Math Geosci (2014) 46:411–427
DOI 10.1007/s11004-014-9523-4

Number 4

More accurate, calibrated bootstrap confidence intervals for estimating the correlation between two time series

K. B. Ólafsdóttir · M. Mudelsee

Received: 12 September 2013 / Accepted: 19 January 2014 / Published online: 18 February 2014

Abstract Estimation of Pearson's correlation coefficient between two time series, in the evaluation of the influences of one time-dependent variable on another, is an often used statistical method in climate sciences. Data properties common to climate time series, namely non-normal distributional shape, serial correlation, and small data sizes, call for advanced, robust methods to estimate accurate confidence intervals to support the correlation point estimate. Bootstrap confidence intervals are estimated in the Fortran 90 program PearsonT (Mudelsee, Math Geol 35(6):651–665, 2003), where the main intention is to obtain accurate confidence intervals for correlation coefficients between two time series by taking the serial dependence of the data-generating process into account. However, Monte Carlo experiments show that the coverage accuracy of the confidence intervals for smaller data sizes can be substantially improved. In the present paper, the existing program is adapted into a new version, called PearsonT3, by calibrating the confidence interval to increase the coverage accuracy. Calibration is a bootstrap resampling technique that performs a second bootstrap loop (it resamples from the bootstrap resamples). It offers, like the non-calibrated bootstrap confidence intervals, robustness against the data distribution. Pairwise moving block bootstrap resampling is used to preserve the serial dependence of both time series. The calibration is applied to standard error-based bootstrap Student's t confidence intervals. The performance of the calibrated confidence interval is examined with Monte Carlo simulations and compared with the performance of confidence intervals without calibration.

K. B. Ólafsdóttir · M. Mudelsee (✉)
Climate Risk Analysis, Schneiderberg 26, 30167 Hannover, Germany
e-mail: mudelsee@climate-risk-analysis.com

K. B. Ólafsdóttir
MARUM, Center for Marine Environmental Sciences,
University of Bremen, 28334 Bremen, Germany

The coverage accuracy is evidently better for the calibrated confidence intervals where the coverage error is acceptably small already (i.e., within a few percentage points) for data sizes as small as 20.

Keywords Pearson's correlation coefficient · Bootstrap resampling · Calibrated confidence interval · Monte Carlo simulations · Climate time series

1 Introduction

Climate time series provide an information about the natural system that generated them. In a geoscientific context, some time series are proxy data used to reconstruct climate fluctuations of the past, output from numerical models which simulate the climate system, or instrumental measurements that cover the most recent time period (Cronin 2010). Statistical inference methods are essential tools providing an estimate of the climate parameter derived from the time series with realistic uncertainty measures. However, a time series is only one indication of the system aimed to be reconstructed. Information about the natural system that produced the time series may be lacking; for example, the statistical distribution behind the data is usually unknown. This requires robust statistical methods, which avoid making strong assumptions about the properties of the process that generated the data. Evidently, other scientific contexts (such as astrophysics) have to deal with non-normal distributions.

In climate sciences, the interest is often in knowing the correlation between two processes to evaluate the influence that one time-dependent variable has on the other. Pearson's (1896) correlation coefficient measures the linear influence as

$$r_{XY} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X(i)-\bar{X}}{S_{n,X}}\right)\cdot\left(\frac{Y(i)-\bar{Y}}{S_{n,Y}}\right), \tag{1}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of two stationary stochastic processes and $S_{n,X}$ and $S_{n,Y}$ are the sample standard deviations calculated with the denominator n (instead of $n-1$). Owing to the selection of this denominator, the coefficient lies in the range $[-1, 1]$ and informs about the strength of the correlation. Applying Pearson's correlation coefficient assumes linearity of the relationship between the two processes. However, the interactions in the climate system do sometimes contain nonlinearities (Pisias et al. 1990; von Storch and Zwiers 1999). In such cases we do not recommend using the Pearson's correlation coefficient without some appropriate transformation prior to the analysis. Granger et al. (2004) consider nonlinear correlation measures. In the application we analyze, however, the scatterplots do not give hints that a transformation is necessary or a nonlinear measure would yield more information.

The climate time series come from various archives (e.g., marine sediment cores, ice cores, speleothems, or climate models), where the measured variables are either direct measurements of the climate variability or approximation of the variability measured through proxies. The timescale is estimated for each individual archive by direct dating or by comparison with other accurately dated archives. Calculation of

Pearson's correlation coefficient requires that the time series are on identical timescales (coevally). This is the case when comparing two time series that come from the same archive; for example, when several variables are measured on the same sediment core. In cases where the two time series are not sampled on identical time points, we recommend different correlation estimators such as the binned correlation coefficient (Mudelsee 2010) or a Gaussian kernel method (Rehfeld et al. 2011).

As r_{XY} is a point estimate of the true theoretical correlation value ρ_{XY}, a statistical inference is needed to evaluate the reliability of the estimate. The two ways are to perform a hypothesis test of $H_0 : \rho_{XY} = 0$, and to estimate a confidence interval to use along with the point estimate. Here, we prefer a confidence interval over a hypothesis test, as it contains more quantitative information (Efron and Tibshirani 1993). It is easier to compare one estimate with another if it is accompanied by a confidence interval instead of P value for a test, which only indicates whether the correlation is significant or not. The width of the confidence interval depends on the data variability and data size. In addition, a confidence interval can be used as a significance test by verifying whether the interval contains zero or not. Nevertheless, many of the existing methods of estimating confidence intervals require strong mathematical assumptions (regarding distributional shape and serial dependence) to be made. These are usually unfulfilled when dealing with climate time series. Correlation analyses that use confidence intervals where these assumptions are not taken into account may provide incorrect results. For example, if the serial dependence is ignored, the confidence intervals are likely too narrow and therefore overestimate the accuracy of the result. The accuracy of the confidence intervals is measured with the coverage performance in Monte Carlo simulations. A 95 % confidence interval implies a nominal probability of 0.95 that the estimated confidence interval covers the true parameter ρ_{XY}. Exact confidence intervals have the actual coverage probability equal to the nominal value. However, this is only the case if all the assumptions of the underlying method are met.

Bootstrap resampling is one way to construct accurate confidence intervals that deal with more complex data properties (regarding data distribution) (DiCiccio and Efron 1996; Efron and Tibshirani 1993; Mudelsee 2010). The bootstrap approach resamples from the data themselves and forms confidence intervals from replicated estimations on the resamples. Therefore, it preserves the data properties and accounts for climate and proxy uncertainties that affect the data. In the method by Mudelsee (2003), Pearson's correlation coefficient between two time series is estimated with a bootstrap confidence interval. The main intention of this method was to obtain an accurate confidence interval for the correlation coefficient between two climate time series by taking into account the serial dependence of the data. However, Monte Carlo experiments (Mudelsee 2003) showed that when data contain persistence, larger data sizes ($n > 200$) are needed to obtain the same coverage accuracy of the confidence intervals than for data without persistence. Yet, especially in paleoclimate sciences, the number of data points is often limited. The goal of this study is to improve the method and increase the coverage accuracy for Pearson's correlation coefficient furthermore, especially for smaller data sizes (n in the order of 20–100).

Bootstrap confidence intervals do usually contain relatively small coverage error (difference between empirical coverage and nominal value) compared to standard

intervals (DiCiccio and Efron 1996; Mudelsee 2010), but by calibrating the confidence intervals the coverage accuracy can be improved remarkably (Mudelsee 2010). Principally, the calibration or iteration methods (Beran 1987; Hall 1986; Hall and Martin 1988; Loh 1987) perform a second bootstrap loop or resample from the bootstrap resamples to improve the coverage accuracy. Here, we apply calibration to standard-error based bootstrap Student's t confidence intervals following Efron and Tibshirani (1993). The performance of the calibrated confidence intervals is examined with Monte Carlo simulations, where we are able to check the actual coverage of the confidence intervals. Examples are taken where the method is applied to paleoceanographic time series from the Cape Basin, southwest off the African coast, which aim to reconstruct the changes in Agulhas Leakage through the past several hundred thousand years. The Fortran 90 software, PearsonT3, which is an improved version of the existing PearsonT (Mudelsee 2003), is introduced (Appendix A).

2 Calibrated Bootstrap Confidence Intervals for the Correlation Coefficient

We have bivariate time series $\{x(i), y(i)\}_{i=1}^{n}$ sampled (unevenly or evenly spaced) at the same timescale $\{t(i)\}_{i=1}^{n}$ from the process $\{X(i), Y(i)\}$ (see Table 1 for notation). This may be called a coevally sampled time series. It is assumed that the process is weakly stationary, that is, mean, variance, and covariance do not change with time. A calibrated bootstrap confidence interval is used to support the correlation point estimate. It offers, such as the non-calibrated bootstrap confidence interval, robustness against the data distribution, as the bootstrap resampling technique replaces the unknown distribution function $F(x, y)$ by a simulation–approximation to the empirical distribution function $F_{\mathrm{emp}}(x, y)$. The calibration requires a second bootstrap loop that further increases the coverage accuracy of the confidence interval.

2.1 Pairwise Moving Block Bootstrap Resampling

The pairwise moving block bootstrap (MBB) resampling technique, advocated by Mudelsee (2010), is used to resample blocks of data from the original time series $\{x(i), y(i)\}_{i=1}^{n}$ to form resamples that preserve the properties of the data-generating process. Block length is chosen oriented on the persistence time, τ, of the time series, which is a quantification of the memory of the climate processes. The persistence times, τ_X and τ_Y, are estimated with a least-squares algorithm (Mudelsee 2002). This fits a first-order autoregressive or AR(1) persistence model to unevenly spaced time series; for example

$$\begin{aligned} X(1) &= \mathscr{E}_{\mathrm{N}(0,\,1)}(1), \\ X(i) &= \exp\left\{-\left[T(i) - T(i-1)\right]/\tau_X\right\} \cdot X(i-1) \\ &\quad + \mathscr{E}_{\mathrm{N}(0,\,1-\exp\{-2[T(i)-T(i-1)]/\tau_X\})}(i), \quad i = 2, \ldots, n. \end{aligned} \tag{2}$$

Table 1 Notation

$\{x(i), y(i)\}_{i=1}^{n}$	Bivariate time series
$\{t(i)\}_{i=1}^{n}$	Time
$\{X(i), Y(i)\}$	Bivariate climate process
n	Data size
$d(i)$	Time spacing
$\bar{d}$	Average time spacing
$F(x, y)$	Distribution function of $\{X(i), Y(i)\}$
$F_{\mathrm{emp}}(x, y)$	Empirical distribution function of $\{x(i), y(i)\}_{i=1}^{n}$
MBB	Moving block bootstrap
l_{opt}	Optimal block length
AR(1)	First-order autoregressive model
a	AR(1) autocorrelation parameter (even spacing)
τ	AR(1) persistence time (uneven spacing)
τ_X	AR(1) persistence time for $\{X(i)\}$ process (uneven spacing)
τ_Y	AR(1) persistence time for $\{Y(i)\}$ process (uneven spacing)
$\hat{\tau}$	AR(1) persistence time estimator (uneven spacing)
$\mathcal{E}_{\mathrm{N}(\mu,\,\sigma^2)}$	Purely random process, normal shape, mean μ, variance σ^2
$S(\widetilde{\tau}_X)$	Least-squares sum
$\widetilde{\tau}_X$	Argument value for τ_X
$\hat{a}$	Estimated AR(1) equivalent autocorrelation parameter (uneven spacing)
$\hat{a}'_X$	Estimated, bias-corrected AR(1) equivalent autocorrelation parameter for $\{X(i)\}$ process
$\hat{a}'_Y$	Estimated, bias-corrected AR(1) equivalent autocorrelation parameter for $\{Y(i)\}$ process
$\hat{a}'_{XY}$	Combined estimated, bias-corrected AR(1) equivalent autocorrelation parameter
ρ_{XY}	True theoretical correlation value
r_{XY}	Pearson's correlation coefficient estimator
$\{x^*(i), y^*(i)\}_{i=1}^{n}$	Bootstrap resample
$\{x^{**}(i), y^{**}(i)\}_{i=1}^{n}$	Bootstrap resample from resample (calibration)
$\{t^*(i)\}_{i=1}^{n}$	Bootstrap time
B	Number of bootstrap resamples
B_2	Number of bootstrap resamples from resample (calibration)
$\{r_{XY}^{*b}\}_{b=1}^{B}$	Bootstrap replications of r_{XY}
$\hat{\mathrm{se}}_{r_{XY}^*}$	Estimated bootstrap standard error
CI	Confidence interval
$\mathrm{CI}_{r_{XY},\,1-2\alpha}$	Confidence interval for r_{XY} of level $1-2\alpha$
$1-2\alpha$	Confidence level
$r_{XY,\mathrm{l}}$	Lower confidence interval bound for r_{XY}
$r_{XY,\mathrm{u}}$	Upper confidence interval bound for r_{XY}
$\mathrm{CI}_{r_{XY}^{*b},\,1-2\lambda}$	Confidence interval for r_{XY}^{*b} of level $1-2\lambda$
$1-2\lambda$	Confidence level (calibration)

Table 1 continued

$r^{*b}_{XY,\mathrm{l}}(\lambda)$	Lower confidence interval bound of $\mathrm{CI}_{r^{*b}_{XY},1-2\lambda}$
$r^{*b}_{XY,\mathrm{u}}(\lambda)$	Upper confidence interval bound of $\mathrm{CI}_{r^{*b}_{XY},1-2\lambda}$
γ	Coverage of confidence interval
$\hat{p}(\lambda)$	Calibration curve
t_v	Percentage point of the t distribution function with v degrees of freedom
n_{sim}	Number of Monte Carlo simulations

$\mathscr{E}_{\mathrm{N}(\mu,\,\sigma^2)}$ is a purely random process with normal shape, mean μ, and variance σ^2; $T(i)$ are the time points. The least-squares estimation uses the sum of squares

$$S(\widetilde{\tau}_X) = \sum_{i=2}^{n} \left[x(i) - \exp\left\{-\left[t(i) - t(i-1)\right]/\widetilde{\tau}_X\right\} \cdot x(i-1)\right]^2, \tag{3}$$

and takes the minimizer as τ_X estimator, $\widehat{\tau}_X = \mathrm{argmin}[S(\widetilde{\tau}_X)]$. The minimization is carried out numerically. The equivalent autocorrelation coefficient is given by $\hat{a} = \exp(-\bar{d}/\hat{\tau})$, where $\bar{d} = [t(n) - t(1)]/(n-1)$ is the average spacing of the time series. We use a block length selector for the univariate case from Sherman et al. (1998), who adapted the formula from Carlstein (1986) to the original MBB (Künsch 1989; Liu and Singh 1992)

$$l_{\mathrm{opt}} = \mathrm{NINT}\left\{\left[6^{1/2} \cdot \hat{a}'_{XY}/(1 - \hat{a}'^2_{XY})\right]^{2/3} \cdot n^{1/3}\right\}, \tag{4}$$

where NINT is the nearest integer function. Here, this equation is used for bivariate case by combining the bias-corrected equivalent autocorrelation coefficients, $\hat{a}'_X$ and $\hat{a}'_Y$, as (Mudelsee 2010)

$$\hat{a}'_{XY} = \left[\hat{a}'_X \cdot \hat{a}'_Y\right]^{1/2}. \tag{5}$$

For the bias correction, an approximation for an AR(1) process with unknown mean on an evenly spaced timescale (Kendall 1954) is used

$$E(\hat{a}) \simeq a - (1 + 3a)/(n-1). \tag{6}$$

The equation is solved for a, which gives the bias-corrected estimator $\hat{a}' = [\hat{a} \cdot (n-1) + 1]/(n-4)$. The choice of block-length selector has some influence on the accuracy of results; Mudelsee (2010) shows by means of Monte Carlo experiments that our choice (Eq. 4) is superior (in terms of confidence interval accuracy) to another selector (Mudelsee 2003) that uses a block length equal to four times the estimated, bias-corrected persistence time.

Next the time series $\{x(i), y(i)\}_{i=1}^{n}$ is split up into a set of overlapping blocks (length l_{opt}) of observations. Random pairs of blocks (same block from $\{x(i)\}_{i=1}^{n}$ and

$\{y(i)\}_{i=1}^{n}$) are drawn from the set, with replacement. The blocks are pasted together to form the bootstrap resample $\{x^*(i), y^*(i)\}_{i=1}^{n}$ of the same length n as the original sample. When the last point, $\{x^*(n), y^*(n)\}$, is inserted, any remaining values of that block are discarded. The resampled timescale is unchanged $\{t^*(i)\}_{i=1}^{n} = \{t(i)\}_{i=1}^{n}$. This means, we assume that timescale uncertainties are negligible. At least for two variables with strongly coupled timescales (e.g., oxygen and carbon isotopic composition measured on the same sediment core), this assumption is not strong. This is repeated many times, until B resamples exist. Typically, $B = 2{,}000$ is used (Efron and Tibshirani 1993).

2.2 Bootstrap Confidence Interval for the Correlation Coefficient

The correlation coefficient is now estimated for each of the resamples, yielding $\{r_{XY}^{*b}\}_{b=1}^{B}$, which is called the replications of the correlation coefficient. The bootstrap standard error is the standard deviation of the replications

$$\hat{\mathrm{se}}_{r_{XY}^*} = \left\{ \sum_{b=1}^{B} \left[r_{XY}^{*b} - \langle r_{XY}^{*b} \rangle \right]^2 / (B-1) \right\}^{1/2}, \tag{7}$$

where $\langle r_{XY}^{*b} \rangle = \sum_{b=1}^{B} r_{XY}^{*b}/B$. For non-calibrated confidence interval, the bootstrap standard error may now be used to construct a bootstrap confidence interval.

2.3 Calibration of the Bootstrap Confidence Interval

Exact equi-tailed confidence intervals do have a coverage γ equal to $\mathrm{prob}\{r_{XY,\mathrm{l}}(\alpha) < r_{XY} < r_{XY,\mathrm{u}}(\alpha)\} = 1 - 2\alpha$, which is called confidence level, where $r_{XY,\mathrm{l}}$ and $r_{XY,\mathrm{u}}$ are the confidence interval bounds. A calibration curve, $\hat{p}(\lambda) = 1-2\alpha$, is used to see if another confidence point, λ, can be used instead of α to obtain the desired coverage. A second bootstrap loop, with typically $B_2 = 1{,}000$ resamplings (Mudelsee 2010) and block length identical to that of the first bootstrap loop, is performed to estimate the calibration curve. That is, if $\{x^*(i), y^*(i)\}_{i=1}^{n}$ denotes a resample (first bootstrap loop, which is repeated B times), then a resample from the resample, $\{x^{**}(i), y^{**}(i)\}_{i=1}^{n}$, is obtained by pairwise moving block bootstrap resampling from the resample. The total number of series $\{x^{**}(i), y^{**}(i)\}_{i=1}^{n}$ produced is, thus, equal to $B \cdot B_2 = 2{,}000 \cdot 1{,}000 = 2{,}000{,}000$. The confidence interval is estimated for each of the bootstrap replications $\{r_{XY}^{*b}\}_{b=1}^{B}$ from the first bootstrap loop, and over a grid of confidence levels λ

$$\mathrm{CI}_{r_{XY}^{*b},1-2\lambda} = \left[r_{XY,\mathrm{l}}^{*b}(\lambda)\,;\, r_{XY,\mathrm{u}}^{*b}(\lambda) \right], \qquad \lambda = 0.001, \ldots, 0.499. \tag{8}$$

The confidence interval is, as below, a Student's t type ($\upsilon = 2n - 5$ degrees of freedom) that uses the bootstrap standard error determined over the second-loop replications. For each λ we compute

$$\hat{p}(\lambda) = \frac{\#\left\{r^{*b}_{XY,\mathrm{l}}(\lambda) < r_{XY} < r^{*b}_{XY,\mathrm{u}}(\lambda)\right\}}{B}, \tag{9}$$

where $\#\{r^{*b}_{XY,\mathrm{l}}(\lambda) < r_{XY} < r^{*b}_{XY,\mathrm{u}}(\lambda)\}$ means the number of replications, where $r^{*b}_{XY,\mathrm{l}}(\lambda) < r_{XY} < r^{*b}_{XY,\mathrm{u}}(\lambda)$. Finally, $p(\lambda) = 1 - 2\alpha$ is solved for λ. The relevant λ value is used to construct a calibrated bootstrap Student's t confidence interval

$$\mathrm{CI}_{r_{XY,1-2\alpha}} = \left[r_{XY} + t_v(\lambda) \cdot \hat{\mathrm{se}}_{r^*_{XY}} \,;\, r_{XY} - t_v(\lambda) \cdot \hat{\mathrm{se}}_{r^*_{XY}}\right], \tag{10}$$

where the bootstrap standard error $\hat{\mathrm{se}}_{r^*_{XY}}$ from the first bootstrap loop is inserted (Eq. 7), and t_v is the percentage point of the t distribution function with $v = 2n - 5$ degrees of freedom (Mudelsee 2010).

3 Monte Carlo Experiments

Monte Carlo simulations were carried out to evaluate the performances of the calibrated confidence intervals and to test if the calibration does indeed improve the coverage accuracy in comparison to confidence intervals without calibration. We are able to determine the actual coverage of confidence intervals by generating many artificial time series with prescribed known properties, where situations typical for climate time series are simulated by including non-normal (skewed) distribution shape and serial dependence of both time series. Confidence intervals are estimated for each of the generated series, and then the actual coverage is defined by counting the fraction of simulations for which the known theoretical value, ρ_{XY}, is within the confidence interval.

For each experiment, $n_{\mathrm{sim}} = 47{,}500$ random, unevenly spaced time series were generated from a bivariate log-normal AR(1) process $\{X(i), Y(i)\}_{i=1}^{n}$ (Appendix B). This particular number of simulations gives the nominal standard error $\sigma = 0.001$ for $2\alpha = 0.05$, as $\sigma = \sqrt{2\alpha(1-2\alpha)/n_{\mathrm{sim}}}$. The uneven time spacing, $d(i)$, was drawn from gamma distribution and the average time spacing set to $\bar{d} = 1$. The log-normal shape was given by taking $\exp[X(i)]$ and $\exp[Y(i)]$. The persistence times were set to $\tau_X = 1$ and $\tau_Y = 2$. The white noise terms of the processes are correlated with the correlation coefficient $\rho_{\mathcal{E}}$, which is given by $\mathrm{CORR}[X(i), Y(i)] = \rho_{XY} = [\exp(\rho_{\mathcal{E}}) - 1]/(e - 1)$, where the correlation operator is defined as

$$\mathrm{CORR}[X(i), Y(i)] = \frac{\mathrm{COV}[X(i), Y(i)]}{\{\mathrm{VAR}[X(i)] \cdot \mathrm{VAR}[Y(i)]\}^{1/2}}. \tag{11}$$

We use two different predefined correlation values $\rho_{XY} \in \{0.3, 0.8\}$. Calibrated confidence intervals were estimated with PearsonT3 as well as bootstrap bias-corrected and accelerated (BCa) confidence intervals estimated with the previous PearsonT software. In addition, non-calibrated bootstrap Student's t confidence intervals were estimated with the same generated time series to evaluate the pure effect of the calibration.

Table 2 Empirical coverages for 95 % confidence intervals (Student's t, calibrated Student's t, and BCa)

n	$\gamma_{r_{XY}}$ True correlation, ρ_{XY}						Nominal
	0.3			0.8			
	Bootstrap CI type			Bootstrap CI type			
	Student's t	Calibrated	BCa	Student's t	Calibrated	BCa	
10	0.902	0.819	0.740	0.863	0.815	0.730	0.950
20	0.903	0.922	0.790	0.898	0.934	0.783	0.950
50	0.902	0.926	0.860	0.902	0.946	0.834	0.950
100	0.900	0.930	0.873	0.900	0.947	0.844	0.950
200	0.892	0.929	0.869	0.898	0.947	0.854	0.950

Monte Carlo simulations were carried out for several different data sizes from 10 to 200 and two different predefined correlation coefficients, $\rho_{XY} = 0.3$ and 0.8

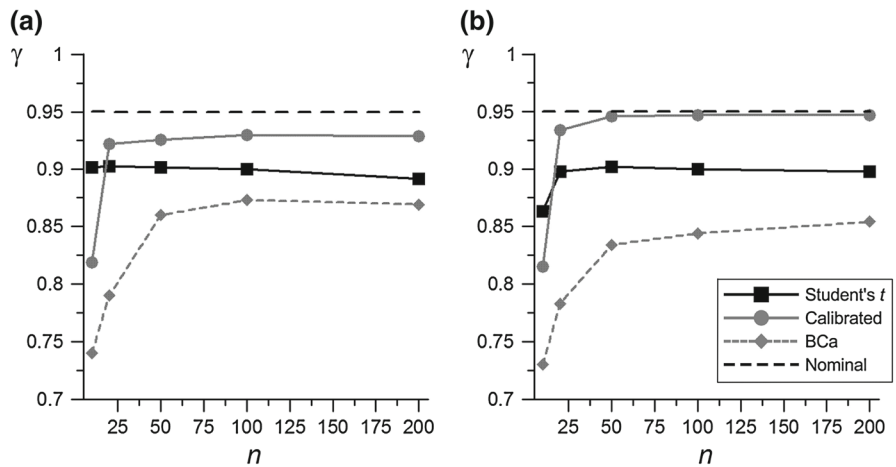

Fig. 1 Coverage performance of 95 % confidence intervals (calibrated Student's t, Student's t, and BCa) from the Monte Carlo simulations for several data sizes, n. **a** Results for the predefined correlation value $\rho_{XY} = 0.3$, **b** results for $\rho_{XY} = 0.8$

The confidence level was 95 % in all cases. Several experiments were done for different data sizes, $n \in \{10, 20, 50, 100, 200\}$.

The empirical coverages of the confidence intervals estimated with the three different methods are shown in Table 2; they are plotted against data size in Fig. 1. The coverage accuracy is evidently best for the calibrated confidence intervals, for which the coverage error is acceptably small (i.e., within a few percentage points) already for data sizes as small as 20. The BCa confidence intervals estimated with PearsonT do not produce acceptable coverage accuracy for the data properties and data sizes tested here, although the coverage error decreases with increasing data size. The coverages of the bootstrap Student's t confidence intervals are, according to the results, not as dependent on the data sizes as the other methods. It shows remarkably good accuracy

Table 3 Average width (i.e., $r_{XY,\mathrm{u}} - r_{XY,\mathrm{l}}$) of 95 % confidence intervals (Student's t, calibrated Student's t, and BCa) over the Monte Carlo simulations ($n_{\mathrm{sim}} = 47,500$).

n	$\mathrm{CI}_{r_{XY},95\,\%}$ length True correlation, ρ_{XY}					
	0.3			0.8		
	Bootstrap CI type			Bootstrap CI type		
	Student's t	Calibrated	BCa	Student's t	Calibrated	BCa
10	1.243	1.054	0.868	0.700	0.673	0.448
20	0.883	1.052	0.676	0.444	0.635	0.309
50	0.594	0.743	0.529	0.293	0.416	0.238
100	0.455	0.579	0.425	0.228	0.317	0.198
200	0.365	0.556	0.347	0.186	0.254	0.169

for data sizes as small as 10, and then progressively shifts toward stable coverage of ∼0.90. However, it does not have nearly as good coverage accuracy as the calibrated confidence intervals for data sizes >20, for which the actual coverage gets very close to the nominal value. Therefore, it seems that the calibration of bootstrap Student's t confidence intervals improves the coverage accuracy considerably for bivariate log-normal AR(1) processes. In addition, the calibrated confidence intervals estimated with PearsonT3 show much better performance for small data sizes than the BCa confidence intervals estimated with the previous PearsonT. Another property of interest is the confidence interval width, $r_{XY,\mathrm{u}} - r_{XY,\mathrm{l}}$, which is preferably small. However, the price for good coverage accuracy is paid with wider intervals. The interval width is shown in Table 3 for each of the three methods. The results show that the calibration increases the width of the confidence interval.

4 Application

As an example, we applied the method to paleoceanographic time series from the Agulhas system based on proxy data. Proxy records from a sediment core in the Cape Basin, off South Africa, are used to reconstruct the changes in the Agulhas Leakage through the past 560 thousand years (ka) (Peeters et al. 2004). The transport of warm, salty water from the Indian to the Atlantic Ocean via the Agulhas Leakage plays a key role in the global ocean circulation (Gordon 1986). We studied two relationships: first between Agulhas Leakage and the position of sub-tropical convergence zone (STC) relative to the core location and, second, between the Agulhas Leakage and ice volume changes. Peeters et al. (2004) show that the Agulhas Leakage decreases during the glacial intervals parallel with a northward movement of the STC. Our method is used to test the results.

The Agulhas Leakage is traced by the so-called Agulhas Leakage Fauna (ALF) (Peeters et al. 2004), which is an estimate of the relative abundance of Agulhas planktic foraminiferal species, typical for Agulhas waters. The relative position of the STC with respect to the sediment core site is based on the ratio between specific planktic

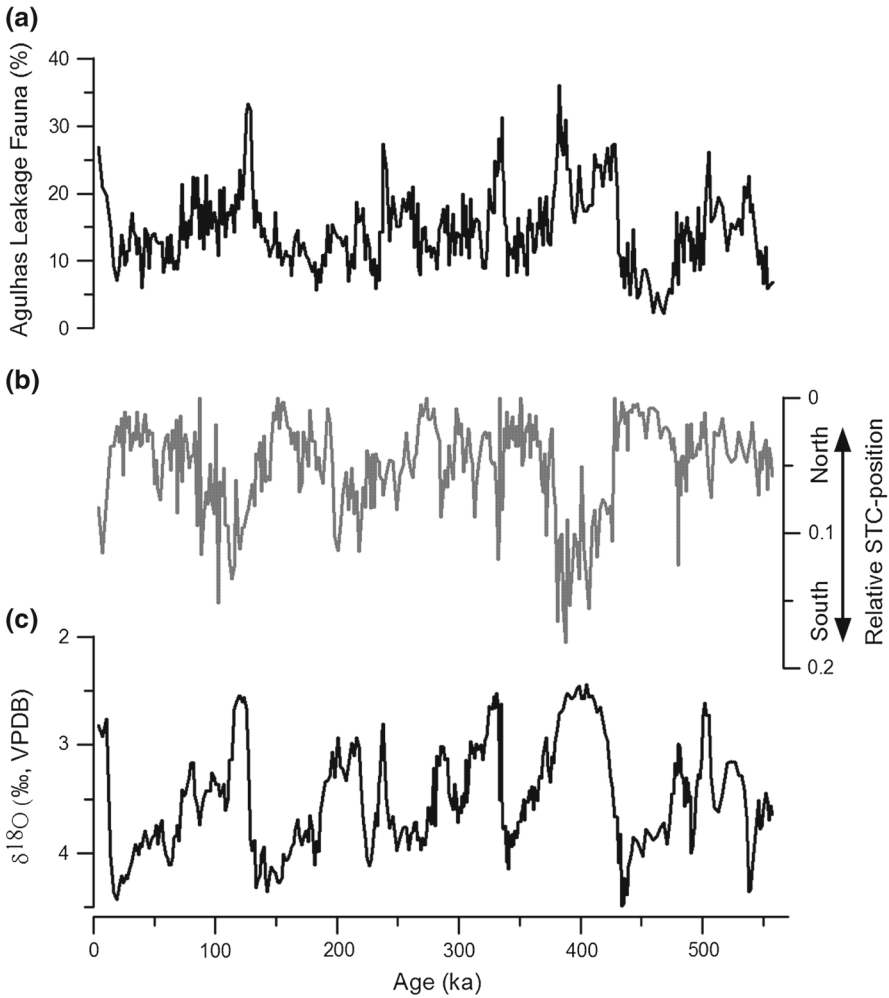

Fig. 2 Proxy time series from a sediment record recovered from the Southern Cape basin (Peeters et al. 2004). **a** ALF, proxy for the leakage of Agulhas waters. **b** STC-position index (arbitrary units). **c** δ^{18}O of *Cibicidoides wuellerstorfi*, indicating changes in ice volume

foraminifera species that reflect subtropical to transitional waters north of the STC, and species that reflect waters of the Southern subtropical to subantarctic zone. High values indicate southward position of STC, while low values indicate northward position. The down-core oxygen isotope ratio (δ^{18}O) of the benthic foraminifera *Cibicidoides wuellerstorfi* is used as a proxy for ice volume changes. High values denote more ice, low values less ice. See Peeters et al. (2004) for further details on the data. These three time series, ALF, STC-position index, and δ^{18}O, are used in the following example.

The three time series are on the same timescale. The time period is from 1.5 to 557 ka before the present, and the time series are unevenly spaced with $\bar{d} = 1.5$ ka and $n = 365$ (Fig. 2). The persistence times are large, $\tau_{\mathrm{ALF}} = 6$ ka, $\tau_{\mathrm{STC}} = 4$ ka, and $\tau_{\delta^{18}\mathrm{O}} = 18$ ka. This results in longer blocks, $l = 17$, for the first example, where

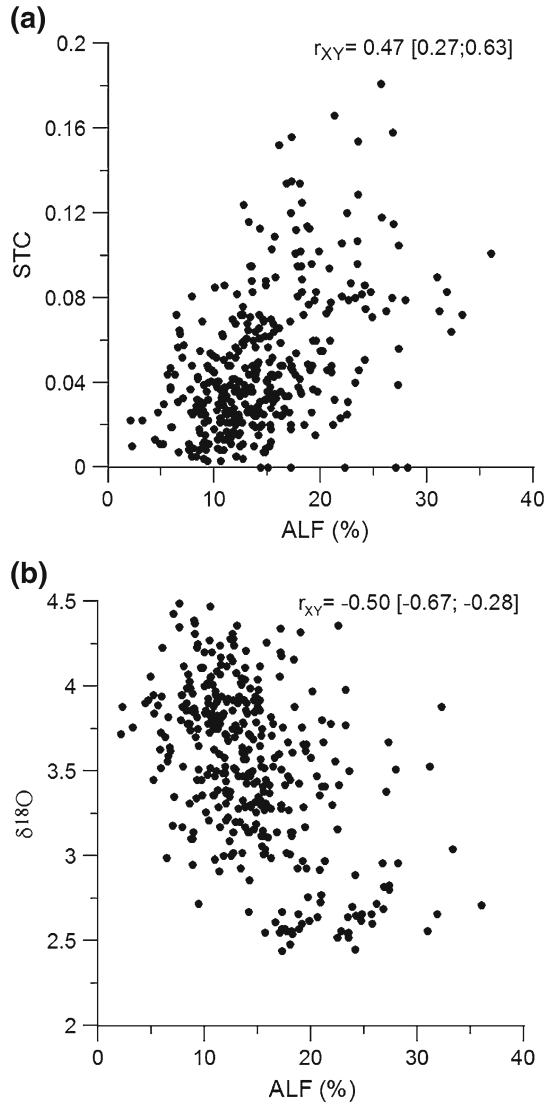

Fig. 3 Scatterplots. **a** The relationship between ALF and the STC-position index over the past 560 ka, $n = 365$ and $r_{XY} = 0.471$ [0.269; 0.633]. **b** The relationship between ALF and δ^{18}O of *Cibicidoides wuellerstorfi*, $n = 365$ and $r_{XY} = -0.497$ [−0.668; −0.275]

the correlation between ALF and STC is estimated. Note that the paleoceanographic timescales are uncertain, which make the persistence time estimates also uncertain. The resulting effects on block length selection (Eq. 4), however, are minor. The scatterplot indicates positive linear correlation (Fig. 3a). Here, the calibration gives a λ value of 0.004 to construct a 95 % confidence interval (Fig. 4a), which results in considerable wider confidence interval than without the calibration (see Table 4). The Pearson's correlation coefficient is estimated as $r_{XY} = 0.471$ with 95 % calibrated confidence interval [0.269; 0.633]. This means a significant correlation between the two variables: increased Agulhas Leakage flow is related to southward shift in the STC.

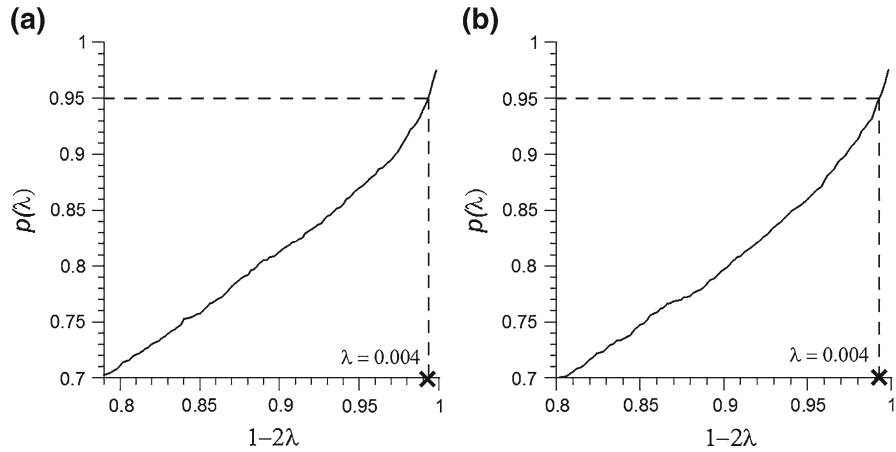

Fig. 4 Calibration curves. **a** The calibration gives a λ value of 0.004 to construct a 95 % confidence interval for r_{XY} between ALF and the STC-position index. **b** A λ value of 0.004 is used to form 95 % confidence interval for r_{XY} between ALF and δ^{18}O of *Cibicidoides wuellerstorfi*

Table 4 Pearson's correlation coefficient with 95 % confidence intervals estimated with Student's *t*, calibrated Student's *t*, and BCa methods

	r_{XY}	$\mathrm{CI}_{r_{XY},95\,\%}$
ALF vs STC		
Student's *t*	0.471	[0.325; 0.595]
Calibrated		[0.269; 0.633]
BCa		[0.322; 0.585]
ALF vs δ^{18}O		
Student's *t*	−0.497	[−0.628; −0.338]
Calibrated		[−0.668; −0.275]
BCa		[−0.618; −0.334]

Upper panel ALF versus the STC-position index. *Lower panel* ALF versus δ^{18}O of *Cibicidoides wuellerstorfi*

In the second example, where the relationship between ALF and δ^{18}O is evaluated, a block length of 26 was used according to the estimated persistence times. Here, the scatterplot indicates a negative linear correlation between ALF and δ^{18}O (Fig. 3b). In addition, a λ value of 0.004 was obtained from the calibration curve to obtain the desired coverage of 0.95 (Fig. 4b), which again results in wider interval than without the calibration (Table 3). There is a negative correlation between the two variables, $r_{XY} = -0.497$ with 95 % calibrated confidence interval [−0.668; −0.275]. This means that high ALF values (increased Agulhas Leakage) covary with low δ^{18}O values (which means less ice volume). The correlation is significant, therefore both exercises confirm the conclusion of Peeters et al. (2004).

5 Discussion

As expected, the calibration increases the accuracy of the confidence interval, but not by unreasonably increasing its width. However, the calibrated confidence intervals are

rather large when the data sizes are small and the correlation coefficients are small in size. It is more difficult to obtain accurate confidence intervals for small data sizes than for larger sizes, since fewer data mean higher uncertainty. This is clearly reflected in the interval width in our case. However, as previously mentioned, the coverage accuracy is increased considerably, which is the main intention of our study. BCa confidence intervals were used in the previous version of PearsonT (Mudelsee 2003). Calibrating the BCa confidence interval would require considerably larger computational costs than the calibration of standard error-based Student's t confidence interval (which already is computational expensive), and therefore we changed the confidence interval method in the software. It is interesting to look at the performance of the calibrated confidence interval when the data have other properties, such as normal distribution, more unevenly spaced timescale, larger persistence times, and so forth. However, the high computational cost resulting from the second bootstrap loop limits the Monte Carlo simulations as they are computationally very expensive. The major results here show that the calibration is worthwhile when the data have properties realistic for the climate world, especially for data sizes >20.

6 Conclusions

It is a challenging task to make accurate statistical inferences from climate time series with complex properties, such as non-normal distribution, uneven spacing, serial correlation, and small data sizes. We present a method for estimating accurate confidence interval for Pearson's correlation coefficient, where the coverage accuracy is increased by applying calibration to standard error-based bootstrap Student's t confidence interval. The method had been adapted into the Fortran 90 program PearsonT, and the new version is now called PearsonT3. It is clear from Monte Carlo simulations that the calibrated bootstrap confidence intervals have better coverage accuracy than the confidence intervals estimated with the previous method, especially when the data sizes are small (n in the order of 20–100). The increase in the coverage accuracy compared to non-calibrated bootstrap Student's t confidence intervals shows the effect of the second bootstrap loop, which clearly improves the performance of the confidence interval. The examples shown here demonstrate the applicability of the method to real-world climate problems. The method fits well for coevally sampled climate time series of all kinds, but should be especially useful for evaluating the reliability of correlation estimation when the data are non-normally distributed and include serial dependence.

Acknowledgments We thank Alexander Gluhovsky and three anonymous persons for constructive review comments. We thank Michael Schulz, Arne Biastoch, Jonathan Durgadoo, Frank Peeters, Conor Purcell, and Gema Martínez Méndez for discussions and helpful comments. The work described in this paper and the research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013), Marie-Curie ITN, under Grant Agreement No. 238512, GATEWAYS project.

Appendix A: Software

The calibration method explained in the paper was adapted into the Fortran 90 software PearsonT3. The software is freely available at http://www.climate-risk-analysis.

com. The installation requires copying the PearsonT3 executable file into an appropriate directory and installing the free graphic program Gnuplot (http://www.gnuplot.info/). The gnuplot executable file, gnuplot.exe, needs to be in the same directory as PearsonT3.exe. The software is command line driven and can be run from the Windows command prompt or simply by double clicking the executable file. After starting PearsonT3 the program asks for a name and path of input data file. The data file should be a simple text file and in the format

$$\begin{array}{ccc} t_1 & x_1 & y_1 \\ t_2 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ t_n & x_n & y_n, \end{array}$$

where t is sampling times and x and y are two equally long time series with data size ≥ 10. The time series are automatically mean detrended (the mean of the data is subtracted from the data). A linear detrend option was included in the old version of the software but is not included in PearsonT3. If the time series samples contain linear or more complex trend, we recommend some detrending prior to the analysis in PearsonT3 to fulfill the weakly stationary assumptions.

The persistence times are estimated with the least-squares algorithm TAUEST Mudelsee (2002) with automatic bias correction. If the bias-corrected equivalent autocorrelation coefficient becomes > 1 (Eq. 6), then the bias correction is not performed, which can occur if n is small and the autocorrelation coefficient is large. After the estimation the time series are plotted up on the screen along with an $x - y$ scatterplot to test for the linear relationship. The results are printed on the screen, which informs about the data file name, the time interval $[t(1); t(n)]$, the number of data points (n), the persistence times (τ_X and τ_Y), and the estimated correlation coefficient (r_{XY}) with 95 % calibrated confidence interval. The results are also written into a result file, along with the data, means, and mean detrended data. The final result file, named PearsonT3.dat, is a plain ASCII file, which is saved in the same directory as the executable file PearsonT3.exe.

Appendix B: Bivariate AR(1) process

The bivariate AR(1) process for uneven spacing is given by (Mudelsee 2010, Ch. 7.6)

$$\begin{aligned} X(1) &= \mathscr{E}^{X}_{\mathrm{N}(0,\,1)}(1), \\ Y(1) &= \mathscr{E}^{Y}_{\mathrm{N}(0,\,1)}(1), \\ X(i) &= \exp\left\{-\left[T(i) - T(i-1)\right]/\tau_X\right\} \cdot X(i-1) \\ &\quad + \mathscr{E}^{X}_{\mathrm{N}(0,\,1-\exp\{-2[T(i)-T(i-1)]/\tau_X\})}(i), \qquad i = 2, \ldots, n, \\ Y(i) &= \exp\left\{-\left[T(i) - T(i-1)\right]/\tau_Y\right\} \cdot Y(i-1) \\ &\quad + \mathscr{E}^{Y}_{\mathrm{N}(0,\,1-\exp\{-2[T(i)-T(i-1)]/\tau_Y\})}(i), \qquad i = 2, \ldots, n, \end{aligned} \tag{12}$$

where the white-noise terms are correlated as

$$CORR\left[\mathscr{E}^{X}_{\mathrm{N}(0,1)}(1), \mathscr{E}^{Y}_{\mathrm{N}(0,1)}(1)\right] = \rho_{\mathscr{E}}, \tag{13}$$

$$\begin{aligned} &CORR\left[\mathscr{E}^{X}_{\mathrm{N}(0,1-\exp\{-2[T(i)-T(i-1)]/\tau_X\})}(i), \mathscr{E}^{Y}_{\mathrm{N}(0,1-\exp\{-2[T(i)-T(i-1)]/\tau_Y\})}(i)\right] \\ &\quad = \left(1-\exp\{-[T(i)-T(i-1)]\cdot(1/\tau_X+1/\tau_Y)\}\right) \\ &\qquad \times\left(1-\exp\left\{-2\left[T(i)-T(i-1)\right]/\tau_X\right\}\right)^{-1/2} \\ &\qquad \times\left(1-\exp\left\{-2\left[T(i)-T(i-1)\right]/\tau_Y\right\}\right)^{-1/2}\rho_{\mathscr{E}}, \quad i=2,\ldots,n, \end{aligned}$$

$$\begin{aligned} &CORR\left[\mathscr{E}^{X}_{\mathrm{N}(0,\,1-\exp\{-2[T(i)-T(i-1)]/\tau_X\})}(i), \mathscr{E}^{Y}_{\mathrm{N}(0,\,1-\exp\{-2[T(j)-T(j-1)]/\tau_Y\})}(j)\right] = 0, \\ &\quad i,j=2,\ldots,n, \qquad i\neq j, \end{aligned}$$

$$\begin{aligned} &CORR\left[\mathscr{E}^{X}_{\mathrm{N}(0,\,1-\exp\{-2[T(i)-T(i-1)]/\tau_X\})}(i), \mathscr{E}^{Y}_{\mathrm{N}(0,\,1)}(1)\right] = 0, \\ &\qquad i=2,\ldots,n, \end{aligned}$$

$$\begin{aligned} &CORR\left[\mathscr{E}^{X}_{\mathrm{N}(0,\,1)}(1), \mathscr{E}^{Y}_{\mathrm{N}(0,\,1-\exp\{-2[T(i)-T(i-1)]/\tau_Y\})}(i),\right] = 0, \\ &\qquad i=2,\ldots,n. \end{aligned}$$

The process is strictly stationary with the properties

$$\mathrm{E}[X(i)] = \mathrm{E}[Y(i)] = 0, \tag{14}$$

$$\mathrm{VAR}[X(i)] = \mathrm{VAR}[Y(i)] = 1, \tag{15}$$

$$\mathrm{CORR}[X(i), Y(i)] = \rho_{XY} = \rho_{\mathscr{E}}. \tag{16}$$

References

Beran R (1987) Prepivoting to reduce level error of confidence sets. Biometrika 74(3):457–468

Carlstein E (1986) The use of subseries values for estimating the variance of a general statistic from a stationary sequence. Ann Stat 14(3):1171–1179

Cronin TM (2010) Paleoclimates: understanding climate change past and present. Columbia University Press, New York

DiCiccio TJ, Efron B (1996) Bootstrap confidence intervals. Stat Sci 11(3):189–212

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, New York

Gordon AL (1986) Interocean exchange of thermocline water. J Geophys Res 91(C4):5037–5046

Granger CW, Maasoumi E, Racine J (2004) A dependence metric for possibly nonlinear processes. J Time Ser Anal 25(5):649–669

Hall P (1986) On the bootstrap and confidence intervals. Ann Stat 14(4):1431–1452

Hall P, Martin MA (1988) On bootstrap resampling and iteration. Biometrika 75(4):661–671

Kendall MG (1954) Note on bias in the estimation of autocorrelation. Biometrika 41(3–4):403–404

Künsch HR (1989) The jackknife and the bootstrap for general stationary observations. Ann Stat 17(3):1217–1241

Liu RY, Singh K (1992) Moving blocks jackknife and bootstrap capture weak dependence. In: LePage R, Billard L (eds) Exploring the limits of bootstrap. Wiley, New York, pp 225–248

Loh W-Y (1987) Calibrating confidence coefficients. J Am Stat Assoc 82(397):155–162

Mudelsee M (2002) TAUEST: a computer program for estimating persistence in unevenly spaced weather/climate time series. Comput Geosci 28(1):69–72

Mudelsee M (2003) Estimating Pearson's correlation coefficient with bootstrap confidence interval from serially dependent time series. Math Geol 35(6):651–665

Mudelsee M (2010) Climate time series analysis: classical statistical and bootstrap methods. Springer, Dordrecht

Pearson K (1896) Mathematical contributions to the theory of evolution—III. Regression, heredity, and panmixia. Philos Trans R Soc Lond Ser A 187:253–318

Peeters FJC, Acheson R, Brummer G-JA, de Ruijter WPM, Schneider RR, Ganssen GM, Ufkes E, Kroon D (2004) Vigorous exchange between the Indian and Atlantic oceans at the end of the past five glacial periods. Nature 430(7000):661–665

Pisias NG, Mix AC, Zahn R (1990) Nonlinear response in the global climate system: evidence from benthic oxygen isotopic record in core RC13-110. Paleoceanography 5(2):147–160

Rehfeld K, Marwan N, Heitzig J, Kurths J (2011) Comparison of correlation analysis techniques for irregularly sampled time series. Nonlinear Process Geophys 18(3):389–404

Sherman M, Speed FM Jr, Speed FM (1998) Analysis of tidal data via the blockwise bootstrap. J Appl Stat 25(3):333–340

von Storch H, Zwiers FW (1999) Statistical analysis in climate research. Cambridge University Press, Cambridge