i.e. no white noise added (yellow line), capture the target very well. However, the more white noise is added to the signal, the more this technique fails to properly reconstruct, and underestimates the amplitude of the target temperature variations. Thus, the difference between negative temperature anomalies of the reconstruction results and the calibration period mean is not as large as that of the target and the calibration period mean, i.e. the reconstruction being to warm. There is a shift from the scenarios with

Table 1 RE as well as CE skill scores for the NCAR CSM 1.4 results shown in Figs. 3, 4, 5 and 6, and the non-filtered reconstruction results (not shown)

	PC reg				RegEM							
	30 year filtered		Non-filtered		30 year filtered				Non-filtered			
	su	wi	su	wi	su		wi		su		wi	
Reduction of error (RE)												
Perfect	0.99	0.97	0.96	0.95	0.96	0.98	0.874	0.91	0.84	0.91	0.77	0.88
SNR1	0.92	0.86	0.81	0.73	0.98	0.96	0.873	0.92	0.65	0.78	0.73	0.75
SNR0.5	0.74	0.66	0.56	0.42	0.97	0.95	0.86	0.84	0.58	0.51	0.4	0.32
SNR0.4	0.65	0.59	0.45	0.32	0.96	0.91	0.8	0.77	0.34	0.31	0.14	0.02
SNR0.25	0.45	0.45	0.22	0.14	0.839	0.82	0.6	0.54	−0.12	−0.24	−0.58	−0.56
SNR1, $\phi = 0.32$	0.86	0.74	0.74	0.6	0.93	0.9	0.86	0.95	0.65	0.63	0.62	0.76
SNR1, $\phi = 0.71$	0.8	0.68	0.67	0.49	0.843	0.91	0.75	0.9	0.66	0.71	0.39	0.6
Coefficient of efficiency (CE)												
Perfect	0.98	0.92	0.85	0.86	0.86	0.94	0.633	0.74	0.45	0.68	0.33	0.66
SNR1	0.73	0.59	0.36	0.2	0.92	0.85	0.630	0.77	−0.19	0.24	0.22	0.28
SNR0.5	0.13	0.003	−0.47	−0.7	0.88	0.81	0.598	0.53	−0.43	−0.66	−0.76	−0.98
SNR0.4	−0.19	−0.21	−0.86	−0.996	0.59	0.703	0.43	0.34	−1.23	−1.35	−1.53	−1.86
SNR0.25	−0.85	−0.62	−1.63	−1.52	0.45	0.4	−0.16	−0.35	−2.8	−3.19	−3.62	−3.55
SNR1, $\phi = 0.32$	0.54	0.24	0.13	−0.18	0.77	0.67	0.605	0.85	−0.18	−0.24	−0.11	0.31
SNR1, $\phi = 0.71$	0.31	0.05	−0.13	−0.5	0.47	0.704	0.263	0.71	−0.14	0.02	−0.77	−0.16

The calibration period is from 1900 to 1990 AD, the verification period from 1001 to 1899 AD. For RegEM RE and CE are shown for two different TTLS parameters (left, TTLS parameters chosen as in Mann et al. (2007), right, as additionally proposed in this paper)

Table 2 As Table 1, but for ECHO-G 4 (results see supplementary online material)

	PC reg				RegEM							
	30 year filtered		Non-filtered		30 year filtered				Non-filtered			
	su	wi	su	wi	su		wi		su		wi	
Reduction of error (RE)												
Perfect	0.99	0.996	0.96	0.93	0.947	0.98	0.986	0.99	0.84	0.92	0.75	0.78
SNR 1	0.93	0.89	0.85	0.7	0.92	0.96	0.95	0.97	0.79	0.81	0.36	0.04
SNR 0.5	0.79	0.66	0.65	0.35	0.89	0.87	0.82	0.85	0.63	0.46	−1.73	−1.45
SNR 0.4	0.7	0.6	0.54	0.25	0.88	0.82	0.77	0.77	0.53	0.27	−2.72	−2.31
SNR 0.25	0.62	0.49	0.42	0.07	0.91	0.92	0.81	0.65	0.12	−0.11	−2.3	−2.36
SNR 1, $\phi = 0.32$	0.94	0.952	0.84	0.78	0.948	0.95	0.95	0.94	0.82	0.72	0.39	0.4
SNR 1, $\phi = 0.71$	0.81	0.953	0.68	0.62	0.97	0.84	0.86	0.85	0.79	0.32	0.18	0.17
Coefficient of efficiency (CE)												
Perfect	0.95	0.991	0.8	0.86	0.760	0.92	0.97	0.98	0.28	0.61	0.47	0.53
SNR 1	0.67	0.76	0.29	0.35	0.64	0.79	0.9	0.93	0.02	0.13	−0.37	−1.04
SNR 0.5	0.01	0.28	−0.62	−0.37	0.51	0.39	0.62	0.68	−0.7	−1.48	−4.81	−4.22
SNR 0.4	−0.38	0.15	−1.11	−0.59	0.45	0.18	0.51	0.52	−1.15	−2.37	−6.91	−6.05
SNR 0.25	−0.75	−0.09	−1.67	−0.97	0.6	0.62	0.59	0.27	−3.07	−4.11	−6.01	−6.15
SNR 1, $\phi = 0.32$	0.72	0.898	0.25	0.53	0.762	0.78	0.88	0.87	0.15	−0.3	−0.29	−0.27
SNR 1, $\phi = 0.71$	0.12	0.899	−0.47	0.2	0.87	0.26	0.71	0.69	0.01	−2.14	−0.74	−0.77

higher SNR, SNR ∞ and SNR 1 (yellow and red lines) to those with lower SNR, SNR 0.5, 0.4, 0.25 (blue and green lines), and a decrease in skill indicated by the RE and CE scores in Tables 1 and 2. The RegEM reconstruction result of the SNR ∞ scenario captures the target well, too. In comparison to PC regression, RegEM captures the target

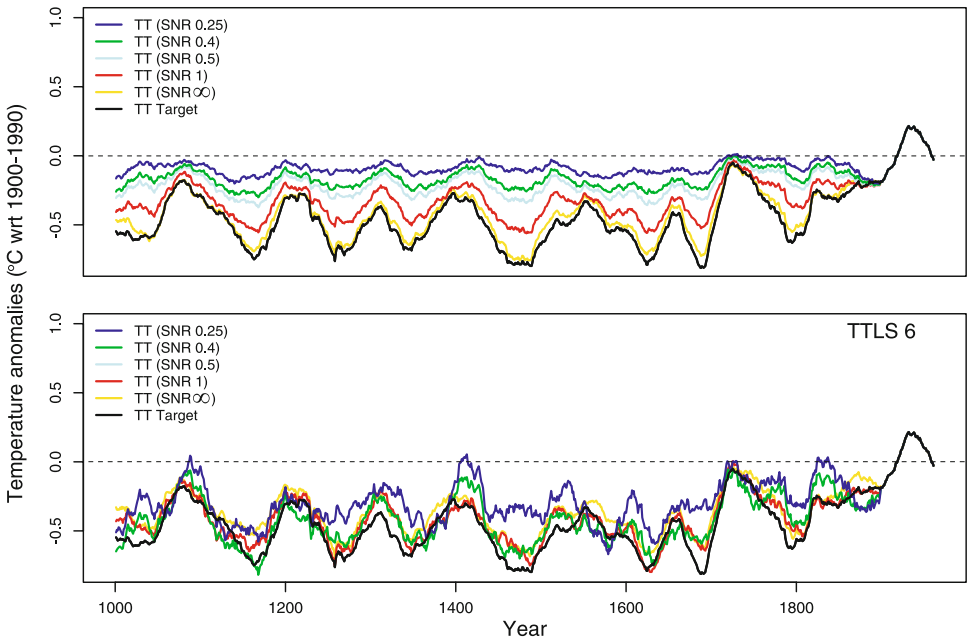

Fig. 3 European summer average temperature anomalies (30-year running mean) wrt 1900–1990 AD, for PC regression (*top*) and RegEM (*bottom*), using 30 pseudoproxies (see Fig. 1) with varying white noise added to the signal. The target (*black line*) is compared to the reconstruction results (*colored lines*). TTLS indicates which truncation parameter is used to reconstruct

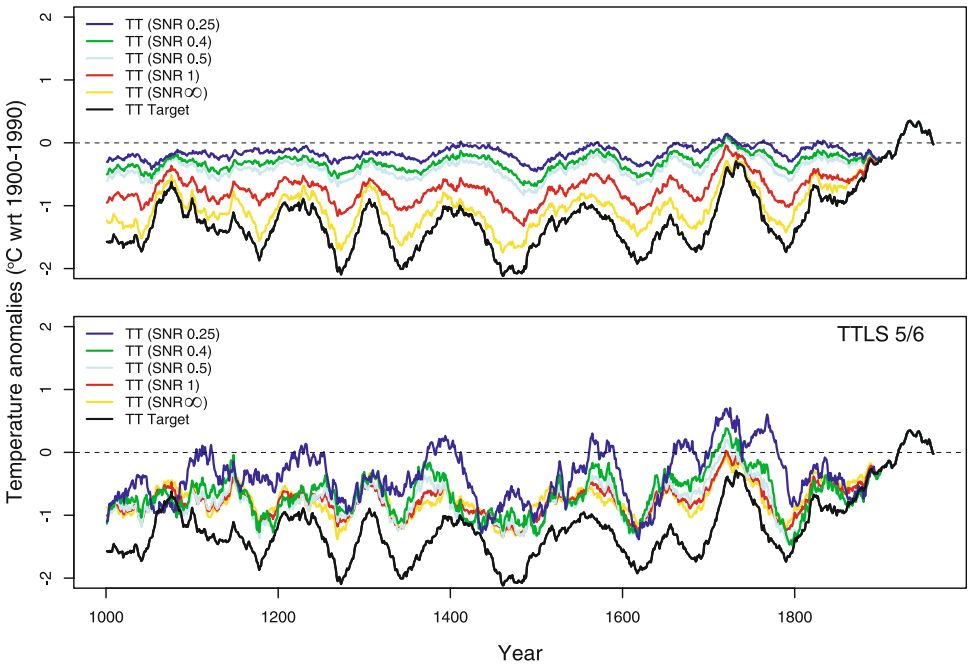

Fig. 4 As Fig. 3, but for winter

summer average temperature (Fig. 3, suppl. Fig. 3, bottom) more adequately for all white noise levels.

After focusing on the performance of the techniques for summer reconstructions, we now turn to the reconstruction results for European winter average temperatures (Fig. 4, suppl. Fig. 4). Figure 4 shows that both techniques capture the target average temperature less accurately for winter than for summer (Fig. 3, suppl. Fig. 3), a finding which is more pronounced for NCAR CSM 1.4 than for ECHO-G 4. In principle, we obtain the same picture for PC regression as described above for the European summer average temperature reconstruction results. However, the RE and CE skill scores are higher for summer than for winter (Tables 1, 2). Overall, RegEM seems to be more robust and less sensitive to the amount of white noise added to the signal than PC regression, although, as seen for winter (Fig. 4, and even more so suppl. Figure 4), it appears that RegEM can 'invent' undesirable, temporal features, such as various spurious quasi-periodic variations, which do not exist in the target data. Nevertheless, the range of the variability of the 30-year filtered results corresponds better to that of the target for RegEM (RE and CE 30-year filtered

in Tables 1 and 2). While both techniques reconstruct the target average temperature less accurately with increasing noise level (Tables 1, 2), RegEM does so to a considerably lesser degree than PC regression.

Figures 5 and 6 (suppl. Figs. 5, 6) show a second comparison of reconstruction results of the summer and winter average temperature anomalies with regard to the 1900–1990 AD calibration period, now with red noise applied in comparison to the corresponding white noise scenario. The middle white noise scenario SNR 1 is displayed together with the two red noise scenarios with the same SNR, but different sample lag-1 autocorrelation coefficients $\rho = 0.32$, 0.71 (as mentioned above, chosen according to Mann et al. (2007)). For PC regression (Figs. 5 and 6, suppl. Figs. 5 and 6, top) the addition of red noise (orange and magenta line) affects the skill of the reconstruction slightly more than the addition of white noise with SNR 1 (red line), both for summer and winter according to RE and CE (Tables 1, 2). For RegEM, the target temperature variations also remain appropriately reconstructed for summer when red noise is added instead of white noise according to RE and CE (Tables 1, 2), although adding red noise with an autocorrelation coefficient $\rho = 0.71$ (magenta line) clearly increases the variability of the reconstruction result in winter.

The RE and CE scores for the 30-year filtered data (Tables 1, 2) quantitatively describe the reconstruction results (Figs. 3, 4, 5 and 6, likewise for ECHO-G 4 in the supplementary online material) and confirm that RegEM performs better than PC regression focusing on the evaluation of the low frequency variations (RE and CE 30-year filtered in Table 1 and 2). Nevertheless, a glance at the RE and CE scores calculated for non-filtered results (figures not shown) reveals differences in the performance of the reconstructions seen in Figs. 3, 4, 5 and 6. Summer average temperature reconstructions using RegEM also produce lower RE scores than those using PC regression under the different white and red noise scenarios (RE non-filtered in Tables 1 and 2). Winter temperature reconstructions based on RegEM and PC regression RE and CE scores are comparable (Table 1), and slightly lower in a few cases for RegEM (Table 2). The SNR 0.25 scenario, in particular, leads to lower skill score values, and the result is generally unsatisfactory. Using rednoise scenarios (Fig. 6, suppl. Fig. 6, bottom), the range of the variability of the SNR 1 scenario with an autocorrelation coefficient of $\rho = 0.71$ (magenta line) is rather somewhat too large compared to the target (Fig. 5, suppl. Fig. 5) for RegEM. Finally, several scenarios for both summer and winter even return negative annual RE and CE scores with RegEM, indicating that these reconstruction results have no skill. With the alternative way of determining the TTLS parameters for RegEM (supplementary online material), equally skilful, and in some cases even more skilful reconstructions can be achieved. For PC regression, RE scores indicate that all reconstruction results have skill; however, this is contradicted (for SNR 0.5, SNR 0.4 and SNR 0.25) by the corresponding CE scores (Figs. 3 and 4, suppl. Figs. 3 and 4).

To summarize: Figs. 3 and 4 (Suppl. Figs. 3 and 4) as well as Tables 1 and 2 indicate that both techniques reconstruct European temperature variability more adequately for summer than for winter. RegEM seems to be

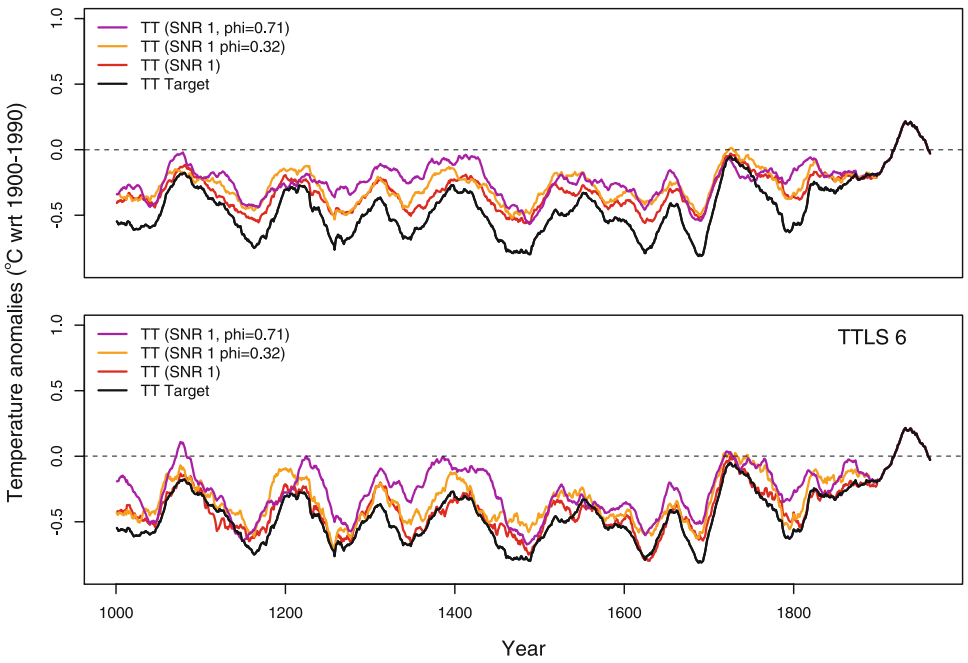

Fig. 5 European summer average temperatures anomalies (30-year running mean) for PC regression (*top*) and RegEM (*bottom*). The white noise scenario SNR 1 (*red line*) is compared with two different red noise scenarios (*orange and magenta lines*); the target is shown in *black*. TTLS indicates which truncation parameter is used

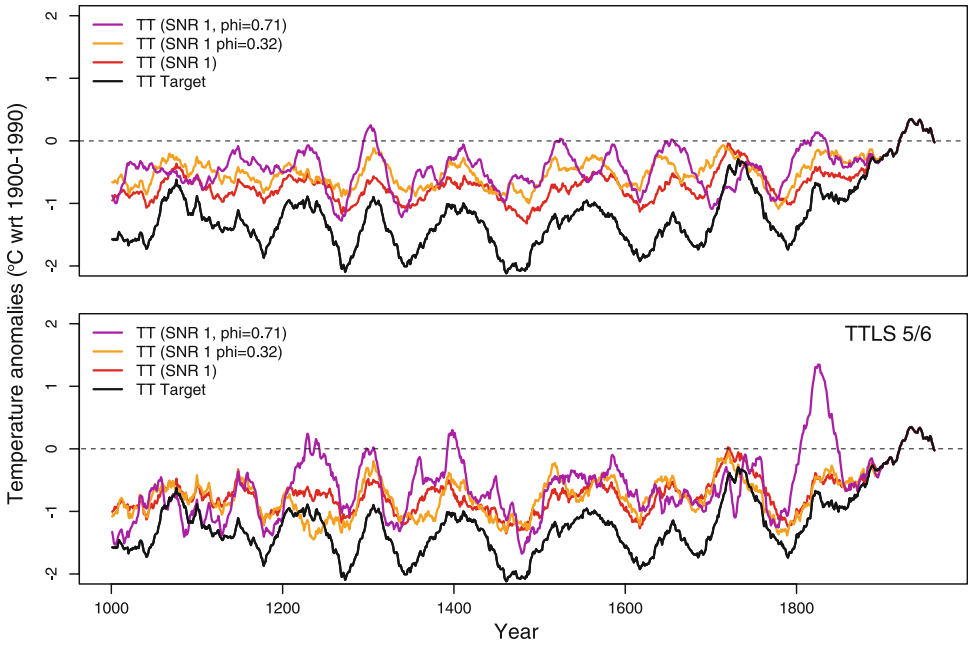

Fig. 6 As Figure 5, but for winter

more robust than PC regression with regard to the effect of noise added to the signal. Figures 5 and 6 (Suppl. Figs. 5 and 6), as well as Tables 1 and 2 display that reconstructions using red noise instead of white noise still retain skill. Nevertheless, the increase in variability in the results affects the reconstruction skill, more so in winter than in summer. Finally, there is a difference in reconstruction skill depending on variability frequency.

3.2 The spatial skill patterns of the reconstructions

Figures 7 and 8 (Suppl. Figs. 7 and 8) show the spatial skill patterns of the summer and winter reconstruction results from Figs. 3 and 4 (Suppl. Figs. 3 and 4) under the three different white noise scenarios, i.e. SNR ∞, SNR 1 and SNR 0.5. Since examining RE, the relation between the squared reconstruction error and the squared anomalies from the calibration average, is somewhat controversial (Buerger and Cubasch 2007), we have chosen to add a more intuitive skill measure, and also to look at the spatial differences of the two techniques, thus making it possible to directly determine the origins of the underestimation of the target temperature variations in the reconstruction results. Accordingly, the spatial skill is defined here as the differences between reconstructed and target temperature anomalies, i.e. the bias, averaged over the verification period, 1001–1899 AD, and the RE skill scores for the 30-year filtered results calculated for each gridpoint. This corresponds to a validation of the whole summer and winter temperature field. Positive bias values indicate that the difference between the average of reconstructed temperature anomalies over the verification period and the calibration period mean is smaller than that between target and calibration mean. Thus the target temperature anomalies are underestimated by the reconstructed anomalies, and overestimated for negative bias values. A lack of predictor model gridpoints (see Fig. 1) in the Atlantic leads to considerable uncertainties over that area both for summer and winter reconstructions (Figs. 7 and 8, suppl. Figs. 7 and 8). This effect is to be expected. However, the smaller the SNR, the larger the area with underestimation of target temperature anomalies becomes for summer and winter. Again, this is less pronounced for RegEM than for PC regression. Thus RegEM seems to be less dependent on the SNR than PC regression. The spatial skill patterns of RegEM are quite similar to those of PC regression. Nevertheless, for PC regression the underestimation of the target temperature variations during the verification period in the field is more clearly indicated. The spatial validation of the two techniques discloses the underestimation of amplitude seen for the European average temperatures in Figs. 3 and 4 (Suppl. Figs. 3 and 4) for PC regression. Focusing on the spatial RE scores for the 30-year filtered reconstruction results, we conclude that no large differences can be seen, despite the fact that RE skill scores are again higher for summer results than for winter.

4 Discussion

The results presented in this comparison of PC regression and RegEM reveal a seasonal dependence of reconstruction skill. Both techniques seem to perform more accurately (Figs. 3 and 5 compared to Figs. 4 and 6, likewise for

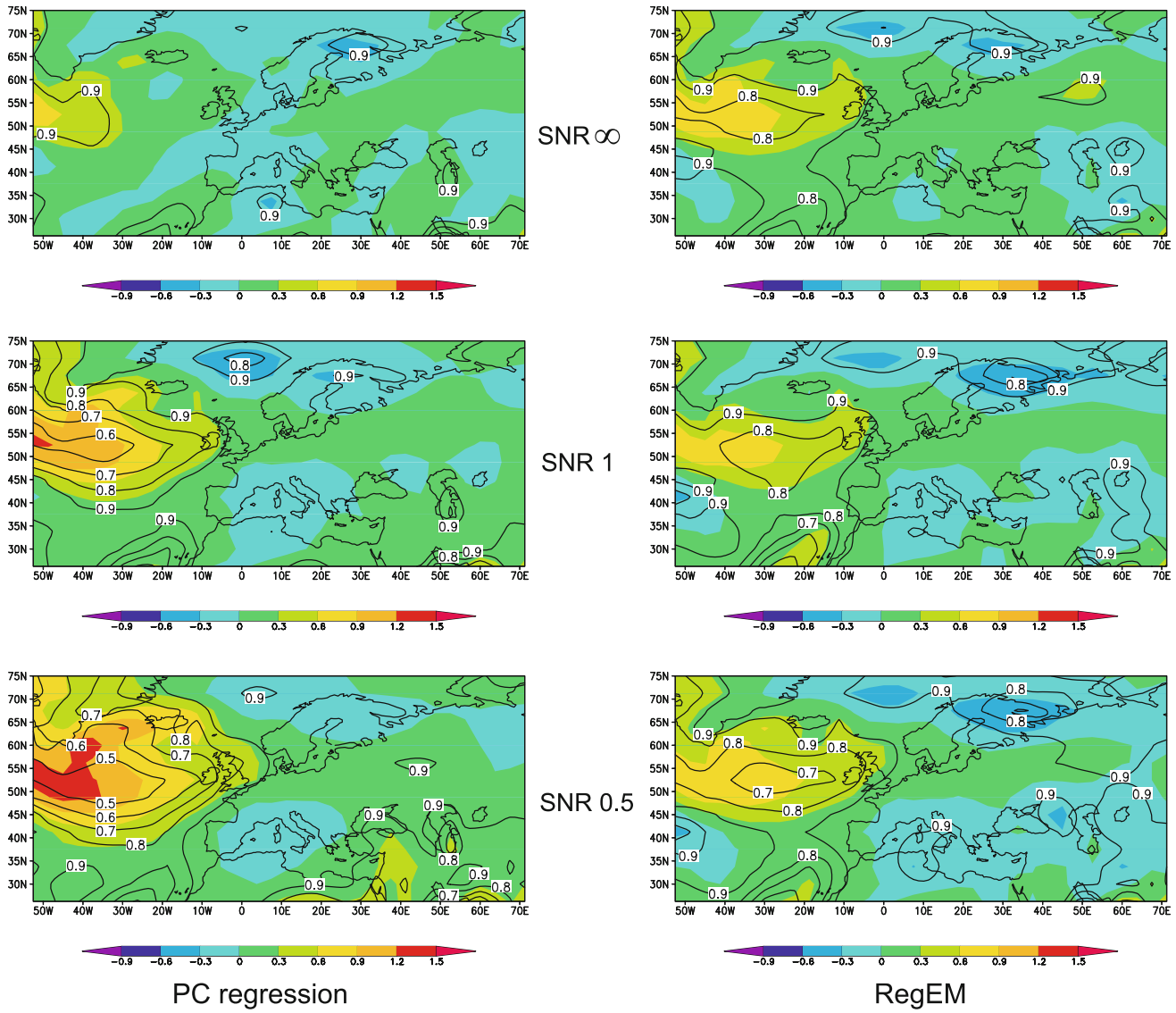

Fig. 7 Spatial skill patterns of the European summer temperature reconstructions using PC regression (*left*) and RegEM (*right*) with white noise scenarios SNR ∞, SNR 1, and SNR 0.5. The skill is defined by the average of the bias (reconstructed values—target values) (*shaded*) and RE (*contours*) calculated for each gridpoint over the verification period from 1001 to 1899 AD. The scale refers to the bias, i.e. differences in temperature anomalies for summer. *Colors* indicate reconstructed values that are about (*greenish blue* and *green*), higher (*light green*, *yellow* to *red*) or lower (*light blue* to *violet*) than the target values

ECHO-G 4 in the supplementary online material) for reconstructing summer average temperatures than winter when focusing on how well the low frequency variability of the target is captured. Testing the techniques with a less dense predictor network (12 gridpoints representing real world proxy series used to reconstruct the late Maunder Minimum (Kuettel et al. 2007), not shown) confirms these findings, although with an additional decrease in reconstruction skill. The more skilful performance in reconstructing European summer temperatures over the last millennium might be explained by the fact that the range of temperature variability is smaller in summer than in winter. Consequently, the impact of adding noise to the signal with smaller standard deviations in summer than in winter is less remarkable. Thus the reconstruction skill is less affected for summer than for winter. Furthermore, this is also potentially related to the spatial distribution of the predictor network used here. Predictor networks which may be optimal for reconstructing summer temperatures are not necessarily optimal for reconstructing winter temperatures (Pauling et al. 2003; Luterbacher et al. 2006; Kuettel et al. 2007).

The performance of reconstructions seems to depend less on the red structure of noise for the SNR 1 scenario with an autocorrelation coefficient $\rho = 0.32$ than with $\rho = 0.71$ (RE and CE 30-year filtered in Tables 1 and 2). For $\rho = 0.71$ the variability of the reconstruction results,

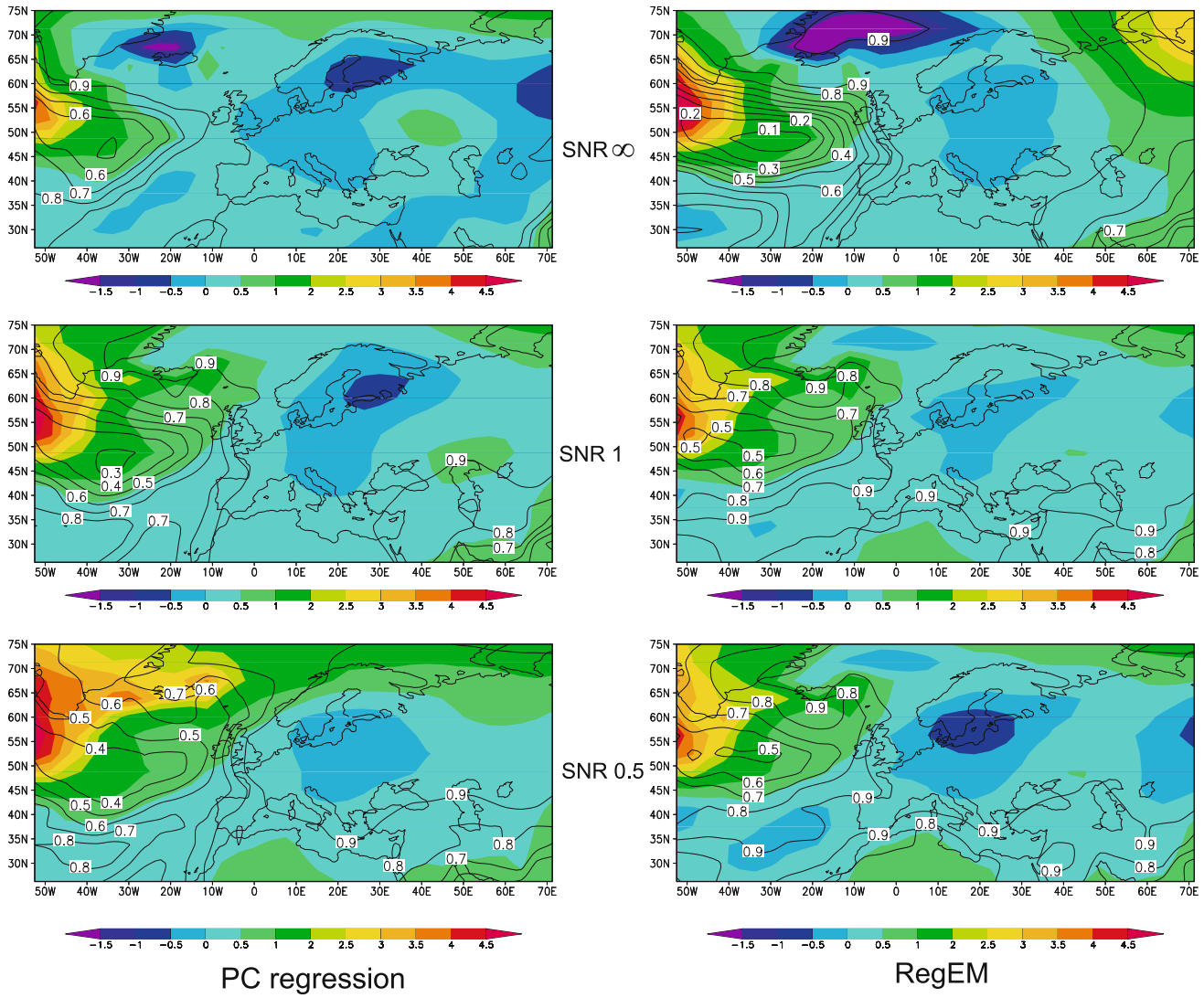

Fig. 8 As Figure 7, but for winter. Colors indicate reconstructed values that are about (*light blue* and *greenish blue*), higher (*light green* to *green*, *yellow*, *red*) or lower (*dark blue* to *violet*) than the target values

especially for winter, is considerably increased using RegEM (Fig. 6, suppl. Figure 6). Furthermore the skill of the reconstruction is generally more affected for the SNR 1 scenario with $\rho = 0.71$ than with $\rho = 0.32$ or for white noise only. However, analyses of typical red noise characteristics of realworld data (in Luterbacher et al. 2004) reveal, that $\rho = 0.71$ is not seen in the data and $\rho = 0.32$ is presumably more indicative of real world proxies. Still it is useful to study a range of autocorrelation coefficients to obtain an understanding of how reconstruction results depend on different types of noises. Nevertheless, the noise scenarios in this paper certainly do not mimic the full range of characteristics of noisy real world predictor series, once again indicating that there is a need to model predictor data and inherent uncertainties more realistically (Moberg et al. 2008). Tables 1 and 2 indicate that both techniques lose skill to an increasing degree as more noise is added to the signal. RegEM is less sensitive to and less affected by the noise addition than PC regression, but applying RegEM instead of PC regression in reconstructing, one of the fundamental statistical problems remains. Furthermore, while for the 30-year filtered data (Tables 1, 2) RE and CE skill scores confirm that RegEM performs more accurately than PC regression, the skill scores for the non-filtered reconstruction results are nevertheless lower for RegEM than for PC regression, especially in winter. One explanation might be, that using RegEM, mean and covariance of the whole input data matrix are iteratively estimated. The fact that the statistical characteristics of the whole input matrix are addressed together over the calibration and verification periods might be a reason for the less accurate inter-seasonal performance of RegEM, as

exhibited by the validation of the non-filtered results. Furthermore, considering the reconstruction results in Figs. 3 and 4 (top, likewise ECHO-G 4 in the supplementary online material) and the RE scores in Tables 1 and 2, it is an alarming sign that the PC regression results still achieve such high RE scores; moreover, the RE scores are put into the right perspective by the negative CE scores. The implication for reconstructions with real world proxy data is that verification has to be conducted very carefully by applying different means of validation. The interpretation of reconstruction skill and the reasonable verification of reconstruction results are delicate and not free from contradictions. Therefore, the development of alternative and more intuitive tools, as well as more thorough validation must be attempted (e.g. Wahl and Ammann 2007).

Why should it be the case that RegEM captures the target average 30-year filtered temperature variations more adequately than PC regression? When applying PC regression, we use OLS to estimate the regression coefficients for the calibration period. By contrast, when applying RegEM we use either the conditional Maximum likelihood method (if no regularization is needed) or TTLS (if the problem is ill-posed). These different estimation techniques, especially TTLS, which takes into account errors in the explanatory variables (Eq. 1, $\mathbf{e}_{\text{proxy}}$), have a crucial impact on the reconstruction skill. Another important difference is the nature of RegEM as an iterative process which is non-linear in general. Finally, RegEM not only provides estimates of the mean with each iteration step, but of the variance as well. We expected RegEM to be better than PC regression prior to this study, but we also expected it to be better than our results now indicate. One expectation for the less pronounced difference is that the reconstruction performance depends not only on the statistical technique chosen, but also on the choice and quality of the predictor network. Therefore, these other factors should be optimized, as well.

However, the use of RegEM also leaves room for future methodological improvements. Mann et al. (2007) recently addressed the problem of choosing truncation parameters. This was also investigated prior to applying RegEM here. The validation of a range of parameters, close to the one proposed by Mann et al. (2007), demonstrated that comparable results can be obtained by using alternative parameters (supplementary online material). We therefore urge the evaluation of several truncation parameters over the verification period.

Despite all this, we prefer RegEM to PC regression in this case, as it captures the multi-decadal variations of the target summer and winter European average temperatures more accurately (Figs. 3, 4, 5 and 6, likewise for ECHO-G 4 in the supplementary online material) than PC regression when focusing on lower frequency variability (RE and CE 30-year filtered in Tables 1 and 2).

5 Conclusions and perspectives

The outcomes regarding the performance of the two reconstruction techniques are restricted to the specific experimental setting used in this paper. As mentioned above the tests are based on NCAR CSM 1.4 and ECHO-G 4 climate model data, a predictand which consists of 476 gridpoints (land and sea), a pseudoproxy network with 30 gridpoints (Fig. 1), and scenarios based on different SNR constant over time. By comparing the two CFR techniques, -PC regression and RegEM,- at a continental and seasonal scale, we have demonstrated that the reconstruction skill differs according to the spatial and temporal scales the techniques are applied to. The fact that RegEM achieves different results for continental and hemispheric reconstructions (Mann et al. 2007) emphasizes the necessity of downscaling to smaller spatial and subannual temporal scales, in order to achieve a better understanding of the robustness and skill of the reconstruction techniques on higher temporal and spatial scales. Furthermore, hemispheric annual temperature reconstructions do not provide information about regional-scale variations, such as the intrinsic seasonal patterns of climate change as they have occurred, for instance, in Europe during past centuries (Mann et al. 2000; Luterbacher et al. 2004, 2007; Xoplaki et al. 2005). We found seasonal differences in the performance of RegEM and PC regression, and we demonstrated that predictor data quality has a crucial impact on reconstruction skill. RegEM has proved that more adequate results can be obtained by better incorporating the errors in the predictor data to reconstruct surface air temperature fields. However, the choice of the right TTLS parameters turned out to be ambiguous, and the procedure for selecting the most accurate ones needs further investigation. If no noise, or noise with a high SNR, is added to the signal, PC regression performs just as well as RegEM for winter and for summer. If noise with a smaller SNR is added to the climatic signal, the performance of RegEM proves to be more robust compared to PC regression. If the variability range is too large, as is the case e.g. for SNR 0.25 and SNR 1 with $\rho = 0.71$, both RegEM and PC regression exhibit deficits: the amplitude of target temperature variations tends to be underestimated by PC regression and overestimated by RegEM. However, overestimation might be adjusted by the choice of more suitable TTLS parameters.

The next step will be to quantify the differences between PC regression and RegEM by applying the two techniques to real world data, given a varying number of predictors and SNR over time. There is still a need and potential for further

optimizations of CFR techniques, such as RegEM, that take better account of errors. PC regression can still be optimized as well, e.g. by restriction to land areas only (Luterbacher et al. 2004; Xoplaki et al. 2005), optimization of PC truncation, or the implementation of different regression coefficient estimation procedures. Certainly other settings, and more realistic real world conditions have to be considered in future. On the one hand CFR techniques need to be better adapted to the specific character of the predictor data, and on the other, the quality of the predictor data has to be better understood, quantified and modeled. Exclusive use of classical multivariate statistics should be expanded to include solutions already developed in other research areas, e.g. econometrics. Time series analysis offers still further solutions, such as state space models and the use of Kalman filters (Lee et al. 2007), that are also worth exploring with regard to climate field reconstructions.

Acknowledgments This work has been supported by the Swiss National Science Foundation (SNSF) through its National Center of Competence in Research on Climate (NCCR Climate) project PALVAREX 2. Publication cost contributions are kindly provided by the Foundation Marchese Francesco Medici del Vascello. We thank Caspar Ammann and Fortunat Joos for making available the NCAR CSM 1.4 results, as well as Fidel Gonzalez Rouco and Eduardo Zorita for providing the ECHO-G 4 simulation results. We also thank Scott Rutherford, Eugene Wahl and Michael E. Mann for the RegEM code and for their helpful comments and inputs. Finally, we wish to thank the reviewers for their constructive criticism and suggestions, which helped to improve the quality of this study.

References

Ammann C, Joos F, Schimel D, Otto-Bliesner B, Tomas R (2007) Solar influence on climate during the past millennium: results from transient simulations with the NCAR Climate System Model. Proc Natl Acad Sci USA 104:3713–3718

Briffa K, Wigley T, Jones P, Pilcher J, Hughes M (1987) Patterns of tree-growth and related pressure variability in Europe. Dendrochronologia 5:35–59

Briffa K, Osborn T, Schweingruber F, Harris I, Jones P, Shiyatov S, Vaganov E (2001) Low-frequency temperature variations from a northern tree ring density network. J Geophys Res 106:2929–2942

Brohan P, Kennedy J, Harris I, Tett S, Jones P (2006) Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. J Geophys Res 111:D12106

Buerger G, Cubasch U (2005) Are multiproxy climate reconstructions robust? Geophys Res Lett 32:L23711

Buerger G, Cubasch U (2007) On the verification of climate reconstructions. Clim Past 3:397–409

Casty C, Handorf D, Sempf M (2005a) Combined winter climate regimes over the North Atlantic/European sector 1766–2000. Geophys Res Lett 32:L13801

Casty C, Wanner H, Luterbacher J, Esper J, Bohm R (2005b) Temperature and precipitation variability in the European Alps since 1500. Int J Climatol 25:1855–1880

Casty C, Raible C, Stocker T, Wanner H, Luterbacher J (2007) European climate pattern variability since 1766. Clim Dyn 29:791–805

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

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

Esper J, Wilson R, Frank D, Moberg A, Wanner H, Luterbacher J (2005) Climate: past ranges and future changes. Q Sci Rev 24:2164–2166

Esper J, Frank D, Buentgen U, Verstege A, Luterbacher J, Xoplaki E (2007) Long-term drought severity variations in Morocco. Geophys Res Lett 34:L17702

Fischer E, Luterbacher J, Zorita E, Tett S, Casty C, Wanner H (2007) European climate response to tropical volcanic eruptions over the last half millennium. Geophys Res Lett 34:L05707

González-Rouco J, Beltrami H, Zorita E, von Storch H (2006) Simulation and inversion of borehole temperature profiles in surrogate climates: Spatial distribution and surface coupling. Geophys Res Lett 33:L01703

Guiot J, Nicault A, Rathgeber C, Edouard J, Guibal F, Pichard G, Till C (2005) Last-millennium summer-temperature variations in western Europe based on proxy data. Holocene 15:489–500

Hegerl G, Crowley T, Hyde W, Frame D (2006) Climate sensitivity constrained by temperature reconstructions over the past seven centuries. Nature 440:1029–1032

Jansen E, coauthors (2007) Paleoclimate. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the International Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York

Jones P, Mann M (2004) Climate over past millennia. Rev Geophys 42:RG2002

Klimenko V, Klimanov V, Sirin A, Sleptsov A (2001) Climate changes in Western European Russia in the late Holocene. Doklady Earth Sci 377:190–194

Kuettel M, Luterbacher J, Zorita E, Xoplaki E, Riedeyl N, Wanner H (2007) Testing a European winter surface temperature reconstruction in a surrogate climate. Geophys Res Lett 34:L07710

Lee TCL, Zwiers FW, Tsao M (2007) Evaluation of proxy-based millennial reconstruction methods. Clim Dyn. doi:10.1007/s00382-007-0351-9

Li B, Nychka D, Ammann C (2007) The hockey stick and the 1990s: a statistical perspective on reconstructing hemispheric temperatures. Tellus 59A:591–598

Luterbacher J, Dietrich D, Xoplaki E, Grosjean M, Wanner H (2004) European seasonal and annual temperature variability, trends, and extremes since 1500. Science 303:1499–1503

Luterbacher J, et al. (2006) Mediterranean climate variability over the last centuries: a review. In: Lionello P, Malanotte-Rizzoli P, Boscolo R (eds) The Mediterranean Climate: an overview of the main characteristics and issues. Elsevier, Amsterdam, pp 27–148

Luterbacher L, Liniger M, Menzel A, Estrella N, Della-Marta P, Pfister C, Rutishauser T, Xoplaki E (2007) The exceptional European warmth of Autumn 2006 and Winter 2007: historical context, the underlying dynamics and its phenological impacts. Geophys Res Lett 34:L12704

Mangini A, Spötl C, Verdes P (2005) Reconstruction of temperature in the Central Alps during the past 2000 yr from a δ18O stalagmite record. Earth Planet Sci Lett 235:741–751

Mann M, Rutherford S (2002) Climate reconstruction using Pseudoproxies. Geophys Res Lett 29:1501

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

Mann M, Bradley R, Hughes M (1999) Northern hemisphere temperatures during the past millennium: inferences, uncertainties, and limitations. Geophys Res Lett 26:759–762

Mann M, Gille E, Bradley R, Hughes M, Overpeck J, Keimig F, Gross W (2000) Global temperature patterns in pastcenturies: an interactive presentation. Earth Interact 4(1234):1–29

Mann M, Rutherford S, Wahl E, Ammann C (2005) Testing the fidelity of methods used in proxy-based reconstructions of past climate. J Clim 18:4097–4107

Mann M, Rutherford S, Wahl E, Ammann C (2007) Robustness of proxy-based climate field reconstruction methods. J Geophys Res 112:D12109

Moberg A, Sonechkin D, Holmgren K, Datsenko N, Karlen W (2005) Highly variable Northern Hemisphere temperatures reconstructed from low-and high-resolution proxy data. Nature 433:613–617

Moberg A, Mohammad R, Mauritsen T (2008) Analysis of the Moberg et al. (2005) hemispheric temperature reconstruction. Clim Dyn (accepted)

Pauling A, Luterbacher J, Wanner H (2003) Evaluation of proxies for European and North Atlantic temperature field reconstructions. Geophys Res Lett 30:1787

Pauling A, Luterbacher J, Casty C, Wanner H (2006) Five hundred years of gridded high-resolution precipitation reconstructions over Europe and the connection to large-scale circulation. Clim Dyn 26(4):387–405

Proctor C, Baker A, Barnes W (2002) A three thousand year record of North Atlantic climate. Clim Dyn 19:449–454

Rutherford S, Mann M, Osborn T, Bradley R, Briffa K, Hughes M, Jones P (2005) Proxy-Based Northern Hemisphere surface temperature reconstructions: sensitivity to method, predictor network, target season, and target domain. J Clim 18:2308–2329

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

Shabalova M, van Engelen A (2003) aluation of a reconstruction of winter and summer temperatures in the low countries, AD 764–1998. Clim Change 58:219–242

Shindell D, Schmidt G, Mann M, Rind D, Waple A (2001) Solar forcing of regional climate change during the maunder minimum. Science 294:2149–2152

Shindell D, Schmidt G, Miller R, Mann M (2003) Volcanic and solar forcing of climate change during the preindustrial era. J Clim 16:4094–4107

Shindell D, Schmidt G, Mann M, Faluvegi G (2004) Dynamic winter climate response to large tropical volcanic eruptions since 1600. J Geophys Res 109:D05104

Smerdon J, Kaplan A (2007) Comments on testing the fidelity of methods used in proxy-based reconstructions of past climate: the role of the standardization interval. J Clim 20(22):5666–5670

Thejll P, Schmith T (2005) Limitations on regression analysis due to serially correlated residuals: application to climate reconstruction from proxies. J Geophys Res 110:D18103

von Storch H, Zorita E, Jones J, Dimitriev Y, Gonzalez-Rouco F, Tett S (2004) Reconstructing past climate from noisy data. Science 306:679–682

von Storch H, Zorita E, Jones J, Gonzalez-Rouco F, Tett S (2006) Response to comment on reconstructing past climate from noisy data. Science 312:529–529

von Storch H, Zorita E, Gonzalez-Rouco F (2008) Assessment of three temperature reconstruction methods in the virtual reality of a climate simulation. Int J Earth Sci (submitted)

Wahl E, Ammann C (2007) Robustness of the Mann, Bradley, Hughes reconstruction of the Northern hemisphere surface temperatures: examination of criticisms based on the nature and processing of proxy climate evidence. Clim Change 85:33–69

Waple A, Mann M, Bradley R (2002) Long-term patterns of solar irradiance forcing in model experiments and proxy based surface temperature reconstructions. Clim Dyn 18:563–578

Wilks D (1995) Statistical methods in the atmospheric sciences. Academic Press, San Diego

Xoplaki E, Luterbacher J, Paeth H, Dietrich D, Steiner N, Grosjean M, Wanner H (2005) European spring and autumn temperature variability and change of extremes over the last half millennium. Geophys Res Lett 32:L15713