Nonlin. Processes Geophys., 21, 929–937, 2014
www.nonlin-processes-geophys.net/21/929/2014/
doi:10.5194/npg-21-929-2014

Nonlinear Processes in Geophysics
Open Access

Estimating time delays for constructing dynamical networks

E. A. Martin and J. Davidsen

Complexity Science Group, Department of Physics and Astronomy, University of Calgary, Calgary, Alberta, T2N 1N4, Canada

Correspondence to: E. A. Martin (eamartin@ucalgary.ca)

Received: 16 November 2013 – Revised: 6 July 2014 – Accepted: 30 July 2014 – Published: 11 September 2014

Abstract. Dynamical networks – networks inferred from multivariate time series – have been widely applied to climate data and beyond, resulting in new insights into the underlying dynamics. However, these inferred networks can suffer from biases that need to be accounted for to properly interpret the results. Here, we report on a previously unrecognized bias in the estimate of time delays between nodes in dynamical networks inferred from cross-correlations, a method often used. This bias results in the maximum correlation occurring disproportionately often at large time lags. This is of particular concern in dynamical networks where the large number of possible links necessitates finding the correct time lag in an automated way. We show that this bias can arise due to the similarity of the estimator to a random walk, and are able to map them to each other explicitly for some cases. For the random walk there is an analytical solution for the bias that is closely related to the famous Lévy arcsine distribution, which provides an upper bound in many other cases. Finally, we show that estimating the cross-correlation in frequency space effectively eliminates this bias. Reanalysing large lag links (from a climate network) with this method results in a distribution peaked near zero instead, as well as additional peaks at the originally assigned lag. Links that are reassigned smaller time lags tend to have a smaller distance between them, which indicates that the new time delays are physically reasonable.

1 Introduction

Complex networks are increasingly being used to study geophysical systems (Baiesi and Paczuski, 2004; Davidsen et al., 2006, 2008; Peixoto and Davidsen, 2008; Peixoto et al., 2010; Gu et al., 2013; Zaliapin and Ben-Zion, 2013; Zanardo et al., 2013; Dodds and Rothman, 2000; Mantilla et al., 2006; Feng and Dijkstra, 2014; Bassett et al., 2012). Climate dynamics in particular have been an object of much attention where dynamical networks (also called functional or interaction networks) have been used to study phenomena such as: long-range correlation in the atmosphere known as teleconnections (Tsonis et al., 2008; Kawale et al., 2011); climate models (Donges et al., 2009a, b); El Niño Southern Oscillation (Yamasaki et al., 2008; Tsonis and Swanson, 2008; Martin et al., 2013); with more analyses continuously being carried out (Steinhaeuser et al., 2012; Ebert-Uphoff and Deng, 2012; Deza et al., 2013). The dynamics are mapped to networks by taking a set of global positions as nodes, and creating a link between any two satisfying a given criteria based on recorded time series. These methods – particularly when used to analyse climate data – can have inherent biases that are increasingly coming to light.

The widespread use of dynamical networks to geophysical systems makes understanding their associated biases an important question. In previous work, we showed how adjusting the timescale of the analysis can reverse the findings of some climate networks (Martin et al., 2013), and Paluš et al. (2011) showed that autocorrelations, inherent in climate data, can result in erroneous links. In addition, the transitivity of many of the metrics used to construct climate networks results in them being biased towards "small world" networks (Hlinka et al., 2012). These analyses are further hindered by the lack of stationarity in the systems (where probability distributions change in time), forcing us to restrict ourselves to smaller timescales where stationarity is a plausible assumption. However, compensating for this effect by adopting new methods and criteria to establish links in a dynamical network can result in new biases, as we show here.

One of the methods often used for constructing dynamical networks is the pairwise cross-correlation. The cross-correlation function, also called the covariance, of two

stationary time series is defined as

$$E\Big[(\omega(t)-\mu_\omega)\left(\xi\,(t+\tau)-\mu_\xi\right)\Big], \tag{1}$$

where $\xi(t)$ and $\omega(t)$ are the time series of interest, μ_ξ and μ_ω are their respective means, τ is the time lag between $\xi(t)$ and $\omega(t)$ and $E[\ldots]$ is the expected value. The Pearson correlation can be calculated by dividing Eq. (1) by the standard deviations of $\xi(t)$ and $\omega(t)$.

Equation (1) is often estimated as

$$C(\tau)=\frac{1}{T-\tau}\sum_{i=1}^{T-\tau}\omega(i)\xi(i+\tau). \tag{2}$$

Here T is the number of points in the time series, and both $\xi(t)$ and $\omega(t)$ have zero mean either by definition or by construction. Unfortunately, for larger absolute values of the time lag there are fewer points in the calculation leading to larger statistical fluctuations (compared to the $\tau=0$ case) that need to be taken into account. This is especially true if one aims to identify the physical time lag as the maximum of the cross-correlation.

One possibility to circumvent this problem of non-uniform statistical fluctuations, and also suitable in a non-stationary setting, is to use the non-stationary version of the cross-correlation function that includes the same number of points for every time lag used. It is calculated as

$$C(\tau,t)=\frac{1}{\Omega}\sum_{i=t}^{t+\Omega-1}\omega(i)\xi(i+\tau). \tag{3}$$

Now the cross correlation is normalized by the number of points used (Ω). We also note that for different values of τ, with t constant, the values of $\omega(t)$ used will not change while the values of $\xi(t)$ will.

This method has been used in various dynamical network analyses, climate (Yamasaki et al., 2008; Tirabassi and Masoller, 2013) and otherwise (Bashan et al., 2012). To build a network in this way, a specific time delay must be chosen at which to evaluate Eq. (3) for a given pair of time series. Typically, the value that maximizes Eq. (3), or its absolute value, is used. As this is the quantity that we will focus on in this paper, we define it here as

$$T_{\tau_i}^{\tau_f}\Big(\omega(t),\xi(t)\Big)=\tau^* \tag{4}$$

$$|T|_{\tau_i}^{\tau_f}\Big(\omega(t),\xi(t)\Big)=\tau', \tag{5}$$

with the property that

$$C(\tau^*,t)=\mathrm{Max}_{\tau=\tau_i,\ldots,\tau_f}C(\tau)$$
$$C(\tau',t)=\mathrm{Max}_{\tau=\tau_i,\ldots,\tau_f}|C(\tau)|.$$

Here τ_i and τ_f are the smallest and largest lags considered respectively. Throughout this paper τ_i and τ_f are equal and opposite, but this need not always be the case. These lags together with Ω are typically determined by the requirement of quasi-stationarity mentioned above.

In the context of climate, Eq. (3) is a reasonable choice of metric since the correlation structure of daily temperature anomaly data, for example, is linear to a good approximation, resulting in linear methods – like Eq. (3) – generally outperforming non-linear ones (Hlinka et al., 2013b, a). However, Eq. (3) has a predilection for reaching a maximum at large absolute values of the time lag $\left(\left|T_{\tau_i}^{\tau_f}(\omega(t),\xi(t))\right|\gg 0\right.$ and $\left.\left||T|_{\tau_i}^{\tau_f}(\omega(t),\xi(t))\right|\gg 0\right)$, which complicates the estimation of the physical time lag. We illustrate this using a climate network example. Specifically, we consider daily average temperature anomaly time series (average temperature of that day minus the average temperature on that day over all years used) at 2 m a.g.l., with an angular resolution of $\approx 2^\circ$ [(1)]. As in Yamasaki et al. (2008), we look at the region from 30° N to 30° S, and from 120 to 285° E; the time range is 1979–2008.

Creating a dynamical network from these temperature data using Eq. (3) as the basic measure (as outlined in detail by Yamasaki et al., 2008) results in a network with links that have a time delay distribution as shown in Fig. 1. The climate network, constructed by the methodology in Yamasaki et al. (2008), uses sophisticated thresholding techniques to identify significant correlations and links, yet still exhibits an unphysical peak at large $\left||T|_{\tau_i}^{\tau_f}(\omega(t),\xi(t))\right|$ [(2)]. The unphysical nature directly follows from the observation that these peaks occur at τ_i and τ_f, independent of their specific values (not shown). For comparison we also show the distribution of $|T|_{\tau_i}^{\tau_f}(\omega(t),\xi(t))$ obtained for all pairs of the temperature anomaly series in Fig. 1, which shows a similar behaviour. The inset in Fig. 1 shows a representative plot of the absolute value of Eq. (3) as a function of lag, for the same data, where the maximum occurs at $\tau=200$.

In this paper, we will explore the origin of this effect in detail and discuss properties affecting it. To this end, we will make extensive use of synthetic data to illustrate expected behaviour in simplified situations, some of which are analytically tractable, by mapping them to a random walk, and in ones more closely approximating the statistics of temperature time series. Finally, we discuss a method for overcoming the resulting bias in the estimation of physical time delays by estimating the cross-correlation in frequency space, and test its applicability for climate networks.

[1]NCEP Reanalysis 2 data (Kanamitsu et al., 2002) provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their web site at http://www.esrl.noaa.gov/psd/.

[2]Because this method requires that a link exceeds a certain correlation over multiple time periods – and the time lag which maximizes Eq. (3) can change from period to period due to non-stationarities and/or statistical fluctuations – we use $|T|_{\tau_i}^{\tau_f}(\omega(t),\xi(t))$ from the final period in Fig. 1.

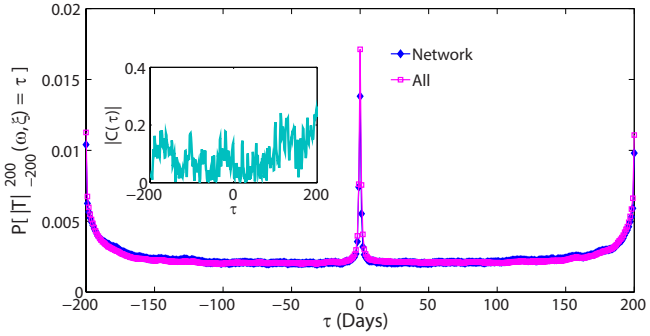

Figure 1. Distribution of links having a given $|T|^{200}_{-200}(\omega(t),\xi(t))$ for $\Omega = 365$ days over all times t (Network). The climate network was constructed from temperature anomaly time series using the method and parameters outlined by Yamasaki et al. (2008)[2], as discussed in the Introduction. We also show $|T|^{200}_{-200}(\omega,\xi)$ over all possible pairs (All) for comparison. In both cases we can see peaks at $\tau = \pm 200$ – which is the bias we are investigating in this paper – as well as an expected peak at about zero. Inset: a sample absolute cross-correlation as a function of time lag with a peak at $\tau = 200$.

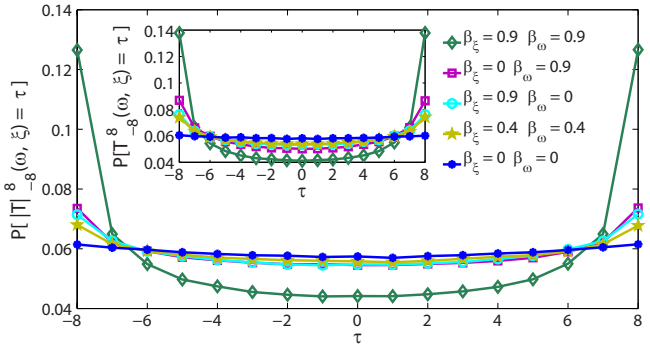

Figure 2. Distribution of $|T|^{8}_{-8}(\omega(t),\xi(t))$ and $T^{8}_{-8}(\omega(t),\xi(t))$ (inset), where $\omega(t)$ and $\xi(t)$ have long-range autocorrelations: β_ξ and β_ω are the exponents of the power spectra of $\xi(t)$ and $\omega(t)$ respectively. Since the time series are stationary, $C(\tau,t) = C(\tau)$. We can see that as β gets larger, increasing the persistence in the time series, the bias also increases. This occurs despite $\xi(t)$ and $\omega(t)$ having Gaussian distributions centred at zero. Fixed parameters are: $\Omega = 16$, with $N = 2^{21}$ realizations.

2 Illustration of bias in the absence of cross-correlations

To better understand the observed phenomenon, we first examine synthetic time series where no cross-correlations are present. Due to the absence of a physical time delay in this case, one would expect that all time delays are equally likely to maximize Eq. (3) and, hence, there should be no bias. The fact that a bias still persists, as we will show below, indicates that it is not necessarily due to real cross-correlations in the data. Instead, the bias can be affected by properties such as power-law autocorrelations in the individual time series. This makes this bias of particular relevance to geophysical data – including climate data – where these types of correlations are common (Pelletier and Turcotte, 1997; Koscielny-Bunde et al., 1998; Huybers and Curry, 2006; Kantelhardt et al., 2006; Vyushin et al., 2009).

2.1 Influence of non-trivial autocorrelations

To create power-law autocorrelated time series we first generate an independent identically distributed (i.i.d.) time series of Gaussian-distributed numbers, i.e. white noise. This is then transformed into a time series with a power spectrum that decays as $f^{-\beta}$ and $0 < \beta < 1$, where f is the frequency, as outlined by Schumann (2011). The final time series has power-law autocorrelations which scale as $C(\tau) = \tau^{\beta-1}$, and is still Gaussian distributed and stationary. We generate one long time series with the desired power spectrum, and then cut it into non-overlapping segments to create different realizations; this results in the time series having non-periodic boundary conditions. The case $\beta = 0$ which we will also consider in the following corresponds to the i.i.d. case.

Figure 2 illustrates that as β increases in one or both series so does the bias in Eq. (5), as well as in Eq. (4) (inset). We can also see that increasing β in $\omega(t)$ strengthens the bias more than when it is increased in $\xi(t)$. Since geophysical data commonly have power-law autocorrelations, as previously discussed, this could make networks based on this data sensitive to false positive links at extremal time lags.

In addition, Runge et al. (2014) showed that for some simple autoregressive processes the theoretical maximum of Eq. (1) does not correspond to the true time lag given by the autoregressive process. This implies that using the maximum cross-correlation is an inappropriate way to determine time delays in these cases. Thus, the bias noted in Runge et al. (2014), has a different origin to the one discussed in this paper.

We also note that increasing the size of Ω will not reduce the bias. If the size of Ω was increased, and the curves in Fig. 2 recomputed, the figure would remain unchanged. The fact that the bias remains the same despite the change in Ω is a general feature of this bias, and is not specific to data with power-law autocorrelations. This is true as long as $\tau_f - \tau_i \leq \Omega$, for reasons that we explain in Sect. 2.4. If we hold Ω fixed and instead compute more (fewer) time lags the curves in Fig. 2 would appear to be stretched (compressed).

2.2 Independent and identically distributed variables

Even for i.i.d. Gaussian variables – where no autocorrelations are present and $\beta = 0$ – the phenomenon persists. This is a result of the time series $\omega(t)$ having a non-zero sample mean. In calculating the cross-correlation we assume that the underlying stochastic process that generates the time series has zero mean. However, any finite portion of the time series will have a non-zero sample mean with probability one. To examine the effect of non-zero means we now look at the

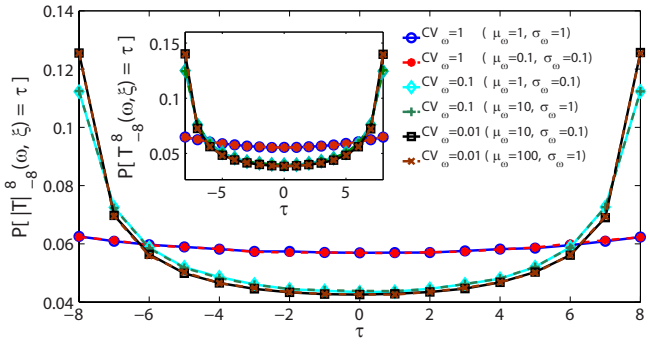

Figure 3. Distribution of $|T|^{8}_{-8}(\omega(t),\xi(t))$ and $T^{8}_{-8}(\omega(t),\xi(t))$ (inset) in the i.i.d. case for different CV (σ/μ) for the time series $\omega(t)$. Both time series are Gaussian distributed, with $\xi(t)$ drawn from a distribution with zero mean and unit variance, and $\omega(t)$ drawn from a distribution with the displayed mean and standard deviation. This means they are stationary and we can again write $C(\tau,t)=C(\tau)$. We can see that where the CV of $\omega(t)$ are the same the curves are identical irrespective of their different means (μ) and standard deviations (σ). This shows that it is the only quantity affecting the bias for Gaussian variables. Fixed parameters are: $\Omega=64$, with $N=2^{21}$ realizations.

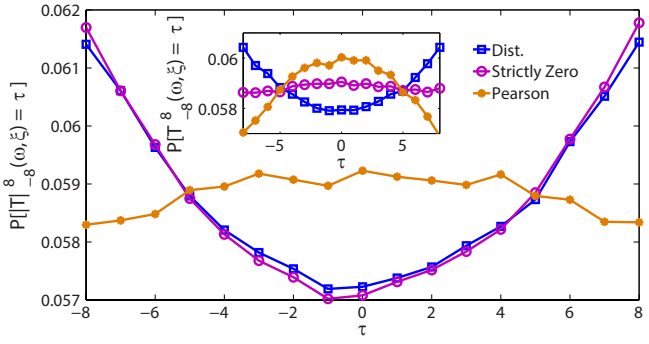

Figure 4. Probability distribution of $|T|^{8}_{-8}(\omega,\xi)$ and $T^{8}_{-8}(\omega(t),\xi(t))$ (inset) for various i.i.d. series, setting the means to zero in different ways. For "Dist." the time series $\omega(t)$ and $\xi(t)$ are drawn from a normal distribution with zero mean; for "Strictly Zero" the sample means ($\bar{\omega}$, $\bar{\xi}$) are then removed from the corresponding series; for "Pearson" the sample means and standard deviations are calculated over the Ω points that enter into the calculation for each τ, and then their means are removed and the series are divided by their standard deviations. Fixed parameters are: $\Omega=16$, with $N=2^{23}$ realizations.

case where $\omega(t)$ is drawn from a distribution with non-zero mean.

We find that the parameter of interest in this case is the coefficient of variation (CV) – the standard deviation divided by the mean. Specifically, it is the CV of the time series $\omega(t)$, and the bias is independent of the CV of $\xi(t)$. The reason for this is due to the similarity of Eq. (3) to a random walk, which we further elaborate on in Sect. 2.4. Figure 3 shows that as the CV of $\omega(t)$ decreases the bias in Eq. (5) increases, as well as in Eq. (4) (inset). For each CV in Fig. 3 two choices of mean and standard deviation are given, and in all cases we see that the same CV in $\omega(t)$ results in the same curve. We find that this result holds for Gaussian and uniformly distributed i.i.d. variables, but not in general (e.g. Student's t distribution with three degrees of freedom).

This would suggest that a simple solution, for i.i.d. time series, is to maximize the CV of $\omega(t)$, i.e. ensure that the mean is set to zero. While setting the mean to zero does drastically reduce the bias it does not eliminate it entirely. In Fig. 4 we show the effect on Eqs. (5) and (4) (inset) of setting the mean to zero in various ways: "Dist.", drawing our time series from a normal distribution (zero mean unit variance); "Strictly Zero", drawing it from a normal distribution, and then removing the sample mean from each series so it is exactly zero; "Pearson", drawing it from a normal distribution, then removing the sample mean and dividing by the sample standard deviation for each series, for each τ.

We can see that bias is still present, though the Pearson correlation shows an inverted behaviour relative to the previous bias, although in this case the bias is minuscule, as can be seen from the y axis. Therefore, when $\omega(t)$ has a very large CV, as in Fig. 4, most of the bias is eliminated in the i.i.d. case. However, this is not true for the case of power-law autocorrelations, which we discuss in Sect. 3. In that case the Pearson correlation gives the best performance of these estimators as well. However, in that case it is also peaked at extreme lags, as well as showing the peak at zero seen in Fig. 4.

2.3 Independent identically distributed case: mapping to a random walk

So why does this bias exist even in the i.i.d. case? In this section we will show that it is due to the similarity between Eq. (3) and a random walk. We will first define a normal random walk, and then show how, in some circumstances, it can be mapped directly to Eq. (3). Finally, we will discuss previous results showing how this translates into the observed bias.

Since the seminal work of Einstein (1905), random walks have been studied extensively (Feller, 1950; Spitzer, 1976), and have provided insight into a wide array of problems (Bouchaud and Georges, 1990; Brockmann et al., 2006). A random walk starts from an initial position $S(0)=S_0$, and the position of the walker at step n is

$$S(n) = S_0 + \sum_{i=1}^{n} x(i) \tag{6}$$

$$= S(n-1) + x(n). \tag{7}$$

Here $x(t)$ denotes the random jump the walker takes at time t, in this work drawn from a continuous distribution symmetric about zero.

When the time series $\omega(t)$ in Eq. (3) is a constant-valued time series it can be mapped directly to Eq. (6). Without loss of generality we use the case that $\omega(t)$ is one for all times, and rewrite Eq. (3) as

$$C(\tau,t) = \frac{1}{\Omega}\sum_{i=t}^{t+\Omega-1} \xi(i+\tau). \tag{8}$$

Since we are comparing different τ with fixed t we can set $t=0$ without loss of generality and rewrite Eq. (8) as

$$C(\tau) = \frac{1}{\Omega}\sum_{i=\tau}^{\tau+\Omega-1} \xi(i). \tag{9}$$

In this form we can map Eq. (3) to a random walk, but first we must define increments $x(t)$. The difference between Eq. (9) at $\tau=0$ and $\tau=1$ is

$$x(1) = \frac{1}{\Omega}\Big(\xi(\Omega) - \xi(0)\Big). \tag{10}$$

From this we can see that in general the difference for $\tau = i-1$ and $\tau = i$ is

$$x(i) = \frac{1}{\Omega}\Big(\xi(i+\Omega-1) - \xi(i-1)\Big). \tag{11}$$

Using these increments, and the initial condition $C(\tau_i) = C_0$, the final mapping is (for $\tau > \tau_i$),

$$C(\tau) = C_0 + \sum_{i=\tau_i+1}^{\tau} x(i). \tag{12}$$

The increments of this equation are all uncorrelated as long as $\tau_f - \tau_i \leq \Omega$. This condition means that no value of $\xi(t)$ will appear in more than one increment, $x(i)$, and therefore all increments will be uncorrelated as long as the values of $\xi(t)$ are themselves uncorrelated. This then means that Eq. (12), with the increments given by Eq. (11), is exactly an i.i.d. random walk as in Eq. (6).

2.4 Independent identically distributed case: comparison with analytical results

In the case of a random walk (with continuous steps, drawn from a symmetric distribution), the solution to the probability of the maximum occurring at a given τ has been solved analytically by Majumdar (2010). He showed that the probability that a random walk of L steps will have reached its maximum at step n is

$$P(n|L) = \binom{2n}{n}\binom{2(L-n)}{(L-n)}2^{-2L}. \tag{13}$$

However, this only holds for the maximum of the random walk, and not the absolute value of the maximum. This equation is closely related to the Lévy arcsine distribution (Lévy, 1939) that describes the probability that $S(n)$ is positive for different fractions of the random walk. Free boundary conditions are assumed here as well. For periodic boundary conditions, one has translational invariance such that Eq. (13) no longer applies and $P(n|N) \equiv$ const. This corresponds to the case where the mean of the increments $x(i)$ is exactly zero over N steps. We utilize this fact to correct for the bias in Sect. 3.

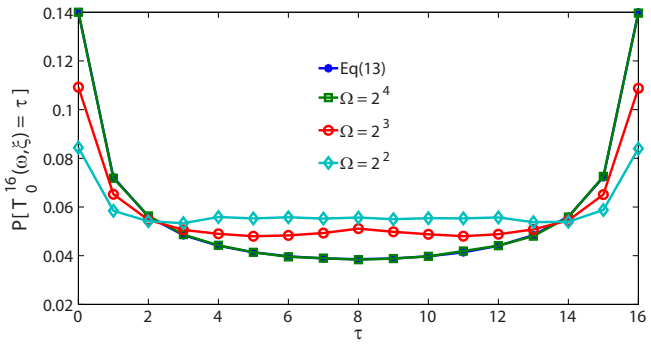

Figure 5. Comparison of Eqs. (13) and (12). We have calculated τ from -8 to 8, then shifted it to start at 0 to compare with the Eq. (13) which starts at 0. We can see that as Ω gets larger the distribution approaches the analytical result for the random walk, becoming exact for $\tau_f - \tau_i \leq \Omega$.

Equation (13) also applies to the random walk we defined in Eq. (12) as long as $\tau_f - \tau_i \leq \Omega$ (the regime where all increments are uncorrelated). In Fig. 5 we show how the random walk deviates increasingly from this result as Ω gets smaller than $\tau_f - \tau_i$.

With this result we can now understand the origin of the bias in the cross-correlation for i.i.d. variables. As we saw previously in Fig. 3, the bias in Eq. (3) increases as the CV of $\omega(t)$ approaches zero; in this limit Eq. (3) can be mapped to a random walk. We can see how this arises by rewriting $\omega(t)$ in terms of its mean value, μ_ω, plus a random fluctuation, $\omega(i) = \mu_\omega + \epsilon(i)$. In the limit that the CV of $\omega(t)$ approaches zero, $\omega(i) = \mu_\omega$ – a constant valued time series – and Eq. (3) can be mapped to a random walk as discussed in Sect. 2.3.

Therefore, as the CV decreases the bias in the cross-correlation should approach the theoretical result of Eq. (13); Fig. 6 shows this is indeed the case. While we have only tested this for Gaussian and uniformly distributed variables, it is very likely that Eq. (13) gives the limiting behaviour of the cross-correlation independent of specific distribution and, hence, is an upper bound on the observable bias for i.i.d. variables.

The bias for long-range autocorrelations can also be (partially) understood now: as β_ω approaches 1, $\omega(t)$ approaches a constant-valued series. In the case that $\xi(t)$ is an i.i.d. series, this can be mapped to a random walk again as before. In Fig. 7 we show how the limit of Eq. (13) is approached from below for increasing β_ω, where $\xi(t)$ is a Gaussian-distributed i.i.d. series. This indicates again that Eq. (13) provides an

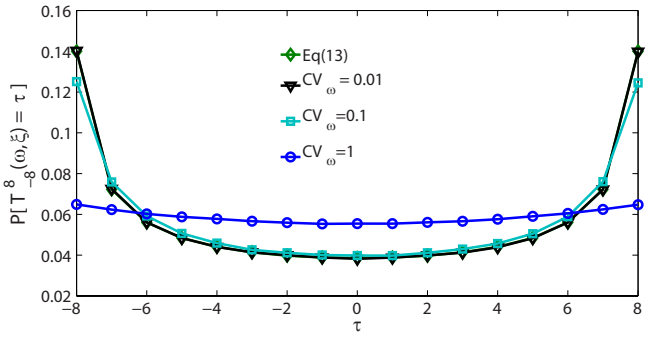

Figure 6. As the CV of $\omega(t)$ approaches zero the statistics of Eq. (3) better approximates the result in Eq. (13). Here we have shifted Eq. (13) to be centred around zero to compare with Eq. (3). Fixed parameters are: $\Omega = 64$, with $N = 2^{21}$ realizations.

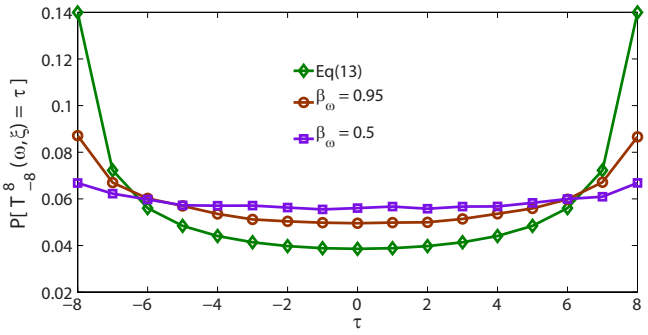

Figure 7. Here we show how Eq. (13) is approached for increasing β_ω. The mapping becomes exact in the limit $\beta_\omega \to 1$, where $\omega(t)$ is a constant-valued series. Fixed parameters are: $\Omega = 32$, with $N = 2^{18}$ realizations.

upper bound on the observable bias. It is important to realize, though, that Eq. (13) is not necessarily the limit when $\xi(t)$ is not i.i.d.. This is particularly true when $0 < \beta_\xi < 1$, which corresponds to a random walk with correlated steps.

3 Estimating the cross-correlation in the frequency domain

As we show in the following, the bias discussed above can be effectively compensated for by measuring the cross-correlation in the frequency domain, i.e. by using a different estimator. This approach is based on the cross-correlation theorem, which states that the inverse Fourier transform of the product of the Fourier transforms of the individual time series is related to the cross-correlation as

$$C(\tau) = \sum_{\nu=0}^{\Omega-1} \widetilde{\omega}^*(\nu)\widetilde{\xi}(\nu) \exp\left(2\pi i \tau \nu/\Omega\right). \qquad (14)$$

Here we use the notation that $\widetilde{\omega}(\nu)$ is the Fourier transform of $\omega(t)$, and $\omega^*(t)$ is the complex conjugate of $\omega(t)$. In order for this to be well defined, $\xi(t)$ and $\omega(t)$ must be stationary

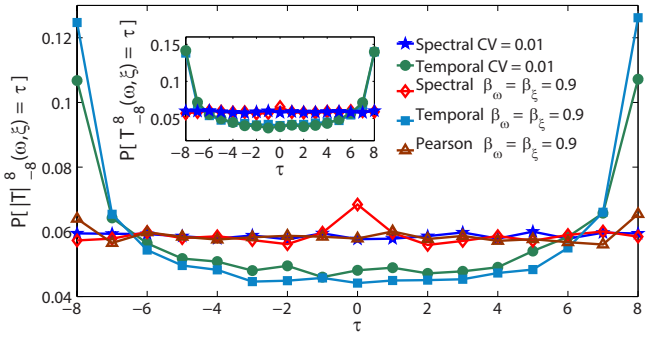

Figure 8. The distribution of $|T|^8_{-8}(\omega(t),\xi(t))$ and $T^8_{-8}(\omega(t),\xi(t))$ (inset) for different time delay estimates when (i) both signals have long-range autocorrelations, $\beta_\xi = \beta_\omega = 0.9$ and (ii) both signals are i.i.d. and $\omega(t)$ has a CV of 0.01. We can see that the frequency domain estimator (Spectral), Eq. (14), is free of the bias seen for the estimator given by (Temporal), Eq. (3). We also show the result of using the Pearson correlation (Pearson) – here the bias is greatly reduced, but still peaked at extreme lags. Fixed parameters are: $\Omega = 17$, with $N = 2^{16}$ realizations.

and periodic over the interval Ω. Equation (14) is a simple extension of the Wiener–Khinchin theorem (Schumann, 2011), which uses the Fourier transform to compute the autocorrelation of a function.

Estimating the cross-correlation in this way results in lags computed over the range $[-\Omega/2,\ \Omega/2-1]$ (Ω even), or $[-(\Omega-1)/2,\ (\Omega-1)/2]$ (Ω odd). Using the Fourier transform to estimate the cross-correlation effectively eliminates the bias in Eq. (3) by enforcing that $C(\tau)$ is periodic over this interval. However, this is only the case when we consider all time lags calculated. It is important to realize that the true cross-correlation need not be periodic and neither are typically the underlying time series, but periodicity is often a reasonable first-order approximation.

We also note that $C(\tau)$ is not necessarily symmetric about 0, so it can still distinguish the directionality of the correlation. Figure 8 shows that estimating the cross-correlation in this way results in Eqs. (5), and (4) (inset), being uniformly distributed when the time series are i.i.d. with small CV. Similarly, for power-law autocorrelations this estimator gives very uniform results, though there is a slight peak about zero. This indicates that Eq. (14) does not have the bias inherent in Eq. (3). For power-law autocorrelation we also compare Eq. (14) to the Pearson correlation. We can see that for the Pearson correlation Eq. (5) is still biased towards extreme delays, and this is also true for Eq. (4) (not shown).

The bias is also eliminated when cross-correlations are present. To show this, we consider again the temperature data discussed in the Introduction. We focus on the pairs assigned an extreme lag (largest or smallest possible lag) by Eq. (3), as that is where the bias is most significant and the time lag most likely to be misclassified. Figure 9 shows the new distribution of lags, originally classified as $\tau = \pm 200$, when

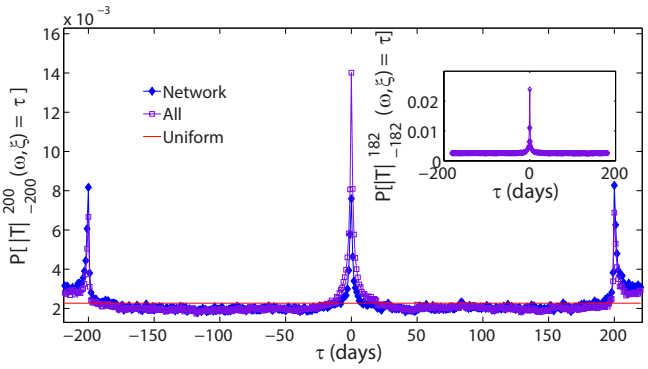

Figure 9. Reanalysis of the ±200-day lag pairs from Fig. 1 using Eq. (14). This was done for 10 % of the ±200-day lag links in the network, and 5 % of the 200-day lag points for all pairs, which were randomly chosen. These are compared to a uniform distribution for reference. Both of these plots show a peak about zero, which is more physically reasonable then a lag of ±200 days, and therefore more likely to be the correct time delay. For these we used a value of $\Omega = 441$ to analyse the peaks about $\tau = \pm 200$. The inset shows a reanalysis of all links using Eq. (14). We can see that there is no bias in that figure, as well as a larger peak about zero than in Fig. 1. Here a value of $\Omega = 365$ was used for comparison with the original results.

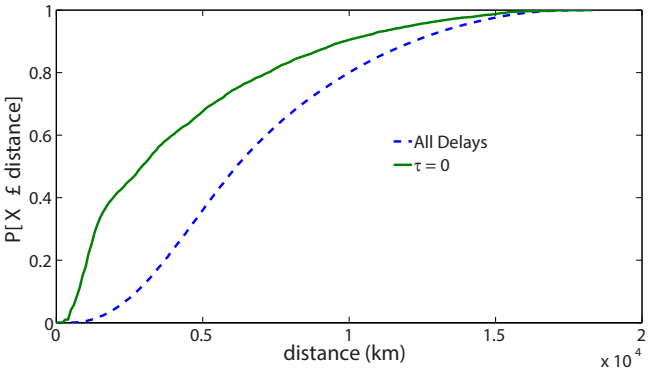

Figure 10. Examining the reanalysed network data in Fig. 9: the cumulative density function (CDF) of the distance between points where the network link has been reassigned a lag $\tau = 0$, and the CDF of the distance between all reanalysed network links (All Delays). We can see that points with zero time lag are closer together on average, indicating that some of these links in Fig. 9 are physically meaningful.

using Eq. (14) (the reanalysis of all links is shown in the inset). Evidently, the distribution is vastly different, no longer being only $\tau = \pm 200$, and indicates clearly that the large lag peaks at the extremes in Fig. 1 are mostly an artifact. Here we have used $\Omega = 441$ in Eq. (14) to calculate time lags over the range from $\tau_i = -220$ to $\tau_f = 220$. This allows us to see that these peaks do not occur at the largest possible lags.

More importantly, there is a clear maximum close to zero lag, which would be a more physically plausible time delay. Equally important is the fact that there are still a significant amount of links assigned extreme time lags, which are unlikely to be physically meaningful, and therefore probably spurious. The height of the peaks at $\tau = |200|$ also tells us that only approximately 0.8 % of these large lag links remain after reanalysis. In the inset we show that reanalysing all points with Eq. (14) is also free of the bias seen in Fig. 1, as well as having a larger peak about zero.

To demonstrate that the time delays estimated by Eq. (14) are physically meaningful we examine how the lag is related to the distance between the locations where the time series were recorded; for small time lags we would typically expect the distance to be shorter compared to large lags. Figure 10 shows that this is indeed the case: pairs of time series which now have an estimated time delay of zero are much more likely to be closer together than for all pairs. The fact that the nodes are close (spatially) with a short time lag makes it likely that their temperatures are related, and that it was correct to originally include them in the network. However, it also shows that Eq. (3) widely misestimated the correct time delay, and any network analysis based on the delays would be dubious. In addition, we also note that many of the pairs still have very large lags, indicating that the links may not be physical.

4 Conclusions

In this work we discussed an, as yet unmentioned, bias in the time lag associated with estimating the maximum cross-correlation. This has the potential to affect dynamical networks through the inclusion of spurious links, and results in the estimated time lags being biased towards extreme values. In addition, the bias is exacerbated when the time series have power-law autocorrelations, which are common in geophysical time series. For the case of i.i.d. time series, we demonstrated that the bias is a result of Eq. (3) approximating a random walk, where the approximation gets better (bias gets worse) as the CV goes to zero. This is similar to the case of long-range autocorrelations; as β_ω approaches one $\omega(t)$ approaches a constant valued series, and we can again map Eq. (3) to a random walk.

This bias can be eliminated by estimating the cross-correlation in the frequency domain, using Eq. (14). In addition, we have shown that this better estimates the time delays; as seen in Fig. 10 it reasonably assigned pairs that were spatially close smaller time delays on average. We can also see from doing this analysis that spurious links are likely being included in the network. Equation (14) still assigned extremal lags to a number of links, as seen in Fig. 9, and these are highly unlikely to be physical, yet are included in the original network analysis.

We also note that we have not discussed the important issue of whether a given time lag corresponds to a significant cross-correlation. In order to determine which relationships

are significant, additional techniques are required. For example in Paluš (2007) significance thresholds are determined using surrogate time series. However, these methods are unlikely to remove the bias discussed here. This is because the probability distribution of the cross-correlation, for the surrogate data, is the same for all τ. Therefore, larger significance levels at larger (absolute) time lags would unfairly penalize large lag correlations. We speculate that information-theoretic-based methods would perform as well or better than the cross-correlation at determining the time delay and direction of interaction between nodes.

Our results for mutual information, not shown here, indicate that it also performs well at eliminating the bias in estimating the time delay. It does not do this as well, or as efficiently as the frequency domain estimate of the cross-correlation however. In general, information-theoretic methods have increased computational costs and some simply indicate the direction of interaction, not a specific time lag. For example, transfer entropy as originally formulated (Schreiber, 2000) does not determine a time lag, though its extension in Runge et al. (2012) and other techniques such as partial mutual information (Frenzel and Pompe, 2007) do. It is our opinion that, for a given application, a variety of methods should be tested on appropriate models to determine which achieves the best and most robust results.

Acknowledgements. The authors would like to thank P. Grassberger and M. Paczuski for helpful discussions. This project was financially supported by Alberta Innovates – Technology Futures.

Edited by: J. Donges
Reviewed by: three anonymous referees

References

Baiesi, M. and Paczuski, M.: Scale-free networks of earthquakes and aftershocks, Phys. Rev. E, 69, 066106, doi:10.1103/PhysRevE.69.066106, 2004.

Bashan, A., Bartsch, R. P., Kantelhardt, J. W., Havlin, S., and Ivanov, P. C.: Network physiology reveals relations between network topology and physiological function, Nature Communications, 3, 702, doi:10.1038/ncomms1705, 2012.

Bassett, D. S., Owens, E. T., Daniels, K. E., and Porter, M. A.: Influence of network topology on sound propagation in granular materials, Phys. Rev. E, 86, 041306, doi:10.1103/PhysRevE.86.041306, 2012.

Bouchaud, J. and Georges, A.: Anomalous diffusion in disordered media: statistical mechanisms, models and physical applications, Phys. Rep., 195, 127–293, 1990.

Brockmann, D., Hufnagel, L., and Geisel, T.: The scaling laws of human travel, Nature, 439, 462–465, 2006.

Davidsen, J., Grassberger, P., and Paczuski, M.: Earthquake recurrence as a record breaking process, Geophys. Res. Lett., 33, L11304, doi:10.1029/2006GL026122, 2006.

Davidsen, J., Grassberger, P., and Paczuski, M.: Networks of recurrent events, a theory of records, and application to finding causal signatures in seismicity, Phys. Rev. E, 77, 066104, doi:10.1103/PhysRevE.77.066104, 2008.

Deza, J., Barreiro, M., and Masoller, C.: Inferring interdependencies in climate networks constructed at inter-annual, intra-season and longer time scales, Eur. Phys. J.-Spec. Top., 222, 511–523, 2013.

Dodds, P. S. and Rothman, D. H.: Geometry of river networks. I. Scaling, fluctuations, and deviations, Phys. Rev. E, 63, 016115, doi:10.1103/PhysRevE.63.016115, 2000.

Donges, J., Zou, Y., Marwan, N., and Kurths, J.: Complex networks in climate dynamics, Eur. Phys. J.-Spec. Top., 174, 157–179, 2009a.

Donges, J., Zou, Y., Marwan, N., and Kurths, J.: The backbone of the climate network, EPL-Europhys. Lett., 87, 48007, doi:10.1209/0295-5075/87/48007, 2009b.

Ebert-Uphoff, I. and Deng, Y.: Causal discovery for climate research using graphical models, J. Climate, 25, 5648–5665, 2012.

Einstein, A.: On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat, Ann. Phys., 17, 549–560, 1905.

Feller, W.: An introduction to probability theory and its applications, Vol. 1, John Wiley & Sons, 1950.

Feng, Q. Y. and Dijkstra, H.: Are North Atlantic multidecadal SST anomalies westward propagating?, Geophys. Res. Lett., 41, 541–546, 2014.

Frenzel, S. and Pompe, B.: Partial mutual information for coupling analysis of multivariate time series, Phys. Rev. Lett., 99, 204101, doi:10.1103/PhysRevLett.99.204101, 2007.

Gu, C., Schumann, A. Y., Baiesi, M., and Davidsen, J.: Triggering cascades and statistical properties of aftershocks, J. Geophys. Res., 118, 4278–4295, doi:10.1002/jgrb.50306., 2013.

Hlinka, J., Hartman, D., and Paluš, M.: Small-world topology of functional connectivity in randomly connected dynamical systems, Chaos, 22, 033107, doi:10.1063/1.4732541, 2012.

Hlinka, J., Hartman, D., Vejmelka, M., Novotná, D., and Paluš, M.: Non-linear dependence and teleconnections in climate data: sources, relevance, nonstationarity, Clim. Dynam., 42, 1873–1886, 2013a.

Hlinka, J., Hartman, D., Vejmelka, M., Runge, J., Marwan, N., Kurths, J., and Paluš, M.: Reliability of inference of directed climate networks using conditional mutual information, Entropy, 15, 2023–2045, 2013b.

Huybers, P. and Curry, W.: Links between annual, Milankovitch, and continuum temperature variability, Nature (London), 441, 329–332, 2006.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S., Hnilo, J. J., Fiorino, M., and Potter, G. L.: NCEP–DOE AMIP-II Reanalysis (R-2), B. Am. Meteorol. Soc., 83, 1631–1643, 2002.

Kantelhardt, J. W., Koscielny-Bunde, E., Rybski, D., Braun, P., Bunde, A., and Havlin, S.: Long-term persistence and multifractality of precipitation and river runoff records, J. Geophys. Res., 111, D01106, doi:10.1029/2005JD005881, 2006.

Kawale, J., Steinbach, M., and Kumar, V.: Discovering dynamic dipoles in climate data, in: SIAM International Conference on Data Mining (SDM), 28–30 April, Phoenix Arizona, USA, 107–108, 2011.

Koscielny-Bunde, E., Bunde, A., Havlin, S., Roman, H. E., Goldreich, Y., and Schellnhuber, H.: Indication of a universal persistence law governing atmospheric variability, Phys. Rev. Lett., 81, 729, doi:10.1103/PhysRevLett.81.729, 1998.

Lévy, P.: Sur certains processus stochastiques homogènes, Compos. Math., 7, 283–339, 1939 (in French).

Majumdar, S. N.: Universal first-passage properties of discrete-time random walks and Lévy flights on a line: Statistics of the global maximum and records, Physica A, 89, 4299–4316, doi:10.1016/j.physa.2010.01.021, 2010.

Mantilla, R., Gupta, V. K., and Mesa, O. J.: Role of coupled flow dynamics and real network structures on Hortonian scaling of peak flows, J. Hydrol., 322, 155–167, 2006.

Martin, E. A., Paczuski, M., and Davidsen, J.: Interpretation of link fluctuations in climate networks during El Niño periods, EPL-Europhys. Lett., 102, 48003, doi:10.1209/0295-5075/102/48003, 2013.

Paluš, M.: From nonlinearity to causality: statistical testing and inference of physical mechanisms underlying complex dynamics, Contemp. Phys., 48, 307–348, 2007.

Paluš, M., Hartman, D., Hlinka, J., and Vejmelka, M.: Discerning connectivity from dynamics in climate networks, Nonlin. Processes Geophys., 18, 751–763, doi:10.5194/npg-18-751-2011, 2011.

Peixoto, T. P. and Davidsen, J.: Network of recurrent events in the Olami-Feder-Christensen model, Phys. Rev. E, 77, 066107, doi:10.1103/PhysRevE.77.066107, 2008.

Peixoto, T. P., Doblhoff-Dier, K., and Davidsen, J.: Spatiotemporal correlations of aftershock sequences, J. Geophys. Res., 115, B10309, doi:10.1029/2010JB007626, 2010.

Pelletier, J. and Turcotte, D. L.: Long-range persistence in climatological and hydrological time series: analysis, modeling and application to drought hazard assessment, J. Hydrol., 203, 198–208, 1997.

Runge, J., Heitzig, J., Petoukhov, V., and Kurths, J.: Escaping the curse of dimensionality in estimating multivariate transfer entropy, Phys. Rev. Lett., 108, 258701, doi:10.1103/PhysRevLett.108.258701, 2012.

Runge, J., Petoukhov, V., and Kurths, J.: Quantifying the Strength and Delay of Climatic Interactions: The Ambiguities of Cross Correlation and a Novel Measure Based on Graphical Models, J. Climate, 27, 720–739, 2014.

Schreiber, T.: Measuring information transfer, Phys. Rev. Lett., 85, 461, doi:10.1103/PhysRevLett.85.461, 2000.

Schumann, A. Y.: Fluctuations and synchronization in complex physiological systems, Logos Verlag Berlin, 2011.

Spitzer, F.: Principles of random walk, vol. 34, Springer, 1976.

Steinhaeuser, K., Ganguly, A., and Chawla, N.: Multivariate and multiscale dependence in the global climate system revealed through complex networks, Clim. Dynam., 39, 889–895, 2012.

Tirabassi, G. and Masoller, C.: On the effects of lag-times in networks constructed from similarities of monthly fluctuations of climate fields, EPL-Europhys. Lett., 102, 59003, doi:10.1209/0295-5075/102/59003, 2013.

Tsonis, A. and Swanson, K.: Topology and Predictability of El Niño and La Niña Networks, Phys. Rev. Lett., 100, 228502, doi:10.1103/PhysRevLett.100.228502, 2008.

Tsonis, A., Swanson, K., and Wang, G.: On the role of atmospheric teleconnections in climate, J. Climate, 21, 2990–3001, 2008.

Vyushin, D. I., Kushner, P. J., and Mayer, J.: On the origins of temporal power-law behavior in the global atmospheric circulation, Geophys. Res. Lett., 36, L14706, doi:10.1029/2009GL038771, 2009.

Yamasaki, K., Gozolchiani, A., and Havlin, S.: Climate Networks around the Globe are Significantly Affected by El Niño, Phys. Rev. Lett., 100, 228501, doi:10.1103/PhysRevLett.100.228501, 2008.

Zaliapin, I. and Ben-Zion, Y.: Earthquake clusters in southern California, I: Identification and stability, J. Geophys. Res., 118, 2847–2864, doi:10.1002/jgrb.50179, 2013.

Zanardo, S., Zaliapin, I., and Foufoula-Georgiou, E.: Are American rivers Tokunaga self-similar? New results on river network topology and its climatic dependence, J. Geophys. Res., 118, 166–183, doi:10.1002/jgrf.20029, 2013.