Science of the Total Environment 860 (2023) 160219

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

The spectral color of natural and anthropogenic time series and its impact on the statistical significance of cross correlation

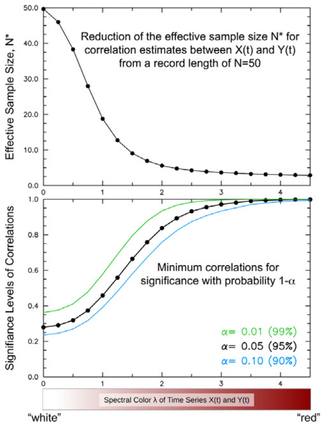
Ismael Núñez-Riboni [a,*], Dudley B. Chelton [b], Valentina Marconi [c]

[a] Thünen-Institut für Seefischerei, Herwigstraße 31, 27572 Bremerhaven, Germany
[b] Oregon State University, USA
[c] Institute of Zoology, Zoological Society of London, United Kingdom of Great Britain and Northern Ireland

HIGHLIGHTS

- Correlation of time series is common to study climate change impact on ecosystems.
- Serial dependence in time series reduces accuracy of correlation significance test.
- Ecology studies generally lack an accurate significance estimate for correlation.
- Monte Carlo simulations allow assessment of performance of significance tests.
- Artificial Skill Method is superior for ecologic, climatic and anthropogenic time series.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Martin Drews

Keywords:
Significance
Pearson correlation
Auto-correlated time series
Climate

ABSTRACT

The cross-correlation between time series is a common tool to study and quantify the impact of climatic and anthropogenic changes on ecosystems. The traditional method for estimating the statistical significance of correlation relies on the assumption that the data are independent, but time series found in nature are often strongly auto-correlated because of low-frequency environmental variability and ecosystem inertia. Previous authors have used Monte Carlo simulations to study the impact of serial auto-correlation on the significance of cross-correlations. Most studies have used random time series that are often a poor representation of those found in nature, e.g., low-order auto-regressive models with normally distributed noise. Moreover, we are not aware of any tests of the applicability of those methods to anthropogenic time series. Here, we study the effect of serial auto-correlation on the performance of two methods for estimating the significance of cross-correlations determined from Monte Carlo simulations with time series that are generated synthetically based on power-law specification of spectral characteristics. Such time series have an autocorrelation structure defined by a single parameter, their spectral "color", and are generally more convenient representations of natural time series than the autoregressive models. Our results show that one of the two methods considered here accurately reproduces prescribed error rates for the wide range of spectral colors representative of climatic, ecological and anthropogenic time series. For this, we characterized roughly 1800 observational records in different categories of spectral colors, including climate variability, abundance of vertebrate species, and pollution. We specifically focus on time series with annual sampling over data records of at least 40 years, which are particularly relevant for

[*] Corresponding author.
E-mail address: ismael.nunez-riboni@thuenen.de (I. Núñez-Riboni).

http://dx.doi.org/10.1016/j.scitotenv.2022.160219
Received 21 April 2022; Received in revised form 11 November 2022; Accepted 12 November 2022
Available online 17 November 2022

climate studies. The methodology advocated in this study provides a simple and realistic assessment of the significance of sample estimates of cross correlation for time series with any sample interval and record length.

1. Introduction

Estimating the statistical significance of a sample estimate of the cross correlation between two time series is important in natural sciences. For instance, calculating the correlation between air temperature and abundance of animal species provides insight into the impact of climate variability on an ecosystem and helps identify potential predictors for ecosystem changes in models of population dynamics (e.g., Iles and Beverton, 1998). For this goal, use of the Pearson correlation coefficient is widespread (Puth et al., 2014; Humphreys et al., 2019; Runge et al., 2019), mainly because of its simplicity and easily understandable scale. Some recent examples of the use of the Pearson correlation coefficient (referred to hereinafter as simply the correlation) in ecology studies include Beaugrand and Reid, 2003; Schindler et al., 2010; Li et al., 2013; Pershing et al., 2015; Akimova et al., 2016; Capuzzo et al., 2018; Bedford et al., 2020.

Estimation of the significance of any statistic is fundamental to distinguish robust outcomes from those occurring merely by chance (Fisher, 1935). Estimation of the statistical significance of correlation is, however, challenging when the time series are serially correlated. In such cases, successive samples are correlated and therefore not independent. The effective number of independent realizations (or effective number of degrees of freedom N*) is thus smaller than the record length N. Failing to take this auto-correlation into account results in an erroneously small critical value for the cross correlation, which can wrongly indicate significant correlation where no real correlation exists (i.e., false positives). Previous studies (Clifford et al., 1989; Pyper and Peterman, 1998; Lennon, 2000) have shown that even modest levels of autocorrelation can cause the traditional significance test to fail in up to 20 % more cases compared to statistically independent time series (e.g., Fig. 1 of Clifford et al., 1989 for an auto-correlation of 0.6 at a lag of 1). This is an important issue in many fields of natural sciences, and is a common weakness in ecological studies (Brown et al., 2011).

There are three approaches to estimate N^* or the significance of a sample estimate of the correlation between auto-correlated the time series: 1) Modification of the test statistic (e.g., Bartlett, 1946; Bayley and Hammersley, 1946; Davis, 1976; Garrett and Petrie, 1981; Chelton, 1983; Kope and Botsford, 1988; Pyper and Peterman, 1998); 2) Cross validation (Michaelsen, 1987) and 3) Randomization (bootstrapping) of the time series (Ebisuzaki, 1997; Lennon, 2000). We focus here only on methods for modifying the significance test. A justification of this choice is given in Section 7 below.

To study the effects of auto-correlation time scale on the statistical significance of a sample estimate of the cross correlation between two time series, it is first necessary to characterize the auto-correlation by some simple parameterization. This has been approached in two different ways. One approach uses auto-regressive models (e.g., Clifford et al., 1989; Pyper and Peterman, 1998; Dale and Fortin, 2002), a commonly used example being the second-order auto-regressive Markov process, AR(2) (e.g., Eq. 3.2.16 of Box and Jenkins, 1970):

$$X(t) = \phi_1 X(t-1) + \phi_2 X(t-2) + a(t), \tag{1}$$

where the time index t spans a prescribed record length $N\Delta t$ consisting of N samples separated by a sample interval of Δt. The parameters ϕ_i are the auto-correlations at lags $i\Delta t$ for $i = 1$ and 2, and $a(t)$ is the noise of the process. Numerical values for $a(t)$ and the first two elements of the AR(2) pro-

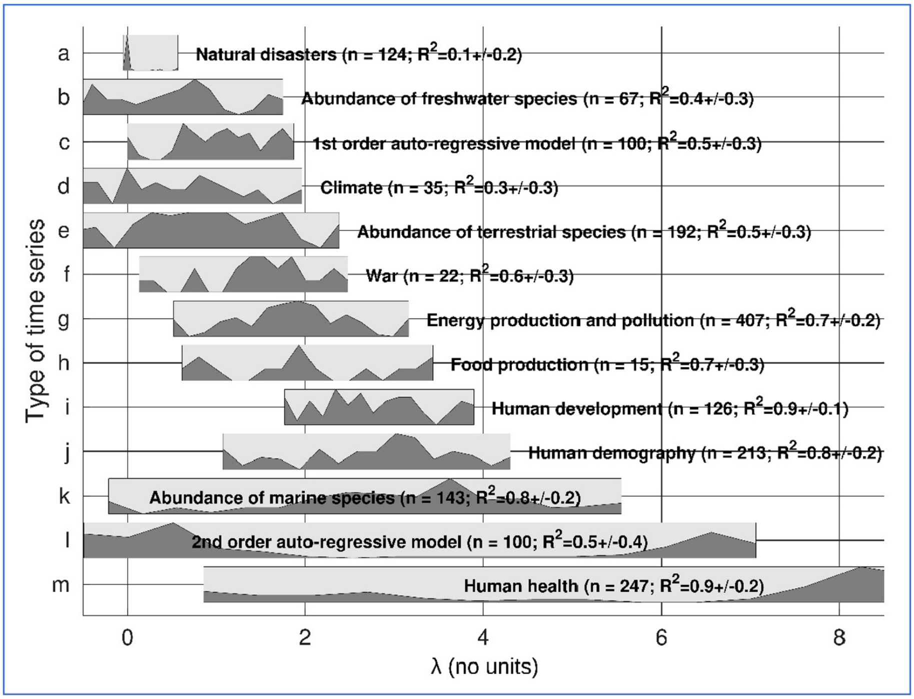


Fig. 1. Range of spectral colors λ (which is a parameterization of the auto-correlation time scale; see Fig. 2) for various types of climatic, ecological and anthropogenic time series. For each type of time series, the horizontal extent of each histogram corresponds to the 0.05 and 0.95 quantiles of the distribution of λ values, sorted from top to bottom according to the upper bound of the range of λ. The relative numbers of time series with a particular λ value are shown with dark gray histograms inside the bars. The number of time series n included in each category, the average coefficient of determination R^2 for the fit of λ, and its STD are also listed.

cess are drawn randomly from a normal distribution with zero mean and a specified standard deviation (STD) σ_a of the noise time series $a(t)$. For present purposes of correlation analysis, σ_a is irrelevant. The AR(2) process is thus defined by two parameters. The simpler first-order Markov process AR(1) is the particular case of Eq. (1) with $\phi_2 = 0$ (Eq. 3.2.10 of Box and Jenkins, 1970) and is thus defined by one parameter.

The second approach to parameterizing the autocorrelation is based on modeling the spectrum of the time series as a power-law dependence on frequency f:

$$S(f) = f^{-\lambda}, \tag{2}$$

(e.g., Lennon, 2000), where S is the power spectral density of the time series and λ is a parameter that effectively defines the auto-correlation time scale of the time series (Lennon, 2000). Except on rare occasions, λ is generally non-negative for time series found in nature. In analogy to spectral analysis of electromagnetic waves, the parameter λ is called the "spectral color" of the time series (e.g., Schroeder, 1991; Lennon, 2000), with positive, negative and zero values of λ corresponding to "red", "blue" and "white" time series, respectively.

A number of previous studies have characterized time series in terms of their spectral color λ, portraying diverse examples of time series which follow the power-law dependence on frequency (Eq. (2)): Acoustic noise, semiconductor devices, natural and human-related disasters such as flooding of rivers, droughts, stock markets and outages of electrical power (Schroeder, 1991). Other examples include traffic flow, DNA sequences and music (Halley, 1996), as well as atmospheric (Pelletier, 1997; Vasseur and Yodzis, 2004) and oceanic (Vasseur and Yodzis, 2004) variations, and abundance of terrestrial and aquatic species (Inchausti and Halley, 2002).

Spectral time series have some advantages over the AR time series, especially for representing ecological phenomena for which λ can be large (Halley, 1996). The role of λ on the traditional (i.e., non-modified) significance test of correlation has been studied by Lennon, 2000. To the best of our knowledge, no previous study has considered the associated effects of serial correlation on the significance test based on spectral time series (Eq. (2)). The present study attempts to fill this gap by comparing the performance of two methods to modify the significance test of cross correlation.

One of the methods considered here was developed by Pyper and Peterman, 1998 (P&P98 hereinafter) based on their comparison of several methods for modifying the significance test from a Monte Carlo (MC) analysis. P&P98 advocated a modification of the procedure they found to be best and that they attribute to Chelton, 1984. They referred to their test as the "modified Chelton method" and it has often been quoted that way in the literature (e.g., Bedford et al., 2020). We will refer to it herein as the "Pyper and Peterman Method" (PPM).

The PPM estimate of the significance test of cross-correlation is effectively based on integration of the auto-correlation functions. It has been widely used in the natural sciences (see for instance Brown et al., 2011). While it was developed for application to fishery time series (e.g., Richardson and Schoeman, 2004; Möllmann et al., 2008; Beaugrand and Kirby, 2010; Kirby and Beaugrand, 2009; Hermant et al., 2010; Bedford et al., 2020), it has also been used in other disciplines such as meteorology (Li et al., 2021; Wang et al., 2021), oceanography (Huo et al., 2021; Xie et al., 2021) and terrestrial ecology (Caldwell et al., 2021), with more than 800 citations as of September 2022, according to Google Scholar.

PPM was developed and tested with auto-regressive time series and has four shortcomings: 1) Its calculation of N^* is based on the assumption that sample estimates of the auto-correlation functions are the true values that are required in the analytically derived formula that is the theoretical basis for the PPM; 2) The theoretical equation on which PPM is based is an equation for the effective number of degrees of freedom N^* (Eq. (1) in P&P98) that was derived analytically with an assumption of a large sample size; this is often not valid, e.g., for ecological time series that are often short (Richardson and Schoeman, 2004); 3) The PPM equation for N^* omits terms involving cross correlations that can be important in some applications; and 4) P&P98 did not compare their recommended method with other methods available at that time. In particular, they overlooked a method based on the average of the sample cross-correlation at long lags. The procedure is referred to herein as the Artificial Skill Method (ASM) and was introduced and favored in the same study that motivated the development of the PPM (Chelton, 1984; see also Davis, 1976 and Chelton, 1983).

The present study has two goals: A) Characterize a comprehensive set of natural and human-related time series in terms of their spectral color λ to study its impact on the performance of modified significance tests like ASM and PPM. To achieve this, we analyzed nearly 1800 natural, anthropogenic and simulated time series in the form of Eq. (2) for a wide range of spectral colors λ (see Section 2); B) Draw attention to the ASM, a method that has been underappreciated for nearly four decades and that outperforms the extensively used PPM that P&P98 incorrectly described as a modification of the Chelton method. In general, we show that PPM is often unable to obtain accurate error rates for time series with spectral characteristics that are typical of natural phenomena. We show in Section 4 that the ASM addresses the shortcomings mentioned above and performs better than PPM, particularly for the very applications for which the PPM was specifically developed, namely for the short record lengths that are common in fishery (and ecological) studies.

2. Data

It can be anticipated that the methods for estimation of correlation significance that are considered in this study will show good performance only for specific ranges of values of spectral color λ. To evaluate the adequacy of the methods, we must first assess what the typical λ values of various types of time series are. To determine the range of λ for time series found in nature, we attempted to gather all available time series related to ecosystems (animal abundance) and their driving mechanisms, i.e., environmental (climatic) and anthropogenic changes. To our knowledge, this is the most comprehensive compilation of time series of this type to date. These time series consist of annual estimates (i.e., $\Delta t = 1$ year; normally observations are averaged throughout the year or during a specific time of the year, e.g., the spawning season) and are thus relatively short (rarely as long as 100 years). Note, however, that the methodology presented here can be applied to time series with any specified Δt and record length $N\Delta t$. If such time series are long enough (Henson et al., 2017 recommend record lengths of at least 40 years), they can be an important source of data for studies of the impact of climate change on nature (e.g., some of the studies quoted in the introduction). Because such studies involve the interaction between atmosphere, hydrosphere, biosphere, and anthroposphere, they are an important contribution to our general understanding of the total environment.

The source of auto-correlation in our collected set of time series is "oversampling" of low-frequency interannual (i.e., year-to-year) and decadal variability associated with long-term climate change. To include as much low-frequency variability as possible in our time series, we have set threshold minima for the record lengths. These thresholds were chosen for each data type considering a trade-off between record length and data availability. Ecological time series were selected only if they had record lengths of at least 50 years, while thresholds for environmental and anthropogenic time series were chosen to be 40 and 60 years, respectively.

We gathered 402 ecological time series from the Living Planet Database (LPD; LPD, 2021), which includes the data from which the Living Planet Index (LPI) was derived (WWF-ZSL, 2020; Loh et al., 2005; Collen et al., 2009; McRae et al., 2017). The LPD data are time series of the population sizes of terrestrial, freshwater and marine vertebrates, in the form of either a direct measure such as population counts, densities, or indices, or as a reliable proxy, e.g., breeding pairs, nests, tracks, capture per unit effort, or biomass for a single species. The data were gathered from a variety of sources, including published scientific literature, online databases, government reports, individual researchers, institutions, and gray literature.

Because of the large number of species, populations and data involved, this data-set is cumbersome to summarize in a table. We therefore refer the reader to WWF-ZSL, 2020 and LPD, 2021 for further details about these time series.

To represent environmental (or climatic) changes, we also gathered 35 time series of atmospheric and oceanic variability (Table S1). A few well-known examples are the El Niño/Southern Oscillation (ENSO), the Pacific Decadal Oscillation (PDO), the North Atlantic Oscillation (NAO), the global-average air temperature, the average sea surface temperature in the German Bight, and the volume of Labrador Sea Water (LSW).

Finally, to represent anthropogenic influences on ecosystems, we gathered 1154 time series of human-related activities from "Our World In Data" (OWID; OWID, 2021). Some examples of these time series are air pollution, energy and food production, number of armed conflicts and demographic growth, all of which are available by country as annual totals (see Table S2). While some of these time series (human health and literacy level, for example) might seem to have little relation to anthropogenic impact on ecosystems, they speak to the level of development of societies and countries, which in turn is an indicator of their capacity to impact nature. These anthropogenic time series are relevant for studies of, for instance, the impact of technology development on fisheries (e.g., Galbraith et al., 2017) or the relation between war and fish abundance (e. g., Engelhard et al., 2014).

In total, we gathered 1591 time series, which we have grouped in the categories listed in Fig. 1.

3. Analysis

3.1. Spectral color λ of the collected time series

To characterize our compilation of time series in terms of their spectral color λ, we tapered each time series with a Hann window (often referred to incorrectly as the Hanning window) and calculated its spectral density *S*(*f*), which in turn was smoothed by non-overlapping band-averages over five adjacent Fourier frequencies. Averaging over a larger number of frequency bands would be desirable but the length of the shortest time series considered here (40 years) results in insufficient frequency resolution of the spectrum. We fitted a straight line by least-squares estimation to the band-averaged spectral values in log-log space with the power-law form of Eq. (2), i.e.:

$$\log\left(S(f)\right) = \lambda \log\left(f\right),$$

obtaining λ as the fitted linear parameter. Our approach is similar to Inchausti and Halley, 2002, but improved through the tapering and band-averaging that is essential to avoid corruption of the spectral estimates by leakage because of the short record lengths and noise in the raw spectral estimates.

The larger the value of the spectral color λ, the steeper the rolloff of the power spectrum with increasing frequency and thus the more dominant the power of low-frequency variability of the time series. As shown in Fig. 2, the auto-correlation time scale increases monotonically with increasing λ. We calculated the value of λ for each time series in each of the 13 categories listed in Fig. 1 in which the range of λ values for each category of time series is characterized by the central 90 % of the distribution of all of the calculated values of λ in that particular category. As a measure of the quality of the fit of λ, and thus of the validity of Eq. (2) as a representation of nature and human influence, we calculated the average coefficient of determination R^2, which is the regression sum of squares divided by the total sum of squares and represents the fraction of the variance of *S(f)* that is explained by the linear fit of λ. We also computed the STD of R^2 within each category. A good fit of λ for a specific category is characterized by an average value of R^2 near 1 with a small STD. Conversely, small values of both the average R^2 and the STD are a sign of a poor λ fit for the corresponding category (because most of its time series would have a small R^2).

Fig. 1 includes application of the spectral procedure to 100 randomly generated AR(1) and AR(2) time series of the form represented by Eq. (1) (200 time series in total) to assess the ranges of values of the spectral color λ that can be represented by these auto-regressive processes. These time series were constructed with record lengths of $N = 100$ for auto-correlation parameters ϕ_i randomly drawn from the uniform distribution *U*(*0,1*). Together with the observed 1591 time series (Section 2), the AR time series yield a total of 1791 analyzed time series.

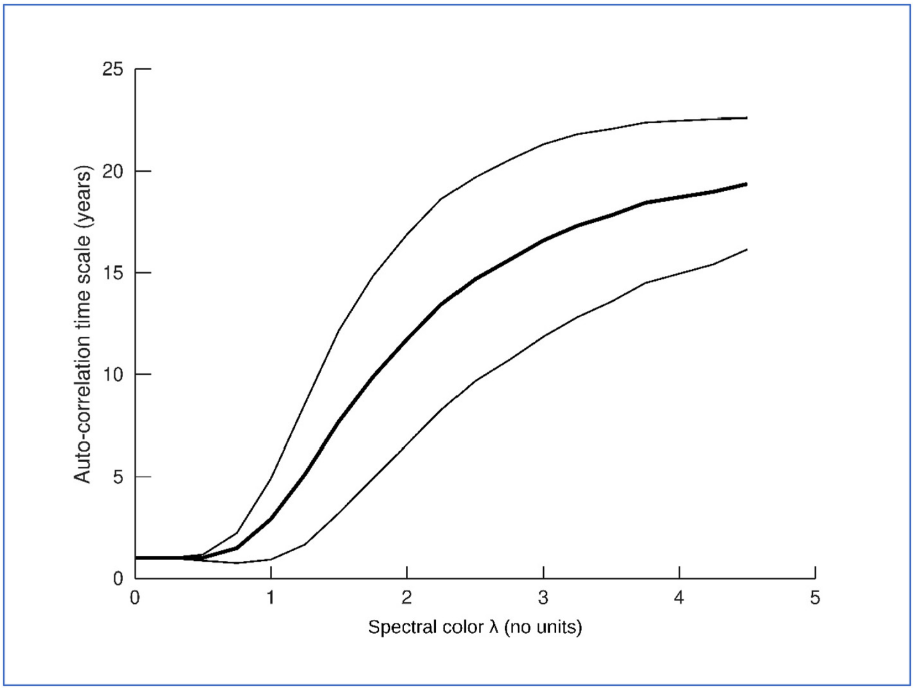


Fig. 2. The average auto-correlation time scale (thick curve) over 5000 synthetic spectral time series with record length N = 100 as function of their spectral color λ (Eq. (2)). The thin curves are plus and minus one STD. The auto-correlation time scale was defined to be the lag for which the auto-correlation decreased to a value of 0.5.

3.2. Comparison of the methods for estimating correlation significance

The two methods tested in this study to modify the significance test for correlation are both described in detail elsewhere: ASM in Chelton, 1983 (see also Appendix A of Dhage and Strub, 2016) and PPM in P&P98. The details of the methods are summarized here in the Supplement (Sections S2 and S3). The two methods are compared in this section by performing MC simulations: We generated random time series by first approximating their power spectra with the power law of Eq. (2) for a given value of λ. The amplitude of the Fourier transform at each Fourier frequency $f_j = j/(N\Delta t)$ is proportional to the square root of the power spectral density at that frequency. The phases of the Fourier transform at these frequencies were drawn randomly from the uniform distribution $U(0,\pi)$.

Random time series were then generated by inverse Fourier transforming the amplitudes and phases of the Fourier transform at each Fourier frequency $f_i = i/(N\Delta t)$. Time series with 512 elements were generated and then subsampled to shorter prescribed lengths of $N \leq 110$ (because our longest observed records are roughly 100 years long). This procedure mimics the real world in which every finite-length time series includes unresolved variability at periods longer than the record length. We generated $n = 5000$ pairs of random time series $X(t)$ and $Y(t)$ in this manner for ranges of N and the spectral color λ that are similar to those of the real time series summarized in Fig. 1. Specifically, we considered N ranging from 10 to 110 years in steps of 2 years and λ ranging from 0 to 4.5 in steps of 0.25.

Because the time series $X(t)$ and $Y(t)$ are generated randomly, they are statistically independent of each other. Mostly, non-significant correlations are expected but correlations that are deemed to be significant with a probability $P = (1\text{-}\ \alpha)$ can arise out of chance. A good method for estimating statistical significance should keep P close to the prescribed value of α that defines the significance P. In the analysis presented here, we used $\alpha = 0.05$, which corresponds to the 95 % significance level. Results for alternative choices of α=0.10 and 0.01, which correspond to the 90 % and 99 % significance levels, respectively, can be inferred graphically from Fig. S1 in the Supplement (see also the graphical abstract).

For the case of ASM, we define the error rate $e_{ASM}(\alpha)$ for the correlation critical value $c_{ASM}(i,\alpha)$ calculated for the *ith* realization by Eq. S5 in the Supplement for a specified value of α to be

$$e_{ASM}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\delta_i,$$

where $n = 5000$ is the total number of realizations of the MC simulation and

$$\delta_i = \begin{cases} 1 & if \quad |\rho(\mathrm{i})| \geq \mathrm{c}_{ASM}(\mathrm{i},\alpha) \\ 0 & if \quad |\rho(\mathrm{i})| < \mathrm{c}_{ASM}(\mathrm{i},\alpha) \end{cases}.$$

Here $\rho = (\rho_1, \rho_2, \ldots, \rho_N)$ is the set of n correlation coefficients between the n MC simulated pairs of time series. The deviation of the error rate $e_{ASM}(\alpha)$ from the true value α is

$$\Delta e_{ASM}(\alpha) = e_{ASM}(\alpha) - \alpha. \tag{3}$$

An analogous set of equations define the error rate deviation $\Delta e_{PPM}(\alpha)$ for the PPM method. The critical value $c_{PPM}(i,\alpha)$ for PPM is given by the same Eq. S5 as $c_{ASM}(i,\alpha)$ but with the different estimate of N^* obtained from Eq. S8 (rather than the Eq. S3 used for ASM). When the method is irrelevant in the discussion that follows, the error rate deviation and critical value of the correlation will be denoted as Δe and c with no subscripts. The quantities Δe_{ASM} and Δe_{PPM} characterize the accuracy of each method, i.e., how well each method yields an error rate close to the true significance level α calculated from the MC simulations.

From the definition (Eq. (3)), it is apparent that positive values of Δe indicate that e is an overestimate of α. This implies that the probability $P = (1\text{-}\ \alpha)$ is lower than intended. Consequently, sample cross correlations could be deemed statistically significant with probability (1- α) when in fact the probability is lower (false positives).

In practice, users normally have only a single pair of time series X and Y. They therefore need the significance test to be as accurate and precise as possible, which we assess here from the n realizations of the MC simulation. In addition to a small error rate deviation Δe, an estimate of the critical value should yield similar error rates e with repeated random tries, i.e., Δe should have a small variance. By averaging the differences of the critical values $c_{PPM}(i,\alpha)$ or $c_{ASM}(i,\alpha)$ calculated by the two methods compared with the true critical value $c_{sim}(\alpha)$ from the MC simulation, we can estimate the variance of each method. Note that a similar metric with the error rates e_{PPM} and e_{ASM} is not possible, since the true error rate is not known (the error rate is itself a statistical concept). We define the critical value $c_{sim}(\alpha)$ for $\alpha = 5$ % as the lower bound of the largest 5 % of all n of the magnitudes $|\rho_i|$ of the correlation estimates from the MC simulation ($c_{sim}(\alpha)$ for the case of $N = 50$ samples and α values of 0.10, 0.05 and 0.01 are shown in the bottom panel of the graphical abstract). We further define the mean absolute difference of critical values as:

$$MADc_{ASM}(\alpha) = \frac{1}{n}\Sigma_{i=1}^{n}|c_{ASM}(i,\alpha) - c_{\mathrm{sim}}(\alpha)|, \tag{4}$$

with an analogous equation for $MADc_{PPM}(\alpha)$. The explicit dependencies of the various statistics above on α can be omitted for clarity and the mean absolute difference of critical values will be denoted as $MADc$ when the method is irrelevant in the discussion that follows.

The ASM and PPM methods are both based on estimation of N^*, which has applications beyond the significance of correlation (e.g., for estimation of confidence intervals for regression coefficients). The estimate of N^* is related to the critical value c as summarized mathematically and graphically in Sections S2 and S3 of the Supplement (see also the top panel of the graphical abstract). A complete comparison of the two methods thus demands studying also the statistical characteristics of N^* (such as its variance) estimated by each method. Use of the ratio $\nu = N^*/N$ (see Eq. S1b) is convenient for graphical reasons as it makes the comparison independent of the record length $N\Delta t$. As a second metric for evaluating the variance of estimates of N^* by each method, we used the STD of the ratio ν:

$$\sigma_\nu = \frac{1}{N}\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(N_i^* - \overline{N^*}\right)^2}. \tag{5}$$

For a few of the MC realizations, one or the other method yielded $N^* > N$, which is clearly erroneous. Kope and Botsford, 1990 and P&P98 constrained N^* to be equal to N in these cases. While this truncation of N^* is reasonable when applying these methods to real data, we explicitly rejected it here in order to include all possible contributions to σ_ν and achieve a better assessment of the two methods.

P&P98 suggest in general using $J = N/5$ lags in the definition of N^* in their Eq. S8 (see also the discussion in Section S3). To assess the effects of this choice on the performance of PPM, we also tested a modified PPM method in which the number of lags involved in the sum in Eq. S8 was extended to $J = N/2$.

We do not advocate detrending for estimates of N^* because that often fundamentally changes the frequency content of a time series, and therefore the lagged autocorrelation values, compared with the original time series. But since detrending is so common in the literature (e.g., Brown et al., 2011), we tested the performance of the methods for two cases: retaining and removing a least-squares fit linear trend.

In addition to the model comparison, the MC simulations were also used to test different configurations for the ASM, which are summarized in Section S4.

4. Results

4.1. Spectral color λ of the collected time series

It can be seen from Fig. 1 that the time series with the smallest spectral color λ (i.e., the shortest auto-correlation time scale, see Fig. 2) are in the

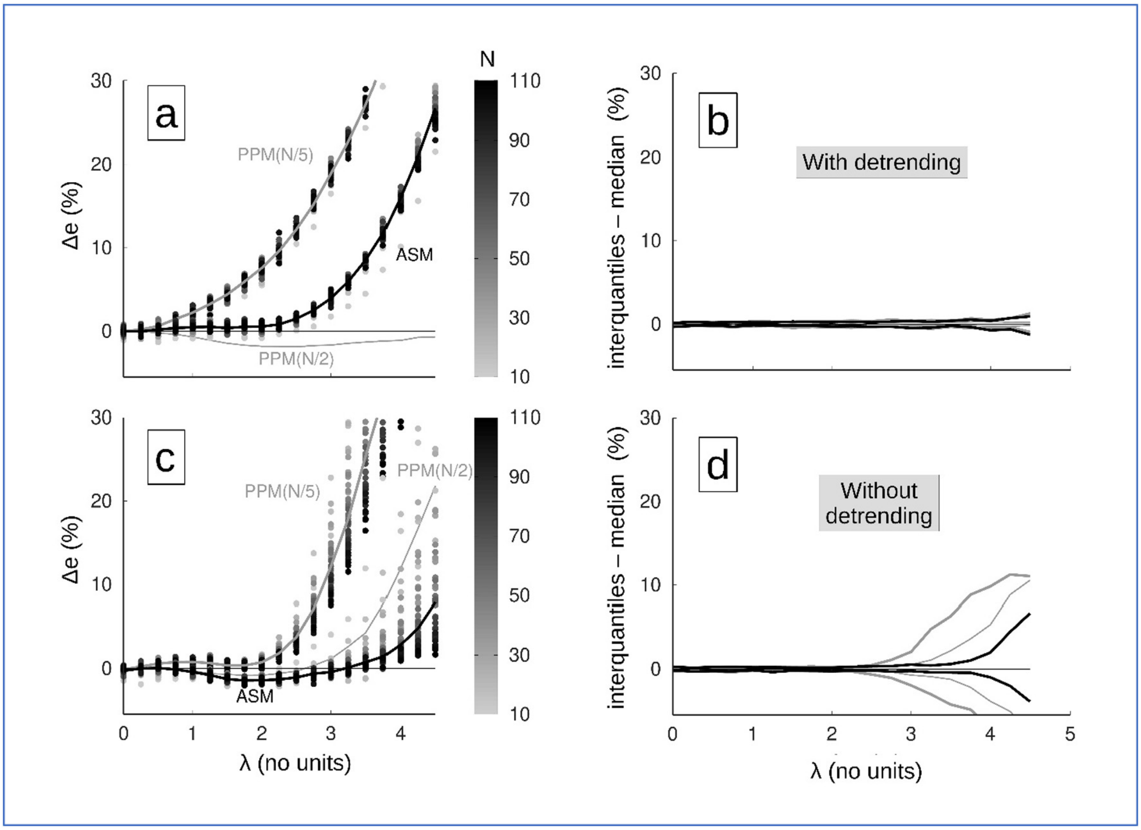


Fig. 3. Error rate deviations Δ*e* (Eq. (3)) expressed in percent as functions of λ and record length *N*Δ*t* from the MC simulations (dots in the left panels) for PPM and ASM with and without detrending of the time series (top and bottom panels, respectively). Different record lengths *N*Δ*t* of the time series are indicated by the dots according to the gray scale. The curves in the left panels are the median Δ*e* (thick gray curve: standard PPM with $J = N/5$; thin gray curve: modified PPM with $J = N/2$; black: ASM). Scatter dots for PPM with $J = N/2$ are omitted for clarity. The curves in the right panels (same formats as in the left panels) are the deviations of the 0.25 and 0.75 quantiles of Δ*e* for each λ from the median of Δ*e*. These curves are, therefore, a measure of the scattering of the dots around the curves in the left panels.

category of natural disasters, for which most time series have near-zero λ, i.e., they are essentially "white noise". These are followed by abundance of freshwater species, first-order autoregressive models and climate, which reach values of roughly λ = 2. Abundance of terrestrial species can have slightly longer auto-correlation time scales (larger λ) than the climatic time series, with a maximum of λ = 2.5. The time series with larger λ values are all anthropogenic, reaching a maximum beyond λ = 8 for the category of human health. The only exception is the abundance of marine species, which have λ values as large as λ = 5.5, and thus considerably longer autocorrelation time scales than their terrestrial and freshwater counterparts.

A few examples of ecological, climatic and anthropogenic time series that closely follow the power law of Eq. (2) are shown in Fig. S5 in the Supplement. More than 1300 of the time series analyzed here (73 % of the total) have a coefficient of determination $R^2 > 0.5$. This indicates that Eq. (2) is a good representation of most natural and anthropogenic time series.

4.2. Performance of the significance tests in terms of spectral color λ

The dependencies of the performances of the significance tests for PPM and ASM on the spectral color λ and length *N* of the time series are shown in Fig. 3. For both methods, the error rate deviations Δ*e* from the true value α (Eq. (3)) are close to zero for short auto-correlation time scales (λ < 0.5), regardless of whether the trend is removed (panel a) or retained (panel c). This is because time series with small λ have very weak trends. For detrended time series, Δe_{ASM} is close to 1 % for λ between 0.5 and 2.5, while Δe_{PPM} increases rapidly from about 5 % for λ = 1.5 to about 20 % for λ = 3, and to more than 30 % for λ > 3.5. The ASM error Δe_{ASM} slowly increases to only 5 % for λ = 3 and then increases more rapidly but does not reach 30 % until λ = 4.5. There is no dependence of Δe_{PPM} or Δe_{ASM} on the record length *N*Δ*t* when the trend is removed (the dots show no particular pattern in terms of their gray scale). This is because detrending shortens the decorrelation time scale of time series with large λ.

Perhaps surprisingly, both methods perform better in terms of Δ*e* without detrending (panel c). The error rate deviations Δe_{PPM} are roughly 1 % for λ between 0 and 2, increasing to more than 10 % for λ > 2 and exceeding 30 % again for λ > 3.5. ASM slightly underestimates α ($\Delta e_{PPM} \approx 1$ %) but is closer to the correct value for all values of λ in comparison to PPM, and exceeds 1 % only for λ larger than about 3.5. In contrast to the weak dependence on record length with detrending, Δe_{ASM} and Δe_{PPM} without detrending both decrease in magnitude with increasing *N* (gray scale), indicating better performance for longer data records. According to *Δe*, the modified PPM with $J = N/2$ (thin gray curve in panel a) performs better with detrending than both the traditional PPM and the ASM, but it will be shown below that it is also much noisier.

The interquartile curves of Δ*e* indicate similar variance of Δ*e* over all values of *N* for both methods with detrending (panel b) and the variance of Δ*e* is larger for PPM than for ASM without detrending (panel d). The modified PPM with $J = N/2$ (thin gray curves) performs considerably better than the traditional PPM, particularly with detrending as indicated also by the median Δe_{PPM} closer to zero (thin gray curve in panel a). The variance of Δe_{PPM} begins to increase when λ > 2.5 for the traditional PPM with $J = N/5$ and when λ > 3 for the modified PPM. The variance of Δe_{ASM} does not begin to increase rapidly until λ exceeds about 3.75.

Fig. 4 shows the mean absolute difference *MADc* of critical values (Eq. (4)) for the two methods. For detrended time series (top panels), the $MADc_{PPM}$ for the traditional PPM (panel a) increases monotonically with increasing values of λ, whereas the $MADc_{PPM}$ for the modified PPM (panel b) increases to a maximum of about 0.16 near λ = 2 and then

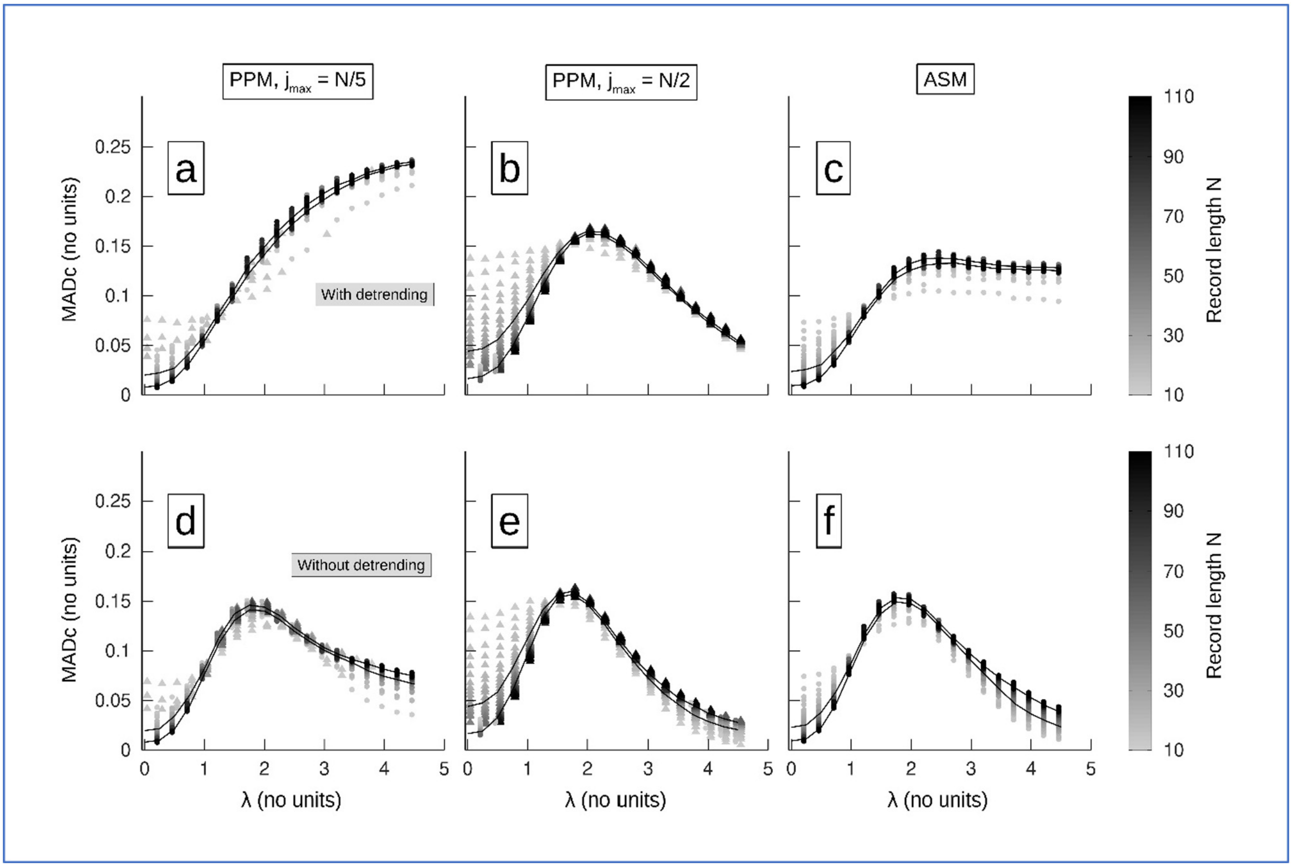


Fig. 4. Mean absolute difference of critical values *MADc* (no units; Eq. 4) for the traditional PPM with J = N/5 (left panels), the modified PPM with J = N/2 (central panels) and the ASM (right panels) as functions of λ and record length *N*Δ*t* with and without detrending (top and bottom panels, respectively). Situations in which at least one critical value was undefined (see text) are represented with triangles. The black and gray lines and the gray scales for the dots are as in Fig. 3.

decreases for λ > 2. $MADc_{ASM}$ increases similarly to a slightly smaller maximum of about 0.13 near λ = 2 and then decreases slowly for larger λ. Without detrending, $MADc_{PPM}$ (panels d and e) and $MADc_{ASM}$ (panel f) are similar to the case with detrending between λ = 0 and λ = 2, where both methods reach their maxima. For λ > 2, both the modified $MADc_{PPM}$ (panel e) and $MADc_{ASM}$ (panel f) slightly outperform the traditional $MADc_{PPM}$. Similar to the case of Δ*e*, both methods perform better without detrending (i.e., they have smaller variance and are thus more precise), as indicated by smaller MADc values for λ > 2 in the bottom panels.

With PPM, many of our MC experiments yielded at least one undefined critical value that arises from the *N** estimate being smaller than 3 (the triangles in Fig. 4; see also Eq. S5). *N** can even be negative with PPM (these particular cases are not distinguished from the case *N** < 3 in Fig. 4). In contrast, all critical values were defined in our MC experiments with ASM (note that it is actually not possible to obtain *N** < 0 with ASM since all of the quantities that define *N** in Eq. S3 are positive). The probability of obtaining *N** < 0 with PPM increases with decreasing *N*, as evidenced by the fact that all of the triangles in the left and central panels are light gray. Similarly, the probability of obtaining *N* < 0* also increases with increasing number of lags *J*, as shown by a larger spreading of *MADc* estimates and a larger number of triangles for the modified PPM with *J* = *N/2* (panels b and e) in comparison to *J* = *N/5* (panels a and d).

Contrary to Δ*e*, there are clear patterns for *MADc* dependence on the record length *N* for both cases, with and without detrending. For λ < 1 (1.5 for the modified PPM), *MADc* decreases with increasing *N* (i.e., the methods are all more precise with long records). For λ > 1, the opposite occurs, i.e., the methods are more precise with short records.

Fig. 5 shows the STDs σ_ν of $\nu = N^*/N$ (Eq. (5)) as functions of λ and record length *N*Δ*t*. Except for the slight increase from λ = 0 to λ = 1, σ_ν mostly decreases with increasing λ for both methods. It also decreases with increasing record length (gray scale), indicating a smaller variance of *N** for long data records. PPM shows a larger variance of *N** in comparison to ASM (note the larger scattering of σ_ν for PPM), especially for the modified PPM with *J* = *N/2* (middle panels). Contrary to ASM, σ_ν with PPM can be quite large also for long records (dark dots), reaching values well beyond 1 in some cases, again particularly for *J* = *N/2* (encircled dots at the tops of both left and middle panels).

5. Discussion

5.1. Spectral color λ of the collected time series

Our results in Fig. 1 are consistent with several previous studies that have characterized environmental time series in terms of their spectral color λ. Halley, 1996 assess that λ = 1 should be common for environmental data. In agreement, Pelletier, 1997 found that atmospheric temperatures from hundreds of stations and ice core records ranged from white noise to λ = 2. Vasseur and Yodzis, 2004 focused not only on atmospheric but also on oceanic time series, estimating lower spectral colors for atmospheric variations, i.e. closer to white noise, and values close to λ = 2 for oceanic variations. These values are all in good agreement with our "climate" category, which includes both atmospheric and oceanic time series (Fig. 1, row d).

Pelletier, 1997 further observed that power spectra of atmospheric temperature from continental stations could be divided into two different frequency regimes with distinct values of λ, while observations from maritime stations closely follow a power-law relation with a single value of λ for all frequencies. Multiple power-law regimes in a single spectrum reduce the quality of the linear fit for λ over the full range of frequencies. This could be a reason that the R^2 values are smaller for terrestrial (row e) and freshwater species (row b) than for marine species (Fig. 1, row k).

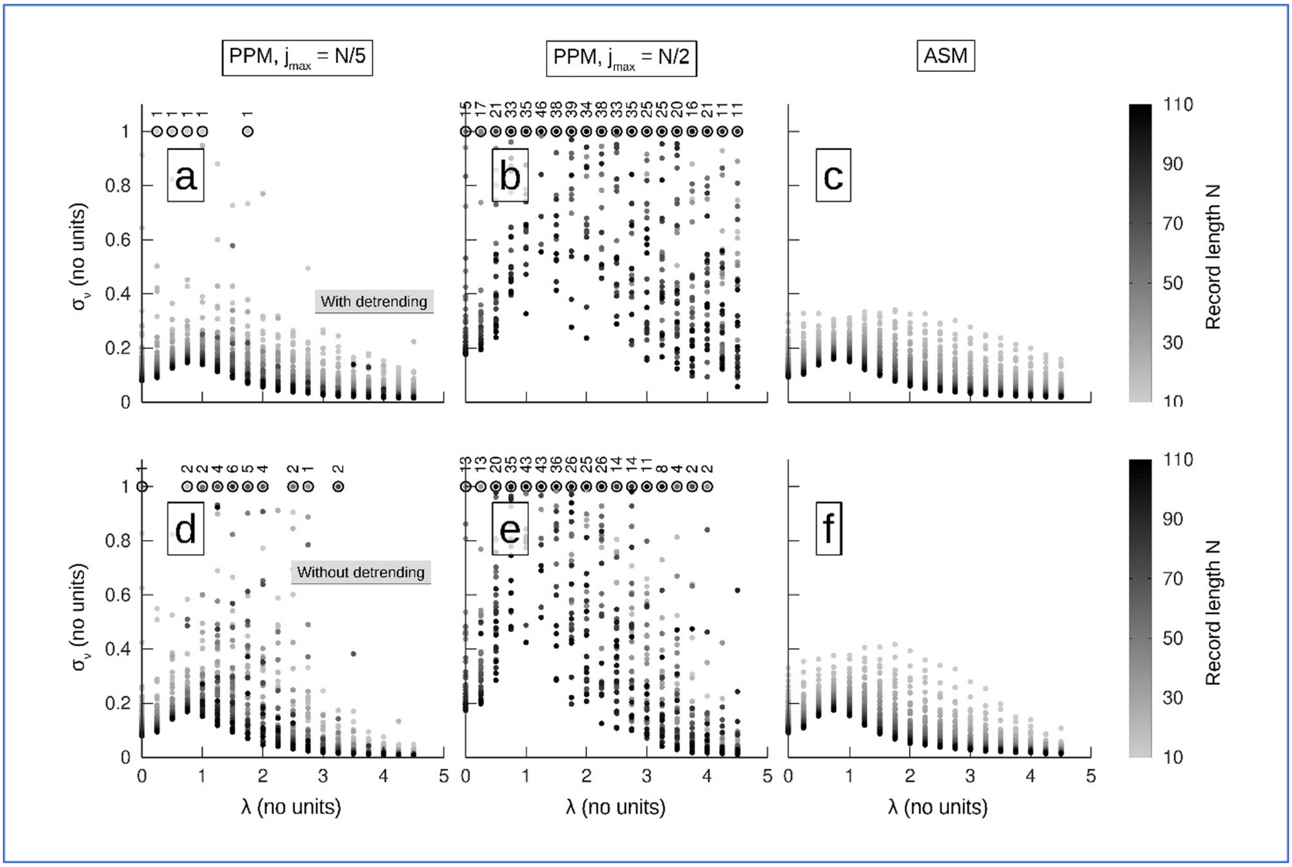


Fig. 5. STD σ_ν of $\nu = N^*/N$ (no units; Eq. (5)). The distribution of panels is as in Fig. 4. The dots inside open circles at the tops of each of the left and middle panels represent PPM values of σ_ν that exceed 1, ranging from 1.2 to 20 beyond the upper limit of the ordinate. The numbers of these extreme PPM values per λ level are labeled above each open circle.

For environmental time series, we have a disagreement with Schroeder, 1991 for the category of natural disasters. That study mentions spectra with $\lambda > 2$, while we find that natural disasters are basically white noise (Fig. 1, row a). Our results seem more plausible because natural disasters occur sporadically in time and are usually short-lived.

Another disagreement occurs in the case of ecological time series. Inchausti and Halley, 2002 observe a median of $\lambda = 1.02$ over a large number of terrestrial and aquatic species. While this value matches our results for freshwater (Fig. 1, row b) and terrestrial species (row e), we observe considerably larger values of λ for marine species (as high as $\lambda = 5.5$; row k). The LPI database includes the Global Population Dynamics Database of Inchausti and Halley, 2002. It also includes considerably more time series. A possible explanation for the differences from the results of our study might thus be that Inchausti and Halley, 2002 underestimated the values of λ because they considered time series for only a limited number of marine species. This is suggested because they find that marine species have particularly small values of λ in comparison to the terrestrial populations, which is not only opposite of what we find (Fig. 1, rows e and k), but also opposite to what has been found by Steele, 1985 and Ariño and Pimm, 2005 (as quoted by Halley, 1996).

The above studies clearly show that time series with power-law spectra of the form in Eq. (2) are ubiquitous in nature. Such time series have broad-banded variability over a continuum of frequencies and thus represent phenomena with no characteristic frequency (Schroeder, 1991). This follows the notion that nature is often fractal in essence (i.e., self-similar on all scales; Halley, 1996), having no preference for cyclical behavior. Exceptions are: a) processes such as tides and annual cycles, which occur at shorter time scales than the annual sample interval Δt considered in this study; and b) cyclical population trends of a few mammals such as lemmings (see Oli, 2019) that are represented by only one time-series in our dataset (the Grey sided vole; *Myodes rufocanus*).

In addition to time series found in nature, our analysis shows that the spectra of time series related to human systems also closely follow the power law of Eq. (2). This is seen from the large average R^2 values for all anthropogenic time series, such as war ($R^2 = 0.6$; row f in Fig. 1), energy production and pollution ($R^2 = 0.7$; row g), food production ($R^2 = 0.7$; row h), as well as human development ($R^2 = 0.9$; row i), demography ($R^2 = 0.8$; row j) and health ($R^2 = 0.9$; row m). In general, human systems seem to produce time series with larger λ values than natural systems. This is also reflected in the λ values of the marine species, which mostly consist of fish and are thus subject to stronger human influence (i.e., commercial exploitation and management) than the terrestrial and freshwater species. This might indicate that human forcing is subject to longer time scales than are generally found in nature, i.e., human-related changes such as modification of fishing policies, improvements in the health care system, and pursuits of world peace that occur secularly over the course of history in comparison to natural variations. These large values of λ are also likely related to the approximately exponential growth of world population.

Anthropogenic time series also have larger R^2 values than the natural time series. There is actually a well-defined and systematic relation between λ and R^2, with R^2 increasing with increasing λ (see Fig. 1 and Fig. S6). This might be interpreted as evidence that human systems produce time series that follow Eq. (2) more closely than natural systems. A simpler and more likely explanation, though, is that R^2 decreases with increasing high-frequency variability (i.e., with decreasing λ) because stochastic noise in the inherently noisy spectral estimates computed from the time series is not adequately reduced with our small amount of band averaging over only five frequency bands, thus yielding a poorer fit of Eq. (2).

5.2. Performance of the methods for estimating correlation significance

A study similar to this one was performed by Lennon, 2000, who computed the significance of the linear coefficient in spatial regression models with the traditionally defined *p*-value that corresponds to the level of probability that a sample estimate of the correlation occurs purely by chance. Lennon, 2000 showed that even modest spectral colors of $\lambda = 1$ result in an error rate of 20 % (his Fig. 4). While Lennon, 2000 appears to be the first to point out this problem in ecology, he did not investigate error rates in modified tests of significance as were considered here.

Among the methods of modifying the significance test to account for serial correlation in the time series, the most commonly used approaches to estimate N^* are, like PPM, based on integration of the auto-correlation functions (Bartlett, 1946; Bayley and Hammersley, 1946; Orcutt and James, 1948; Davis, 1976; Garrett and Petrie, 1981; Kope and Botsford, 1988; Clifford et al., 1989; Pyper and Peterman, 1998). Here we have shown that the ASM methodology that is based on averages of squared *cross*-correlation of the time series at long lags is considerably more reliable, particularly for record lengths and values of λ typical of natural and anthropogenic time series. With and without detrending, ASM consistently yields a smaller error rate than PPM for estimates of N^* (Fig. 3), as well as the same or smaller variances of the estimates of the correlation critical value for a given choice of confidence level α (Fig. 4) and N^* (Fig. 5). The superiority of ASM over PPM for the purpose for which the methods are specifically intended, i.e., estimation of N^*, is readily apparent from Fig. 5, regardless of how that impacts the critical value or any other statistic.

The better performance of ASM for estimation of N^* is particularly evident for the highly auto-correlated time series (Fig. 3a and c), for which the error rate of the traditionally defined PPM with $J = N/5$ increases to about 10 % for $\lambda = 2.5$ and then increases rapidly to more than 30 % for $\lambda > 3.5$. The reason for this rapid increase is that $J = N/5$ lags in Eq. S8 are too few to account for the slow decay of the autocorrelations for such large λ. This issue has been discussed by P&P98, who state that the definition of N^* Eq. S8 should include more than $J = N/5$ lags for strongly auto-correlated time series (P&P98, page 2133). Our results show that, while increasing the number of lags to $J = N/2$ indeed improves the error rate (thin gray curves in Fig. 3), the precision and accuracy of the method considerably decrease as shown by *MADc* (Fig. 4b and e) and σ_ν (Fig. 4b and e). This is further discussed in Section S3.

The poor performance of PPM for $\lambda > 2.5$ is striking because the method was specifically designed for fishery time series, which belong to the marine species category and can thus have spectral colors as high as $\lambda = 5.5$ (Fig. 1). Some examples of fish species from the LPI data with $\lambda > 2$ are Arrowtooth flounder (*Atheresthes stomias*; $\lambda = 4.9$), Sablefish (*Anoplopoma fimbria*; $\lambda = 4.5$) and Indo-Pacific Sailfish (*Istiophorus platypterus*; $\lambda = 4$). ASM shows a modest error rate for values of λ as large as 4. While there are a few time series of marine species with $\lambda > 4$, most of them have $\lambda < 4$ (Fig. 1, row k; note the histogram). ASM is thus the preferred method for studying relations between marine species and their forcing mechanisms.

While trend removal is often used to obtain more trustworthy significance estimates using the traditional test (Brown et al., 2011), our results show that it is detrimental to the performance of *modified* significance methods like PPM and ASM. The traditional PPM and the ASM both perform better without detrending, with generally poor performance for large λ when a trend is removed (Figs. 3 and 4, top panels). The likely explanation for this behavior is that the linear trend contributes to the low frequency variability. Detrending thus decreases the auto-correlation time scale and increases N^*. As a consequence, the critical value decreases (see Eqs. S7) and both the number of significant correlations and the error rate deviation increase.

Related to accuracy, there are two major advantages of ASM in comparison with PPM: 1) With ASM, N^* is defined as a function of positive quantities only (Eq. S3), but PPM can sometimes yield $N^* < 0$ (see Eq. S8 and the triangles in Fig. 4), which is obviously incorrect. This situation occurred infrequently (roughly 0.5 % of the 5000 MC iterations), mostly when the data records were short, but it is a serious issue because critical values are undefined whenever $N^* < 3$ (see Eq. S5). 2) While both methods can yield $N^* > N$ with similar probability, the differences $|N^*\text{-}N|$ are often considerably smaller with ASM than with PPM (Fig. 5). The larger uncertainties of PPM estimates of the correlation critical value and N^* are an important issue because, contrary to the MC simulations from synthetically generated time series analyzed here, most users have only a single pair of time series to work with. While the situation $N^* > N$ can reasonably be handled by adjusting N^* to be N (as done by Kope and Botsford, 1990 and P&P98), there is no similar alternative for an undefined critical value when $N^* < 3$.

Similar to P&P98, we also checked the performances of the methods when one of the time series X or Y has a specified fixed λ value. We performed four additional MC simulations in which X had $\lambda = 0$, 1, 2 and 4, while λ for Y varied as usual in all cases (figure not shown here). Excluding a slightly better error rate deviation for PPM in comparison to ASM for the case of $\lambda \leq 1$, these results confirmed a general better performance of ASM. The reason why PPM performs well if one time series has small λ (i.e., the time series is nearly white with short decorrelation time scale) is that the choice of $N/5$ in PPM is sufficiently large for the product $\rho_X\rho_Y$ of autocorrelations to reach zero, thus accounting for the decorrelation time scale.

ASM and PPM are both generally more accurate (Fig. 3, panel c) and precise (Figs. 4 and 5) when data records are long, as expected. For the time series considered here with a sample interval of $\Delta t = 1$ year, Figs. 3 and 4 can be used to obtain a qualitative guideline for the minimum N that should be considered for correlation analysis. A threshold minimum N can be defined based on adoption of some value of the probability distribution of the ordinate variables. For instance, for the upper quartile value of 75 % for ASM in Figs. 3 and 4, the associated minimum value of N over most of the range of λ is approximately $N = 30$. Working with data records shorter than 30 years should thus be a matter of concern, particularly for PPM. This is especially relevant to applications of the methods based on short time series with large λ and hence long auto-correlation time scales (Fig. 2), as is often the case in ecology applications (Richardson and Schoeman, 2004).

With any method for estimating N^*, users should be careful with applications to very short time series, especially when λ is large (for an example of such usage, see Aschan and Ingvaldsen, 2009). For data records with $N = 10$, for instance, estimates of N^* from the traditional PPM with $J = N/5$ are based on sample correlations at only two lags (see Eq. S8). While ASM is better because it is computed from 10 lags (5 positive lags and 5 negative lags, for our recommended implementation of ASM; see Section S4) over which the long-lag skills are averaged, this is still a small number of lags for sample correlations that are inherently noisy. Statistics with such small numbers cannot be expected to be precise. We therefore recommend that users refrain from computing cross correlations from short data records. It is far better to wait for longer data records to accumulate before attempting correlation analysis, and limit assessment of the relationships between short time series to visual inspection (e.g., Núñez-Riboni et al., 2012). Although such analysis is only qualitative and does not allow predictions (Zimmermann and Werner, 2019), this is preferable to basing predictions on correlations that are not statistically significant.

Finally, it is perhaps important to mention that the bulges of the performance metrics (Figs. 3 to 5) are not related to modes of variability in the random time series (which should not exist given the method used to create them). Rather, they are a consequence of the distributions of N^*, and critical value (for each level of λ) as well as of the transformation between them (i.e., Fig. S1).

5.3. Limitations of auto-regressive time series

To assure that our implementation of PPM was accurate, we confirmed that our analysis reproduced results reported by P&P98 with AR(1) time series. By using the same combination of auto-correlation ϕ_1 at lag 1 and record length N, we were able to reproduce the mean (0.047) minimum (0.043) and maximum (0.051) error rates for their Case L to the 2nd decimal place. Moreover, our results for PPM were the same as Fig. 1F of P&P98 (with random variations of 0.5 % around α).

Previous studies have discussed the abilities of synthetically generated time series from AR models and the spectral method to characterize observed time series. Halley, 1996 advocated the spectral method as a better representation of ecological time series, arguing that real ecological time series keep increasing in variance with increasing record length, while the variances of AR time-series quickly converge to some finite value. In agreement with this notion, Cuddington and Yodzis, 1999 show that low-frequency variations of AR models contribute less variance to the total signal, which is not the case with spectral time series (their Fig. 1). These authors also show that AR(1) and AR(2) models are not able to capture the unpredictability associated with red noise, which could lead to an overly optimistic view of the ability to predict the effects of environmental noise on ecosystems.

In disagreement with a statement of P&P98, we found that AR (1) models are often a poor representation of an important type of ecological time series, namely those of marine species abundance (mostly fish): While AR(1) time series have maximum spectral color values of about $\lambda = 2$ (Fig. 1, row c), time series of marine species abundance can have values as high as $\lambda = 5.5$ (row k). The spectral method is able to reproduce the low-frequency variability and long auto-correlation time scales that characterize many time series of marine species. Such long auto-correlations cannot be represented with the single auto-correlation parameter of AR(1) models. Consistent with this conclusion, Cuddington and Yodzis, 1999 pointed out that AR(1) models fail to simulate large values of λ found in some natural systems. AR(2) models (Fig. 1, row l) are required to reproduce the range of λ values similar to those found for marine species abundance.

AR(q) times series may be able to reproduce all of the important characteristics of natural time series mentioned above if q is sufficiently large. The spectral method, however, which characterizes the auto-correlation structure with single parameter λ, is a simpler and more convenient method than AR(q) models that require specification of q parameters (the case of $q = 2$ is defined by Eq. (1)). Furthermore, regardless of their order q, AR time series, can only reproduce normally distributed random data (Eq. (1)), but ecological and anthropogenic data often have skewed distributions. Random spectral time series are not constrained to normal distributions. The more complex version of the spectral method used by Schreiber and Schmitz, 1996 can even reproduce many statistical properties of a particular type of data, such as their skewness, or their auto-correlation and probability distribution functions. For this goal, the method sets some restrictions on the synthetic time series which, however, also make them more similar to each other when the lengths of the real time series are short, as is often the case. The method used here yields simulated time series with statistical properties (for instance, different distributions) that can differ somewhat from those of the real time series, but it retains their independence, which is required in MC simulations.

6. Conclusions

Estimating the spectral color λ of time series (like those found in nature) is important for a number of reasons, including studies of the probability of occurrence of events (Halley, 1996) or studying the time scales and variability of natural phenomena (Pelletier, 1997). The results obtained from our analysis contribute to that general knowledge by characterizing various types of time series in terms of their spectral color λ (Fig. 1). A highlight in comparison to previous similar studies (like Inchausti and Halley, 2002) is our focus on time series important to inestigate the effect of climate change on ecosystems. For instance, in addition to a large number of ecological and climatic time series, we also estimated λ (for the first time, to our knowledge) for time series of human activities representing potential anthropogenic influences on nature (like CO_2 emissions; Table S2). Our results show that time series of natural disasters have the smallest λ values, being almost white noise (Fig. 1, row a). Other time series with relatively small λ are those of freshwater species (row b), of climatic changes (row d), and of terrestrial species (row e). Time series with larger λ values are almost all anthropogenic (rows b to m), excluding only time series of marine species (row k). The large values of λ for these time series should be of particular relevance for investigations of the influence of climate change on marine ecosystems.

In this study, we characterized nearly 1800 time series in terms of their spectral color λ with the specific goal of studying the effect of λ on the significance of cross correlations between two time series. The effect of λ on the performance of the traditional method of correlation significance has been studied previously by Lennon (2000), showing that increasing λ considerably degrades the performance of the significance test. We showed that such decreases of performance with increasing λ also affects significance tests that have been modified to account for data interdependence (e.g., PPM and ASM). The error rate of the methods for correlation significance tested in this study remains stable and close to zero only until a particular value of $\lambda = 2$ (for PPM) and $\lambda = 3$ (for ASM) and it increases considerably for increasingly large values of λ (e.g., Fig. 3c).

Roughly half of the published studies of the effects of climate change on ecosystems from time-series analysis have neglected the effects of serial correlation of observations on their statistical results (Brown et al., 2011). Additionally, the most commonly used method of modifying the significance test, i.e., PPM, yields results that are less accurate than ASM for the broad range of spectral colors λ and record lengths N found in climatic, ecological and anthropogenic time series. In particular, PPM overestimates α (Fig. 3), which implies that the probability value $P = (1\text{-}\ \alpha)$ of the critical value of the correlation is lower than intended. This can lead to inferences of statistically significant correlation when none really exists.

The overestimation of statistical significance with PPM is particularly large for $\lambda > 2$ (Fig. 3), which encompasses a large number of marine and anthropogenic time series (Fig. 1). This raises the question of whether there is a bias towards "significant relation" when none actually exists in the many published studies in the marine biology and marine ecology literature that have been based on PPM. Use of ASM should substantially improve the assessment of the statistical significance of sample estimates of the cross correlation in studies of the effects of climate change on ecosystems. To help facilitate such analyses, we have made our ASM codes available in the programming languages R, Octave and Fortran through the Internet platform gitlab (https://gitlab.com/ismael_diego/artificial-skill-method).

7. Future work

None of the methods considered here are well-suited to application to some of the time series considered in Fig. 1, e.g., those of human health which can have values of spectral color λ of nearly 8 (Fig. 1, row m). The performance of ASM in terms of error rate deviation degrades for λ values larger than roughly 3.5 (Fig. 3c). For such large values of λ, it is possible that randomization (e.g., Ebisuzaki, 1997) or cross-validation (Michaelsen, 1987) procedures might yield better results than those obtained here with the ASM. We have not considered these alternative methods, mainly because they are computationally intensive, not necessarily in a single, practical application but when evaluating their performance in an MC simulation. To test the performance of a randomizing method, each of the 5×10^3 significance estimates from the MC simulation would arise itself from a large number (order 1×10^3) of iterations. This results in the requirement of computations over 1×10^6 iterations. We therefore leave a comparison with such methods to a future study.

Another possible extension of the techniques presented in this study is related to tests of "field significance". The systematic search for correlations between a (fixed) time series X and N gridded time series Y_i can inflate the probability of finding significantly correlated time series when the time series Y_i are correlated with each other. A field significance test evaluates whether a "patch" of significantly correlated time series is large enough to preclude the possibility that it arises only by chance. A widespread method is Livezey and Chen, 1983, which is based on MC simulations where the probability distribution of the sizes of all patches of significant correlations between X and Y_i is used to set a threshold for the significance of the patch sizes. However, that study modified the serial auto-correlation of X (by creating random time series in their MC simulations) and used a poor significance test (based on integration of the auto-correlation functions; Davis, 1976) for the individual tests between X and Y_i. These shortcomings could be improved with the spectral random time series of the present study and by using ASM for the significance tests between X and Y_i.

CRediT authorship contribution statement

Ismael Núñez-Riboni: Conceptualization, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Dudley B. Chelton:** Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Valentina Marconi:** Writing – review & editing.

Data availability

All data used are in the public domain and are available in the sources described in the manuscript

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank the data providers for making their data-sets freely available. Data from the Living Planet index were provided by Stefanie Deinet (Zoological Society of London). Craig Risien of Oregon State University gathered most of the climatic indices. Discussions with members of the modeling group of the Thünen-Institut für Seefischerei, particularly with Bernhard Kühn and Anna Akimova, are appreciated. The contributions of DBC to this analysis were supported by NASA Grant 80NSSC19K1117. Comments from five anonymous reviewers considerably improved the manuscript.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.scitotenv.2022.160219.

References

Akimova, A., Núñez-Riboni, I., Kempf, A., Taylor, M.H., 2016. Spatially-resolved influence of temperature and salinity on stock and recruitment variability of commercially important fishes in the North Sea. PLOS ONE 11, e0161917.

Ariño, A.H., Pimm, S.L., 2005. On the nature of population extremes. Evol. Ecol. 9, 429–443.

Aschan, M.M., Ingvaldsen, R.B., 2009. Recruitment of shrimp (Pandalus borealis) in the Barents Sea related to spawning stock and environment. Deep-Sea Res. II Top. Stud. Oceanogr. 56, 2012–2022.

Bartlett, M.S., 1946. On the theoretical specification and sampling properties of autocorrelated time-series. Suppl. J. R. Stat. Soc. 8, 27–41.

Bayley, G.V., Hammersley, J.M., 1946. The "effective" number of independent observations in an autocorrelated time series. Suppl. J. R. Stat. Soc. 8, 184–197.

Beaugrand, G., Kirby, R.R., 2010. Climate, plankton and cod. Glob. Chang. Biol. 16, 1268–1280.

Beaugrand, G., Reid, P.C., 2003. Long-term changes in phytoplankton, zooplankton and salmon related to climate. Glob. Chang. Biol. 9, 801–817.

Bedford, J., Ostle, C., Johns, D.G., Atkinson, A., Best, M., Bresnan, E., Machairopoulou, M., Graves, C.A., Devlin, M., Milligan, A., Pitois, S., Mellor, A., Tett, P., McQuatters-Gollop, A., 2020. Lifeform indicators reveal large-scale shifts in plankton across the North-West European shelf. Glob. Chang. Biol. 26, 3482–3497.

Box, G.E.P., Jenkins, G.M., 1970. Time Series Analysis: Forecasting and Control. Holden-Day.

Brown, C.J., Schoeman, D.S., Sydeman, W.J., Brander, K., Buckley, L.B., Burrows, M., Duarte, C.M., Moore, P.J., Pandolfi, J.M., Poloczanska, E., Venables, W., Richardson, A.J., 2011. Quantitative approaches in climate change ecology. Glob. Chang. Biol. 17, 3697–3713.

Caldwell, J.M., LaBeaud, A.D., Lambin, E.F., Stewart-Ibarra, A.M., Ndenga, B.A., Mutuku, F.M., Krystosik, A.R., Ayala, E.B., Anyamba, A., Borbor-Cordova, M.J., Damoah, R., Grossi-Soyster, E.N., Heras, F.H., Ngugi, H.N., Ryan, S.J., Shah, M.M., Sippy, R., Mordecai, E.A., 2021. Climate predicts geographic and temporal variation in mosquito-borne disease dynamics on two continents. Nat. Commun. 12, 1233.

Capuzzo, E., Lynam, C.P., Barry, J., Stephens, D., Forster, R.M., Greenwood, N., McQuatters-Gollop, A., Silva, T., van Leeuwen, S.M., Engelhard, G.H., 2018. A decline in primary production in the North Sea over 25 years, associated with reductions in zooplankton abundance and fish stock recruitment. Glob. Chang. Biol. 24, e352–e364.

Chelton, D.B., 1983. Effects of sampling errors in statistical estimation. Deep Sea Res. Part A Oceanogr. Res. Pap. 30, 1083–1103.

Chelton, D.B., 1984. Commentary: short-term climatic variability in the Northeast Pacific Ocean. In: Pearcy, W. (Ed.), The Influence of Ocean Conditions on the Production of Sahnonids in the North Pacific. Oregon State Univ. Press, Corvallis, Oregon, pp. 87–99.

Clifford, P., Richardson, S., Hemon, D., 1989. Assessing the significance of the correlation between two spatial processes. Biometrics 45, 123–134.

Collen, B.E.N., Loh, J., Whitmee, S., McRae, L., Amin, R., Baillie, J.E.M., 2009. Monitoring change in vertebrate abundance: the living planet index. Conserv. Biol. 23, 317–327.

Cuddington, K.M., Yodzis, P., 1999. Black noise and population persistence. Proc. R. Soc. B Biol. Sci. 266, 969.

Dale, M.R.T., Fortin, M.-J., 2002. Spatial autocorrelation and statistical tests in ecology. Écoscience 9, 162–167.

Davis, R.E., 1976. Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. J. Phys. Oceanogr. 6, 249–266.

Dhage, L., Strub, P.T., 2016. Intra-seasonal sea level variability along the west coast of India. J. Geophys. Res. Oceans 121, 8172–8188.

Ebisuzaki, W., 1997. A method to estimate the statistical significance of a correlation when the data are serially correlated. J. Clim. 10, 2147–2153.

Engelhard, G.H., Righton, D.A., Pinnegar, J.K., 2014. Climate change and fishing: a century of shifting distribution in North Sea cod. Glob. Chang. Biol. 20, 2473–2483.

Fisher, R., 1935. The Design of Experiments. Hafner Press 248 pp.

Galbraith, E.D., Carozza, D.A., Bianchi, D., 2017. A coupled human-earth model perspective on long-term trends in the global marine fishery. Nat. Commun. 8, 14884.

Garrett, C., Petrie, B., 1981. Dynamical aspects of the flow through the Strait of Belle Isle. J. Phys. Oceanogr. 11, 376–393.

Halley, J.M., 1996. Ecology, evolution and 1/f -noise. Trends Ecol. Evol. 11, 33–37.

Henson, S.A., Beaulieu, C., Ilyina, T., John, J.G., Long, M., Séférian, R., Tjiputra, J., Sarmiento, J.L., 2017. Rapid emergence of climate change in environmental drivers of marine ecosystems. Nat. Commun. 8, 14682.

Hermant, M., Lobry, J., Bonhommeau, S., Poulard, J.-C., Le Pape, O., 2010. Impact of warming on abundance and occurrence of flatfish populations in the Bay of Biscay (France). J. Sea Res. 64, 45–53.

Humphreys, R.K., Puth, M.-T., Neuhäuser, M., Ruxton, G.D., 2019. Underestimation of Pearson's product moment correlation statistic. Oecologia 189, 1–7.

Huo, W., Xiao, Z., Wang, X., Zhao, L., 2021. Lagged responses of the tropical Pacific to the 11-yr solar cycle forcing and possible mechanisms. J.Meteorol.Res. 35, 444–459.

Iles, T.C., Beverton, R.J.H., 1998. Stock, recruitment and moderating processes in flatfish. J. Sea Res. 39, 41–55.

Inchausti, P., Halley, J.M., 2002. The long-term temporal variability and spectral colour of animal populations. Evol. Ecol. Res. 4, 1033–1048.

Kirby, R.R., Beaugrand, G., 2009. Trophic amplification of climate warming. Proc. R. Soc. B Biol. Sci. 276, 4095–4103.

Kope, R.G., Botsford, L.W., 1988. Detection of environmental influence on recruitment using abundance data. Can. J. Fish. Aquat. Sci. 45, 1448–1458.

Kope, R., Botsford, L.W., 1990. Determination of factors affecting recruitment of Chinook Salmon Oncorhynchus tshawytscha in Central California. Fish. Bull. 88.

Lennon, J.J., 2000. Red-shifts and red herrings in geographical ecology. Ecography 23, 101–113.

Li, J., Sun, C., Jin, F.-F., 2013. NAO implicated as a predictor of Northern Hemisphere mean temperature multidecadal variability. Geophys. Res. Lett. 40, 5497–5502.

Li, B., Zhou, L., Qin, J., Murtugudde, R., 2021. The role of vorticity tilting in northward-propagating monsoon intraseasonal oscillation. Geophys. Res. Lett. 48, e2021GL093304.

Livezey, R.E., Chen, W.Y., 1983. Statistical field significance and its determination by Monte Carlo techniques. Mon. Weather Rev. 11, 46–59.

Loh, J., Green, R.E., Ricketts, T., Lamoreux, J., Jenkins, M., Kapos, V., Randers, J., 2005. The Living Planet Index: using species population time series to track trends in biodiversity. Philos. Trans. R. Soc. B Biol. Sci. 360, 289–295.

LPD, 2021. Living Planet Database. Downloaded on 13 June 2021 www.livingplanetindex.org.

McRae, L., Deinet, S., Freeman, R., 2017. The diversity-weighted living planet index: controlling for taxonomic bias in a global biodiversity indicator. PLOS ONE 12, e0169156.

Michaelsen, J., 1987. Cross-validation in statistical climate forecast models. J. Appl. Meteorol. Climatol. 26, 1589–1600.

Möllmann, C., Müller-Karulis, B., Kornilovs, G., St John, M.A., 2008. Effects of climate and overfishing on zooplankton dynamics and ecosystem structure: regime shifts, trophic cascade, and feedback loops in a simple ecosystem. ICES J. Mar. Sci. 65, 302–310.

Núñez-Riboni, I., Bersch, M., Haak, H., Jungclaus, J.H., Lohmann, K., 2012. A multi-decadal meridional displacement of the subpolar front in the Newfoundland Basin. Ocean Sci. 8, 91–102.

Oli, M.K., 2019. Population cycles in voles and lemmings: state of the science and future directions. Mammal Rev. 49, 226–239.

Orcutt, G.H., James, S.F., 1948. Testing the significance of correlation between time series. Biometrika 35, 397–413.

OWID, 2021. Data from "Our World in Data". https://ourworldindata.org/ about Access: July, 2021.

Pelletier, J.D., 1997. Analysis and modeling of the natural variability of climate. J. Clim. 10, 1331–1342.

Pershing, A.J., Alexander, M.A., Hernandez, C.M., Kerr, L.A., Le Bris, A., Mills, K.E., Nye, J.A., Record, N.R., Scannell, H.A., Scott, J.D., Sherwood, G.D., Thomas, A.C., 2015. Slow adaptation in the face of rapid warming leads to collapse of the Gulf of Maine cod fishery. Science 350, 809–812.

Puth, M.-T., Neuhäuser, M., Ruxton, G.D., 2014. Effective use of Pearson's product–moment correlation coefficient. Anim. Behav. 93, 183–189.

Pyper, B.J., Peterman, R.M., 1998. Comparison of methods to account for autocorrelation in correlation analyses of fish data. Can. J. Fish. Aquat. Sci. 55, 2127–2140.

Richardson, A.J., Schoeman, D.S., 2004. Climate impact on plankton ecosystems in the Northeast Atlantic. Science 305, 1609–1612.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M.D., Muñoz-Marí, J., van Nes, E.H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., Zscheischler, J., 2019. Inferring causation from time series in Earth system sciences. Nat. Commun. 10, 2553.

Schindler, D.E., Hilborn, R., Chasco, B., Boatright, C.P., Quinn, T.P., Rogers, L.A., Webster, M.S., 2010. Population diversity and the portfolio effect in an exploited species. Nature 465, 609–612.

Schreiber, T., Schmitz, A., 1996. Improved surrogate data for nonlinearity tests. Phys. Rev. Lett. 77, 635–638.

Schroeder, M., 1991. Fractals, Chaos, Power Laws: Minutes From an Infinite Paradise: Manfred Schroeder (W.H. Freeman, New York, NY, 1991). Elsevier 429 pp.

Steele, J.H., 1985. A comparison of terrestrial and marine ecological systems. Nature 313, 355–358.

Vasseur, D.A., Yodzis, P., 2004. The color of enfironmental noise. Ecology 85, 1146–1152.

Wang, J., Liu, Y., Ding, Y., Yang, Y., Xu, Y., Li, Q., Zhang, Y., Gao, M., Yang, J., Wu, Q., Li, C., Li, M., 2021. Future changes in the meteorological potential for winter haze over Beijing during periods of peak carbon emissions and carbon neutrality in China projected by Coupled Model Intercomparison Project Phase 6 models. Int. J. Climatol. 42 (4), 2065–2082.

WWF-ZSL, 2020. In: Almond, R.E.A., Grooten, M., Petersen, T. (Eds.), Living Planet Report 2020 - Bending the Curve of Biodiversity Loss. WWF, Gland, Switzerland.

Xie, T., Li, J., Chen, K., Zhang, Y., Sun, C., 2021. Origin of Indian Ocean multidecadal climate variability: role of the North Atlantic Oscillation. Clim. Dyn. 56, 3277–3294.

Zimmermann, F., Werner, K.M., 2019. Improved management is the main driver behind recovery of Northeast Atlantic fish stocks. Front. Ecol. Environ. 17, 93–99.