Nonlin. Processes Geophys., 18, 389–404, 2011
www.nonlin-processes-geophys.net/18/389/2011/
doi:10.5194/npg-18-389-2011

Nonlinear Processes in Geophysics

Comparison of correlation analysis techniques for irregularly sampled time series

K. Rehfeld[1,2], **N. Marwan**[1], **J. Heitzig**[1], **and J. Kurths**[1,2,3]

[1]Potsdam-Institute for Climate Impact Research, P.O. Box 60 12 03, 14412 Potsdam, Germany
[2]Department of Physics, Humboldt-University of Berlin, Newtonstr. 15, 12489 Berlin, Germany
[3]Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB243UE, UK

Received: 5 April 2011 – Revised: 27 May 2011 – Accepted: 10 June 2011 – Published: 23 June 2011

Abstract. Geoscientific measurements often provide time series with irregular time sampling, requiring either data reconstruction (interpolation) or sophisticated methods to handle irregular sampling. We compare the linear interpolation technique and different approaches for analyzing the correlation functions and persistence of irregularly sampled time series, as Lomb-Scargle Fourier transformation and kernel-based methods. In a thorough benchmark test we investigate the performance of these techniques.

All methods have comparable root mean square errors (RMSEs) for low skewness of the inter-observation time distribution. For high skewness, very irregular data, interpolation bias and RMSE increase strongly. We find a 40 % lower RMSE for the lag-1 autocorrelation function (ACF) for the Gaussian kernel method vs. the linear interpolation scheme,in the analysis of highly irregular time series. For the cross correlation function (CCF) the RMSE is then lower by 60 %. The application of the Lomb-Scargle technique gave results comparable to the kernel methods for the univariate, but poorer results in the bivariate case. Especially the high-frequency components of the signal, where classical methods show a strong bias in ACF and CCF magnitude, are preserved when using the kernel methods.

We illustrate the performances of interpolation vs. Gaussian kernel method by applying both to paleo-data from four locations, reflecting late Holocene Asian monsoon variability as derived from speleothem δ^{18}O measurements. Cross correlation results are similar for both methods, which we attribute to the long time scales of the common variability. The persistence time (memory) is strongly overestimated when using the standard, interpolation-based, approach. Hence, the Gaussian kernel is a reliable and more robust estimator with significant advantages compared to other techniques and suitable for large scale application to paleo-data.

Correspondence to: K. Rehfeld
(rehfeld@pik-potsdam.de)

1 Introduction

Paleoclimate proxy data sample past regional and global climate variation. Through their analysis we can attempt to understand past environmental conditions and changes. In order to separate local from global effects, measures of association like linear correlation and cross spectral density estimation are traditionally employed to analyze these records. A crucial problem with these records it their irregular sampling in time due to the complex sedimentation/ accumulation rate. However, standard methods can not be applied when timescales and resolutions are different. This is not a problem in the geosciences only, as irregular observation of continuous-time processes also occurs in the detection of biomedical rhythms (Schimmel, 2001), astronomy (Edelson and Krolik, 1988; Scargle, 1981, 1982, 1989) or turbulence research, where the velocity of the flow can only be measured if seeding particles pass a measurement volume (Broersen et al., 2000; Harteveld et al., 2005). When our aim is to reconstruct the linear auto- or mutual dependencies of the underlying processes from the observations, we can estimate either (cross-) power spectra or correlation functions, as both are related to each other by the Fourier transform (Chatfield,

2004). The irregular sampling of the time series makes direct use of the standard estimation techniques of association measures impossible, as they rely on regular observation times. For (cross-) power spectral density estimation, standard linear interpolation of these irregular observations onto a regular sampling causes an additional bias towards low frequencies in power spectral density (PSD) estimation (Schulz and Stattegger, 1997).

Historically, there are several approaches to overcome this problem. The concepts can be classified into four categories: (a) direct transform methods, (b) slotting techniques, (c) model-based estimators, and (d) time series reconstruction methods (Broersen et al., 2000).

The Lomb-Scargle (LS) periodogram, introduced for use in astronomy (Scargle, 1981, 1982), is a well-known direct transform method that computes a least squares fit of sine curves to the data. The obtained least squares spectrum detects peaks at high frequencies but turned out to be severely biased for turbulence spectra (Broersen et al., 2000) which do not possess periodic components. If the underlying assumption of least squares optimization, that the noise in the data is normally distributed, is fulfilled, then LS is equivalent to the Maximum-Likelihood estimate. Like all least squares techniques, the estimator is not robust in the presence of outliers. This is illustrated by the limitations of the method in the application to bimodal rhythms and signals with isolated outliers (Schimmel, 2001).

Standard slotting techniques determine the correlation function by binning all available products in the lag domain, so that observations only contribute to the correlation function at a lag if their observation time difference deviates less than half the lag bin width from the considered lag. This technique was proposed by Mayo in 1978 and further elaborated by Edelson and Krolik (1988). It has become popular in velocimetry (Broersen et al., 2000) and is frequently applied in astronomy (Böttcher and Dermer, 2010; Fan et al., 2010; Nieppola et al., 2009; Zhang et al., 2010). The disadvantage of this technique is that, without post-processing, the correlation function estimates are not necessarily positive semidefinite and the spectra computed from their Fourier transform can show negative power. Stoica et al. (2008), therefore, proposed a weighting technique for autocorrelation estimation which weighed observations based on a sinc kernel and claimed that it yielded positive semidefinite results. In their review, Babu and Stoica (2009) also showed the application of other kernels in the time domain, including Laplacian and Gaussian kernels. The distribution of sampling time errors in time series reconstruction from paleo-archives is often assumed to be Gaussian, which, we believe, intuitively supports its use in time domain analysis. Mudelsee (2010) proposed two techniques to estimate the correlation coefficient that he terms "binned correlation" and "synchrony correlation". "Synchrony correlation" consists of using the percentage of pairs of observations in the different time series that have the smallest measurement time difference, treat them as if they were observed coevally and calculate the correlation coefficient. "Binned correlation" essentially resamples the data into time bins on a regular grid that are assigned the mean values of the observations within these bins. Using these regular, reconstructed time series, the standard correlation estimator can be applied. We do not employ these two techniques because both do not utilize all available observations individually, which means loss of information. Also, since the standard estimator is used for calculation of the correlation coefficient, binning – or resampling – is problematic when data gaps are present and we want to estimate the correlation function.

Model-based estimators fit a model to the time series, the spectra or the ACF, which requires prior knowledge about the actual process (cf. Harteveld et al., 2005 and references therein), a prerequisite we typically cannot meet due to the heterogeneity and complexity of geophysical processes.

The fourth group of estimators resamples the data (through some kind of interpolation) in order to create time series on a regularly spaced grid, which then can be analyzed using the standard FFT-based estimators. The most frequently used technique in geophysical time series analysis is linear interpolation. Paleo data often has rather large data gaps and it is controversial if, when and how missing observations can be appropriately approximated. For standard interpolation (e.g. linear, akima-spline and cubic-spline) a significant reduction in variance toward the high-frequency range of the estimated power spectrum occurs in the analysis of irregularly sampled data (Schulz and Stattegger, 1997). When we are interested in phenomena on short timescales (compared to the mean sampling interval), such effects should be considered, and if possible, avoided.

Without objective performance tests of these estimators, application of specific methods is a matter of taste, but the chosen routine may not be the optimal method available. Therefore benchmark tests comparing various methods are crucial. One study, conducted for the estimation of power spectral density from flow velocimetry data in an engineering background, has been performed by Benedict et al. (1998). The test cases exhibited flat or simple exponentially decreasing spectra or contained a single deterministic sinusoidal component. They are therefore not nearly as complex as spectra in geophysical time series analysis typically are. Furthermore, they used a Poisson sampling scheme, which is reasonable in measurements with detector dead time, but less justified for paleo records.

In this paper we first review the methods that are or could reasonably be applied in the estimation of correlation functions of geophysical time series. This encompasses the standard approach, re-sampling by means of (linear) interpolation followed by a FFT-based routine, the LS periodogram, the slotting technique and kernel-weighted estimators.

We then – for the first time to our knowledge – compare and evaluate systematically the performance of methods suitable for estimating correlation functions of geophysical time

series under the presence of varying sampling schemes, and we specifically quantify the extent and direction of estimator variance and bias due to sampling irregularity. We do this using a newly developed testing scheme, based on simulated time series with increasing inter-sampling time irregularity but constant mean sampling rate. In a last step we apply the methods to real proxy data from the Asian summer monsoon region, we evaluate the consistency of the results with respect to the synthetic tests and validate our ACF results further by the application of an independent least squares estimator for the persistence time of autoregressive processes of order 1 (AR(1)) (Mudelsee, 2002).

2 Methods

Assuming that two time series x_t and y_t were observed from stationary stochastic processes at unit time intervals, their sample CCF $\hat{\rho}(k)$ gives an estimate of the strength of a possible linear association between the processes behind the observations at each possible lag number k. It is defined as

$$\hat{\rho}_{xy}(k) = \hat{\rho}_{xy}(k\Delta\tau) = \hat{\gamma}_{xy}(k)/\hat{\sigma}_x\hat{\sigma}_y \quad (1)$$

$$= \frac{1}{\hat{\sigma}_x\hat{\sigma}_y(N-k)} \sum_{t=1}^{N-k} (x_t - \bar{x})(y_{t+k} - \bar{y}) \; . \quad (2)$$

Here, $\hat{\gamma}_{xy}(k)$ is the sample cross-covariance at lag k, N is the number of observations, $\hat{\sigma}_x$, $\hat{\sigma}_y$ the sample standard deviations of the processes and $\bar{x}$, $\bar{y}$ are the estimated mean values of the time series (Chatfield, 2004). The spacing of the CCF lags, $\Delta\tau$, equals – in this standard definition – that of the time series x_t and y_t, $\Delta\tau = t_i^{x,y} - t_{i+1}^{x,y}$.

The discrete Fourier transform of the sample CCF is the sample cross spectral density function or cross spectrum and vice versa. The power spectrum can thus be estimated in two ways, either by computing the discrete Fourier transforms of the input time series and multiplying them after complex-conjugating one of them, or by estimating the CCF and Fourier transforming it (cf. Chatfield, 2004 for more details). We denote all estimators in the definitions in their respective sections by $\hat{\rho}$, for the sake of simplicity.

2.1 The resampling approach for irregular time series

In the case of irregularly sampled time series, the classical definition, as illustrated in Fig. 1a, can not be readily applied. An irregularly spaced time series is a pair (t^x, x) of tuples of common length N^x, where $t_1^x < t_2^x < \ldots < t_{N^x}^x$ are the time points and x_i is the value at time t_i^x. For simplicity we have transformed the time variable to get a normalized mean increment of 1 by dividing by the mean sampling period: $t_i^x = t_i^{\mathrm{orig}}/\Delta t^x$ and we will use this notation in the following. The differences between observation times $\Delta t_i^x = t_i^x - t_{i-1}^x$ are not any more constant and the mean of their distribution is the mean sampling time Δt^x. When we consider irregularly sampled time series (t^x, x), (t^y, y) of second-order stationary processes with zero mean, these have to be resampled onto a common regular time grid $(t^{x,y})$ with constant time increments $t^{x,y}(n) - t^{x,y}(n-1) = \Delta t_x$ for all $n = 1, 2, \ldots N^{x,y}$. The grid spacing we will use is the larger of the mean sampling intervals of the time series.

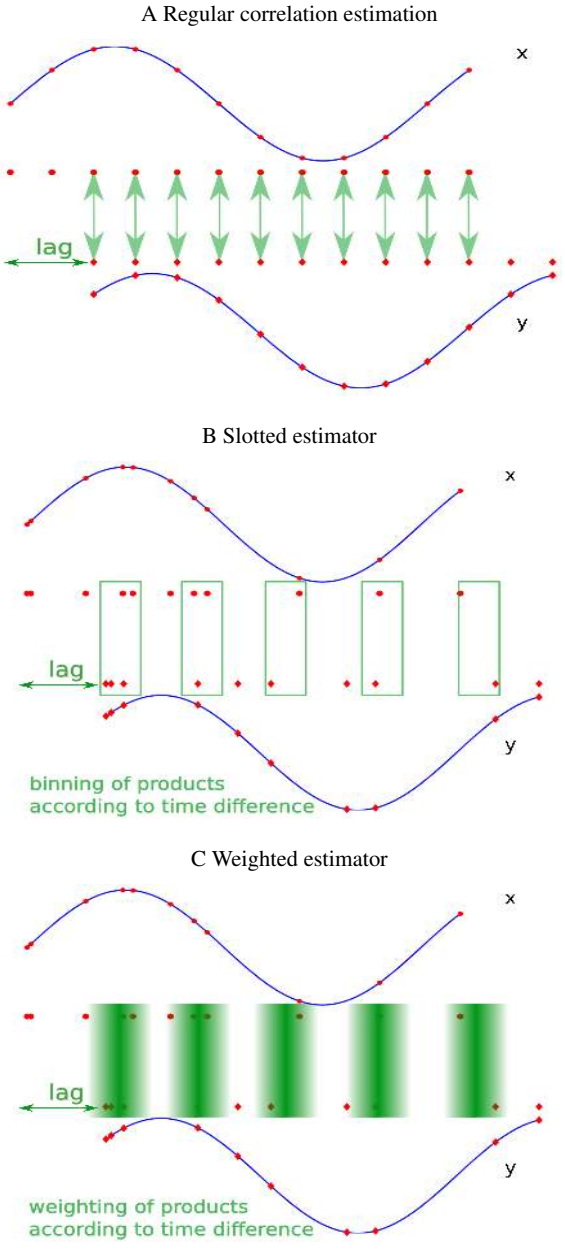

Fig. 1. Principles of correlation function estimation: **(A)** shows the classical estimator, where the correlation $\hat{\rho}_{xy}(k)$ is given by a mean over products of zero-mean observations a lag k apart. **(B)** For irregularly sampled time series, the slotted estimator computes $\hat{\rho}_{xy}(k)$ as the mean over all products in bins whose centers are a lag k apart. **(C)** Non-rectangular correlation uses the weighted mean over all available products with the weight maxima a lag k apart.

We restrict ourselves in this analysis to the linear interpolation technique, as the effects of other standard routines are not much different in their variance reduction towards the high-frequency end of the spectrum (Schulz and Stattegger, 1997). A resampling method which does not result in a reduction in variance is the *nearest neighbor technique*, where the function is approximated at the desired grid points by the value of the observation closest in time. This leads to a shifting bias (Broersen, 2009) which, in the presence of large gaps in the data, can be rather large. We therefore do not employ this scheme. After resampling, the standard FFT-based routines can be employed.

2.2 Lomb-Scargle approach

The Lomb-Scargle approach to the spectral estimation of irregularly sampled data can be understood as a least squares fitting of sinusoids to data (Scargle, 1981). The Lomb-Scargle Fourier transform (LSFT)

$$\mathrm{LSFT}_x(\omega) = F_0(\omega)\sum_{i=1}^{N^x}(Ax_i \ \cos\,\omega\,\hat{t}_i^x + iBx_i \ \sin\,\omega\,\hat{t}_i^x), \quad (3)$$

uses the explicit observation times $\hat{t}_i^x = t_i^x - \tau^x(\omega)$ shifted by the constant (complex) phase shift

$$\tau^x(\omega) = \frac{1}{2\,\omega}\tan^{-1}\left(\sum_i \ \sin 2\,\omega t_i^x \Big/ \sum_i \ \cos 2\,\omega t_i^x\right), \quad (4)$$

to ensure time invariance of the LSFT (Schulz and Stattegger, 1997). The coefficient F_0

$$F_0(\omega) = \frac{1}{\sqrt{2}}\exp(-i\ \omega t_1^x - \tau^x(\omega)) \quad (5)$$

allows for a time shift in the alignment of the two time series in bivariate spectral analysis. The amplitudes A and B are defined as

$$A(\omega) = \left(\sum_i \cos^2\,\omega\hat{t}_i^x\right)^{-1/2}, \quad B(\omega) = \left(\sum_i \sin^2\,\omega\hat{t}_i^x\right)^{-1/2}. \quad (6)$$

In the univariate case, the well-known Lomb-Scargle periodogram is then given by

$$\hat{P}_x(\omega) = \mathrm{LSFT}_x(\omega)\mathrm{LSFT}_x^*(\omega) \quad (7)$$

The (bivariate) cross spectrum can be estimated as

$$\hat{P}_{xy}(\omega) = \mathrm{LSFT}_x(\omega)\mathrm{LSFT}_y^*(\omega) \quad (8)$$

which can be inverted, using the Fourier transform (Scargle, 1989), to get the cross correlation coefficient estimate

$$\hat{\rho}_{xy}(k) = \mathfrak{F}^{-1}[\hat{P}_{xy}(\omega)]. \quad (9)$$

The squared absolute value of the LSFT gives the widely known and used LS periodogram (Schulz and Stattegger, 1997). The choice of the frequencies ω is described in Scargle (1989) and we adopt the recommended values for the fundamental frequency $\omega_0 = \omega_{\min} = \frac{\pi(N^{xy}-1)}{(t_{\max}-t_{\min})N^{xy}}$ and maximum frequency $\omega_{\max} = \frac{2\pi}{\Delta t^{xy}}$. In the bivariate case we define the observation times $t_{\min}$ and $t_{\max}$ as the lower and upper bounds of the overlapping part of both time series x_t and y_t, otherwise, in the univariate case, minimum and maximum observation time are used. $\Delta t^{xy} = \max(\Delta t^x, \Delta t^y)$ is the common sampling rate we define in the bivariate case. The number of frequencies $N_\mathrm{f} = \mathrm{ofac}\cdot N^{xy}$ determines the spacing of the frequency vector. According to Hocke and Kämpfer (2009) there is no principal limit, the oversampling factor $\mathrm{ofac} > 1$ is regarded as a smoothing factor, although the number of independent frequencies is constant. We use $\mathrm{ofac} = 2$, unless otherwise stated.

A thorough introduction to bivariate Lomb-Scargle spectral estimation was given by Schulz and Stattegger (1997). The use of the technique for correlation function estimation, however, has not yet been explored, though it was already proposed in Scargle (1989).

2.3 Correlation slotting

The sample correlation function $\hat{\rho}_{xy}(k)$ at a lag k is calculated by averaging over the lagged products of the standardized observations. For irregular time series the inter-sampling times vary, and without resampling Eq. (1) cannot be applied. An alternative is the *slotting* or *Edelson and Krolik* technique (Edelson and Krolik, 1988; Mayo, 1993). Its key idea is to calculate the cross-products of all available, standardized, observations and discretize them into bins according to their sampling time differences as can be seen in Fig. 1b. The technique was developed in fluid mechanics and applied in astrophysics. $\hat{\rho}(k\Delta\tau)$ at the lag $k\Delta\tau$ is then defined as

$$\hat{\rho}(k\cdot\Delta\tau) = \frac{\sum_{i=1}^{N^x}\sum_{j=1}^{N^y} x_i\, y_j\, b_k(t_j^y - t_i^x)}{\sum_{i=1}^{N}\sum_{j=1}^{N} b_k(t_j^y - t_i^x)} \quad (10)$$

and the *kernel* $b_k(t_j^y - t_i^x)$ selects the products whose time lag is not further than half the bin width from $k\Delta\tau$:

$$b_k(t_i - t_j) = \begin{cases} 1 & \text{for} |(t_j - t_i)| - k| < \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

Note that the observations have to be standardized to zero mean and unit variance before the analysis. We set the lag bin width $\Delta\tau$ to be equal to Δt^{xy}, and since we divide the observation times by this mean sampling interval, we can omit it in the formulae above, for easier readability (cf. Sect. 2.1). We do not choose this width arbitrarily but rather in the context of the desired time resolution of the CCF, more on this in Sect. 2.4.

There are, however, several disadvantages of this technique, primarily a high variance of the estimator (Babu and Stoica, 2009; Benedict et al., 2000; Harteveld et al., 2005)

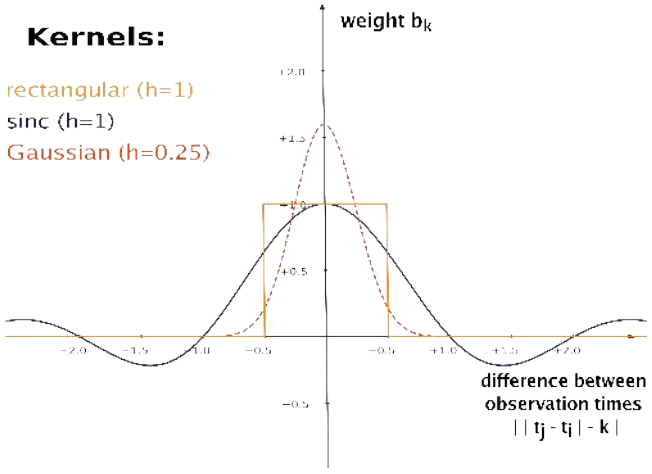

Fig. 2. Kernel-based estimators effectively "use" observations whose inter-sampling time difference is close to the lag for which linear correlation is estimated. Slotting (the rectangular kernel) chooses observations within an interval, Gaussian and sinc kernel weigh the products smoothly according to the difference between observation interval and desired lag. Kernels were scaled to the standard choice for width parameter h (cf. Table 1, Fig. 3).

due to which we will not use this method in the following, but rather apply related, non-rectangular kernels. It also does not always provide positive semidefinite covariance matrix estimates, a problem which can be overcome by "fourier filtering". We discuss this further in Sect. 2.5.

2.4 Non-rectangular kernels

In analogy to the slotting approach, and taking it further, weighted averaging of the observations can be performed using symmetric, smooth density functions that tend to zero for time differences much larger or smaller than the desired lag k (Hall et al., 1994). The similarity is illustrated in Fig. 1c. These requirements are for example met by the sinc kernel (Stoica and Sandgren, 2006) but also the Gaussian kernel (cf. Table 1) as can be seen in Fig. 2. Instead of binning the observations into discrete sets, the weights prevent a sudden cutoff in the time domain.

There is no theoretical definition of the effective width of the weight functions. We decide to scale them to a kernel width of the mean sampling rate for two reasons. (i) This choice ensures that – for non-rectangular kernels – observations at (near-)regular times are rated higher than those that are further away, but are still included in cases where little information is available. (ii) In a trade-off between the loss of resolution and control of estimator variance, the desired resolution of the correlation function also plays a role, as a kernel width choice larger than the lag spacing would result in mixing information for adjacent lags. The width parameters for the kernels and their relation to the mean sampling rate were confirmed as empirical optima in case of irregular

Table 1. Kernels $b(d)$ used in this paper. d denotes the distance between the inter-observation time Δt_{ij}^{xy} and $k\Delta\tau$, k denotes the k-th lag. The standard width parameter h is chosen to result in a main lobe width of Δt^{xy}, the mean sampling interval or common sampling period in the bivariate case.

Kernel (reference)	$b(k-\Delta t_{ij}^{xy}) = b(d)$	Standard choice for h
Rectangle; Edelson and Krolik (1988)	$\begin{cases} 1 & \text{if } d \le h/2, \\ 0 & \text{if otherwise.} \end{cases}$	$\Delta t^{xy}/2$
Sinc; Stoica and Sandgren (2006)	$\frac{1}{N}\frac{\sin(\pi h d)}{\pi h d}$	Δt^{xy}
Gaussian; Bjoernstad and Falck (2001)	$\frac{1}{\sqrt{2\pi}h} e^{-\lvert d\rvert^2/2h^2}$	$\Delta t^{xy}/4$

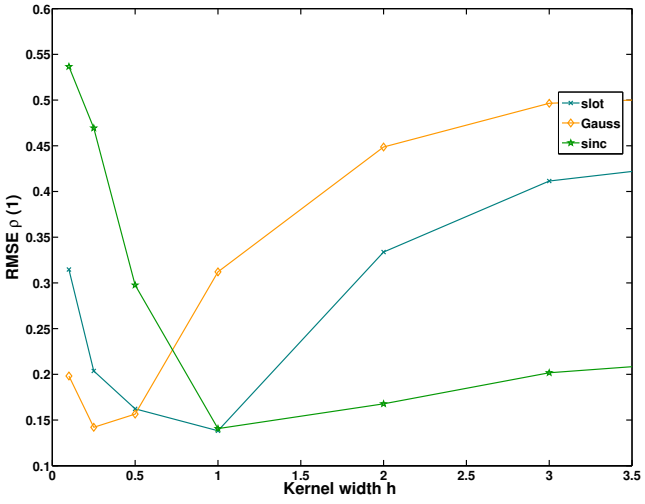

Fig. 3. Influence of varying kernel width h on the RMSE of $\rho(1)$, using the kernel estimators (cf. Table 1, Fig. 2). 100 Realizations of sinusoids with random phase in colored noise (30 %) were sampled using $\Gamma-$distributed sampling intervals (sk = 2.85). (cf. Sect. 3.3).

time series (cf. Fig. 3). Other parameter choices might, however, also be sensible, depending on the nature of the time series and the statistic to be estimated.

2.5 Positive semidefiniteness of the estimated function

In connection with the slotting-based covariance estimation, the issue with the possible lack of positive semidefiniteness of the correlation estimates has been discussed in Broersen (2002), Harteveld et al. (2005) and Stoica and Sandgren (2006). By Bochner's theorem, positive semidefiniteness of the correlation function is necessary and sufficient to ensure non-negativity of the Fourier transform estimate of $\hat{\rho}(t)$. A function $\hat{\rho}(k)$ is positive semidefinite if

$$\int\int \hat{\rho}(l-t)w(t)w(l)dt\ dl \geq 0 \tag{12}$$

for all integrable functions w, and only if this holds true $\hat{\rho}(k)$ is a possible correlation function. For discrete, short, and regularly sampled time series, using Eq. (10) and a simple, integrable function for w, we can find this condition violated for all kernel methods. This problem can, amongst others, be solved by a technique called "Fourier filtering", which involves Fourier-transforming the correlation function estimate, setting any negative power estimates to zero and applying an inverse FFT afterwards to obtain a positive semidefinite correlation function estimate (Babu and Stoica, 2009; Hall et al., 1994). Another routine could involve using the absolute value of the power spectrum, instead of setting negative estimates to zero. Also, positive semidefinite matrices have non-negative eigenvalues, which is another means to test this property, and the same modifications as for the power spectra could be applied here. It should be kept in mind, however, that, due to numerical problems, even the "unbiased" $1/(N-1)$ correlation estimator can result in negative power estimates. When the positive semidefiniteness of the correlation matrix is essential, Fourier filtering should be performed and/or the eigenvalues of the matrix should be checked.

2.6 Quality of performance measures

Our aim is to evaluate which of the approaches listed above yields the best results for the estimation of ACFs and CCFs for geophysical time series. The performance of the estimators can be evaluated with respect to the "true" *expected functions*. This can of course only be done for modeled or synthetic time series where we can calculate ACFs and CCFs exactly.

To evaluate the different estimators we calculate the *root mean square error* (RMSE) of the estimator $\hat{\theta}$ for a statistic θ. θ can be e.g. the cross correlation function at lag k, $\rho_{xy}(k)$. The RMSE is given by

$$\mathrm{RMSE}(\hat{\theta}) = \sqrt{E[(\hat{\theta}-\theta)^2]} = \sqrt{\mathrm{var}(\hat{\theta}) + \mathrm{bias}(\hat{\theta})^2} \tag{13}$$

and incorporates both variance and bias of the estimator, i.e. its variability and its systematic deviation from the true value. To estimate the RMSE we generate a large number of time series of a given signal type and sampling scheme and compute the "target statistic" $\hat{\theta}$ for each. The deviation between the mean of these many estimates and the 'true' function is the approximate *bias* of the estimator, together with the variance around this mean we can estimate the RMSE.

To evaluate the contribution of the sampling irregularity to the estimation error, we perform the analysis for different sampling schemes, first for regular sampling and then for more and more irregular sampling. This we do by drawing inter-sampling-time intervals from the Gamma-distribution and concatenating them into a time line for which we then generate a corresponding signal. Given the shape parameter α and the scale parameter β, the mean μ of the $\Gamma(\alpha,\beta)$-distribution is given by $\mu = \alpha\beta$, the variance by $\sigma^2 = \alpha\beta^2$ and the skewness by $\mathrm{sk} = 2/\sqrt{\alpha}$. For low skewness (in our case the lowest value was 0.1) the distribution is close to normal (cf. Fig. 7b). Since the higher order moments depend only on the shape parameter α, we can vary the scale parameter β in a way to keep the mean constant while increasing skewness and variance. We will only give the skewness parameter in the following, as the variance $\sigma^2 = (2\beta/\mathrm{sk})^2 = (2\mu/(\mathrm{sk}\ \alpha))^2$ is uniquely determined in our parameter configuration. A distribution with a skewness of 2.85 (Fig. 7b) results in a time series with large gaps, as large values become more likely in more and more skewed sampling interval distributions.

3 Comparison for synthetic records

To assess the adaptability and suitability of the different estimators, we perform a number of tests on artificially generated discrete signals for which we know the "true" ACFs and/or CCFs of the underlying processes. For each signal type we first estimate the RMSE in the case of regular sampling. Then we create time series with Γ-distributed inter-observation times with increasing skewness. Since the time vectors are artificial, they do not need to have an actual unit, but we assume that time is measured in years.

3.1 Sinusoids with random phase

Using techniques that are not (yet) fully established, our first concern is to make sure that the results for the standard, regularly sampled case are consistent with those from the standard estimators. Therefore we sample a simple signal, a superposition of three sinusoids:

$$x(t) = \sum_{i=1}^{3} \sin(\omega_i t + \Theta_{i,n}) \tag{14}$$

with $\omega_i = \frac{2\pi}{T_i}$, $T_i = (18, 21, 41)$ yr at a regular rate of 1/4 years. The phase variable $\Theta_{i,n}$ is randomly drawn from a uniform distribution on $(0, 2\pi)$, making this a sample from a stationary stochastic process. The true ACF is then a superposition of cosine functions $\rho_{xx} = 1/2\sum_{i=1}^{3} \cos(\omega_i)$, irrespective of the relative phases of the signal components. The length of the simulated time series is 1000 yr and we evaluate the function for 200 lags. Sample time series, mean ACF and power spectral density (PSD) of the mean ACF are depicted in Fig. 4. The kernel estimators, the LS periodogram as well as the "classical" method perform comparably well with a RMSE below 2 % (see Fig. 5, left columns) in the regularly sampled case.

We now use irregularly sampled observation times and perform a stepwise increase in sampling distribution

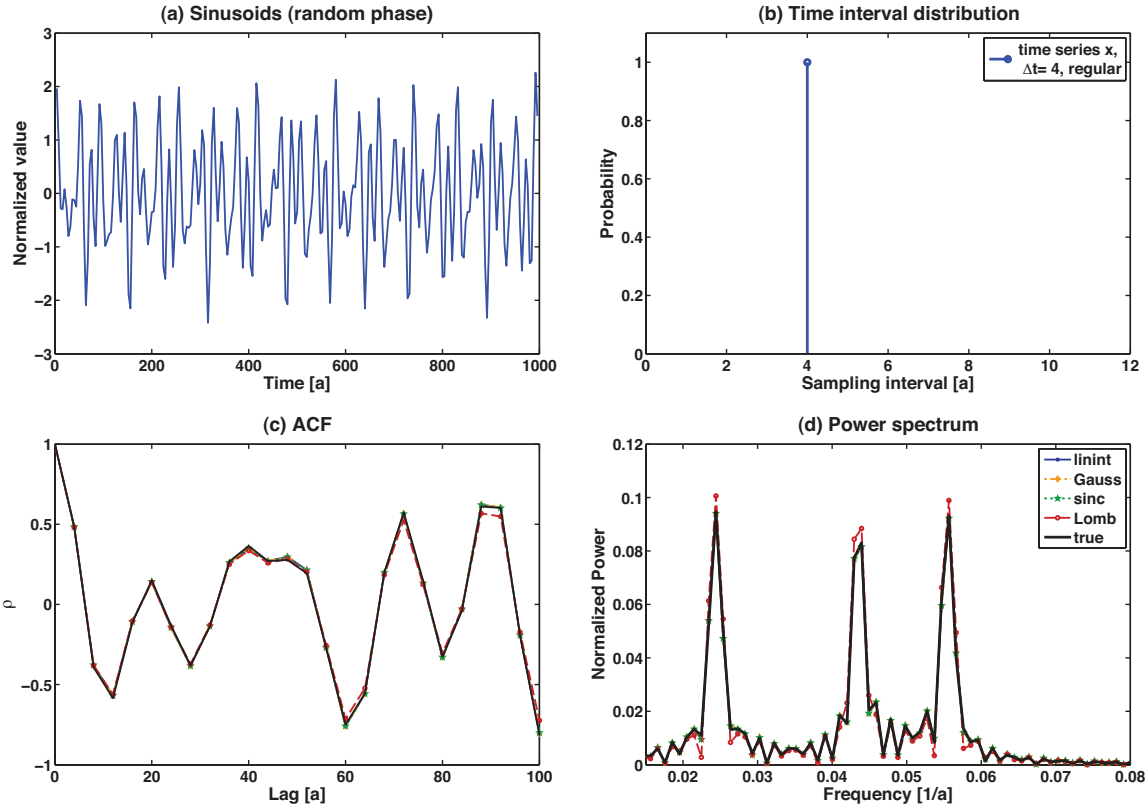

Fig. 4. Autocorrelation analysis of synthetic signals: for a regularly sampled combination of sinusoids (cf. Eq. 14) we give a sample time series **(a)**, the sampling interval probability density **(b)**, the expected correlation function **(c)** and the corresponding power spectrum **(d)** determined from 100 realizations of sinusoid time series with random phase arguments. Legends for each row are given in the right panels. All estimators perform equally well.

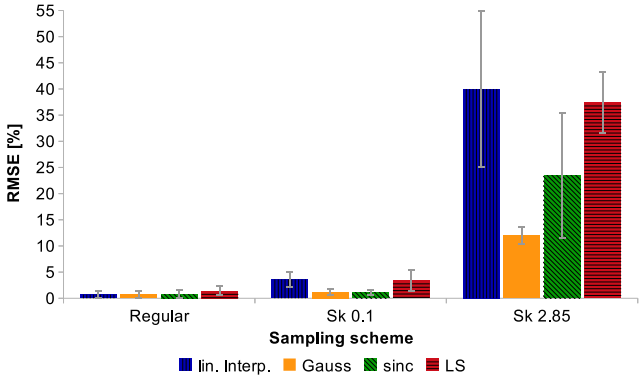

Fig. 5. Mean RMSE for the ACF estimation (lags 1–3) using linear interpolation, Gaussian or sinc kernel or the inversion of the Lomb-Scargle periodogram of noise-free sinusoids given for regular, gamma-distributed and mildly irregular (skewness sk = 0.1) resp. very irregular (sk = 2.85) sampling. Errorbars give the standard deviation of the estimate, calculated using 1000 bootstrap iterations.

skewness (as described in Sect. 2.6). For skewness sk = 0.1 the RMSEs are only slightly higher (Fig. 5, middle columns), but for a skewness sk = 2.85 the RMSE is as high as 40 % for interpolation and 35 % for the LS method. The estimated RMSE for the Gaussian kernel method, is rather small compared to that, with an approximate 12 %, lower than that of the sinc kernel method (23.5 %). We have increased the skewness in steps of 0.25 from sk = 0.1 to sk = 2.85 and note that the RMSE of the ACF seems to be increasing almost linearly for all the methods. For the LS estimate it jumps in the beginning, from 5 % to ≈ 20 %, and continues to increase at a rate of 9 % per unit skewness, with the breakpoint occurring at a skewness of 0.35. The RMSE of the interpolation followed by the FFT-based estimator (denoted "linint" in the figure legends) increases at a faster overall rate than all the other methods (6.5 % per unit skewness). The Gaussian kernel method has the lowest RMSE at high skewness and the lowest increase with respect to the estimate for regular sampling.

To investigate the reason for the differences between the methods further, we evaluate the RMSE of the power spectra obtained from the Fourier-transformed ACFs at the highest

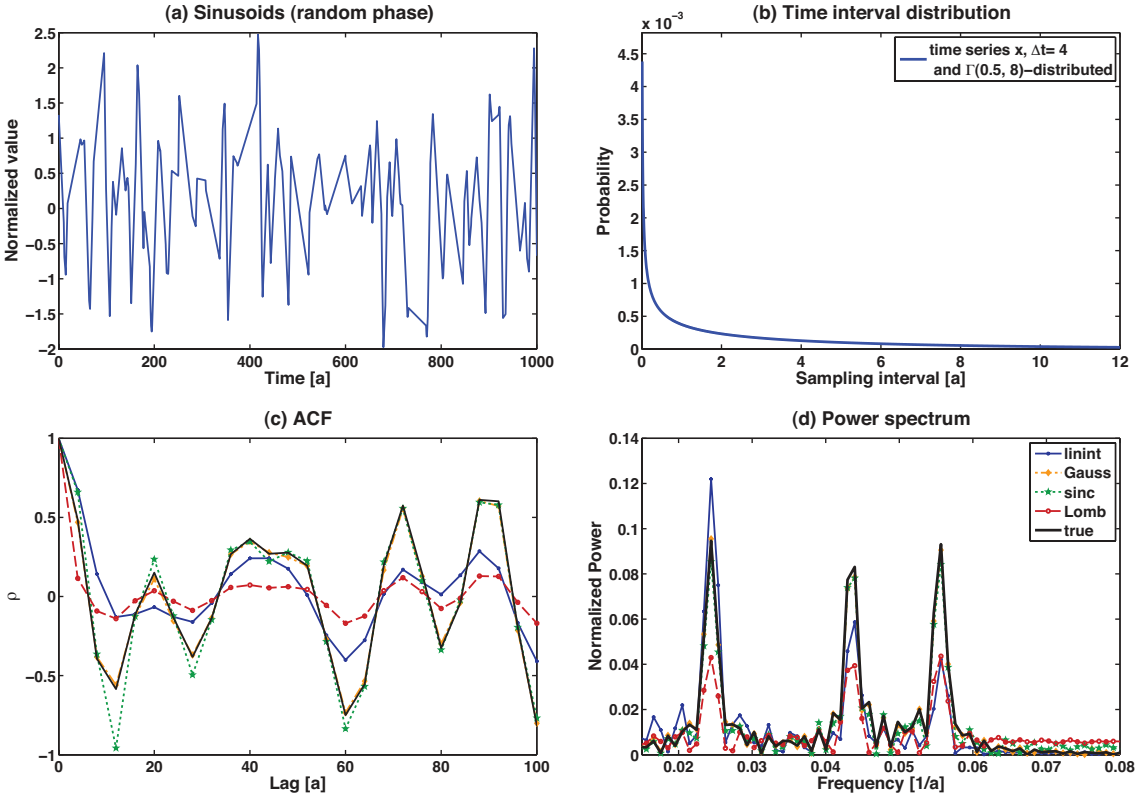

Fig. 6. Autocorrelation analysis of synthetic signals: for an irregularly sampled combination of sinusoids (cf. Eq. 14) we give a sample time series **(a)**, the sampling interval probability density **(b)**, the expected correlation function **(c)** and the corresponding power spectrum **(d)** determined from 100 realizations of sinusoid time series with random phase arguments. Legends for each row are given in the right panels. High sampling irregularity leads to a variance reduction in the ACF for LS and interpolation.

input signal frequency $\omega = 2\pi/18$ (c.f. Figs. 4d and 6d). We find, that with increasing skewness, the RMSE of this peak increases from around 3 % to 10 % for interpolation and the LS correlation function estimate, while for sinc and Gaussian kernel it goes from < 1 % to approximately 2 %. Estimating the bias of this peak, we observe that the comparatively high RMSE for interpolation and LS method corresponds to a negative bias increasing linearly from 5 % to > 50 % with respect to the expected peak power at the high-frequency component. In contrast to that, the bias is nearly constant for the kernel methods, the slight increase in RMSE must therefore be due to an increase in variance. This lack of power in the high frequency component of the estimated spectrum is accompanied by a positive bias for the lowest frequency component $\omega = 2\pi/100$ (results not shown).

3.2 Autoregressive processes

To understand the quantitative and qualitative effect of the different estimation techniques on the short-term correlative properties (e.g. the persistence time, the lag at which the ACF has dropped to $\Delta t/e$), we use AR(1) processes generated at high time resolution and then re-sample the observations onto the desired irregular sampling times. We perform the same simulations as before, first evaluating for regular sampling and then, for gamma-distributed inter-sampling intervals, where we subsequently increase the skewness of the interval distribution. The driving process is given by

$$X(t_i) = \phi X(t_{i-1}) + \xi_i = e^{-\Delta t/\tau} X(t_{i-1}) + \xi_i \tag{15}$$

and for bivariate correlation analysis we sample a second process driven by the first at lag ℓ:

$$Y(t_i) = \alpha X(t_{i-\ell}) + \varepsilon_i \ . \tag{16}$$

ξ and ε are uncorrelated Gaussian distributed noise processes with a variance σ_ξ^2, σ_ε^2 such that the overall process variances $\sigma_x^2 = \sigma_\xi^2/(1-\phi^2)$ and $\sigma_y^2 = \sigma_\varepsilon^2 + (\alpha^2\sigma_x^2)$. We choose the AR(1) coefficient as $\phi = 0.7$, corresponding to a persistence time $\tau = -\Delta t/\ln\phi$, the coupling strength $\alpha = 0.5$, coupling lag $\ell = 1$ and generate our time series (e.g. Fig. 7a) following the different sampling schemes (e.g. Fig. 7b). Then we set out to estimate ϕ and α from the time series.

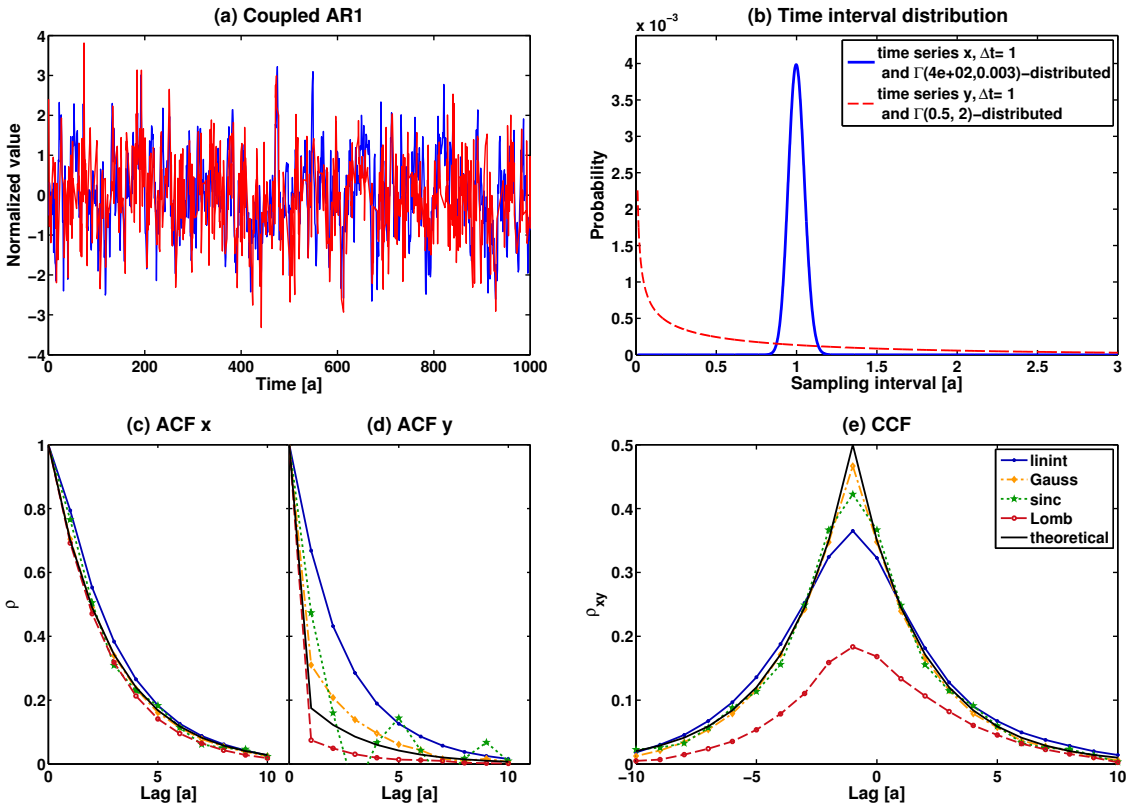

Fig. 7. Cross correlation analysis for two irregularly sampled signals (cf. Eqs. 15, 16) from different sampling schemes: Sample time series **(a)** and sampling time interval histograms **(b)**, the mean ACFs out of 100 realizations **(c)** and the mean estimated CCF **(d)**. Legends for each row are given in the right panels. A positive bias in interpolation ACF estimates and a negative bias in the interpolation and LS CCF estimates is observable for increased sampling irregularity.

In the estimation of the AR(1) coefficient ϕ, the RMSE for interpolation increases from 2 % to 17 % and the error for the sinc-kernel increases from 6 % to 13.5 %. The LS technique results in the largest increase for high skewness with a RMSE of 52 %. The Gaussian kernel method remains more accurate with an increase from 2 % to 8.5 %.

The coupling strength α is the true value of the CCF at the coupling lag ℓ. A typical application in the geoscience context is the estimation of the degree of similarity for time series from different sources, with different sampling properties. Analyzing two time series of inter-sampling time distribution skewnesses sk_x =0.1 sk_y =2.85, we find that the CCF estimation at lag $\ell = -1$ has a negative bias for all techniques. The bias of the LS technique is strongly negative, underestimating the true correlation by more than 65 %. Linear interpolation results in a 30 % lower estimated coupling strength, the sinc kernel method in 15 % and the Gaussian kernel estimate is negatively biased by 8 % with respect to the "true" coupling strength of 0.5 (Fig. 7e).

Looking at the performance under the increasing sampling time distribution skewness of time series y_t (keeping sk_x constant at 0.1), we find that the RMSE of the estimated α

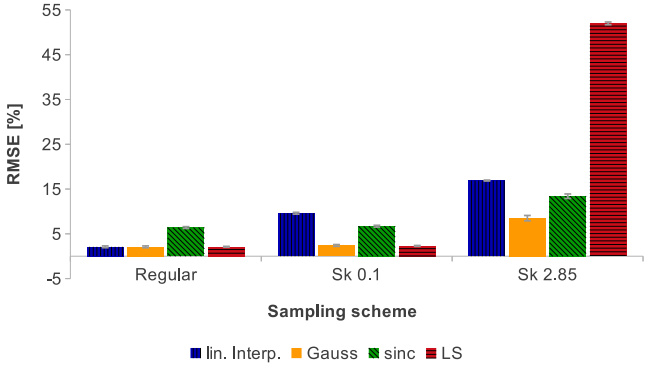

Fig. 8. Mean RMSE for the ACF estimation (lag 1) using linear interpolation, Gaussian or sinc kernel or the inversion of the LS Periodogram of time series from AR(1) processes (cf. Eqs. 15), given for regular, gamma-distributed and mildly irregular (skewness sk = 0.1) resp. very irregular (sk = 2.85) sampling. Errorbars give the standard deviation of the estimate, calculated using 1000 bootstrap iterations.

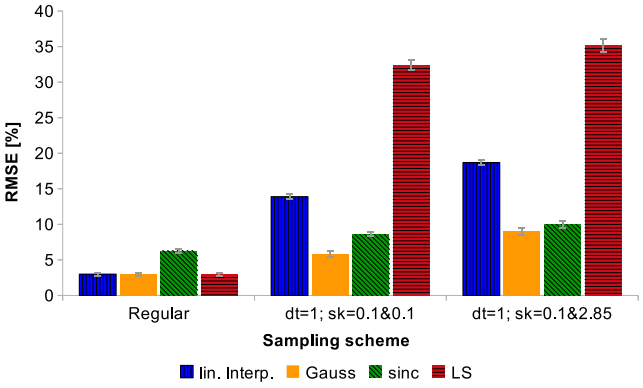

Fig. 9. RMSE for the CCF estimation (at the lag of coupling) using linear interpolation, Gaussian or sinc kernel or the inversion of the LS Periodogram of time series from coupled AR(1) processes (cf. Eqs. 15, 16) – given for regular and two gamma-distributed samplings with mildly irregular (skewness sk_x and $\mathrm{sk}_y = 0.1$) and very irregular (skewness $\mathrm{sk}_x = 0.1$, $\mathrm{sk}_y = 2.85$) inter-observation-times. Errorbars give the standard deviation of the estimate, calculated using 1000 bootstrap iterations.

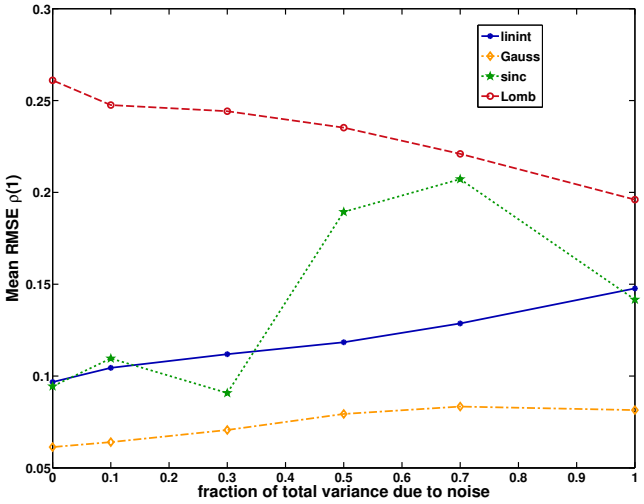

Fig. 10. Effect of the Signal-to-Noise ratio on the RMSE of the ACF for skewed (sk = 2) inter-observation times. The share of the noise variance in the overall process variance increases from left to right (cf. Eq. 17).

increases for all methods, but least for the Gaussian kernel (Fig. 9).

3.3 Sinusoids with random phase in colored noise

For irregular time series, the effect of interpolation on the ACF estimation of noise-free sinusoids is that it seems to suppress high-frequency variability. For red-noise signals we find that it, similarly, leads to an overestimation of autocorrelation. To generate more "realistic" signals, we synthesize the above-mentioned sinusoidal signals (Eq. 14) with varying amounts of additive red (AR) noise:

$$x(t) = \frac{1-s}{3} \sum_i \sin(\omega_i t + \Theta_{i,n}) + s\zeta_i \ . \tag{17}$$

The sinusoidal components vary with the frequencies $\omega_i = \frac{2\pi}{T_i}$, $T_i = [18, 21, 41]$ years. The time vector t is concatenated into a time line from random variables drawn from a gamma-distribution with $\mu = 4$ and $\mathrm{sk} = 0.1$. The phase variable $\Theta_{i,n}$ is, for each realization n, randomly drawn from a uniform distribution on $(0, 2\pi)$. This makes the time series samples from stationary stochastic processes. ζ_i represents a red noise process (cf. Eq. 15) whose variance s we vary in the range $[0, 1]$. The persistence time τ is, for this intercomparison, fixed at $\tau = 4$ (corresponding to $\phi \approx 0.78$). Since we adjust the overall variance of the process to equal unity, the signal-to-noise ratio varies in proportion with s.

The "true" ACF is then given by

$$\rho(k) = (1-s)/3 \sum_i \cos(\omega_i |k|) + s \cdot \exp(-|k|/\tau) \ . \tag{18}$$

Varying s and using irregular time series (sk = 2) we find that the mean RMSE of $\hat{\rho}_x(1)$ estimated for the Gaussian kernel method increases slightly from 5 % for sinusoidal signals ($s = 0$, cf. Fig. 10), to 7 % for pure red noise ($s = 1$). At the same time, the RMSE for the interpolation-based routine rises from 10 % to 15 %, that for the LS-technique decreases from 27 % to 19 %. The sinc kernel performs similar to the interpolation routine for sinusoidal signals with up to 30 % of noise, but has a higher RMSE for noise-dominated signals. For irregular time series with low inter-sampling-time distribution skewness (sk = 0.1) we find that the RMSE is maximal for medium signal-to-noise ratios, i.e. it is lower for purely deterministic and purely random time series than for the mixture of both (results not shown). For mostly deterministic time series, $s \leqslant 0.5$, the LS technique has then the highest RMSE, while sinc and Gaussian kernel-based methods give more accurate results. For dominant red noise $s \geqslant 0.5$, the LS technique gives good results with low RMSE, where at the same time the performance of the sinc kernel deteriorates. The interpolation-based FFT-routine is not the best choice for irregular time series, irrespective of the signal-to-noise ratios of the processes generating the time series. The increased RMSE for interpolation observable for the ACF estimates is due to a positive bias for $\rho_x(1)$. The RMSE of the kernel-based methods is lower and the ACF bias is constant and negligible. The high-frequency variability is systematically underestimated when using interpolation. The higher the persistence time τ in the AR(1) component, the lower are the advantages of the Gaussian-kernel based estimator, since the high-frequency variability in the signal is lower.

3.4 Summary of the synthetic tests

In all tests we performed in this section, we find that linear interpolation comes with two systematic effects. Firstly, it has a positive bias for ACF estimation and secondly, it has a negative bias in CCF estimation. Both effects become more severe with increasing sampling time distribution skewness. The LS technique performed well for the ACF estimation of slightly irregular autocorrelated time series but not for sinusoids. We find the opposite pattern for the sinc kernel: its RMSEs are low in the application to sinusoidal data – but high for the ACF of autocorrelated noise processes. The Gaussian kernel estimates are consistent and have the, or close to the, lowest RMSEs in all tests. Therefore we recommend the use of the Gaussian kernel-based estimator instead of – or in addition to – the standard interpolation routine for irregular time series with positive inter-sampling time distribution skewness, and especially in the presence of observation gaps.

Fig. 11. Map showing the location of the paleo records and the main wind directions of the Indian and East-Asian summer monsoon systems. Presently there are three major inflow corridors into Southern China, through the Bay of Bengal and over Indo-China, through the South China Sea and from the south east (Liu et al., 2008; Clemens et al., 2010).

4 Comparison for paleo data

We will now apply the Gaussian kernel estimator and interpolation followed by the standard FFT-routine to paleo records from the Asian Monsoon domain, to evaluate possible differences between the CCF/ACF estimates of these datasets, depending on the analysis technique.

The Asian monsoon system (cf. Fig. 11) affects a large share of today's world population. Zhang et al. (2008) find its strength in the past 1800 yr to be correlated with agricultural and cultural prosperity, its weakening with periods of unrest and instability. It can be divided into the Indian and the East Asian monsoon subsystems (ISM and EASM), that transport moisture from different sources. Oxygen isotope ratios (δ^{18}O) from cave records have been used to study the Holocene variability of monsoonal precipitation over China and India. While most of them show a millennial-scale trend, believed to be linked with the decreasing solar irradiation through the Holocene (Maher, 2008; Wang et al., 2005), sources for variability on shorter time scales are debated (Berkelhammer et al., 2010). In an inter-comparison of four published, acclaimed records of monsoonal precipitation from four different geographical locations we want to investigate the spatial and temporal consistency of linear dependencies among these time series. Cross-correlation analysis of monsoon records could give clues to the interrelationships between the different monsoon branches and their development with time. Autocorrelation analysis can, amongst other methods, give insights into the persistence inherent to the time series and is believed to increase before certain dynamical transitions (Scheffer et al., 2009). Persistence time (cf. Eq. 15) is a characteristic parameter for the time scales on which these climate processes operate.

For the late Holocene time span of 387–1100 BP, we estimate cross correlation and persistence time of four speleothem δ^{18}O records (cf. Fig. 11), reconstructed from Dongge cave in southern China (Wang et al., 2005), Heshang cave in central China (Hu et al., 2008), Wanxiang cave in north-central China (Zhang et al., 2008) and from Dandak cave in southern India (Berkelhammer et al., 2010). The sample locations lie in different branches of the Asian monsoon and therefore enable us to assess spatial variability of the monsoon system. The data sets have quite different inter-sampling time distributions, with rather high time resolution (0.5a–3.9a) and considerable time uncertainties. The details of the overlapping part of the records, which we will use, are given in table 2. For all four records, δ^{18}O variations are interpreted as mainly dominated by precipitation amount changes, thus reflecting summer monsoon strength (Wang et al., 2005; Hu et al., 2008; Zhang et al., 2008; Berkelhammer et al., 2010).

Prior to correlation analysis we subtract (nonlinear) trends from the records, that we estimate using a 500a wide Gaussian kernel smoother (high-pass filter), adapted for irregular sampling. For the standard approach of CCF (ACF) estimation the time series are then interpolated linearly to a regular grid with spacing of the larger of the mean sampling intervals (a spacing equalling the mean sampling period) of the respective two time series involved. This means that in case of the CCF comparison of Dandak and Wanxiang records, this CCF has a lag resolution of 3.31a, in case we compare the Wanxiang to the Dongge record the resolution is at 3.92 a.

Table 2. Mean sampling intervals,variances and skewnesses of the inter-sampling time distributions and number of observations in the overlapping section of the used paleo proxy records (625 AD–1563 AD).

Record	Mean sampling rate $\mu_{\Delta t}$ [a]	Skewness $SK_{\Delta t}$	No. of observations N	Reference
Dandak	0.50	2.95	1874	Berkelhammer et al. (2010)
Wanxiang	3.31	−0.96	284	Zhang et al. (2008)
Heshang	2.34	1.45	402	Hu et al. (2008)
Dongge (DA)	3.92	0.41	241	Wang et al. (2005)

4.1 Results from ACF analysis

First we look at the individual ACFs (e.g. Fig. 12c and d) and find that the Gaussian estimate shows a much stronger initial decline than that resulting from interpolation. To investigate whether this more pronounced decline, this lower persistence time τ (cf. Eq. 15), is due to a negative bias of the kernel method or to a positive bias of the interpolation we perform the additional least squares analysis (LSq). The estimator, implemented similar to that in Mudelsee (2002), fits a simplified Ohrnstein-Uhlenbeck process, a continuous-time AR(1) analog, to the time series. Its estimates are robust with respect to variations in sampling rates, irregularity and persistence time and show a small, but constant, bias (−10 %) and variance. We compare four results: from interpolation, followed by ACF estimation involving the FFT; from interpolation, followed by the LSq estimation; from the Gaussian kernel ACF estimate and from LSq analysis of the original record (cf. Fig. 13). We find a pronounced overestimation, up to a factor of two, when interpolation is involved. This is irrespective of whether τ was estimated via ACF or the LSq fit. The Gaussian kernel estimate is generally lower than that of LSq analysis, but differs by not much, except in the estimate for the Heshang cave record where it is 50 % lower. We could relate this to the differences in the respective sampling time distributions: The Heshang sampling time distribution shows high skewness and a large mean sampling period, both combining into a source of estimation error. Neither high skewness (Dandak) nor a lower sampling rate (Dongge) alone lead to such a deviation between the LSq and Gaussian kernel estimates, which is in agreement with the results from Sect. 3.

It follows from this, that interpolation causes a strongly positive bias on persistence and the kernel-based estimate is slightly negatively biased. Thus, if in an analysis the two estimates coincide we could assume the result to be unbiased. On the other hand, we should exercise caution when the results from different methods disagree. Persistence times give a measure of memory in processes and are thus important to characterize time scales on which climate processes operate. As we see in this section, interpolation leads to a strong overestimation of persistence for irregular time series, with a bias changing also in relation to the skewness of the observation time distribution. Caution should therefore be exercised and additional methods employed when performing autocorrelation analysis of irregular time series.

4.2 Results from cross correlation analysis

Next, pairwise cross correlation functions were calculated for all four records. Only two combinations resulted in significant correlation at zero lag (Fig. 14a). The correlation coefficient of 0.29 (−0.17,0.21) for interpolation resp. 0.295 (−0.19,0.27) for the Gaussian kernel at lag zero between the Wanxiang and Dandak records is significant to the 95 % level in the two-sided test for zero correlation under the null hypothesis of the time series being sampled from autocorrelated red noise processes. In the brackets we give the estimated critical values of the test that were determined using AR(1) processes (with persistence times based on the LSq estimate) on the original time axis of the records.

The late Holocene section (387–1325 BP) of the record from Wanxiang cave correlates also significantly with that from Heshang cave with a lag zero correlation coefficient of 0.28 (−0.2, 0.23) based on interpolation and 0.28 (−0.2,0.19) from the kernel estimator. We find that the high-frequency variability of the estimated correlation function is more pronounced in the kernel estimate. However, the overall shapes of the functions agree well.

The lack of significant correlation between the other records could have several reasons. One may be that our estimators did fail to capture the "true" underlying monsoon variability common to all records. This is not unlikely, since there are time uncertainties and local influences to be taken into account, especially when analyzing records that are spaced so far apart and reconstructed over such a time span. It may well be that the strongest commonality between the records are trends on centennial to millennial time scales that cannot be reconstructed from a less than 1000 yr long overlapping section and that possible links operating on shorter time scales were obscured in the generation process. On the other hand, our time section includes the Northern hemispheric Medieval Warm Period (MWP, ca. 700 BP–1000 BP) and parts of the Little Ice Age (LIA, ca. 100 BP–400 BP), periods where monsoonal circulation seemed to be stronger (MWP) or weaker (LIA) according to Zhang et al. (2008).

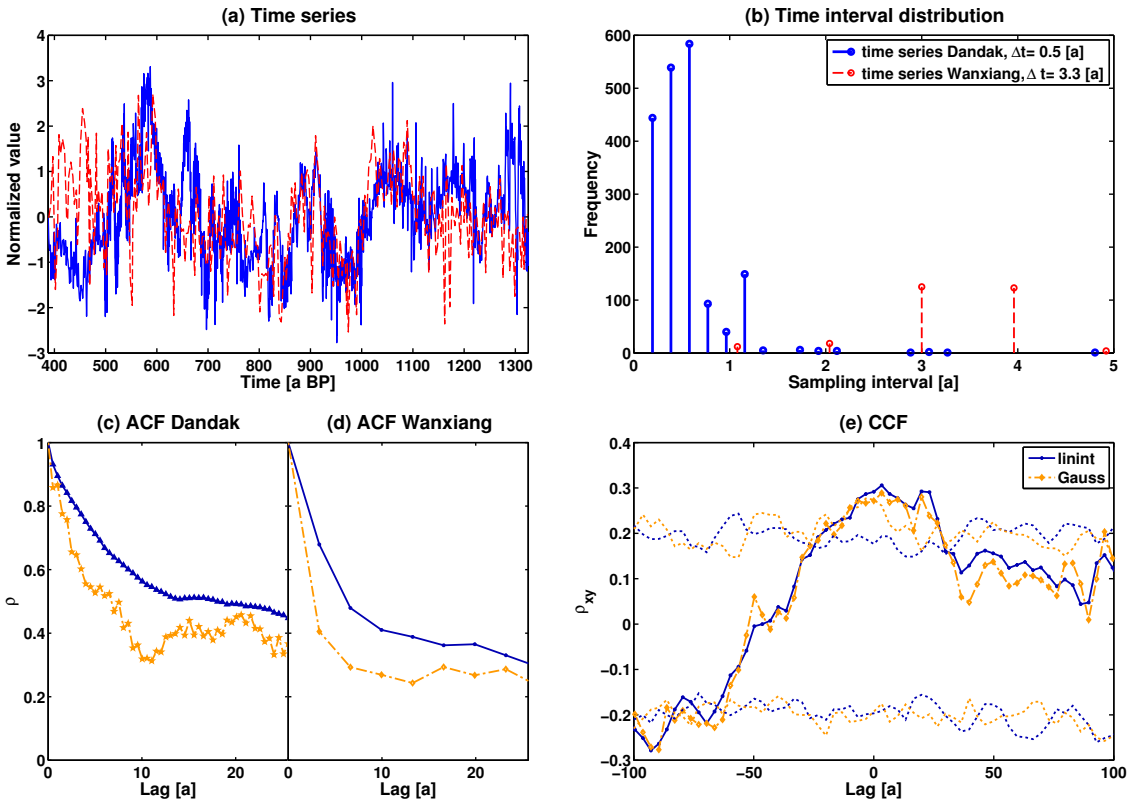

Fig. 12. Exemplary cross correlation analysis of δ^{18}O records from Dandak and Wanxiang caves: Standardized time series **(a)**, the sampling interval distributions **(b)**, the estimated ACFs **(c, d)** and the corresponding CCF with the dashed lines representing the estimated 95 % critical values of a two-sided test for the null hypothesis of time series being sampled from a red noise process **(e)**. Legends for each row are given in the right panels.

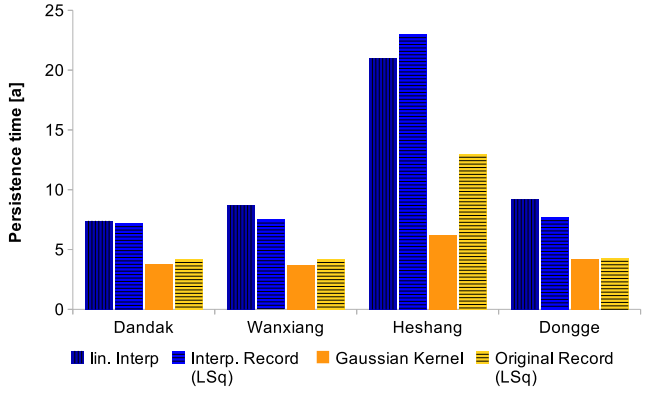

Fig. 13. Persistence times of the δ^{18}O records estimated using linear interpolation ACF estimate, Gaussian kernel ACF estimate and the least squares fitting of AR(1) processes (denoted by LSq) on interpolated and original record.

The Asian summer monsoon is a large-scale atmospheric circulation phenomenon. During northern hemisphere cold phases, less energy available for its generation might have led to a weakening of the monsoon. In contrast, warm phases should have led to a strong circulation which results in an increased influence of the ISM on Chinese precipitation. This should be observable in an increased correlation between Indian and Chinese rainfall variation and, at the same time, an increased correlation between the δ^{18}O records.

We therefore analyze two time slices (389 BP–700 BP and 700 BP–1100 BP) of the records separately. After significance testing – and considering lags of 0 to 30 yr absolute value –, we find a contrasting picture: while the Northern Chinese records correlate with the Indian Dandak record during the warm phase (MWP), this correlation is insignificant in the colder phase (towards the LIA) that followed (cf. Fig. 14c). On the other hand, while the southern Chinese Dongge record correlates with the more northern records from Heshang and Wanxiang caves during the LIA, this correlation is not significant during the MWP.

This points us towards a more differentiated interpretation of these correlations, emphasizing the geographical origins of these cave records. According to the "isotopic zones" shown in Maher (2008), Feng et al. (1999) and references therein, Wanxiang cave is located in a zone that is, at present, dominated by the ISM. Heshang and Dongge cave lie in an

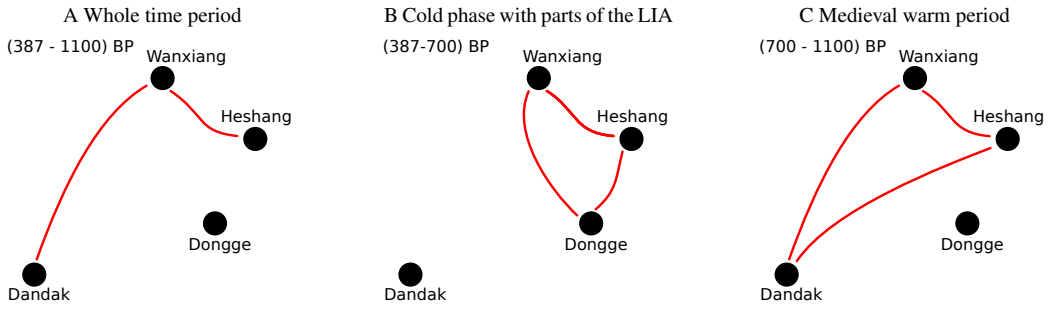

Fig. 14. Results from pairwise cross correlation analysis for all records: Red links indicate significant positive cross correlation at or close to zero lag for the respective records. While in the warm phase of the MWP the Northern Chinese records correlate with the Indian Dandak record and not with the southernmost Dongge cave record **(B)**, this is reversed in the cold phase after the MWP. The Chinese records then correlate amongst each other, but not with the Indian Dandak record **(C)**.

"isotopic zone" where both monsoonal branches are influential. However, Dongge cave lies closer to the southern zone that is, at present, dominated by monsoonal precipitation from the south east (South China Sea) but not from the south-west (ISM). Recent investigations show, that even within southern China, moisture sources and their isotopic signature, differ orthogonally to these "isotopic zones" (Liu et al., 2008), pointing at a stronger influence of the South East monsoon in direction of Dongge cave. We conclude that during warm phases our results are consistent with these isotopic zones, since the records from central China correlate with the Indian Dandak record. In the cold phases, the atmospheric circulation might have been different, emphasizing the south east moisture source for allover China, evident through a correlation between the Dongge cave record and the more northern Chinese records and, since we observe no significant correlation with the Dandak record, less ISM impact.

Interpolation and kernel-based estimation give similar results. The CCF estimates at and around lag zero were not – or not significantly – lower for interpolation where a significant correlation was detected. We believe that this is due to the long time scales on which these correlations are recorded, as the advantages of the kernel-based method are larger for low persistence (cf. Sect. 3).

5 Conclusions

Comparing different methods for analyzing correlations from irregularly sampled time series, we have found that the kernel-based method is robust and has a comparable – and often even lower – RMSE and bias than the traditionally employed schemes using interpolation in the application to synthetic records, for regular and irregular sampling.

For the interpolation and FFT-based routine we find a four to seven times increase in RMSE, predominantly caused by an increase in the absolute value of the bias. This bias is positive for ACF estimation but negative for CCF quantification and its magnitude scales linearly with sampling irregularity.

In all synthetic test cases we studied the Gaussian kernel was close to or was the estimator with the lowest RMSE. Its performance was slightly inferior to that of the sinc kernel for sinusoidal time series but significantly better for red noise ACF and CCF estimation, especially in the application to records with disparate sampling rates.

We find that the sinc-kernel performs well for ACF estimation of sinusoidal signals. It shows, however, alternating bias patterns in the ACF for red noise time series, resulting in a high RMSE comparable to the FFT-based result. This might be due to the shape of the kernel with its positive and negative weights, thus emphasizing regular, deterministically recurrent structures that are not present in stochastic processes. Another reason for the mixed performance could be cutoff effects, since the kernel effectively presents a rectangular filter in the frequency domain.

The performance of the Lomb-Scargle periodogram-based routine showed advantages over interpolation for low skewness time sampling. For very irregular time series, in ACF as in CCF tests, we found a strong sampling effect resulting in a large bias.

In all tests we performed on synthetic data, we have found that linear interpolation comes with two systematic effects: It shows a positive bias for ACF estimation, and it has a negative bias in CCF estimation, emphasizing low-frequency variability at the cost of high-frequency components. Both effects become more severe with increasing sampling time distribution skewness and lower persistence in the processes from which we generate the time series. The Gaussian kernel estimates are consistent with those from interpolation for regular sampling and have the, or close to the, lowest RMSEs in all tests.

The estimated persistence time using the Gaussian kernel shows generally only a small negative bias with respect to the least squares estimate on the original record. The least

squares persistence time estimator, which is, fitting an AR(1) process to the observations, has a constant and low bias for varying sampling irregularity. Compared to this, interpolation leads to an overestimation of this persistence time by a factor of two. This difference is especially unnerving as the frequency of observations recorded through paleo archives varies in dependence on climatic parameters (e.g. lower accumulation rates through less precipitation). A change in the inter-observation time distribution could then lead to an artificial change in the estimated persistence.

In the cross correlation analysis of the paleo records, the kernel-based lag-zero cross correlation functions are consistent for interpolation and cross correlation. We believe that this is due to the short time scales on which the interactions of the monsoon systems are recorded, as the advantages of the kernel-based method are not as pronounced for records with high persistence. The kernel-based cross correlation functions show more high-frequency variability which could be investigated through cross spectral analysis. We do not attempt to characterize it here, since this is outside the scope of this paper. The bias effects from interpolation could cause problems in the evaluation of phenomena emerging on time scales close to the actual mean sampling rate. This is where the kernel methods show significant advantages and especially the Gaussian kernel correlation method can provide high-resolution, robust estimates of time-dependent cross correlation coefficients.

In our cross correlation analysis of four Asian monsoon records in the time interval of 387 BP–1100 BP, we have found significant evidence that the Indian summer monsoon circulation influenced Chinese rainfall variability during the northern hemispheric MWP, as then the δ^{18}O record from Dandak cave in India correlates with the central China records from Wanxiang and Heshang caves. During the colder phase after the MWP and into the LIA, significant cross correlation coefficients are found amongst the Chinese records, indicating a spatially more homogeneous moisture source. At the same time these records do not correlate with the Indian record during the LIA cold phase, pointing to less ISM impact on Chinese precipitation.

To summarize, we have shown that in correlation estimation of irregularly sampled time series, caution should be exercised when these records have an inter-observation time distribution that is strongly skewed. In the CCF estimation we found a strongly negative bias for the standard approach with interpolation for processes with little persistence. The advantages of the kernel-based estimators are higher for coupling on short time scales, compared to the sampling rate. This is especially interesting for the investigation of proxy data with low resolution. Our results indicate that the bias properties of the Gaussian kernel and the interpolation techniques have different signs in ACF estimation, indicating when sampling irregularity causes problems in the analysis.

Acknowledgements. This research was financially supported by the the German Federal Ministry of Education and Research (BMBF project PROGRESS, 03IS2191B), the German Science Foundation (DFG graduate school 1364) and by the Leibniz association (project ECONS). The authors would like to thank Jeffrey D. Scargle and one anonymous referee as well as Sebastian Breitenbach for their valuable and insightful advice. Software to analyze irregularly sampled time series using the methods in this paper can be found on www.tocsy.pik-potsdam.de.

Edited by: S. Barbosa
Reviewed by: J. Scargle and another anonymous referee

References

Babu, P. and Stoica, P.: Spectral analysis of nonuniformly sampled data - a review, Digit. Signal Process., 20, 359–378, doi:10.1016/j.dsp.2009.06.019, 2009.

Benedict, L., Nobach, H., and Tropea, C.: Benchmark tests for the estimation of power spectra from LDA signals, in: Proc. 9th Int. Symp. on Applications of Laser Technology to Fluid Mechanics, p. 32.6, Lisbon, Portugal, 1998.

Benedict, L., Nobach, H., and Tropea, C.: Estimation of turbulent velocity spectra from laser Doppler data, Meas. Sci. Technol., 11, 1089, doi:10.1088/0957-0233/11/8/301, 2000.

Berkelhammer, M., Sinha, A., Mudelsee, M., Cheng, H., Edwards, R. L., and Cannariato, K.: Persistent multidecadal power of the Indian Summer Monsoon, Earth Planet. Sc. Lett., 290, 166–172, doi:10.1016/j.epsl.2009.12.017, 2010.

Bjoernstad, O. N. and Falck, W.: Nonparametric spatial covariance functions: Estimation and testing, Environ. Ecol. Stat., 8, 53–70, doi:10.1023/A:1009601932481, 2001.

Böttcher, M. and Dermer, C. D.: Timing Signatures of the Internal-Shock Model for Blazars, The Astrophysical Journal, 711, 445, doi:10.1088/0004-637X/711/1/445, 2010.

Broersen, P. M.: Autoregressive spectral estimation by application of the Burg algorithm to irregularly sampled data, IEEE T. Instrum. Meas., 51, 1289–1294, doi:10.1109/TIM.2002.808031, 2002.

Broersen, P. M.: Five Separate Bias Contributions in Time Series Models for Equidistantly Resampled Irregular Data, IEEE T. Instrum. Meas., 58, 1370, doi:10.1109/TIM.2009.2012928, 2009.

Broersen, P. M., de Waele, S., and de Waele, S.: The Accuracy of Time Series Analysis for Laser-Doppler Velocimetry, in: Proceedings of the 10th International Symposium on Applications of Laser Techniques to Fluid Dynamics, Lisbon, Portugal, 2000.

Chatfield, C.: The analysis of time series: an introduction, Texts in statistical science, CRC Press, Florida, US, 6th Edn., 2004.

Clemens, S. C., Prell, W. L., and Sun, Y.: Orbital-scale timing and mechanisms driving Late Pleistocene Indo-Asian summer monsoons: Reinterpreting cave speleothem δ^{18}O, Paleoceanography, 25, 4207–4223, doi:10.1029/2010PA001926, 2010.

Edelson, R. and Krolik, J.: The discrete correlation function – A new method for analyzing unevenly sampled variability data, Astrophys. J., 333, 646–659, doi:10.1086/166773, 1988.

Fan, J.-H., Liu, Y., Qian, B.-C., Tao, J., Shen, Z.-Q., Zhang, J.-S., Huang, Y., and Wang, J.: Long-term variation time scales in OJ 287, Res. Astron. Astrophys., 10, 1100, doi:10.1088/1674-4527/10/11/002, 2010.

Feng, X., Cui, H., Tang, K., and Conkey, L. E.: Tree-Ring δD as an Indicator of Asian Monsoon Intensity, Quaternary Res., 51, 262–266, doi:10.1006/qres.1999.2039, 1999.

Hall, P., Fisher, N. I., and Hoffmann, B.: On the Nonparametric Estimation of Covariance Functions, Ann. Stat., 22, 2115–2134, 1994.

Harteveld, W. K., Mudde, R. F., and Van den Akker, H. E. A.: Estimation of turbulence power spectra for bubbly flows from Laser Doppler Anemometry signals, Chem. Eng. Sci., 60, 6160–6168, doi:10.1016/j.ces.2005.03.037, 2005.

Hocke, K. and Kämpfer, N.: Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram, Atmos. Chem. Phys., 9, 4197–4206, doi:10.5194/acp-9-4197-2009, 2009.

Hu, C., Henderson, G. M., Huang, J., Xie, S., Sun, Y., and Johnson, K. R.: Quantification of Holocene Asian monsoon rainfall from spatially separated cave records, Earth Planet. Sci. Lett., 266, 221–232, doi:10.1016/j.epsl.2007.10.015, 2008.

Liu, J., Song, X., Yuan, G., Sun, X., Liu, X., Wang, Z., and Wang, S.: Stable isotopes of summer monsoonal precipitation in southern China and the moisture sources evidence from δ^{18}O signature, J. Geogr. Sci., 18, 155–165, doi:10.1007/s11442-008-0155-9, 2008.

Maher, B.: Holocene variability of the East Asian summer monsoon from Chinese cave records: a re-assessment, The Holocene, 18, 861–866, doi:10.1177/0959683608095569, 2008.

Mayo, W.: Spectrum measurements with laser velocimeters (in: Proceedings of the Dynamic Flow Conference 1978), in: Selected Papers on Laser Doppler Velocimetry, edited by: Adrian, R. J., Society of Photo-Optical Instrumentation Engineers: SPIE milestone series; 78, chap. 3, 222–235, SPIE Press, 1993.

Mudelsee, M.: TAUEST: a computer program for estimating persistence in unevenly spaced weather/climate time series, Comput. Geosci., 28, 69–72, doi:10.1016/S0098-3004(01)00041-3, 2002.

Mudelsee, M.: Climate Time Series Analysis: Classical Statistical and Bootstrap Methods (Atmospheric and Oceanographic Sciences Library), Springer, 2010.

Nieppola, E., Hovatta, T., Tornikoski, M., Valtaoja, E., Aller, M. F., and Aller, H. D.: Long-Term Variability of Radio-Bright BL Lacertae Objects, Astron. J., 137, 5022, doi:10.1088/0004-6256/137/6/5022, 2009.

Scargle, J.: Studies in Astronomical Time Series Analysis, I. Modeling Random Processes in the time domain, Astrophys. J. Suppl. S., 45, 1–71, 1981.

Scargle, J.: Studies in Astronomical Time Series Analysis. II. Statistical Aspects of Spectral Analysis of unevenly spaced data., Astrophys. J., 263, 835–853, 1982.

Scargle, J.: Studies in astronomical time series analysis, III – Fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data, Astrophys. J., 343, 874–887, doi:10.1086/167757, 1989.

Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., Van Nes, E. H., Rietkerk, M., and Sugihara, G.: Early-warning signals for critical transitions, Nature, 461, 53–9, doi:10.1038/nature08227, 2009.

Schimmel, M.: Emphasizing difficulties in the detection of rhythms with Lomb-Scargle periodograms, Biol. Rhythm. Res., 32, 341–5, doi:10.1076/brhm.32.3.341.1340, 2001.

Schulz, M. and Stattegger, K.: SPECTRUM: spectral analysis of unevenly spaced paleoclimatic time series, Comput. Geosci., 23, 929–945, doi:10.1016/S0098-3004(97)00087-3, 1997.

Stoica, P. and Sandgren, N.: Spectral analysis of irregularly-sampled data: Paralleling the regularly-sampled data approaches, Digit. Signal Process., 16, 712–734, doi:10.1016/j.dsp.2006.08.012, 2006.

Stoica, P., Li, J., and He, H.: Spectral Analysis of Nonuniformly Sampled Data: A New Approach Versus the Periodogram, IEEE T. Signal Process., 57, 843–858, doi:10.1109/TSP.2008.2008973, 2008.

Wang, Y., Cheng, H., Edwards, R. L., He, Y., Kong, X., An, Z., Wu, J., Kelly, M. J., Dykoski, C. A., and Li, X.: The Holocene Asian Monsoon: Links to Solar Changes and North Atlantic Climate, Science, 308, 854–857, doi:10.1126/science.1106296, 2005.

Zhang, B.-K., Dai, B.-Z., Zhang, L., and Cao, Z.: Multi-band optical variability of BL Lac object OQ 530, Res. Astron. Astrophys., 10, 653, doi:10.1088/1674-4527/10/7/004, 2010.

Zhang, P., Cheng, H., Edwards, R. L., Chen, F., Wang, Y., Yang, X., Liu, J., Tan, M., Wang, X., Liu, J., An, C., Dai, Z., Zhou, J., Zhang, D., Jia, J., Jin, L., and Johnson, K. R.: A Test of Climate, Sun, and Culture Relationships from an 1810-Year Chinese Cave Record, Science, 322, 940–942, doi:10.1126/science.1163965, 2008.