Dendrochronologia xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Dendrochronologia

journal homepage: www.elsevier.de/dendro

Persistence matters: Estimation of the statistical significance of paleoclimatic reconstruction statistics from autocorrelated time series

Marc Macias-Fauria[a,b,*,1], Aslak Grinsted[c,1], Samuli Helama[d], Jari Holopainen[e]

[a] Biodiversity Institute, Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, Oxford OX1 3PS, United Kingdom
[b] Biogeoscience Institute, University of Calgary, Calgary, AB, Canada T2N 1N4
[c] Niels Bohr Institute, University of Copenhagen, Blegdamsvej, 172100 Copenhagen, Denmark
[d] Arctic Centre, University of Lapland, PL122, 96100 Rovaniemi, Finland
[e] Department of Geosciences and Geography, P.O. Box 64, 00014, University of Helsinki, Finland

ARTICLE INFO

Article history:
Received 7 July 2010
Accepted 4 August 2011

Keywords:
Paleoclimatology
Dendroclimatology
Verification
Calibration
Temporal autocorrelation

ABSTRACT

Proxy data forms natural time series used to lengthen instrumental climatic records, and may contain a significant portion of autocorrelation. Increased serial correlation limits the number of independent observations, not satisfying the assumptions of conventional statistical methods. We estimate the significance of calibration and verification statistics used in dendroclimatic reconstructions by combining Monte-Carlo iterations with frequency (Ebisuzaki) or time (Burg) domain time series modelling. Significance tests are presented for Coefficient of Determination (R^2), Coefficient of Correlation (r^2), Reduction of Error (RE) and Coefficient of Error (CE) for time series ranging from very low to very high autocorrelation. Increased autocorrelation implies higher occurrences of relatively high but spurious reconstruction statistics. Ebisuzaki time series modelling shows greater robustness and its use is recommended over Burg's method, which penalizes the restriction in the number of autocorrelation coefficients imposed by the Akaike Information Criterion. Positive RE and CE values, traditionally viewed as successful reconstruction statistics, are not necessarily significant and depend on the temporal structure of the time series used. This approach is further implemented successfully to compute confidence intervals based on the temporal structure of the residuals of the transfer function. A Matlab® package and a Windows executable file for non-Matlab® users are provided to perform the described analyses.

Introduction

Paleoclimatic research employs indirect estimates of past climate (proxy records) that are time series of physical or chemical properties of geological, glaciological and paleontological archives. Proxy series elongate the instrumental records retrospectively over the past centuries and millennia and thus strengthen our understanding about climate variability prior to any direct weather observation. Typical paleoclimatic proxies are tree-rings, corals, ice-cores, boreholes and various other kinds of sedimentary and fossil evidence (Bradley, 1999). Similarly to many other types of natural time series, proxy records are often highly autocorrelated. A principal constituent of this autocorrelation originates from climate, which shows persistence via fluctuations and trends (Karl, 1988; Trenberth, 1984). Additional constituents of the autocorrelation may originate from factors attached to each proxy. In tree-rings, for example, the physiology of trees serves as a basis for growth resources from previous years to be carried over a number of forthcoming years, resulting in tree-ring values that are depending on temporally adjacent values (Fritts, 1976). Alternatively, methodological aspects may require that the series become 'low-pass' filtered. This practice produces artificial autocorrelation to data. Filtering may be applied to improve the paleoclimatic reconstructions by timescale-dependent calibrations due to either physical limitations of the proxy at certain frequencies or frequency-dependent correlation between the proxy and climate (*e.g.* Guiot, 1985; Holopainen et al., 2009; Macias Fauria et al., 2010; Moberg et al., 2005; Osborn and Briffa, 2000; Rutherford et al., 2005; Timm et al., 2004). In any case, serially uncorrelated time series (*i.e.* white noise) do not serve for most climatic reconstruction purposes, being unable to capture climatic variations at timescales longer than the proxy data resolution (generally 1 year in dendrochronology). Tree-ring series standardization

* Corresponding author at: Biodiversity Institute, Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, Oxford OX1 3PS, United Kingdom. Tel.: +44 1865281878; fax: +44 01865271249.
E-mail address: marc.maciasfauria@zoo.ox.ac.uk (M. Macias-Fauria).
[1] Contributed equally to this work.

1125-7865/$ – see front matter
doi:10.1016/j.dendro.2011.08.003

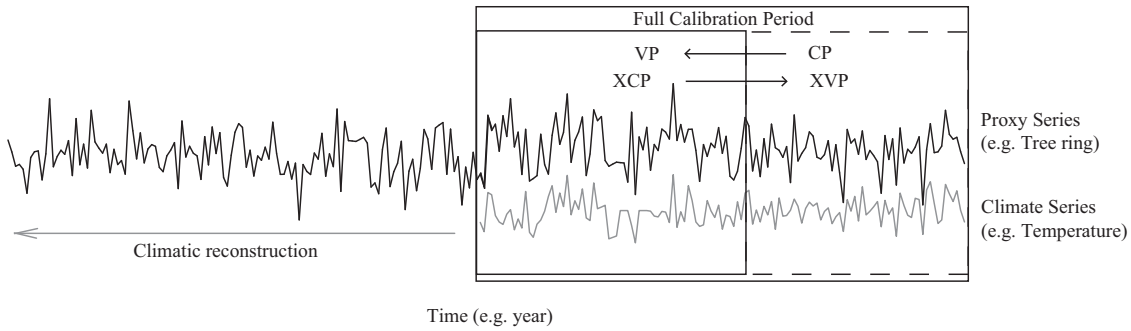

Fig. 1. Schematic representation of the general calibration/verification procedure prior to computing a dendroclimatic reconstruction. Full Calibration Period (large rectangle) defines the time over which proxy and instrumental data overlap, and the time which will optimally be used to compute the transfer function. Such period is divided into two sub-periods in order to test the performance of the relationship between proxy and instrumental data with independent data. Thus, the proxy/instrumental data regression computed over the Calibration Period (CP, dashed-edged rectangle) is used to predict (arrow) climate over the Verification Period (VP, smaller continuous-edged rectangle), and is thus validated against instrumental data. During Cross-Calibration–Verification, the same procedure is applied but the original Verification Period becomes the Cross-Calibration Period (XCP), which is used to predict (arrow) climate over the Cross-Verification Period (XVP), and is again validated against instrumental data. Once the validation is performed, a transfer function is computed over the Full Calibration Period which is applied to the proxy data beyond the instrumental record to reconstruct past climate (grey arrow).

procedures currently used in climate reconstructions, such as Regional Curve Standardization (RCS, Erlandsson, 1936) or the signal-free approach (Melvin and Briffa, 2008), seek to retain the largest possible climate-related autocorrelation.

An important caveat in this respect is that autocorrelation may complicate any interpretation of statistical analyses. The high incidence of spurious associations between highly autocorrelated time series was recognized already by early statisticians (Bartlett, 1935; Quenouille, 1952; Yule, 1926). Moreover, conventional methods to estimate statistical significance (*e.g.* Henkel, 1976) may be analytically intractable as their assumptions are not satisfied. More flexible bootstrap and Monte-Carlo techniques (Efron and Tibshirani, 1986) can often perform better than classical methods with special regard to climatic (von Storch and Zwiers, 1999) or paleoclimatic (*e.g.* Mudelsee, 2003; Young et al., 2000) persistence.

This paper aims to quantify the effect of autocorrelation in the statistics commonly used to assess the validity of dendroclimatic reconstructions. We provide tests of significance for various statistics typically used in paleoclimatic calibration–verification procedures, namely Coefficient of Determination (R^2), Coefficient of Correlation (r^2), Reduction of Error (RE), and Coefficient of Efficiency (CE). A combination of Monte-Carlo iterations and time series modelling in the frequency (Ebisuzaki, 1997) or temporal domains (Burg, 1978) is used to simulate the influence of autocorrelation in the statistical associations between the independent (proxy) and the dependent (observed climate) variables in the calibration/verification procedures. In addition to overcoming the problems related to autocorrelation, to our knowledge a methodology to perform tests of statistical significance for the calibration–verification statistics commonly used in various types of paleoclimatic studies (Briffa et al., 1988; Cook et al., 1994; Fritts, 1976; Woodhouse, 1999) is provided for the first time. Finally, the same approach is used to produce confidence intervals of the climatic reconstruction, based on the temporal structure of the residuals of the transfer function. Software is provided to perform the described analyses.

Calibration and verification statistics

A paleoclimatic reconstruction is based on the transfer function (Fritts, 1976), a model of the relationship between the proxy and the instrumental data, applied to the proxy data for the reconstruction period. In most cases, this involves a simple or multiple linear regression analysis, which is the focus of our study. The time span over which such relationship is modelled depends on the period of common overlap between proxy and instrumental data, and is called Full Calibration Period (Fig. 1). The quality of a paleoclimatic calibration model is commonly tested using the Coefficient of Determination (R^2), which is defined as the squared correlation between the model and the predictand over the Full Calibration Period. However, the validity of the calibration should be tested using independent data not used in the training process. Hence, the Full Calibration Period, which will ultimately be used to compute the transfer function, is typically divided into two sub-periods, the Calibration and Verification Periods (CP and VP, respectively; Fig. 1). The model parameters are optimized over CP and the predictive skill is tested over VP. Verification of the model using withheld data is of special importance, as the comparison between the calibration and verification statistics may reveal potential over-fit of the calibration model or decreased sensitivity to climatic variations over the non-calibrated period.

A number of tests (verification statistics) are performed that aim to check how well the values predicted by the calibration model fit the observed values in the VP. Commonly these are Squared Pearson correlation (r^2), Reduction of Error (RE), and Coefficient of Error (CE) (Briffa et al., 1988; Cook et al., 1994; Fritts, 1976; Woodhouse, 1999). Pearson product-moment correlation coefficient between two series x and y consists of:

$$r_{xy} = \frac{\sum_{t=1}^{N}(x_t - \bar{x})(y_t - \bar{y})}{(n-1)S_xS_y} \quad (1)$$

where N is the length of the time series, $\bar{x}$ and $\bar{y}$ are the mean value of each time series, and s_x and s_y are their standard deviations. RE and CE are defined as follows:

$$\mathrm{RE} = 1.00 - \frac{\sum_{t=1}^{N}(x_t - \hat{x}_t)^2}{\sum_{t=1}^{N}(x_t - \bar{x}_c)^2} \quad (2)$$

and

$$\mathrm{CE} = 1.00 - \frac{\sum_{t=1}^{N}(x_t - \hat{x}_t)^2}{\sum_{t=1}^{N}(x_t - \bar{x}_v)^2} \quad (3)$$

where x_t is the observed climate in year t, $\hat{x}_t$ is the predicted (reconstructed) climate in year t, and $\bar{x}_c$ and $\bar{x}_v$ are the arithmetic means of the actual climate over CP and VP, respectively. N is the number of years in VP. In order to seek for further temporal stability of the relationship between climatic and proxy data, this process can be re-done by considering the old VP as a new CP, and the old CP as a

Please cite this article in press as: Macias-Fauria, M., et al., Persistence matters: Estimation of the statistical significance of paleoclimatic reconstruction statistics from autocorrelated time series. Dendrochronologia (2012), doi:10.1016/j.dendro.2011.08.003

new VP: such a procedure is called Cross-Calibration–Verification (Fig. 1).

RE and CE are used to estimate the strength of the linear relationship between the observed and reconstructed series. Their values range from 1 to $-\infty$, where the maximum value of 1indicates perfect fit between observed and reconstructed time series. Importantly, RE > 0 indicates that the reconstruction is better than the CP average. CE differs from RE only in that $\bar{x}_c$ in Eq. (2) is replaced by the mean of VP ($\bar{x}_v$) in Eq. (3). CE is thus a more restrictive verification statistic than RE, as it is a true measure of the variance in common between the real and the estimated data over VP (Briffa et al., 1988). Conventionally, >0 outcomes for RE and CE have been interpreted to mean that the reconstruction bears predictive skill and thus considered as acceptable without any further tests of statistical significance (*e.g.* Briffa et al., 1988; Cook et al., 1994; Fritts, 1976; Woodhouse, 1999). However, RE and CE should be interpreted cautiously if data contains high autocorrelation or trends (Cook et al., 1994).

Methods

This section describes the methodology employed to obtain tests on the significance of the reconstruction statistics described above. A Monte-Carlo-based test of significance consists of the generation of surrogate (virtual) data in order to be able to produce an empirical probability density function (PDF) for a given statistic, and analysing where the value of the statistic we are interested in lays within such distribution, hence, knowing its significance.

Selection of a model able to generate appropriate surrogate time series

The first step is to select a model able to reproduce surrogate data of the same characteristics as the original data. In our case, a model has to be chosen able to generate time series with the same autocorrelation structure as the original normalized data. This study used two approaches in this respect: a frequency-domain and a time-domain model.

- The frequency-domain method (Ebisuzaki, 1997) aims to generate a number of random time series that keep the same power spectrum as the original time series but with a random phase. It basically consists of three steps: first, the discrete Fourier transform is computed for the time series; second, a Fourier series with random phases and the same power spectrum as the original series is calculated; third, new synthetic series are obtained by the inverse Fourier transform.
- The time-domain method (Burg, 1978) is based on the application of autoregressive (AR) models to the time series in order to simulate their autocorrelation structure. AR coefficients are directly computed from the data by estimating the reflection coefficients (partial autocorrelations) at successive orders. The model's AR coefficients are the harmonic mean of the forward and backward partial autocorrelation estimates. Burg's method minimizes the forward and backward (least squares) prediction errors. In this case, AR models will be computed for as many orders as we define, and thus, a best model (*i.e.* a best model order) for each time series must be selected. To do so we used the Akaike Information Criterion (AIC; Akaike, 1974). AIC trades off the complexity of an estimated model (number of parameters) against how well the model fits the data. The preferred model is the one with the lowest AIC value. In our study we used 30 as the default maximum model order, and we found in no case the need for a higher model order.

In both approaches, the employed noise-generating algorithms assume normal distribution in the data (*i.e.* surrogate time series were calculated using normally distributed noise). Thus, the Jarque–Bera statistic (JB statistic; Jarque and Bera, 1987) for normally distributed data was calculated for each time series (original and surrogate). JB statistic is based on the sample Skewness and Kurtosis of the time series.

Monte-Carlo iterations

The chosen models were used to generate a number of surrogate time series (autocorrelated noise series) with the same autocorrelation as the original data. The statistics for the calibration/verification periods were then computed for each group of surrogate time series. That is, in the case of 1000 Monte-Carlo iterations, 1000 different R^2 are calculated by default for the full periods, as well as 1000 r^2 for the calibration periods, and 1000 RE, CE, and r^2 for the verification periods.

Empirical probability density functions

The empirical PDF of each statistic can then be calculated and, hence, its significance for the single-tailed distribution. In the case of 1000 Monte-Carlo iterations, the maximum degree of significance that can be given is for $p < 0.001$. Minimum p-values will depend on the number of Monte-Carlo iterations chosen (which will considerably change the computing time requirements), following:

$$Minimum \quad p\text{-}value < \left(\frac{1}{mccount}\right) \tag{4}$$

where *mccount* is the number of Monte-Carlo iterations.

Reconstruction and generation of confidence intervals

Using the same approach described above, the temporal structure of the residuals of the reconstruction regression (*i.e.* transfer function) can be used to calculate confidence intervals for the whole reconstructed period. The autocorrelation structure of the residuals can be computed by either Burg's or Ebizusaki's approach, and surrogate residuals obtained. When added to the regression estimates, they create intervals around the reconstruction which depict its expected error.

Empirical time series

Dataset

A tree-ring width chronology made of living and sub-fossil *Pinus sylvestris* (L.) from the forest limit of Finland and Norway (68–70°N, 20–30°E; period 752–1998 AD) was used to assess the performance of the methods explained in the previous section (Fig. 2b; Helama et al., 2009). The chronology was computed using the Regional Curve Standardization procedure (*e.g.* Briffa et al., 1992, 1996; Erlandsson, 1936; Fritts, 1976) and calibrated against Karasjok (northern Norway, 69°28′N, 25°31′E) July mean temperature for the period 1876–1998.The corresponding transfer function for the normalized data is

$$T_t = -0.19x_{t-2} - 0.15x_{t-1} + 0.914x_t - 0.1x_{t+1} + 0.142x_{t+2} \tag{5}$$

where T_t is July Karasjok temperature in year t and x_t is tree-ring index in year t.

As it is seen in the transfer function, predictors are clearly not independent (*i.e.* they consist of a combination of the same time

Please cite this article in press as: Macias-Fauria, M., et al., Persistence matters: Estimation of the statistical significance of paleoclimatic reconstruction statistics from autocorrelated time series. Dendrochronologia (2012), doi:10.1016/j.dendro.2011.08.003

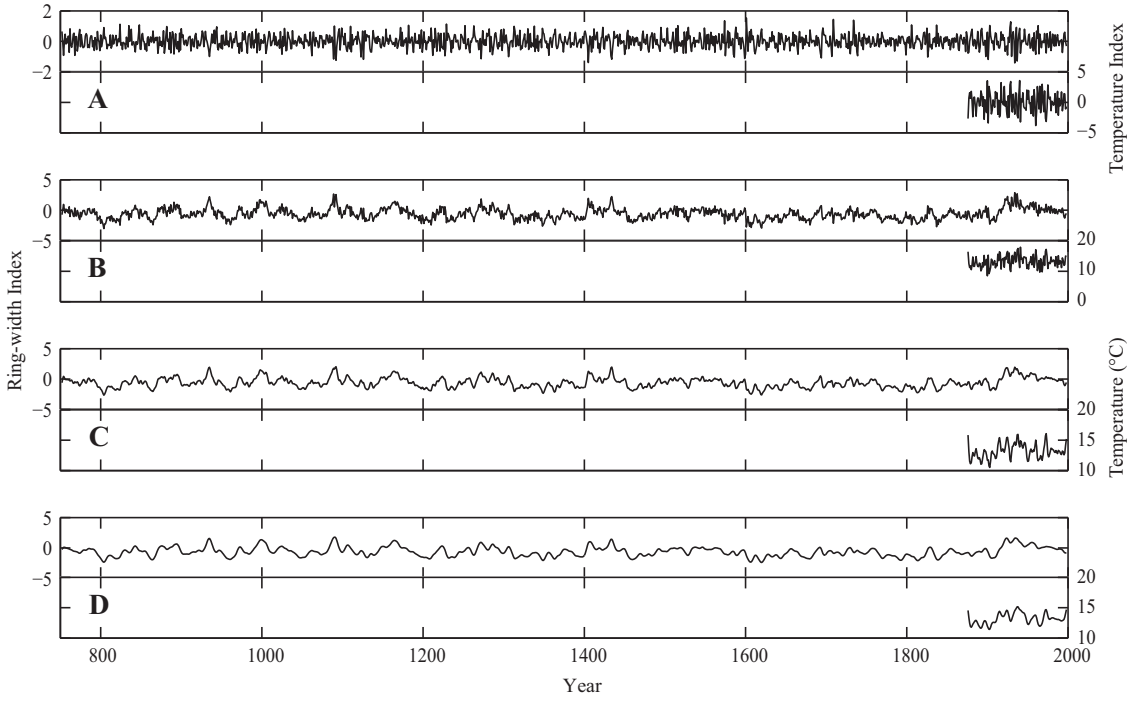

Fig. 2. Northern Lapland ring-width chronology (as in Helama et al., 2009) for the period 752–1998 (upper) *vs.* summer temperature for Karasjok (northern Norway, 69°28′N, 25°31′E) for the period 1876–1998 (lower). (A) Residual (white noise) time series; (B) standard time series (unfiltered in temperature); (C) 5-year spline time series; (D) 10-year spline time series. Note the increasing smoother lines from A to D, reflecting increasing temporal autocorrelation.

series with different lags). This is the case in many multiple regressions to some degree. Dependence between predictors tends to give conservative significance estimates, as in practice the data has fewer degrees of freedom than what we are testing against. Thus, it does not represent a concern.

Reconstruction statistics as a function of persistence in the time series

In order to assess the influence of increasing autocorrelation in calibration/verification statistics, 5 and 10-year cubic spline smoothing functions (Cook and Peters, 1981) were fit to the original (standard) time series. Further, residuals from a 15-year spline fit were taken to obtain largely non-autocorrelated time series (Fig. 2). Fig. 3 displays correlograms of the proxy chronology and of its filtered counterparts. As expected, increased autocorrelation coefficients appear with longer spline filters, affecting several autocorrelation orders.

100,000 Monte-Carlo iterations were performed in order to maximize the robustness of our results and to generate enough samples as to be able to clearly visualize the structure and shape of the resulting empirical PDFs for each of the analysed statistics.

Results and discussion

Empirical probability density functions

We discuss in this section the sensitivity of our methods to increased autocorrelation in the time series, as well as the differences in the performance of Burg's and Ebisuzaki's approaches. Three statistics will be discussed, namely: RE, CE, and R^2. The Coefficient of Determination (R^2) is the square of the correlation coefficient between the constructed predictor and the response variable: the analysis of the PDF of this statistic is thus analogous to the analyses of the Pearson correlation statistic. Note that for R^2 all values are positive.

The PDFs (*i.e.* the distribution of values of statistics resulting from the Monte-Carlo iterations) from which the statistics' significances were computed were very sensitive to the degree of autocorrelation in the time series (Fig. 4 and Table 1).

- RE and CE: the range of PDF values expanded with increasing autocorrelation in the series, especially to the left (negative values), but also and importantly for assessing the quality of dendroclimatic reconstructions, towards positive values (Table 1). This occurred together with a strong decrease in the *peakedness* of the distribution (*i.e.* a less distinct and sharp mode); the

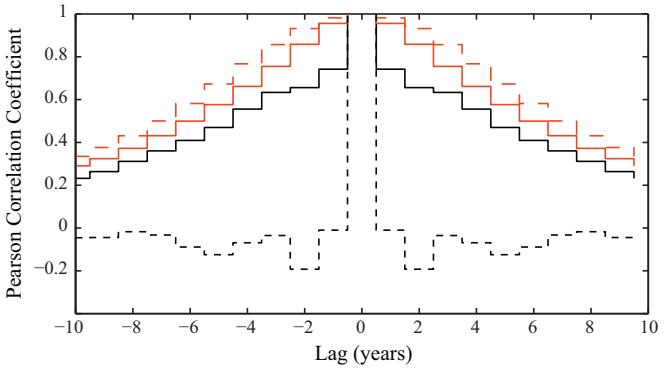

Fig. 3. Correlograms of the Northern Lapland *P. sylvestris* ring width chronology (Helama et al., 2009) for: black dashed line: residual chronology; black continuous line: standard chronology; red continuous line: 5-year spline chronology; red dashed line: 10-year spline chronology. Autocorrelations are shown for up to 10-year lags, forward and backward. At lag zero, all correlations are 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: Macias-Fauria, M., et al., Persistence matters: Estimation of the statistical significance of paleoclimatic reconstruction statistics from autocorrelated time series. Dendrochronologia (2012), doi:10.1016/j.dendro.2011.08.003

Table 1
Statistics of the empirical density distributions for a set of reconstruction statistics, namely: RE: Reduction of Error; CE: Coefficient of Efficiency; R^2: squared Pearson Correlation Coefficient computed for the full calibration period (*i.e.* 1876–1998). Distributions obtained after 100,000 Monte-Carlo iterations. Results are shown for Res: residual (white noise); Std: standard; Spl 5-yr: 5-year spline; and Spl 10-yr: 10-year spline series. A (upper block): time series modelled using Ebisuzaki's approach; B (lower block): time series modelled using Burg's approach. Descriptor statistics of the distributions are: x_{min}: minimum value; x_{max}: maximum value; Δx: range; $\bar{x}$: arithmetic mean; Md: mode; $\tilde{x}$: median; s^2: variance; s: standard deviation; Υ_1: Skewness. Data source: northern Lapland *P. sylvestris* ring-width chronology *vs.* summer temperature for Karasjok (northern Norway, 69°28′N, 25°31′E).

	x_{min}	x_{max}	Δx	$\bar{x}$	Md	$\tilde{x}$	s^2	s	Υ_1
A									
RE									
Res	−2.95	0.43	3.38	−0.12	−0.05	−0.09	0.03	0.18	−1.49
Std	−5.33	0.38	5.71	−0.22	−0.04	−0.13	0.11	0.34	−2.67
Spl 5-yr	−35.47	0.67	36.13	−0.78	−0.09	−0.43	1.50	1.22	−4.16
Spl 10-yr	−105.33	0.85	106.18	−1.82	−0.06	−0.86	10.76	3.28	−6.04
CE									
Res	−2.96	0.43	3.39	−0.12	−0.05	−0.09	0.03	0.18	−1.50
Std	−6.44	0.28	6.71	−0.34	−0.12	−0.23	0.15	0.38	−2.73
Spl 5-yr	−45.73	0.52	46.25	−1.24	−0.29	−0.78	2.46	1.57	−4.29
Spl 10-yr	−221.98	0.71	222.69	−3.18	−0.32	−1.76	22.85	4.78	−6.04
R^2									
Res	0.00	0.42	0.42	0.09	0.07	0.09	0.00	0.05	0.85
Std	0.00	0.29	0.29	0.08	0.06	0.07	0.00	0.04	0.68
Spl 5-yr	0.00	0.51	0.50	0.19	0.18	0.19	0.00	0.07	0.26
Spl 10-yr	0.01	0.77	0.76	0.35	0.36	0.35	0.01	0.11	−0.04
B									
RE									
Res	−3.32	0.33	3.65	−0.17	−0.08	−0.13	0.04	0.20	−1.73
Std	−5.82	0.39	6.21	−0.19	−0.07	−0.12	0.06	0.24	−3.32
Spl 5-yr	−76.95	0.68	77.62	−0.85	−0.09	−0.48	1.67	1.29	−6.43
Spl 10-yr	−230.56	0.88	231.44	−2.16	−0.19	−1.06	14.90	3.86	−8.79
CE									
Res	−3.32	0.33	3.65	−0.18	−0.08	−0.13	0.04	0.20	−1.73
Std	−5.94	0.28	6.22	−0.23	−0.10	−0.16	0.06	0.25	−3.32
Spl 5-yr	−86.23	0.52	86.74	−1.12	−0.25	−0.68	2.32	1.52	−6.20
Spl 10-yr	−239.47	0.86	240.33	−3.15	−0.41	−1.69	25.18	5.02	−7.53
R^2									
Res	0.00	0.38	0.38	0.07	0.05	0.07	0.00	0.04	1.05
Std	0.00	0.25	0.25	0.04	0.03	0.04	0.00	0.03	1.12
Spl 5-yr	0.00	0.56	0.56	0.14	0.09	0.12	0.01	0.08	0.86
Spl 10-yr	0.00	0.84	0.84	0.25	0.18	0.24	0.02	0.13	0.54

Skewness of the distributions largely shifted towards more negative values. Logically, these transformations were accompanied by increases in the variances of the PDFs. The shift of the mode and median towards negative values slightly compensated the increased spurious high values (longer and thicker distribution tails) in the determination of significance thresholds, which was insufficient in the RE statistic, very sensitive to autocorrelation.

Table 2 shows the significances at $p < 0.05$ and $p < 0.01$ of the studied statistics. Whereas threshold values for RE did not change for standard and residual time series, they increased with autocorrelation, being far from the traditional threshold value of 0. In the case of CE, the overall shift towards negative values of the whole distribution might have compensated the more dispersed distribution and thus the larger occurrence of spurious values.

Table 2
Subset of reconstruction statistics and their 95% and 99% significance thresholds, namely: R^2: squared Pearson Correlation Coefficient computed for the full calibration period (*i.e.* 1876–1998); RE: Reduction of Error; CE: Coefficient of Efficiency. Results are shown for Res: residual (white noise); Std: standard; Spl 5-yr: 5-year spline; and Spl 10-yr: 10-year spline series. A (left block): time series modelled using Ebisuzaki's approach; B (right block): time series modelled using Burg's approach. Note the overall increase in significance threshold values with increasing autocorrelation in the series. Significance thresholds computed after 100,000 iterations. Data source: northern Lapland *P. sylvestris* ring-width chronology *vs.* summer temperature for Karasjok (northern Norway, 69°28′N, 25°31′E).

	A				**B**			
	Res	Std	Spl 5-yr	Spl 10-yr	Res	Std	Spl 5-yr	Spl 10-yr
R^2	0.40	0.40	0.38	0.45	0.40	0.40	0.38	0.45
R^2 95%	0.19	0.14	0.31	0.53	0.16	0.09	0.28	0.48
R^2 99%	0.24	0.18	0.37	0.59	0.21	0.12	0.35	0.58
RE	0.28	0.26	0.03	−0.06	0.28	0.26	0.03	−0.06
RE 95%	0.11	0.11	0.22	0.33	0.06	0.04	0.12	0.20
RE 99%	0.18	0.18	0.36	0.52	0.14	0.09	0.29	0.45
CE	0.28	0.24	−0.07	−0.26	0.28	0.24	−0.07	−0.26
CE 95%	0.11	0.03	0.04	0.05	0.06	0.01	0.00	−0.04
CE 99%	0.18	0.10	0.20	0.33	0.14	0.05	0.14	0.23
XRE	0.35	0.38	0.34	−0.19	0.35	0.38	0.34	−0.19
XRE 95%	0.11	0.11	0.22	0.33	0.06	0.03	0.12	0.19
XRE 99%	0.19	0.18	0.35	0.52	0.14	0.09	0.28	0.44
XCE	0.35	0.36	0.30	−0.33	0.35	0.36	0.30	−0.33
XCE 95%	0.11	0.03	0.04	0.05	0.06	0.00	−0.01	−0.04
XCE 99%	0.19	0.10	0.19	0.32	0.14	0.05	0.14	0.22

Please cite this article in press as: Macias-Fauria, M., et al., Persistence matters: Estimation of the statistical significance of paleoclimatic reconstruction statistics from autocorrelated time series. Dendrochronologia (2012), doi:10.1016/j.dendro.2011.08.003

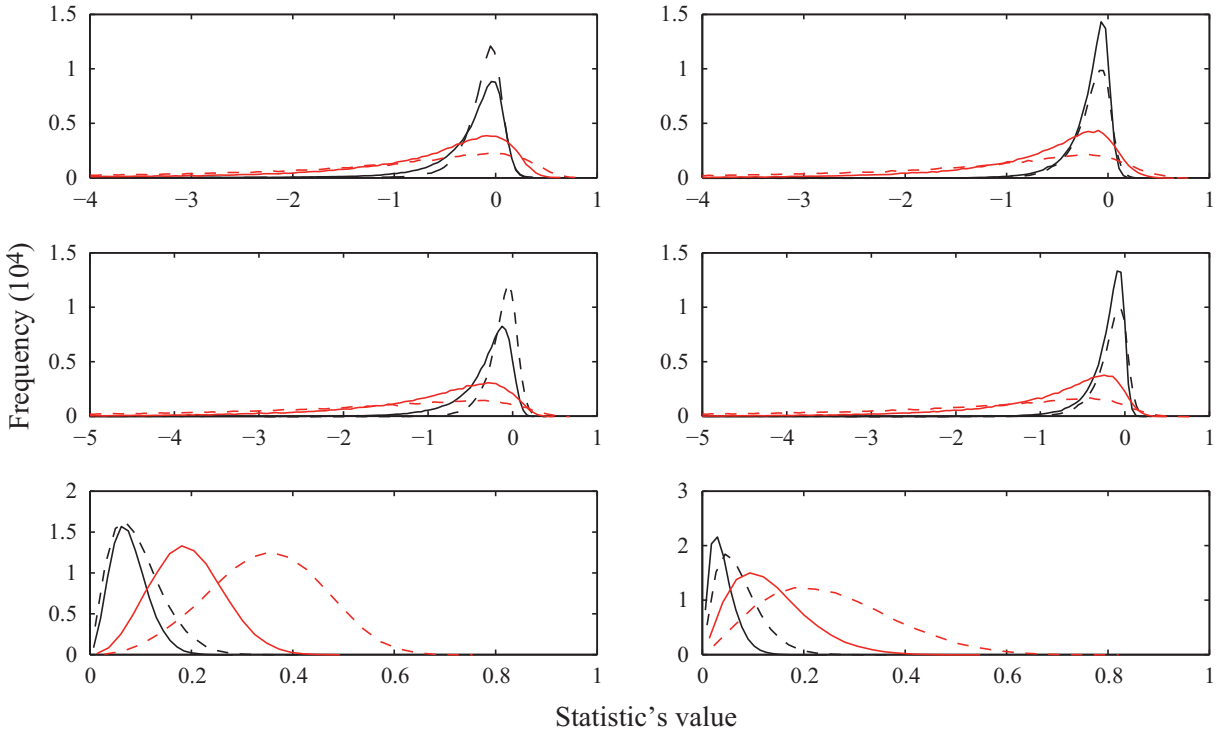

Fig. 4. Empirical probability density functions resulting from 100,000 Monte Carlo iterations for Reduction of Error (upper row), Coefficient of Efficiency (middle row) and squared Pearson Correlation Coefficient for the Full Calibration period (1876–1998 AD; lower row). Source data: northern Lapland *P. sylvestris* ring-width chronology *vs.* summer temperature for Karasjok (northern Norway, 69°28′N, 25°31′E). Left and right columns show the distributions resulting from modelling the autocorrelation structure of the original series using Ebisuzaki's and Burg's approaches, respectively. Black dashed line: residual series; black continuous line: standard series; red continuous line: 5-year spline series; red dashed line: 10-year spline series. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Threshold values remained close to 0 at all autocorrelation levels for $p < 0.05$. Significance at $p < 0.01$ was much more sensitive to increased autocorrelation, with threshold values jumping to 0.33 (Ebisuzaki's) or 0.23 (Burg's) for the 10-year spline dataset. However, our data revealed a strong tendency towards smaller and negative CE values with smoother filters, and thus CE values were not significant with the 5 and 10-year spline time series.

- R^2: the range of the PDFs also increased. R^2 PDFs were very sensitive to increased autocorrelation: the distributions' mean, median, and mode shifted towards larger values, with the consequent increase of spurious high Coefficients of Correlation (Table 1 and Fig. 4). This resulted in higher measured R^2 values with the original data (0.4–0.45), but even higher significance thresholds at both $p < 0.05$ and $p < 0.01$, such that when using 10-year splines the relationships between tree-ring index and temperature were not significant.

Ebisuzaki's vs. Burg's approaches

Statistics' PDFs and significance thresholds changed in the same qualitative way with increasing time series autocorrelation for both Ebisuzaki's and Burg's method, in agreement with the notion of a decrease in the number of independent observations with increasing serial correlation. However, Ebisuzaki's approach offered a higher overall performance. Although surrogate time series appeared to be visually similar and successfully created using both methods (see Supplementary Figs. 1 and 2 for characteristic surrogate time series), a closer inspection of their autocorrelation structure as compared to the original data revealed a much higher performance of Ebisuzaki's method (Fig. 5): Burg's method consistently underestimated autocorrelation levels, especially at higher orders, also creating models with a large degree of negative autocorrelation when modelling residual time series. This explains the higher significance thresholds found in the Ebisuzaki method with highly autocorrelated data (Table 2) and the more sensible results of Ebisuzaki's method with non autocorrelated time series (Fig. 4).

That is, Ebisuzaki's approach generated surrogate time series which mimicked better the structure of the original input data than Burg's. This might be a consequence of the procedure chosen in order to select the best model order in the Burg's method. AIC trades off parsimony and goodness of fit. That is, it penalizes over parametrization. This might result in the selection of lower model orders than needed in order to have the best possible fit to the data, and thus in the ultimate generation of surrogate data with different persistence than the original one (Fig. 5).

Finally, Ebisuzaki's model also offers faster computation times than Burg's method, which might be important when computing a large number of Monte Carlo iterations (*i.e.* >10,000) but insignificant if using a smaller but already reliable number of iterations (*e.g.* 1000, that is, significance resolution at $p < 0.001$).

Overall, short time series and high persistence have a parallel effect on the statistics' values and significances: they both decrease the amount of independent observations in the data, increasing the chance of spurious high statistic values.

Reconstruction confidence intervals

Fig. 6 shows the reconstruction of July Karasjok temperature using the northern Lapland *P. sylvestris* ring-width chronology (Helama et al., 2009), and its 95% confidence intervals based on the structure of the residuals of the transfer function computed using Ebizusaki's method (Burg's method results are shown in

Please cite this article in press as: Macias-Fauria, M., et al., Persistence matters: Estimation of the statistical significance of paleoclimatic reconstruction statistics from autocorrelated time series. Dendrochronologia (2012), doi:10.1016/j.dendro.2011.08.003

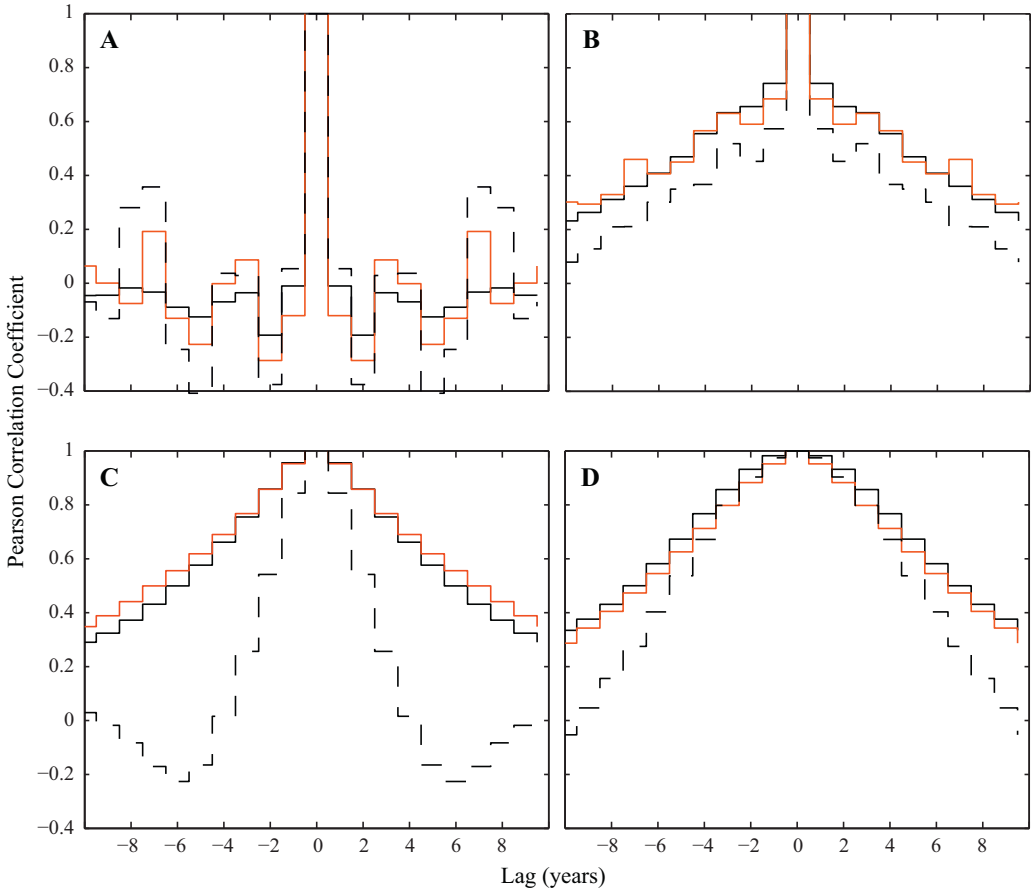

Fig. 5. Correlograms of the Northern Lapland *P. sylvestris* ring width chronology (black continuous line) and surrogate time series modelled following Ebisuzaki's (red continuous line) and Burg's (black dashed line) approaches. (A) residual; (B) standard; (C) 5-year spline; (D) 10-year spline time series. Autocorrelations are shown for up to 10-year lags, forward and backward. At lag zero, all correlations are 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

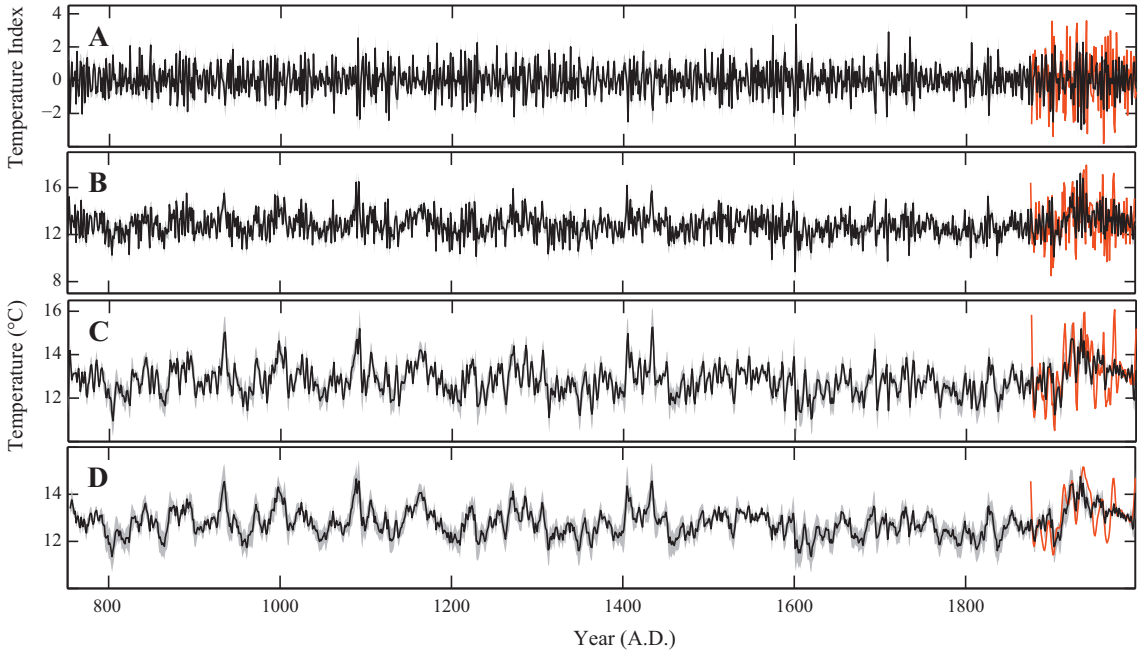

Fig. 6. Black continuous line: Karasjok (northern Norway, 69°28′N, 25°31′E) summer temperature reconstruction using Northern Lapland *P. sylvestris* ring-width chronology (period 752–1998, as in Helama et al., 2009); red continuous line: instrumental data (period 1876–1998); grey shaded area: area within the 95% confidence interval of the reconstruction using Ebisuzaki's approach (see 'Methods' section for computation). (A) Residual (white noise) series; (B) standard series; (C) 5-year spline series; (D) 10-year spline series. Note the increasing autocorrelation from A to D. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: Macias-Fauria, M., et al., Persistence matters: Estimation of the statistical significance of paleoclimatic reconstruction statistics from autocorrelated time series. Dendrochronologia (2012), doi:10.1016/j.dendro.2011.08.003

Table 3
Statistics of the empirical density distributions of the reconstruction residuals of the Karasjok summer temperature reconstruction over the period 752–1998 AD. Results are shown for Res: residual (white noise); Std: standard; Spl 5-yr: 5-year spline; and Spl 10-yr: 10-year spline series. A (upper block): time series modelled using Ebisuzaki's approach; B (lower block): time series modelled using Burg's approach. Distributions obtained after 100,000 Monte-Carlo iterations. Descriptor statistics of the distributions are: x_{min}: minimum value; x_{max}: maximum value; Δx: range; $\bar{x}$: arithmetic mean; Md: mode; $\tilde{x}$: median; s^2: variance; s: standard deviation; Υ_1: Skewness; Υ_2: Kurtosis. Data source: northern Lapland *P. sylvestris* ring-width chronology *vs.* summer temperature for Karasjok (northern Norway, 69°28′N, 25°31′E).

	x_{min}	x_{max}	Δx	$\bar{x}$	Md	$\tilde{x}$	s^2	s	Υ_1	Υ_2
A										
Res	−4.31	5.07	9.38	0.01	0.08	0.00	0.80	0.90	0.04	3.44
Std	6.74	18.85	12.11	12.80	13.05	12.77	1.14	1.07	0.22	3.64
Spl 5-yr	9.28	17.04	7.76	12.81	12.73	12.80	0.52	0.72	0.23	3.51
Spl 10-yr	10.06	16.15	6.09	12.79	12.76	12.77	0.38	0.61	0.28	3.47
B										
Res	−4.53	5.23	9.76	0.01	0.09	0.00	0.81	0.90	0.04	3.46
Std	6.18	19.23	13.05	12.80	12.89	12.77	1.17	1.08	0.21	3.64
Spl 5-yr	9.21	17.24	8.02	12.81	12.78	12.80	0.54	0.74	0.20	3.48
Spl 10-yr	9.98	16.17	6.20	12.79	12.77	12.78	0.39	0.62	0.25	3.41

Supplementary Fig. 3). The statistics of the confidence interval values (Table 3) show smaller ranges with increasing autocorrelation (from standard to 10-year spline series), in accordance with progressively smoother series. Most importantly, the mean, median and mode of the distributions are very close to each other, consistent with the idea of the residuals being equally distributed around the reconstructed values. Further, Skewness values close to zero and Kurtosis values close to 3 for all types of series and for both Ebisuzaki's and Burg's methods indicate a distribution of errors close to normal and robust regarding the degree of autocorrelation present in the original time series used in the reconstruction (Fig. 7). In any case, and as seen in the previous section, Ebisuzaki's approach will outperform Burg's method in modelling the structure of the residual series and is thus recommended in this step too.

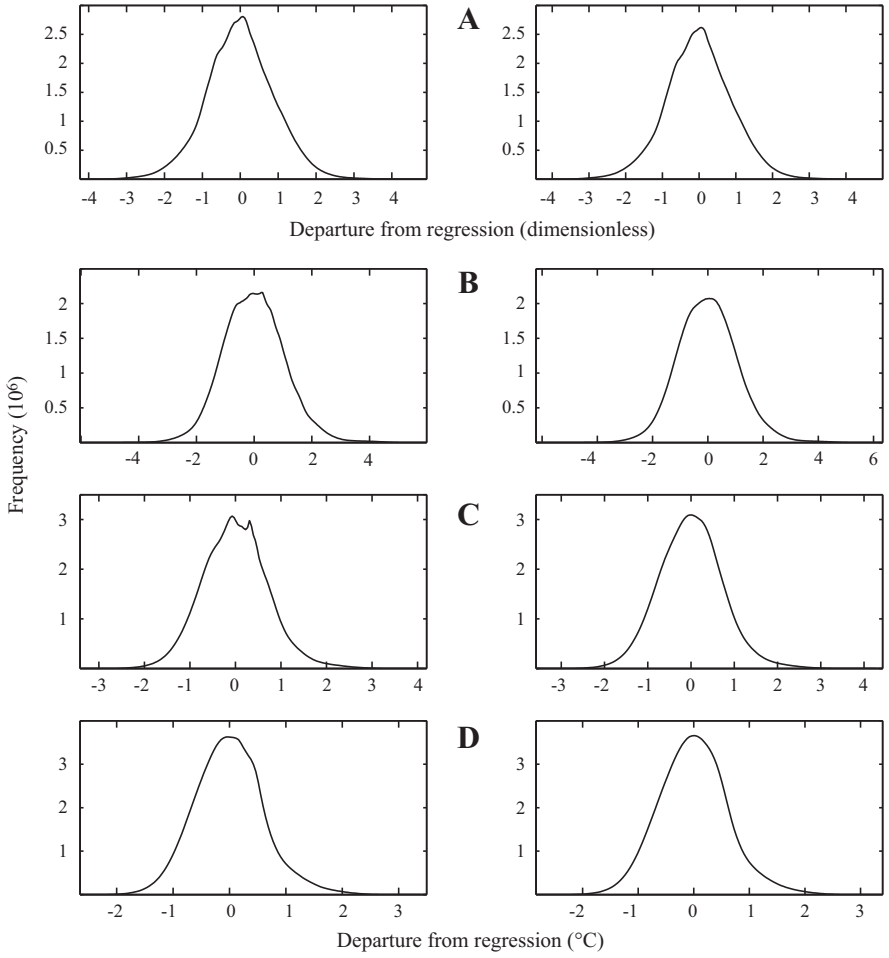

Fig. 7. Empirical probability density functions for the modelled residuals of the reconstruction of Karasjok (northern Norway, 69°28′N, 25°31′E) summer temperature using Northern Lapland *P. sylvestris* ring-width chronology (period 752–1998, as in Helama et al., 2009). Left column: residual time series modelled using Ebisuzaki's approach; right column: residual time series modelled using Burg's approach (see 'Methods' section for description). Reconstructions based on: (A) residual (white noise); (B) standard; (C) 5-year spline; (D) 10-year spline time series.

Please cite this article in press as: Macias-Fauria, M., et al., Persistence matters: Estimation of the statistical significance of paleoclimatic reconstruction statistics from autocorrelated time series. Dendrochronologia (2012), doi:10.1016/j.dendro.2011.08.003

Conclusions and significance

1. We present a simple method that enables a robust calculation of the value and significance of climatic or environmental reconstruction statistics, as well as reconstructions' confidence intervals, taking into account autocorrelation within time series, and based on a combination of time series modelling and Monte-Carlo iterations.
2. Highly autocorrelated time series show an increased occurrence of relatively high but spurious RE and R^2 values. CE is a more robust statistic, but shows the same limitations at higher significances (*i.e.* $p < 0.01$). Threshold values of 0 for RE and CE, traditionally used to distinguish between successful and unsuccessful reconstructions, are not necessarily valid and depend on the temporal structure of the time series analysed.
3. Ebisuzaki's (frequency domain) outperformed Burg's (time domain) method by better mimicking the original structure of the modelled time series and it is thus the recommended procedure.
4. Burg's approach limitations may reflect goodness of fit being more important than over parametrization in the area of statistical inference. Model selection by a method other than AIC might improve its performance. In any case, and due to limited testing, we encourage testing both methods by anyone attempting the method with other proxy data.
5. Confidence intervals for the reconstruction based on the temporal structure of the residuals of the transfer function were successfully created and showed robustness as related to varying levels of autocorrelation level in the original time series.
6. The same procedure has been successfully tested on proxy time series of different nature (*e.g.* plant phenological time series and schlerochronologies; Helama et al., 2010; Holopainen et al., 2006).
7. We offer a Matlab-based program with a user interface which allows the user to perform such analyses, as well as a Windows executable file for non-Matlab users. Please visit the following webpage to freely download it: http://oxlel.zoo.ox.ac.uk/reconstats.

Acknowledgements

Marc Macias-Fauria thanks Professor E.A. Johnson at the University of Calgary for funding his work in this study. Marc Macias-Fauria is currently funded by an EC Marie Curie Fellowship (No. 254206). Because one of the authors of this article, Samuli Helama, is a Guest Editor of this Special Issue, this manuscript was handled and edited completely independently by another Guest Editor, Margaret Devall. Finally, we would like thank two anonymous reviewers for their insightful comments on this manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.dendro.2011.08.003.

References

Akaike, H., 1974. New look at statistical-model identification. IEEE Transactions on Automatic Control AC19, 716–723.
Bartlett, M.S., 1935. Some aspects of the time correlation problem in regard to tests of significance. Journal of the Royal Statistical Society 98, 536–543.
Bradley, R.S., 1999. Paleoclimatology – Reconstructing Climates of the Quaternary. Academic Press, London, 613 pp.
Briffa, K.R., Jones, P.D., Bartholin, T.S., Eckstein, D., Schweingruber, F.H., Karlen, W., Zetterberg, P., Eronen, M., 1992. Fennoscandian summers from AD-500 – temperature-changes on short and long timescales. Climate Dynamics 7, 111–119.
Briffa, K.R., Jones, P.D., Pilcher, J.R., Hughes, M.K., 1988. Reconstructing summer temperatures in northern Fennoscandinavia back to ad 1700 using tree-ring data from Scots pine. Arctic and Alpine Research 20, 385–394.
Briffa, K.R, Jones, P.D., Schweingruber, F.H., Karlen, W., Shiyatov, S.G., 1996. Tree-ring variables as proxy-climate indicators: problems with low-frequency signals. In: Jones, P.D., Bradley, R.S., Jouzel, J. (Eds.), Climatic Variations and Forcing Mechanisms of the Last 2000 Years. , 1st ed. Springer-Verlag, Berlin, pp. 9–41.
Burg, J.P., 1978. A new analysis technique for time series data. In: Childers, D.G. (Ed.), Modern Spectrum Analysis. IEEE Press, New York, pp. 42–48.
Cook, E.R., Briffa, K.R., Jones, P.D., 1994. Spatial regression methods in dendroclimatology – a review and comparison of 2 techniques. International Journal of Climatology 14, 379–402.
Cook, E.R., Peters, K., 1981. The smoothing spline: a new approach to standardizing forest interior tree-ring width series for dendroclimatic studies. Tree-Ring Bulletin 41, 45–53.
Ebisuzaki, W., 1997. A method to estimate the statistical significance of a correlation when the data are serially correlated. Journal of Climate 10, 2147–2153.
Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science 1, 54–75.
Erlandsson, S., 1936. Dendrochronological studies. Stockholm's College Geochronological Institute Rep. 23, Uppsala, Sweden, 116 pp.
Fritts, H.C., 1976. Tree Rings and Climate. Academic Press, New York, NY, 567 pp.
Guiot, J., 1985. The extrapolation of recent climatological series with spectral canonical regression. Journal of Climatology 5, 325–335.
Helama, S., Läänelaid, A., Tietäväinen, H., Macias Fauria, M., Kukkonen, I.T., Holopainen, J., Nielsen, J.K., Valovirta, I., 2010. Late Holocene climatic variability reconstructed from incremental data from pines and pearl mussels – a multi-proxy comparison of air and subsurface temperatures. Boreas 39 (4), 734–748.
Helama, S., Timonen, M., Holopainen, J., Ogurtsov, M.K.M., Eronen, M., Lindholm, M., Merilainen, J., 2009. Summer temperature variations in Lapland during the Medieval Warm Period and the Little Ice Age relative to natural instability of thermohaline circulation on multi-decadal and multi-centennial scales. Journal of Quaternary Science 24, 450–456.
Henkel, R.E., 1976. Tests of Significance. Sage Publications, Thousand Oaks, CA, 92 pp.
Holopainen, J., Helama, S., Kajander, J.M., Korhonen, J., Launiainen, J., Nevanlinna, H., Reissell, A., Salonen, V.P., 2009. A multiproxy reconstruction of spring temperatures in south-west Finland since 1750. Climatic Change 92, 213–233.
Holopainen, J., Helama, S., Timonen, M., 2006. Plant phenological data and tree-rings as palaeoclimate indicators in south-west Finland since AD 1750. International Journal of Biometeorology 51, 61–72.
Jarque, C.M., Bera, A.K., 1987. A test for normality of observations and regression residuals. International Statistical Review/Revue Internationale de Statistique 55, 163–172.
Karl, T.R., 1988. Multi-year fluctuations of temperature and precipitation – the gray area of climate change. Climatic Change 12, 179–197.
Macias Fauria, M., Grinsted, A., Helama, S., Moore, J., Timonen, M., Martma, T., Isaksson, E., Eronen, M., 2010. Unprecedented low twentieth century winter sea ice extent in the Western Nordic Seas since AD 1200. Climate Dynamics 34, 781–795.
Melvin, T.M., Briffa, K.R., 2008. A "signal-free" approach to dendroclimatic standardisation. Dendrochronologia 26, 71–86.
Moberg, A., Sonechkin, D.M., Holmgren, K., Datsenko, N.M., Karlen, W., 2005. Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data. Nature 433, 613–617.
Mudelsee, M., 2003. Estimating Pearson's correlation coefficient with bootstrap confidence interval from serially dependent time series. Mathematical Geology 35, 651–665.
Osborn, T.J., Briffa, K.R., 2000. Revisiting timescale-dependent reconstruction of climate from tree-ring chronologies. Dendrochronologia 18, 9–25.
Quenouille, M.H., 1952. Associated Measurements. Butterworth Scientific, London, 242 pp.
Rutherford, S., Mann, M.E., Osborn, T.J., Bradley, R.S., Briffa, K.R., Hughes, M.K., Jones, P.D., 2005. Proxy-based Northern Hemisphere surface temperature reconstructions: sensitivity to method, predictor network, target season, and target domain. Journal of Climate 18, 2308–2329.
Timm, O., Ruprecht, E., Kleppek, S., 2004. Scale-dependent reconstruction of the NAO index. Journal of Climate 17, 2157–2169.
Trenberth, K.E., 1984. Some effects of finite-sample size and persistence on meteorological statistics. 1. Autocorrelations. Monthly Weather Review 112, 2359–2368.
von Storch, H., Zwiers, F.W., 1999. Statistical Analysis in Climate Research. Cambridge University Press, Cambridge, 484 pp.
Woodhouse, C.A., 1999. Artificial neural networks and dendroclimatic reconstructions: an example from the Front Range, Colorado, USA. Holocene 9, 521–529.
Young, R., Walanus, A., Goslar, T., 2000. Auto-correlation analysis in search of short-term patterns in varve data from sediments of Lake Gosciaz, Poland. Boreas 29, 251–260.
Yule, G.U., 1926. Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series. Journal of the Royal Statistical Society 89, 1–69.

Please cite this article in press as: Macias-Fauria, M., et al., Persistence matters: Estimation of the statistical significance of paleoclimatic reconstruction statistics from autocorrelated time series. Dendrochronologia (2012), doi:10.1016/j.dendro.2011.08.003