An exactly valid and distribution-free statistical significance test for correlations between time series

Alex E. Yuan[1,2*] and Wenying Shou[3*]

January 25, 2022

1 Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, 2 Molecular and Cellular Biology PhD program, University of Washington, Seattle, Washington, United States of America, 3 Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, London, the United Kingdom

* alexericyuan@gmail.com (AY); wenying.shou@gmail.com (WS)

Abstract

Computing a correlation between a pair of time series is a routine task in disciplines from biology to climate science. How do we test whether such a correlation is statistically significant (i.e. unlikely under the null hypothesis that the time series are independent)? This problem is made especially challenging by two factors. First, time series typically exhibit autocorrelation, which renders standard statistical tests invalid. Second, researchers are increasingly turning to nonlinear correlation statistics with no known analytical null distribution, thus rendering parametric tests inviable. Several statistical tests attempt to address these two challenges, but none is perfect: A few are valid only under restrictive data conditions, others have differing degrees of correctness depending on user-supplied parameters, and some are simply inexact. Here we describe the truncated time-shift procedure, which can be used with any correlation statistic to test for dependence between time series. We show that this test is exactly valid as long as one time series is stationary, a minimally restrictive requirement among nonparametric tests. Using synthetic data, we demonstrate that our test performs correctly even while other tests suffer high false positive rates. We apply the test to data sets from climatology and microbiome science, verifying previously discovered dependence relationships.

Introduction

Researchers routinely look for correlations between variables to identify potentially important relationships, or to use as starting points for downstream modeling and experiments. In disciplines such as climatology, ecology, and physiology, data are often collected as time series, and so correlation analyses using time series are common.

Interpreting a correlation between time series can be tricky because it is easy to obtain a seemingly high correlation between two time series that have no causal relationship [1, 2]. To avoid being fooled by such spurious correlations, it helps to distinguish between the concepts of "correlation" and "dependence", and how each relates to causation. In research involving time series, the term "correlation" is often used procedurally [1, 3, 4]. That is, a correlation function is any function that takes two time series and produces a statistic, which is usually interpreted as a measurement of similarity or relatedness. Examples include Pearson's correlation coefficient, local similarity [5], and cross-map skill [6]. Whereas a correlation statistic is a *summary* of an observed dataset, statistical dependence (or independence) is a *hypothesis* about the relationship between variables.

Two variables x and y are (statistically) dependent if the probability distribution of x while statistically controlling for y (the conditional distribution of x given y) differs from the distribution of x while not controlling for y (the marginal distribution of x). The concept of dependence applies not only to pairs of univariate variables, but also to pairs of vectors such as time series. Importantly, dependence is linked to causality (as defined in the usual sense: x causes y if perturbations in x can alter y). This link is due to Reichenbach's common cause principle, which states that if two variables are dependent, then they are "causally related": Either they share a common cause, or one variable causes the other (possibly indirectly) [7, 8]. Thus, before be we infer that one such relationship is present, we ought to first test whether the observed correlation indicates the presence of dependence.

In the simpler (non-temporal) case where one computes a pairwise correlation from measurements that are independent and identically distributed (iid), a highly general test of dependence is available in the permutation test. To describe it, suppose that measurements can be ordered in some way, so that (x_i, y_i) is the ith pair of measurements. The permutation test repeatedly computes the correlation after randomly shuffling the measurement order in one of the variables. Under the null hypothesis that x and y are independent, the original ordering and all randomized orderings are equally probable. Thus, a p-value can be calculated as (see, for example, section 6.2.5 of [9]):

$$\frac{N_{\geq} + 1}{N_{randomized} + 1} \tag{1}$$

where $N_{\geq}$ is the number of randomized orderings that produce a correlation at least as strong as the original (unrandomized) correlation, and where $N_{randomized}$ is the total number of randomized orderings. The "+1" terms account for the unrandomized correlation. This test has three especially admirable properties. First, the test is valid: If we infer dependence only when p is less than some significance level α, then our false positive rate (i.e. the chance of erroneously reporting dependence) will be no more than α [9, 10]. Second, the test is distribution-free: It does not require that the variables or the correlation statistic follow a particular probability distribution [11]. Lastly, the test lacks critical parameters: Its validity does not depend on any parameters that must be estimated or chosen by the user. If possible, we would like a test that maintains these properties, but which is appropriate for time series.

Testing for dependence between a single pair of time series is more difficult. Since time series are often autocorrelated (e.g. what occurs today influences what occurs tomorrow), a single pair of time series contains information from only a single trial. That is, there is an "n-of-one problem". In such cases, using correlation to infer dependence requires a great deal of care.

Parametric tests that account for autocorrelation represent one class of potential remedies for the n-of-one problem [12]. However, these tests are limited to a particular correlation statistic (such as Pearson's correlation coefficient) [13, 12] because each test relies on the availability of analytical results tailored to a specific statistic under the null hypothesis of independence. Consequently, the parametric testing approach may not be applicable to several statistics that have become popular within particular disciplines (e.g. local similarity in microbiology [5, 14, 15, 16, 17, 18]; cross-map skill in environmental sciences [6, 19, 20, 21, 22, 23]). These tests also require parametric assumptions (e.g. the sample correlation coefficient

follows a prespecified distribution) and/or require that the user correctly estimate some parameter (e.g. the autocovariance matrix of the time series) [24, 13, 12].

When parametric tests are unavailable or inappropriate, dependence between time series is typically tested by an approach called surrogate data testing [25]. Specifically, one begins with two time series $\{x_1, x_2, ...x_n\}$ and $\{y_1, y_2, ...y_n\}$ (abbreviated $\{x_t\}$ and $\{y_t\}$), and computes some measure of correlation between them. Next, one uses a computer to simulate how independent replicates of $\{y_t\}$ might have looked. These simulated $\{y_t\}$ time series are called "surrogate" $\{y_t\}$ series. Finally, one computes the correlation between $\{x_t\}$ and each surrogate $\{y_t\}$. A p-value is then given by the proportion of surrogate $\{y_t\}$ series that produce a correlation equal to or larger than the real $\{y_t\}$. More precisely, the p-value is given by Eq. 1, but where where $N_{\geq}$ is now the number of surrogates that produce a correlation statistic at least as large as the original $\{y_t\}$ series and $N_{randomized}$ is the total number of surrogates [26]. The permutation test is a special case of surrogate testing that is valid when data are iid instead of time series.

Several procedures are used to generate time series surrogates, each with different strengths and limitations. For instance, the random phase procedure decomposes the $\{y_t\}$ time series into sine waves and randomly shifts these sine waves horizontally before summing the sine waves back together to produce surrogates [24, 20]. This procedure is valid when $\{y_t\}$ is a Gaussian, linear, and stationary process [25, 27]. (See [28] for a discussion of precise validity conditions.) There is also a more general version called the iterative amplitude-adjusted Fourier transform (IAAFT) procedure, which is valid when $\{y_t\}$ is a Gaussian, linear, and stationary process that is observed through an invertible but possibly nonlinear observation function [26, 25]. Yet even this more general version would exclude processes with underlying nonlinear dynamics. Surrogates can also be produced by a block bootstrap method in which random subsequences are selected from $\{y_t\}$ and joined together [29, 30]. However, junctions between the blocks can produce disruptions, rendering the test inexact [29]. A sophisticated variant of the block bootstrap method, called the twin method, attempts to carefully position blocks so that the disruptions are minimized [31]. Yet even in this method, performance depends on parameters called "embedding parameters" [32] that must be appropriately chosen by the user, a potentially difficult task [33]. Overall, such surrogate procedures do not appear to fully translate the three desired properties of the permutation test (being exactly valid, distribution-free, and without critical parameters) to the case of time series.

Our procedure is most closely related to time-shifted surrogates [34, 35, 36, 21, 30, 27]. Although different variant procedures exist, the essential idea is to produce surrogates of $\{y_t\}$ by somehow shifting the original $\{y_t\}$ in time. One can then calculate a p value according to Eq. 1. Time-shifting was considered in an early article by Bartlett [37], who noted that the test is generally invalid because the surrogates are statistically dependent on each other.

Here we introduce the truncated time-shift (TTS) test, an exactly valid test for dependence between two time series. Like the permutation test applied to iid data, the TTS test is compatible with essentially any correlation statistic and its validity does not require the assumption of a particular probability distribution, nor does it require that a user correctly select some parameter. Specifically, the TTS test is exactly valid as long as one time series is strict-sense stationary. We first illustrate our procedure and mathematically prove (see Appendix 1) that it correctly controls the false positive rate. We then compare our test with other surrogate data tests in numerical experiments. We also consider how to set test parameters to increase the power with which it detects genuine dependence. Lastly, we demonstrate some possible use cases of our test with real data sets from climatology and microbiome science. This test is very general in the sense that it can evaluate the statistical significance of essentially any correlation statistic and only requires that one time series be stationary.

Results

The truncated time shift test

Since our test is based on time-shifted surrogates, we first need a way to shift the original $\{y_t\}$ series in time without altering its length. One way to achieve this is to use cyclic permutations [21, 30]. That is, if the original $\{y_t\}$ series were $\{1, 2, 3, 4\}$, then there would be 3 surrogates, given by $\{2, 3, 4, 1\}$, $\{3, 4, 1, 2\}$, and $\{4, 1, 2, 3\}$. However, these surrogates artificially force the first and final points of the original $\{y_t\}$ series to become neighbors, which can distort the dynamics [25].

Instead, we will use truncation to generate time-shifted surrogates [25]. Starting with $\{x_1, x_2, ...x_n\}$, we define the truncated x time series:

$$x^{trunc} = \{x_{1+r}, x_{2+r}, ..., x_{n-r}\}$$

where r, called the "truncation radius", is the number of time points that are removed on either side of x^{trunc}. We then define a collection of truncated y time series, which all have the same length as x^{trunc}, but can be shifted horizontally:

$$y^{trunc}(\delta) = \{y_{1+r+\delta}, y_{2+r+\delta}, ..., y_{n-r+\delta}\}$$

where the shift δ can take on integer values between $-r$ and r (Fig 1, step 1). Note that when $\delta = 0$, x^{trunc} and $y^{trunc}(\delta)$ are aligned. Thus, we can think of $y^{trunc}(0)$ as our original time series and $y^{trunc}(\delta)$ for $\delta \neq 0$ as our surrogate time series. For each value of δ between $-r$ and r, we compute the correlation between x^{trunc} and $y^{trunc}(\delta)$ (Fig 1, step 2). We then define B as the number of shifts δ that produce a correlation at least as large as when $\delta = 0$ (Fig 1, step 3). B is bounded between 1 and $2r + 1$: We have $B = 1$ if the strictly greatest correlation is obtained when $\delta = 0$. Conversely, B is equal to $2r + 1$ if the lowest correlation (or a correlation that is tied for lowest) is obtained when $\delta = 0$.

If we were to naively apply the traditional logic of surrogate data testing here, we could then write down a p-value as the proportion of correlations (shifted or not) at least as large as as the unshifted correlation. In our notation, this is:

$$p_{naive} = B/(2r + 1). \tag{2}$$

As Bartlett noted, p_{naive} is not a valid p-value because the surrogate y series are not independent of each other (e.g. two consecutive shifts are nearly identical). Instead, we show that these time-shifted surrogates can be used to test for independence in a different way. Our approach relies upon the following statistic:

$$u = B/(r + 1). \tag{3}$$

Although u is not a p-value in the usual sense (e.g. $u > 1$ is possible), u can be used in the same way to reject a null hypothesis: For a given significance level α, if $u \leq \alpha$, reject the null hypothesis of independence. Otherwise, do not. Using this procedure, as long as the y series is stationary, the chance of falsely rejecting the null hypothesis will not exceed the significance level. Appendix 1 gives a proof of this claim. We refer to this overall procedure as the truncated time shift (TTS) test.

Our mathematical result may also provide a touch of comfort to analyses performed using the naive test: Comparing the formulas for u and p_{naive} (Eq. 3 and Eq. 2), we can see that as long as the requirement of the TTS test is satisfied (i.e. the time series used to generate surrogate data is stationary), the false positive rate of the naive test will not be inflated above the significance level by more than a factor of 2.

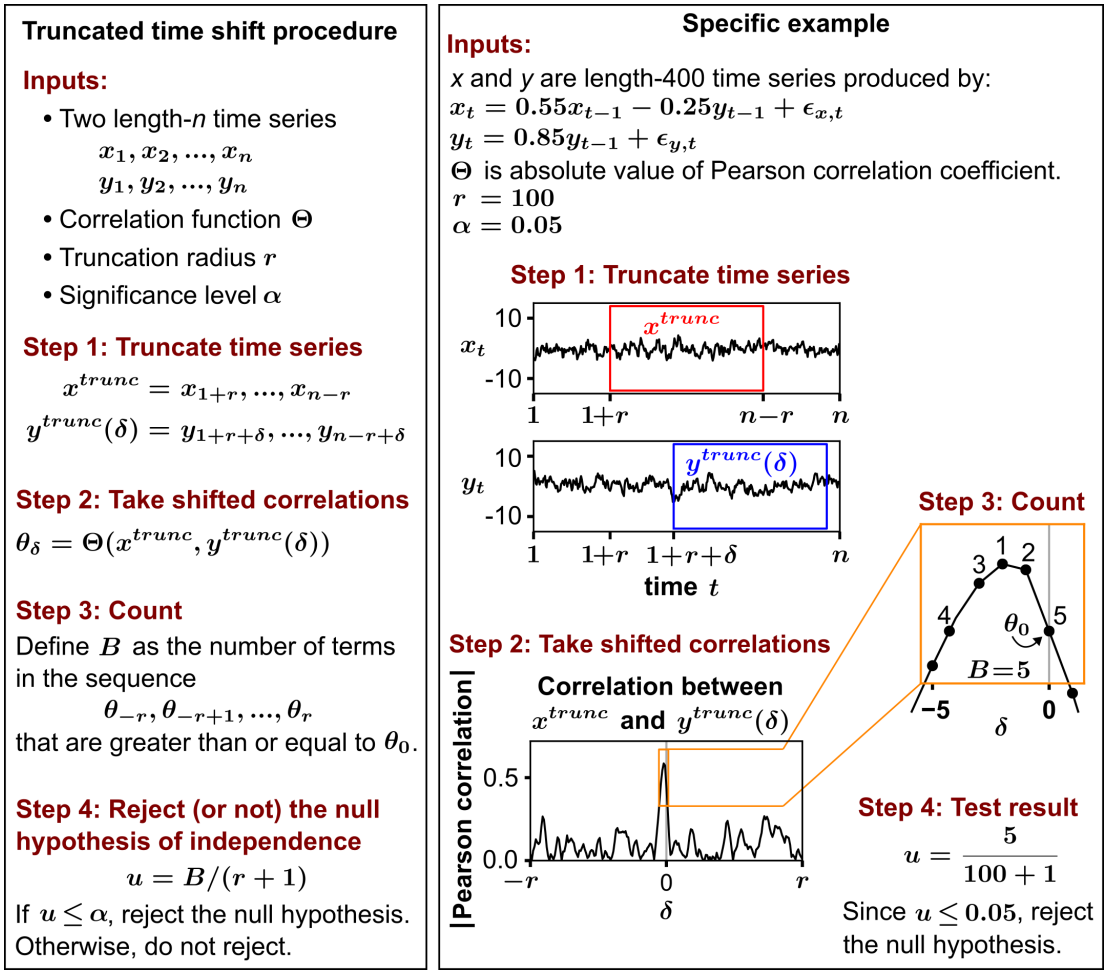

Figure 1: The truncated time shift procedure. The left panel states the procedure and the right panel shows a specific example. In the example, the process noise terms $\epsilon_{x,t}$ and $\epsilon_{y,t}$ are independent and identically distributed normal random variables with zero mean and variance of 1.

The TTS test correctly controls the false positive rate when other surrogate data tests fail

Here we look at how well the TTS test and several existing surrogate data tests for dependence control the false positive rate using different simulated systems. For each system, two independent time series were generated, their correlation was computed, and the significance of this correlation (at the 0.05 level) was determined using one of several surrogate data tests. Empirical false positive rates were then determined as the proportion of 15,000 trials in which a test reported a significant correlation. Whereas all other tests fail in at least one stationary system, the TTS test performs correctly in all stationary systems as expected from our proof. The TTS test also performs well in the two nonstationary systems considered here.

Fig 2A-J shows example dynamics of eight benchmark systems. Fig 2A-B shows a first order autoregressive process and the logistic map, which are two stationary systems commonly used for benchmarking (e.g. [24, 6]). Fig 2C-D shows two stationary systems with a combination of periodic dynamics and noise designed to challenge existing tests: Fig 2C shows two sine waves, one of which has additive noise whose strength varies with a low-frequency sawtooth wave. Fig 2D shows a linear superposition of a sine wave and measurement noise where the signal-to-noise ratio can plummet intermittently. Fig 2E-F show two biologically-inspired nonlinear systems: a stochastic FitzHugh-Nagumo neuronal model [38] and a competitive Lotka-Volterra system with chaotic behavior [39]. These two systems are likely to be stationary, although formal proofs are generally difficult for multivariate nonlinear systems [40]. Fig 2G-H show two systems known to be

nonstationary: a random walk and a first order autoregressive process (same as A) with an additive term that increases over time. In all cases, the two time series are independent. See Appendix 3 for mathematical details.

Two different correlation statistics were used: the sample Pearson correlation coefficient (Fig 2I, left half), and an estimator of mutual information, which is a popular nonlinear form of correlation (Fig 2I, right half).

We compared the following surrogate data tests: the IAAFT procedure [26], the stationary block bootstrap [41, 29, 30], the twin surrogate procedure [31], cyclic permutation time-shifted surrogates, the naive TTS test (Eq. 2) and the TTS test (Fig 1). The TTS test has a false positive rate below 5% for all stationary systems, and even for the two nonstationary systems. All procedures other than the TTS test mistakenly detect dependence at rates above 5% in one or more stationary systems (see the third row of Fig 2I).

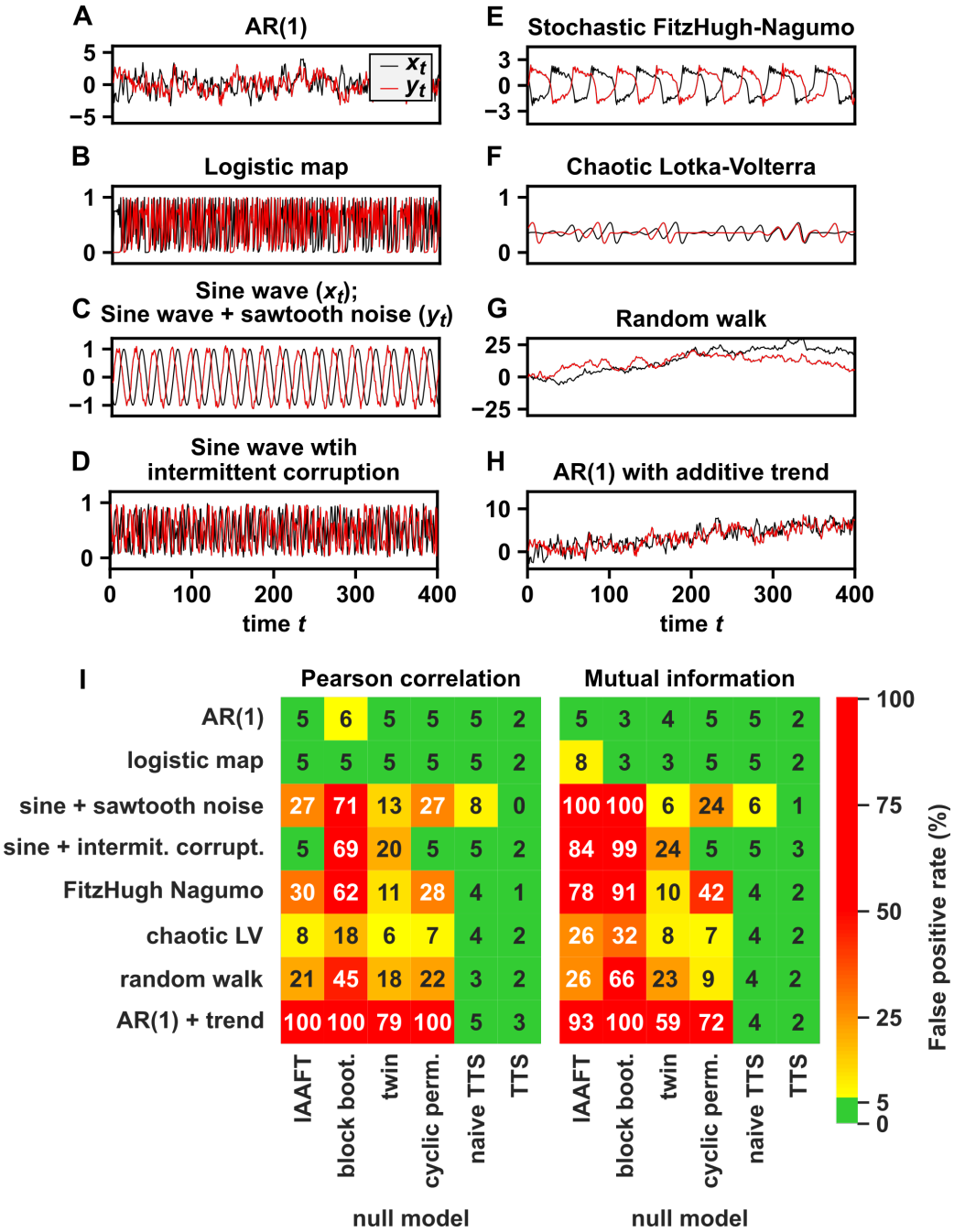

Figure 2: The truncated time shift (TTS) test controls the false positive rate in stationary systems and some nonstationary systems. (A-K) Example dynamics from different benchmark systems, some of which (i.e. A-D) can be proven to be stationary (Appendix 3). E-F are probably stationary, but it is in general difficult to prove strict stationarity for multivariate nonlinear stochastic systems [40]. G-H are not stationary. (A) First-order autoregressive process: current values depend linearly on values one step ago. (B) Logistic map: a deterministic discrete-time model of population dynamics with growth and a carrying capacity [42]. (C) Two sine waves, one with measurement noise whose strength varies with a slow sawtooth wave. (D) A linear superposition of a sine wave and additive noise where the signal-to-noise ratio fluctuates. (E) A stochastic discrete-time FitzHugh-Nagumo model: a nonlinear oscillator inspired by neural voltage dynamics. (F) Chaotic Lotka-Volterra model: an ecological model where species engage in intra- and inter-species competition. (G) A random walk (e.g. the trajectory of a particle whose movements follow a zero-mean normal distribution). (H) The same process as in (A), but with an additive temporal trend. (I) False positive rates of dependence tests. Two independent time series were generated from each system and correlated using either the sample Pearson correlation coefficient or an estimate of mutual information. Each surrogate data test was then used to evaluate the significance of the correlation under the null hypothesis of independence. We show the percentage of 15,000 trials in which each test (falsely) detected dependence at 0.05 significance level, which should ideally be $\leq 5\%$. The labels "block boot." and "cyclic perm." are shorthand for stationary block bootstrap and cyclic permutation surrogates. For the TTS and naive TTS tests, the truncation radius r was set to 59 or 50 respectively. See Appendix 3 for further mathematical details about data-generating processes, correlation statistics, and surrogate data procedures.

Despite the low false positive rates from the TTS test with the two nonstationary systems examined here (Fig 2G-H), we stress that in general, the TTS procedure is not guaranteed to give a valid test when surrogates are generated from a nonstationary process. For instance, Fig S7 shows a nonstationary process for which the TTS test fails to correctly control the false positive rate. One strategy that has previously been used to produce surrogates for some types of nonstationary time series is detrending [43]. This strategy is appropriate for time series that can be decomposed into a deterministic "trend" component and a stationary component. The basic idea is to: (1) extract the stationary component by removing the trend, (2) generate surrogates from the stationary component, and then (3) add the trend back to the surrogates. In Appendix 4, we demonstrate that the TTS test can be modified to incorporate this strategy of detrending and retrending, thus producing a valid surrogate data test for nonstationary time series that can be made stationary via detrending.

Increasing the power of the TTS test

We have so far discussed the probability that the TTS test erroneously rejects the null hypothesis of independence (i.e. the false positive rate). We now consider the probability that the TTS test correctly rejects the null hypothesis (i.e. detection power or the true positive rate).

The truncation radius r simultaneously determines the length of the truncated time series (which is $2r$ less than the total number of time points) and the number of time-shifted surrogates (which is $2r$). Although the false positive rate is always upper-bounded by the significance level α for any choice of r, the true positive rate can be sensitive to r. Since a rearrangement of Eq. 3 gives $r = B/u - 1$, it follows that to detect dependence at the 0.05 level requires an r value of at least 19 ($= 1/0.05 - 1$). If r is exactly 19, significance would be detected at the 0.05 level only when $B = 1$, meaning that the unshifted correlation would need to be strictly greater than all shifted correlations. If $20 \leq r \leq 38$, significance at the 0.05 level still requires $B = 1$ (since if $B = 2$ and $r = 38$, then $u = 2/39 > 0.05$). Thus, setting $r = 19$ always achieves greater detection power than setting $20 \leq r \leq 38$. In general, for a significance level α, power is maximized when r is one less than an integer multiple of $1/\alpha$. As r grows larger, the TTS test will be able to detect dependence with progressively greater values of B, but at the same time, the truncated time series will become smaller.

What is the minimum length of time series required to detect dependence at the 0.05 significance level? As we noted just now, the minimum truncation radius r is 19. Thus, if we use the sample Pearson correlation coefficient as our correlation statistic (which requires at least 3 points), the bare minimum time series length would be $2 \times 19 + 3 = 41$. In practice, many more than three points are desired to quantify correlation.

Different systems will require different time series lengths in order for the TTS test to achieve good power, but simulation results indicate that there are systems where over 80% detection power can be achieved with as few as 100 time points (Fig S4-S5).

In time series with delayed coupling, pre-shifting data can increase the detection power of the TTS test. For example, suppose that an environmental pollutant affects a population of insects by arresting the development of newborns. Then we might expect that a transient increase in pollutant levels will correspond to a dip in insect population size after a delay of about one generation. The TTS test might fail to detect this dependence because the correlation between unshifted time series would lower than when the pollutant time series is advanced by 1 insect generation. Pre-shifting one of the series by one generation can remedy this problem (Fig 3). If the optimal coupling delay is uncertain, a simple approach is to shift a time series by several plausible delays and perform a Bonferroni multiple testing correction (Fig 3).

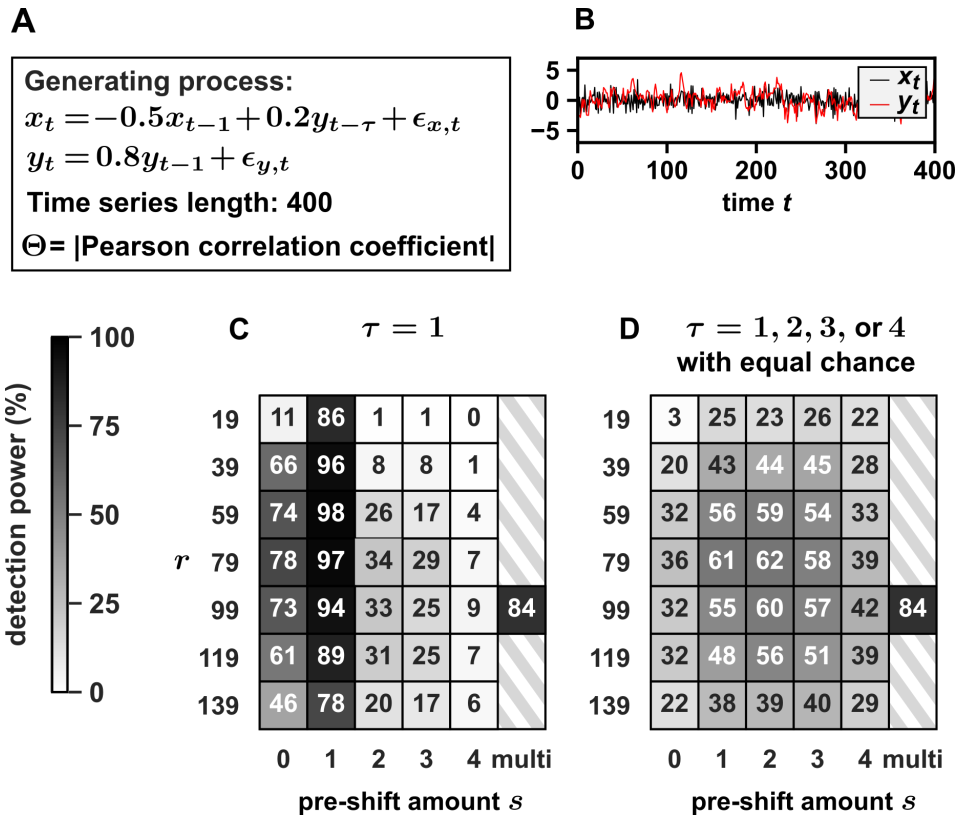

Figure 3: Detection power varies with the truncation radius r and the pre-shift amount s. (A, B) The data generating process and a typical example of time series. Pairs of length-400 time series were generated by coupled autoregressive processes in which y influences x with a lag of τ, whereas x does not influence y. The absolute value of the Pearson correlation between the two time series was computed, and the TTS test was used to evaluate statistical significance at the 0.05 level. (C, D) Prior to performing the TTS test, the x series was shifted s time steps into the future, where s is $0, 1, 2, 3,$ or 4. That is, we tested for dependence between $\{x_{1+s}, ..., x_n\}$ and $\{y_1, ..., y_{n-s}\}$. We also tried using all five of these s values and performing a Bonferroni multiple test correction (the "multi" column). In all cases, detection power was computed as the proportion of 1000 simulations in which dependence was detected at the 0.05 significance level. Note that in the test with multiple pre-shifts, the Bonferroni-corrected significance level is $0.05/[\#\text{ of tests}] = 0.05/5 = 0.01$. Thus, we chose $r = 1/0.01 - 1 = 99$. For the multiple pre-shift test, if the correlation was strictly maximized at $\delta = 0$ for any of the five pre-shifts, then a u-value of 0.01 was obtained, and thus we rejected the null hypothesis that the x and y series are independent. When the true coupling delay τ was 1 (C) or uncertain (chosen with equal chance from between $1, 2, 3,$ or 4; D), the multiple pre-shift test gave good ($> 80\%$) detection power.

An example from climate science

Pre-industrial climatological change is widely believed to have been driven largely by variability in the Earth's orbit around the sun. The Earth's rotation around the Sun is characterized by parameters known as eccentricity, obliquity, and precession. Eccentricity describes the shape of the orbit (which. varies from nearly circular to slightly elliptical over approximately $96,000$-year cycles); obliquity is the angle between Earth's rotational axis and the normal of the orbital plane (which cycles over roughly $41,000$ years in a band roughly bounded between $22°$ and $24.5°$); and precession describes how the rotational axis of the earth rotates around the line normal to the orbital plane (roughly $21,000$ years/cycle) [44, 45]. Each of these parameters is thought to play a role in Earth's climate, although some parameters may be more influential than others, and the extent of a parameter's influence may change over time [46, 47, 45]. The climate record is characterized by repeated episodes of cooling followed by warming events called deglaciations. Until about one million years ago, deglaciations occured with a period of about 41 kiloyears, which is the period of obliquity cycles. Because of this, obliquity is often said to "pace" glacial cycles [48, 49]. Yet, two time series with shared periodic elements can be statistically independent (e.g. Fig 2C).

Using the TTS procedure, we tested for dependence between orbital parameters and deglaciations with only the assumption that the orbital parameters are stationary. We converted a list of 36 deglaciation events in the last two million years (2 Myr) calculated by Huybers [49] into a deglaciation time series (Fig 4A). Specifically, we used a "sampling frequency" of 1000 years by assigning a 1 to the deglaciation variable if a kiloyear contained a deglaciation event and a 0 otherwise. Our deglaciation time series thus has 2000 time points. We did not use a higher sampling frequency due to uncertainty in deglaciation timing, and avoided a lower frequency (e.g. 10 kiloyears) in order to adequately capture the shapes of the obliquity and precession cycles. For estimates of obliquity, precession, and eccentricity, we used the numerical solution from [44]. Since this model provides accurate estimates of orbital parameters over at least tens of millions of years (and also predicts future values), we used the entire past 2 Myr as our "truncated" time series and generated surrogates using orbital parameter time series with a truncation radius of 10 Myr (Fig 4A).

We used the absolute value of the Pearson correlation coefficient as the correlation function (Fig 4B, left column). To address the possibility of a primarily nonlinear dependence between deglaciations and orbital parameters (as in Fig S5), we also tried a prediction-based correlation function similar to the cross-map skill statistic used in other largely deterministic nonlinear systems [6, 20] (Fig 4B, right column). For both correlation functions, we detected a significant ($u < 0.05$) correlation between deglaciation times and obliquity, but not the other two orbital parameters (Fig 4B). A qualitatively equivalent result was also obtained by the original period-based analysis of Huybers [49].

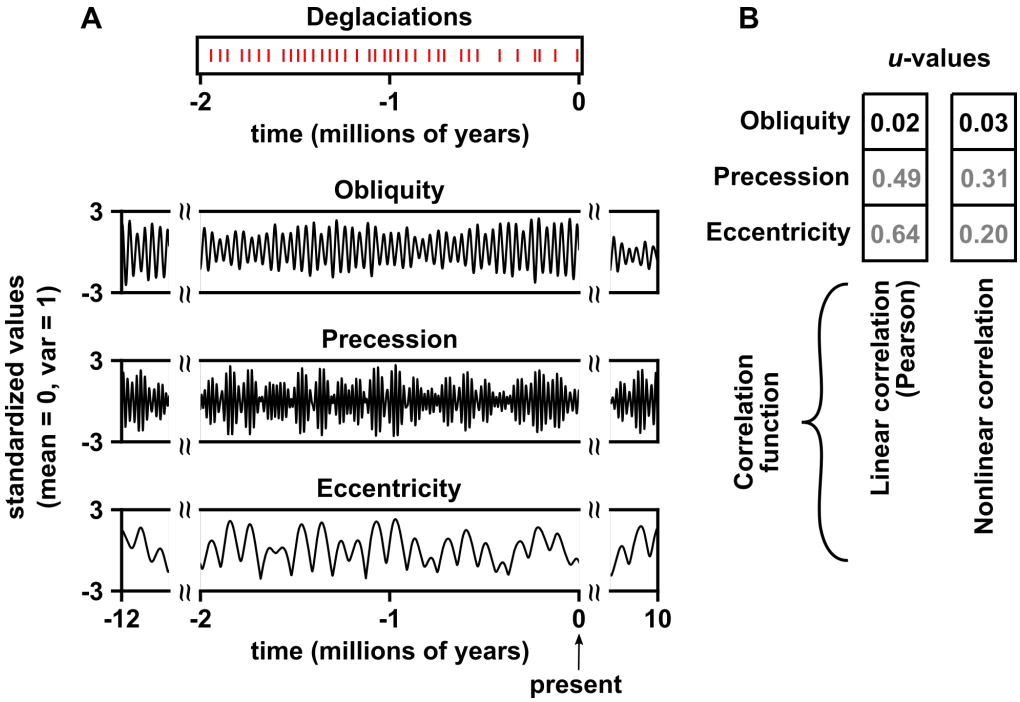

Figure 4: The TTS test detected dependence between deglaciation and obliquity, but not precession or eccentricity. (A) Time series of orbital parameters (estimated from the model in [44]) and deglaciation events (from [49]). Orbital parameter time series were used to generate time-shifted surrogates. (B) TTS test u-values for linear correlation strength (i.e. the absolute value of the Pearson correlation coefficient) and a prediction-based nonlinear correlation function. In the latter case, we predicted deglaciation events based on contemporaneous and previous orbital parameter values, computed the resulting mean squared prediction error (MSE), and used $-1 \times MSE$ as the nonlinear correlation statistic. See Appendix 5 for methodological details. We made surrogate data by shifting the orbital parameter time series and used a truncation radius of 10 million years (i.e. $10,000$ time points at the sampling frequency of 1 kiloyear). The null hypothesis that obliquity and deglaciation are independent can be rejected at the 0.05 significance level.

An example from human microbiome science

The human microbiome is highly spatial, with different body sites playing host to distinct microbial communities [50]. Caporaso et al. performed a longitudinal study in which daily microbial surveys were conducted for over a year on a male subject at four body sites: the left palm, right palm, tongue, and gut (feces) [51]. Since this study measured relative abundance of microbial taxa (not absolute population size), it is difficult to rigorously test for dependence between taxa. However, testing for overall dependence between the microbial communities of two body sites remains unproblematic. Intuitively, this is because the compositional transformation does not mix measurements from different body sites (see Appendix 6.1 for a formal argument). With the time series from [51] we used the TTS test to look for dependence between the microbial communities living on the left palm, right palm tongue, and gut.

Data were prepared (Fig 5A) as follows: We obtained OTU (operational taxonomic units) relative abundance tables from the data in [51] using the online Qiita platform [52] (see Appendix 6.4). OTUs that were absent in over half of measurements were discarded. We used measurements from day 42 to day 418 to avoid data gaps of 7 or more days. Shorter gaps were filled by linear interpolation. (Results did not qualitatively change if gaps were instead filled with randomly chosen abundance measurements; see Appendix 6.4). Next, any OTUs not considered stationary by an augmented Dickey-Fuller (ADF) test (at the 0.05 significance level) were removed. After these pre-processing steps, the number of remaining OTUs ranged from 180 (tongue) to 507 (right palm).

We then performed a TTS test between each pair of body sites. Although we have so far focused on

examples in which time series are univariate, the TTS test and the theorem by which it is justified (Appendix 1) are also valid if time series are multivariate, as in this case study (Fig 5C). To correlate datasets from two body sites (Fig 5B), we first listed all of their shared OTUs. Then for each shared OTU, we computed the sample Pearson correlation coefficient between the relative abundances of that OTU in the two sites. Our correlation statistic Θ was the median of these correlation coefficients across all shared OTUs. We also tried substituting Pearson correlation with local similarity, which is a correlation statistic that is designed to detect transient temporal correlations and is popular in microbiome science [5, 14]. We emphasize that we did not perform a separate TTS test for each OTU since the correlation between body sites was summarized by a single univariate statistic (Fig 5B).

We detected dependence between the microbial communities on the left and right palms, and between the palms and the tongue (Fig 5D). This result was the same regardless of whether Pearson correlation or local similarity was used in the step of Fig 5B. These results reveal more cross-site dependences than the original analysis of [51], which examined correlations among body sites using a different approach and detected dependence only between the left and right palms. Caporaso et al. tested for dependence among body sites by first computing the phylogenetic distance between temporally adjacent microbiomes within body sites, and then calculating correlations between between the phylogenetic distances of different body sites [51]. Unlike the analysis of Caporaso et al., our analysis relies on correlation statistics that lack a well-known sampling distribution (i.e. the median of many Pearson correlation coefficients or local similarities). Yet, due to the flexibility of the TTS test, we are nevertheless able to perform an exact statistical test, assuming that abundance time series are stationary.

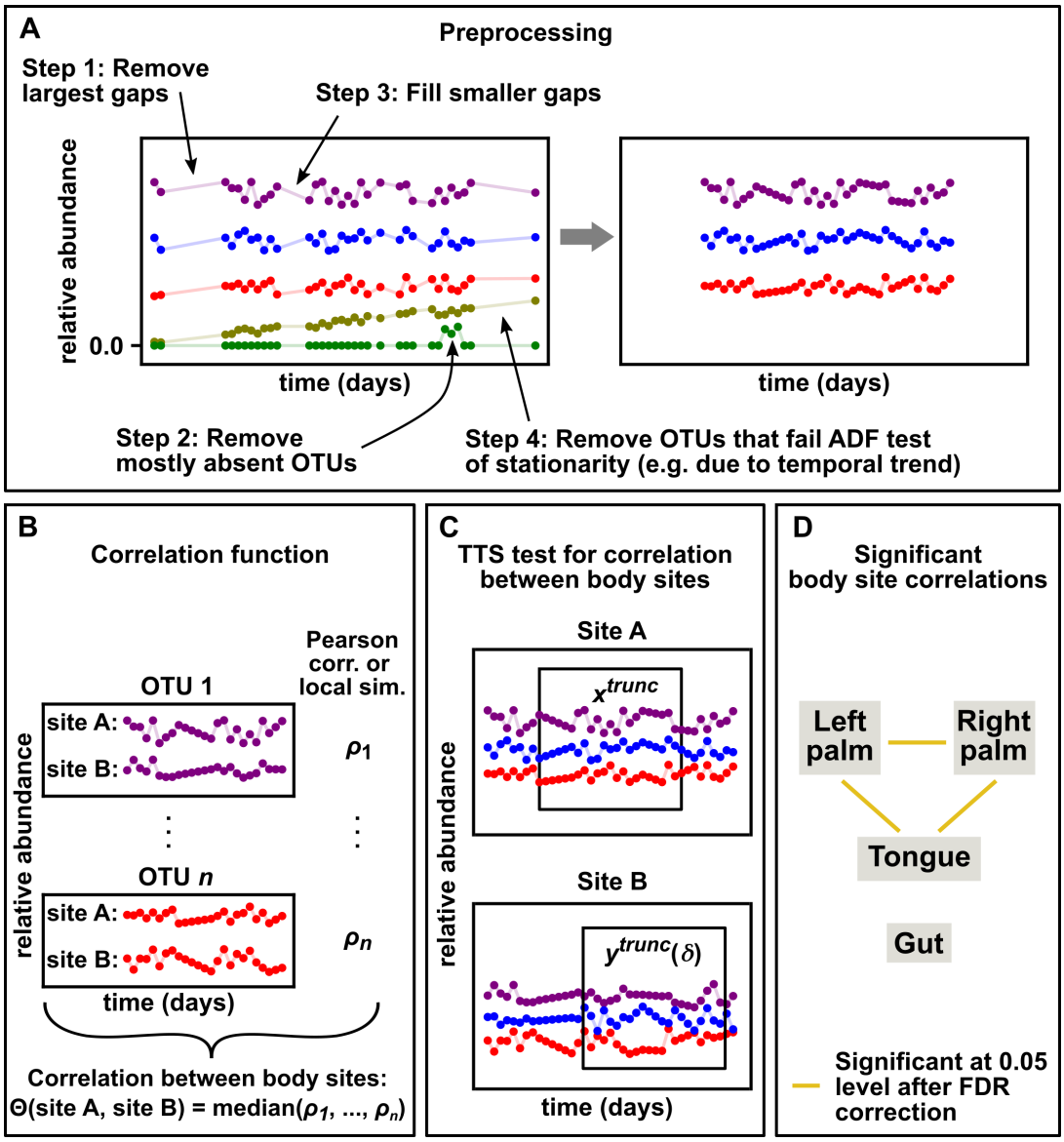

Figure 5: The TTS test applied to longitudinally sampled microbiomes from four body sites detects dependence between body sites. (A) Data preprocessing. To remove long (> 6-day) gaps at the beginning and end of time series, we only used measurements from day 42 to day 418. Remaining gaps were filled by linear interpolation or by randomly resampling abundance values. OTUs were removed if they were absent from most measurements or if they were not considered stationary (at the 0.05 significance level) by an ADF test implemented in the statsmodels Python package [53]. OTUs removed from one body site were not necessarily removed from others. (B) Correlation between body sites. If an OTU ("OTU i") was shared by two sites, we computed the Pearson correlation or local similarity score between OTU i abundances in one site and OTU i abundances in the other (call this ρ_i). We defined the correlation between any two body sites (i.e. our "Θ" in the notation of Fig 1) to be the median of $(\rho_1, ..., \rho_n)$, where n is the number of shared OTUs. (C) We assessed the statistical significance of the correlations between body sites using a TTS test. Note that this setup avoids the need to perform a separate test for each OTU. We used an intermediate truncation radius r of 79 days (inspired by Fig 3C). (D) The test detected dependence between the two palms and between palms and the tongue. All links shown were detected with a significance level of 0.05 after a Benjamini-Hochberg false discovery rate adjustment for performing 6 tests [54, 55]. Results did not qualitatively depend on either the gap-filling method in (A) or whether Pearson correlation versus local similarity was used in (B). Additionally, for each correlation, we obtained the same result regardless of which body site was used to generate surrogates. See Appendix 6.4 for more details.

Discussion

A statistical hypothesis test for dependence between two time series usually requires at least two key ingredients: (1) a correlation statistic that quantifies the strength of the relationship between the two time series, and (2) a null model that can be used to determine the probability distribution of that statistic if the two time series are in fact independent. These two ingredients seem to have received different levels of attention over the past couple of decades. Recent years have witnessed the development and rapid adoption of new correlation statistics that can detect transient or nonlinear forms of dependence, some of which even attempt to infer the direction of causation [5, 14, 6, 56, 57]. In practice however, these correlation statistics have often been paired with surrogate data null models which assume a Gaussian linear process (i.e. random phase surrogates), or even independent and identically distributed data (e.g. the permutation test). Pairing a correlation statistic that is intended for nonlinear processes with a null model that requires a linear process seems intellectually inconsistent (and indeed dangerous; see Fig 2). Yet one might sympathize with this choice as other tests do not appear to be panaceas (Fig 2). Indeed, the general problem of assigning statistical significance to nonlinear correlations between time series does not appear to have a broadly-accepted solution [58, 59, 8].

Here, we have introduced the TTS test, a statistical hypothesis test of dependence between two time series which can be used with any correlation function and which is valid as long as one of the time series is stationary (or can be made stationary by detrending; Appendix 4). This is a minimally restrictive requirement among exactly valid nonparametric tests of dependence between time series. We showed that this test was able to verify the previously observed dependences between obliquity and deglaciation timing (Fig 4) as well as between the microbiome compositions of the left and right palms (Fig 5). In the microbiome dataset, it even identified additional relationships that went undetected by the original analysis of Caporaso et al. [51]. Since surrogate data tests are presently used in several disciplines [24, 20, 60, 22, 61] (with time-shift surrogates in particular being an established method in neuroscience [34, 62, 35, 27]), we expect the TTS test to find utility in diverse application domains.

The main limitation of the TTS test is the amount of data it requires for high detection power. Under the TTS test, the required length of the time series increases with the reciprocal of the desired significance level. Even more data are required if the optimal coupling delay is uncertain. Moreover, whereas other surrogate data tests often specify the desired false positive rate (at least approximately), the TTS test only specifies an upper bound on the false positive rate, and the actual false positive rate is often substantially lower (Fig 2). We can therefore expect other tests for dependence to typically achieve high detection power with fewer data than the TTS test. Indeed, this is what we have found in numerical experiments (Fig S10). Thus, the TTS test is ideally used when a nonparametric test is required, when time series are long, when the coupling lag can be pinned down to a small range, and when one is unable to verify the assumptions necessary for existing surrogate data tests (which typically include stationarity as well as additional requirements [25, 27]).

Can an exactly valid and distribution-free surrogate data test for independence be performed even when we cannot assume that a temporal process is stationary (and do not know how to transform it to make it stationary)? Perhaps, but this seems difficult without some other similarly restrictive modeling assumption. Without knowing whether a system is stationary, many potential data sets can be described similarly well by either independent or dependent processes (Fig 6).

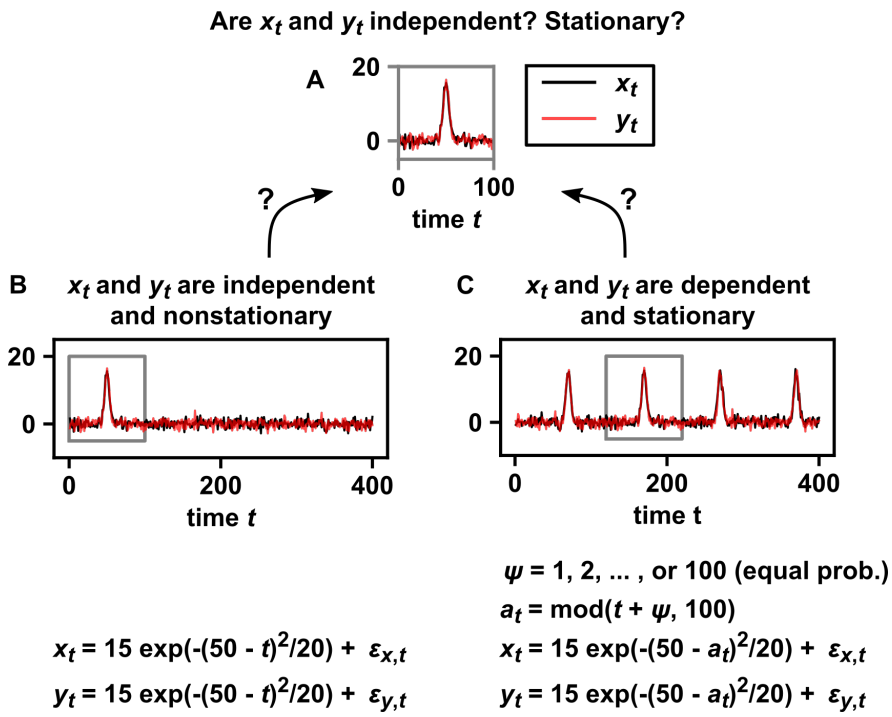

Figure 6: Without knowledge about stationarity, a pair of temporal processes could be either independent or dependent. (A) 100 time points from two time series with a shared spike centered at $t = 50$. (B) These data are described well by a process in which the time series are independent and nonstationary. (C) The data are also described well by a process where the time series are dependent and stationary. In both processes, the noise terms $\epsilon_{x,t}$ and $\epsilon_{y,t}$ are independent and identically distributed normal random variables with mean of zero and variance of one. With the data in (A), the TTS test with $r = 19$ found a significant Pearson correlation strength at the 0.05 significance level, thus rejecting the null hypothesis of independence. If (B) were the ground truth, we would consider this a "false positive", whereas if (C) were the ground truth, we would consider this a "true positive" result.

Since the stationarity assumption seems so central to dependence testing, it would be convenient to have a statistical test for stationarity. However, a single time series (e.g. Fig 6A) can often be plausibly described by either a nonstationary process (e.g. Fig 6B) or a stationary one (e.g. Fig 6C). Nevertheless, there are many statistical methods that attempt to test for stationarity or a similar property in a single time series, albeit with various modeling assumptions or other caveats [63, 64, 65, 66]. For example, although the popular augmented Dickey Fuller (ADF) test [63, 66] is sometimes used as a pragmatic means of assessing stationarity (e.g. Fig 4; [67, 68]), its null hypothesis is not exactly nonstationarity. In fact, a significant p-value from the ADF test indicates that a time series is free from some sources of nonstationarity (e.g. a random walk), but other sources of nonstationarity (e.g. time-varying parameters) may in principle still be present. Overall, statistical tests cannot guarantee that a time series is stationary, but can provide supporting evidence. Background knowledge of the process can also be used to support the stationarity assumption: Mathematical work has shown that the stationarity condition is often met by stochastic processes that tend to relax toward a stable equilibrium [69]. Periodic processes, measurement noise processes, and chaotic processes (e.g. Fig 2B), can also be stationary, as can combinations of these processes.

In sum, an important and longstanding problem is that of nonparametrically testing for the statistical significance of a correlation between autocorrelated time series. The TTS test provides an approach that imposes relatively minimal requirements onto both the correlation statistic and the data-generating process. This gives researchers the freedom to apply a large arsenal of correlation statistics to a wide array of processes without sacrificing rigor. This freedom will become more valuable in the future as both correlation techniques and data availability continue to proliferate across diverse fields of research.

References

[1] G. U. Yule, "Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series," *Journal of the royal statistical society*, vol. 89, no. 1, pp. 1–63, 1926.

[2] C. Granger and P. Newbold, "Spurious regressions in econometrics," *Journal of Econometrics*, vol. 2, no. 2, pp. 111–120, 1974.

[3] S. Weiss, W. Van Treuren, C. Lozupone, K. Faust, J. Friedman, Y. Deng, L. C. Xia, Z. Z. Xu, L. Ursell, E. J. Alm, *et al.*, "Correlation detection strategies in microbial data sets vary widely in sensitivity and precision," *The ISME journal*, vol. 10, no. 7, pp. 1669–1681, 2016.

[4] A. R. Coenen and J. S. Weitz, "Limitations of correlation-based inference in complex virus-microbe communities," *mSystems*, vol. 3, no. 4, pp. e00084–18, 2018.

[5] Q. Ruan, D. Dutta, M. S. Schwalbach, J. A. Steele, J. A. Fuhrman, and F. Sun, "Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors," *Bioinformatics*, vol. 22, no. 20, pp. 2532–2538, 2006.

[6] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems," *Science*, vol. 338, no. 6106, pp. 496–500, 2012.

[7] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

[8] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, *et al.*, "Inferring causation from time series in earth system sciences," *Nature communications*, vol. 10, no. 1, pp. 1–13, 2019.

[9] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1997.

[10] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[11] W. Conover, "Distribution-free methods in statistics," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 2, pp. 199–207, 2009.

[12] S. Afyouni, S. M. Smith, and T. E. Nichols, "Effective degrees of freedom of the pearson's correlation coefficient under autocorrelation," *NeuroImage*, vol. 199, pp. 609–625, 2019.

[13] B. J. Pyper and R. M. Peterman, "Comparison of methods to account for autocorrelation in correlation analyses of fish data," *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 55, no. 9, pp. 2127–2140, 1998.

[14] L. C. Xia, J. A. Steele, J. A. Cram, Z. G. Cardon, S. L. Simmons, J. J. Vallino, J. A. Fuhrman, and F. Sun, "Extended local similarity analysis (elsa) of microbial community and other time series data with replicates," in *BMC systems biology*, vol. 5, pp. 1–12, Springer, 2011.

[15] A. Eiler, F. Heinrich, and S. Bertilsson, "Coherent dynamics and association networks among lake bacterioplankton taxa," *The ISME journal*, vol. 6, no. 2, pp. 330–342, 2012.

[16] A. Shade, P. S. McManus, and J. Handelsman, "Unexpected diversity during community succession in the apple flower microbiome," *MBio*, vol. 4, no. 2, 2013.

[17] F. Zhang, F. Sun, and Y. Luan, "Statistical significance approximation for local similarity analysis of dependent time series data," *BMC bioinformatics*, vol. 20, no. 1, p. 53, 2019.

[18] V. Cyriaque, A. Géron, G. Billon, J. Nesme, J. Werner, D. C. Gillan, S. J. Sørensen, and R. Wattiez, "Metal-induced bacterial interactions promote diversity in river-sediment microbiomes," *FEMS Microbiology Ecology*, vol. 96, no. 6, p. fiaa076, 2020.

[19] E. Brookshire and T. Weaver, "Long-term decline in grassland productivity driven by increasing dryness," *Nature communications*, vol. 6, no. 1, pp. 1–7, 2015.

[20] A. A. Tsonis, E. R. Deyle, R. M. May, G. Sugihara, K. Swanson, J. D. Verbeten, and G. Wang, "Dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature," *Proceedings of the National Academy of Sciences*, vol. 112, no. 11, pp. 3253–3256, 2015.

[21] E. H. Van Nes, M. Scheffer, V. Brovkin, T. M. Lenton, H. Ye, E. Deyle, and G. Sugihara, "Causal feedbacks in climate change," *Nature Climate Change*, vol. 5, no. 5, pp. 445–448, 2015.

[22] S.-i. S. Matsuzaki, K. Suzuki, T. Kadoya, M. Nakagawa, and N. Takamura, "Bottom-up linkages between primary production, zooplankton, and fish in a shallow, hypereutrophic lake," *Ecology*, vol. 99, no. 9, pp. 2025–2036, 2018.

[23] M. Wang, C. Yoshimura, A. Allam, F. Kimura, and T. Honma, "Causality analysis and prediction of 2-methylisoborneol production in a reservoir using empirical dynamic modeling," *Water research*, vol. 163, p. 114864, 2019.

[24] W. Ebisuzaki, "A method to estimate the statistical significance of a correlation when the data are serially correlated," *Journal of Climate*, vol. 10, no. 9, pp. 2147–2153, 1997.

[25] R. G. Andrzejak, A. Kraskov, H. Stögbauer, F. Mormann, and T. Kreuz, "Bivariate surrogate techniques: necessity, strengths, and caveats," *Physical review E*, vol. 68, no. 6, p. 066202, 2003.

[26] T. Schreiber and A. Schmitz, "Surrogate time series," *Physica D: Nonlinear Phenomena*, vol. 142, no. 3-4, pp. 346–382, 2000.

[27] G. Lancaster, D. Iatsenko, A. Pidde, V. Ticcinelli, and A. Stefanovska, "Surrogate data for hypothesis testing of physical systems," *Physics Reports*, vol. 748, pp. 1–60, 2018.

[28] K.-S. Chan, "On the validity of the method of surrogate data," *Fields Inst. Commun*, vol. 11, pp. 77–97, 1997.

[29] C. Diks and J. DeGoede, "A general nonparametric bootstrap test for granger causality," *Global analysis of dynamical systems*, pp. 391–403, 2001.

[30] A. Papana, C. Kyrtsou, D. Kugiumtzis, and C. Diks, "Assessment of resampling methods for causality testing: A note on the us inflation behavior," *PloS one*, vol. 12, no. 7, p. e0180852, 2017.

[31] M. Thiel, M. C. Romano, J. Kurths, M. Rolfs, and R. Kliegl, "Twin surrogates to test for complex synchronisation," *EPL (Europhysics Letters)*, vol. 75, no. 4, p. 535, 2006.

[32] M. C. Romano, M. Thiel, J. Kurths, K. Mergenthaler, and R. Engbert, "Hypothesis test for synchronization: twin surrogates revisited," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 19, no. 1, p. 015108, 2009.

[33] Z. Jia, Y. Lin, Y. Liu, Z. Jiao, and J. Wang, "Refined nonuniform embedding for coupling detection in multivariate time series," *Physical Review E*, vol. 101, no. 6, p. 062113, 2020.

[34] T. I. Netoff and S. J. Schiff, "Decreased neuronal synchronization during experimental seizures," *Journal of Neuroscience*, vol. 22, no. 16, pp. 7297–7307, 2002.

[35] L. Faes, A. Porta, and G. Nollo, "Mutual nonlinear prediction as a tool to evaluate coupling strength and directionality in bivariate time series: comparison among different strategies based on k nearest neighbors," *Physical Review E*, vol. 78, no. 2, p. 026201, 2008.

[36] I. Vlachos and D. Kugiumtzis, "Nonuniform state-space reconstruction and coupling detection," *Physical Review E*, vol. 82, no. 1, p. 016207, 2010.

[37] M. Bartlett, "Some aspects of the time-correlation problem in regard to tests of significance," *Journal of the Royal Statistical Society*, vol. 98, no. 3, pp. 536–543, 1935.

[38] R. FitzHugh, "Impulses and physiological states in theoretical models of nerve membrane," *Biophysical journal*, vol. 1, no. 6, pp. 445–466, 1961.

[39] J. Vano, J. Wildenberg, M. Anderson, J. Noel, and J. Sprott, "Chaos in low-dimensional lotka–volterra models of competition," *Nonlinearity*, vol. 19, no. 10, p. 2391, 2006.

[40] D. Tjøstheim, "Non-linear time series and markov chains," *Advances in applied probability*, vol. 22, no. 3, pp. 587–611, 1990.

[41] D. N. Politis and J. P. Romano, "The stationary bootstrap," *Journal of the American Statistical association*, vol. 89, no. 428, pp. 1303–1313, 1994.

[42] R. M. May, "Simple mathematical models with very complicated dynamics," *The Theory of Chaotic Attractors*, pp. 85–93, 2004.

[43] J. Lucio, R. Valdés, and L. Rodríguez, "Improvements to surrogate data methods for nonstationary time series," *Physical Review E*, vol. 85, no. 5, p. 056202, 2012.

[44] J. Laskar, P. Robutel, F. Joutel, M. Gastineau, A. Correia, and B. Levrard, "A long-term numerical solution for the insolation quantities of the earth," *Astronomy & Astrophysics*, vol. 428, no. 1, pp. 261–285, 2004.

[45] M. Maslin, "Forty years of linking orbits to ice ages," *Nature*, vol. 540, no. 7632, pp. 208–209, 2016.

[46] J. D. Hays, J. Imbrie, N. J. Shackleton, *et al.*, "Variations in the earth's orbit: pacemaker of the ice ages," *science*, vol. 194, no. 4270, pp. 1121–1132, 1976.

[47] C. Lorius, J. Jouzel, C. Ritz, L. Merlivat, N. Barkov, Y. S. Korotkevich, and V. Kotlyakov, "A 150,000-year climatic record from antarctic ice," *Nature*, vol. 316, no. 6029, pp. 591–596, 1985.

[48] P. Huybers and C. Wunsch, "Obliquity pacing of the late pleistocene glacial terminations," *Nature*, vol. 434, no. 7032, pp. 491–494, 2005.

[49] P. Huybers, "Glacial variability over the last two million years: an extended depth-derived agemodel, continuous obliquity pacing, and the pleistocene progression," *Quaternary Science Reviews*, vol. 26, no. 1-2, pp. 37–55, 2007.

[50] C. Huttenhower, D. Gevers, R. Knight, S. Abubucker, J. H. Badger, A. T. Chinwalla, H. H. Creasy, A. M. Earl, M. G. FitzGerald, R. S. Fulton, *et al.*, "Structure, function and diversity of the healthy human microbiome," *nature*, vol. 486, no. 7402, p. 207, 2012.

[51] J. G. Caporaso, C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, *et al.*, "Moving pictures of the human microbiome," *Genome biology*, vol. 12, no. 5, pp. 1–8, 2011.

[52] A. Gonzalez, J. A. Navas-Molina, T. Kosciolek, D. McDonald, Y. Vázquez-Baeza, G. Ackermann, J. DeReus, S. Janssen, A. D. Swafford, S. B. Orchanian, *et al.*, "Qiita: rapid, web-enabled microbiome meta-analysis," *Nature methods*, vol. 15, no. 10, pp. 796–798, 2018.

[53] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[54] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[55] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of statistics*, pp. 1165–1188, 2001.

[56] H. Ye, E. R. Deyle, L. J. Gilarranz, and G. Sugihara, "Distinguishing time-delayed causal interactions using convergent cross mapping," *Scientific reports*, vol. 5, p. 14750, 2015.

[57] D. Harnack, E. Laminski, M. Schünemann, and K. R. Pawelzik, "Topological causality in dynamical systems," *Physical review letters*, vol. 119, no. 9, p. 098301, 2017.

[58] S. Cobey and E. B. Baskerville, "Limits to causal inference with state-space reconstruction for infectious disease," *PloS one*, vol. 11, no. 12, p. e0169050, 2016.

[59] C.-W. Chang, M. Ushio, and C.-h. Hsieh, "Empirical dynamic modeling for beginners," *Ecological Research*, vol. 32, no. 6, pp. 785–796, 2017.

[60] B. Hannisdal, K. A. Haaga, T. Reitan, D. Diego, and L. H. Liow, "Common species link global ecosystems to climate change: dynamical evidence in the planktonic fossil record," *Proceedings of the Royal Society B: Biological Sciences*, vol. 284, no. 1858, p. 20170722, 2017.

[61] M. Ushio, C.-h. Hsieh, R. Masuda, E. R. Deyle, H. Ye, C.-W. Chang, G. Sugihara, and M. Kondoh, "Fluctuating interaction network and time-varying stability of a natural fish community," *Nature*, vol. 554, no. 7692, pp. 360–363, 2018.

[62] R. Q. Quiroga, A. Kraskov, T. Kreuz, and P. Grassberger, "Performance of different synchronization measures in real data: a case study on electroencephalographic signals," *Physical Review E*, vol. 65, no. 4, p. 041903, 2002.

[63] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American statistical association*, vol. 74, no. 366a, pp. 427–431, 1979.

[64] A. Witt, J. Kurths, and A. Pikovsky, "Testing stationarity in time series," *physical Review E*, vol. 58, no. 2, p. 1800, 1998.

[65] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?," *Journal of econometrics*, vol. 54, no. 1-3, pp. 159–178, 1992.

[66] R. Davidson, J. G. MacKinnon, *et al.*, *Econometric theory and methods*, vol. 5. Oxford University Press New York, 2004.

[67] S. M. Gibbons, S. M. Kearney, C. S. Smillie, and E. J. Alm, "Two dynamic regimes in the human gut microbiome," *PLoS computational biology*, vol. 13, no. 2, p. e1005364, 2017.

[68] N. M. Odogwu, C. A. Onebunne, J. Chen, F. A. Ayeni, M. R. Walther-Antonio, O. O. Olayemi, N. Chia, and A. O. Omigbodun, "Lactobacillus crispatus thrives in pregnancy hormonal milieu in a nigerian patient cohort," *Scientific reports*, vol. 11, no. 1, pp. 1–19, 2021.

[69] D. A. Jones and D. R. Cox, "Nonlinear autoregressive processes," *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, vol. 360, no. 1700, pp. 71–95, 1978.

[70] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *The annals of statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[71] A. E. Yuan and W. Shou, "Data-driven causal analysis of observational time series in ecology," *bioRxiv*, pp. 2020–08, 2021.

[72] J. S. Rosenthal, *A First Look At Rigorous Probability Theory.* World Scientific Publishing Company, 2006.

[73] H. White, *Asymptotic theory for econometricians.* Academic press, 1984.

[74] D. L. Cohn, *Measure theory.* Springer, 2013.

[75] W. Greene, *Econometric Analysis.* Pearson, 2018.

[76] A. Lasota and M. C. Mackey, *Chaos, fractals, and noise: stochastic aspects of dynamics*, vol. 97. Springer Science & Business Media, 2013.

[77] J. F. Donges, J. Heitzig, B. Beronov, M. Wiedermann, J. Runge, Q. Y. Feng, L. Tupikina, V. Stolbova, R. V. Donner, N. Marwan, *et al.*, "Unified functional network and nonlinear time series analysis for complex systems science: The pyunicorn package," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 25, no. 11, p. 113101, 2015.

[78] G. Sugihara and R. M. May, "Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series," *Nature*, vol. 344, no. 6268, pp. 734–741, 1990.

[79] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[80] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.