Data-driven causal analysis of observational time series in ecology

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

February 25, 2021

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 Department of Genetics, Evolution and Environment, University College London, London, the United Kingdom

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

Authorship: AY wrote and ran all simulations. AY and WS wrote all versions of the manuscript together. AY performed the formal analysis contained in the Appendix.

Data accessibility: Simulation results can be reproduced using the source code, which will be made available on Github.

Short title: Data-driven causal analysis of time series

Keywords: time series, causal inference, model-free, correlation, Reichenbach's common cause principle, Granger causality, state space reconstruction, convergent cross mapping, surrogate data, Takens's theorem

Abstract

Natural ecosystems pose a major scientific challenge as they are complex and often defy manipulative experiments for practical or ethical reasons. In response, several different fields have developed parallel approaches to infer causality from passively measured time series. Yet these methods are easy to misunderstand and often controversial. Here we offer an accessible and critical review of three popular statistical approaches that do not require a priori mechanistic knowledge: pairwise correlation, Granger causality, and state space reconstruction. For each, we ask what information a method tells us, what causal statement it might imply, and when it could lead us astray. We devise new ways of visualizing key concepts, describe some novel pathologies of causal inference methods, and point out how so-called "model-free" causality tests are not assumption-free. We hope that our synthesis will facilitate thoughtful application of causal inference approaches and encourage explicit statements of assumptions.

Introduction

Microscopic and macroscopic ecological communities play key roles in diverse environments, from industrial food fermentors, to global cycling of elements, to the human body. Communities are often highly complex, with many species engaging in diverse interactions such as the release and consumption of multiple chemical compounds. To control or design communities, it is critical to identify causal relationships within ecosystems. In other words, we want to know whether perturbing the abundance of one species might perturb the abundance of another.

Ideally, biologists discover such causal relationships from manipulative experiments. However, manipulative experiments can be infeasible or inappropriate: Natural ecosystems may not offer enough replicates for comprehensive manipulative experiments and perturbations can be impractical at large scales and may have unanticipated negative consequences. Moreover, it may be difficult to faithfully recreate communities in lab experiments, for instance when bacterial species are unculturable. On the other hand, there exists an ever-growing plethora of population dynamic time series which are observational in nature (i.e. without intentional perturbations). The idea that one might wrest reasonably accurate causal predictions from these time series is tantalizing.

Causal inference can become more straightforward if one already knows, or is willing to assume, a model that captures key aspects of the underlying process. For example, the Lotka-Volterra model popular in mathematical ecology assumes that an individual receives additive fitness effects from interacting with each species in the community in a pairwise way and that pairwise interactions can be represented by a single equation with parameters reflecting signs and strengths of fitness effects. By fitting such a model to time series of species abundances and environmental factors, one can make specific predictions about, for instance, which species interact or how a community might respond to certain perturbations [1, 3, 2]. However, the Lotka-Volterra equations often fail to describe complex ecosystems and chemically-mediated interactions [61, 60, 6].

When prior knowledge is insufficient to support knowledge-based modeling, how might we formulate causal hypotheses? There is a large and rapidly growing literature of advanced methods that attempt to infer causal statements from time series data, either explicitly [80, 78, 77, 76, 84, 82] or implicitly by predicting the effects of new perturbations [81, 83]. Here we focus on three approaches that are commonly used in ecology research: pairwise correlation, Granger causality, and state space reconstruction. Each of these approaches is "*model-free*" in that it does not use a system-specific model. For each, we ask (1) what information does the method tell us, (2) what causal statement might that information imply, and (3) when might it lead us astray?

We found that answering these seemingly basic questions was at first surprisingly challenging for several reasons. First, modern causal inference approaches have intellectual roots in several communities including philosophy, statistics, econometrics, and chaos theory, which sometimes use different words for the same idea, and the same word for different ideas. The word causality itself is a prime example: Many philosophers (and scientists) would say that X causes Y if an intervention upon X would result in a change in Y [35, 19]. Conversely, Granger's original works instead defined causality to be about how important the history of X is in predicting Y [20, 30], and in the nonlinear dynamics field, the term is sometimes used to mean that the trajectories of X and Y have certain shared geometric or topological properties [17]. Such language, while unproblematic when confined to a single community, can nevertheless obscure important differences between methods from different communities. A second challenge is that in methodological articles, key assumptions are sometimes hidden in algorithmic details, or simply not mentioned. Finally, some methods deal with nuanced or advanced mathematical ideas that can be difficult even for those with quantitative training. Given these challenges, it is no surprise that research using causal inference from observational time series has been highly controversial, with an abundance of "letters to the editor", often followed by impassioned dialogue [23, 28, 29, 27].

We strive to balance precision and readability. To accomplish this, we devised new ways to visualize key concepts. We also provide refreshers of basic concepts and extended discussions of nuanced mathematical notions in the Appendices. Lastly, we compare all methods to a common definition of causality that is useful to experimental scientists. Our goals are to inform biologists who analyze temporal data and to inspire thoughtful methodological development.

Causality

Since we have seen that the term "causality" can take on different meanings in different communities, we define it now to avoid confusion. In this article, whenever we use the term "causality", we mean the definition that is common in statistics and intuitive to scientists: X has a causal effect on Y ("X causes Y" or "X is a causer of Y") if some externally applied perturbation of X can result in a perturbation in Y (Fig 1A). We say that X and Y are *causally related* if X causes Y, Y causes X, or some other variable causes both. Otherwise, X and Y are *causally unrelated*. Additionally, one can talk about direct versus indirect causality (Fig 1B; see legend for definitions). Systems for which one can correctly reveal all direct causal relationships between all pairs of variables are called causally identifiable and various theoretical identifiability conditions are known [11, 85]. Yet practical time series obtained from natural systems often do not contain enough information for easy causal identifiability [78]. Indeed, as we will see, methods that try to go beyond mere

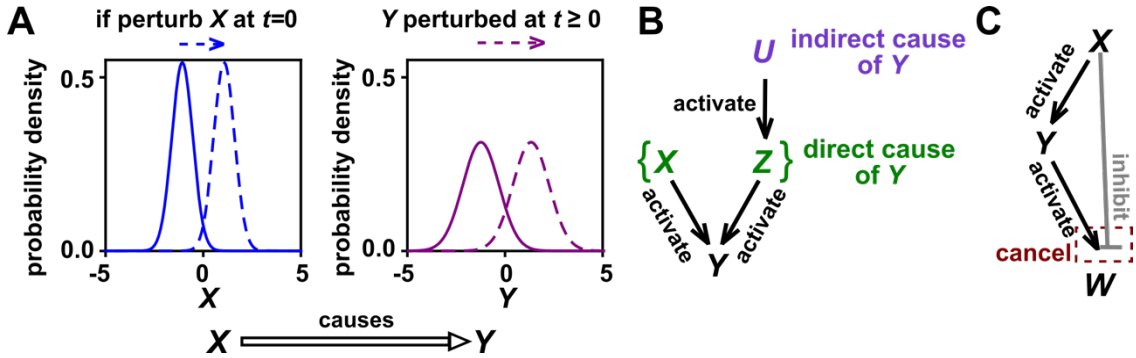

Figure 1: Causality. (**A**) If a perturbation in X causes future (or current) Y to change, then X causes Y. This definition does not require that *any* perturbation in X will perturb Y. For example, if the effect of X on Y has saturated, then a further increase in X will not affect Y. To embody probabilistic thinking (e.g. drunk driving increases the chance of accidents) [19], X and Y are depicted as histograms. Perturbing X can perturb the current value of Y if, for example, X and Y describe the same variable or if X and Y are linked by conservation law (e.g. conservation of energy). (**B**) By holding some variables constant, we can tease apart direct and indirect causality. The direct cause of Y is the minimal set of variables such that once the entire set is fixed, no other variables can cause Y. Here, three players X, Z, and U activate Y. The set $\{X, Z\}$ constitutes the direct cause of Y (or Y's "Markovian parents" in statistical jargon [34, 19]), since if we fix both X and Z, then Y becomes independent of U. Both X and Z are called direct causers of Y. If a cause is not direct, we say that it is indirect. Note that direct and indirect causes should be considered within the scope of included variables. For example, suppose that yeast releases acetate, and acetate inhibits the growth of bacteria. If acetate is not in our scope, then yeast is a direct causer of bacteria (e.g. increasing yeast density will reduce bacteria growth). Conversely, if acetate is included in our scope, then acetate (but not yeast) is a direct causer of bacteria since fixing acetate concentration would fix bacterial growth regardless of the presence or absence of yeast. (**C**) Counterintuitively, direct causality does not always imply causality. Here, $\{X, Y\}$ is the direct cause of W. While X is a direct causer of W, X does not cause W due to cancellation through Y. Also note that causality is not always transitive: X causes Y and Y causes W, but X does not cause W In this article, causality is represented by a hollow arrow, while activation and inhibition are represented by filled arrow and blunt-head arrow, respectively.

causal relation by inferring the direction of causality often require conditions that can be difficult to verify.

Dependence, correlation, and causality

The adage "correlation is not causality" is well-known to the point of being cliché. Indeed, there is no shortage of works pointing out how misplaced correlative analyses can lead investigators astray [5, 13, 12, 33]. Yet, to dismiss correlative evidence altogether seems too extreme. Applied scientists often need some way to make use of correlative evidence, without being reckless. To do so, it helps to distinguish between the terms "correlation" and "dependence". When applied to time series, the term "correlation" is often used procedurally [75, 13]. In this sense, a correlation is any procedure that takes two time series and generates a statistic to quantify how similar their dynamics are. Examples include the Pearson correlation formula and local similarity analysis [38]. On its own, correlation does not tell us much about causation (Fig 2). Conversely, statistical dependence is a rigorous property with close connections to causality.

In statistics, dependence is most easily described for two binary events. For instance, if the incidence of vision loss is higher among diabetics than among the general population, then vision loss and diabetes are statistically dependent. In general, events A and B are dependent if across many independent trials (e.g. patients), the probability that A and B both occur (e.g. both vision loss and diabetes) is different from the background probability that A occurs (e.g. background incidence of vision loss).

It is widely accepted that *if two variables are dependent, then they are causally related* [49, 11]. The concept of dependence is readily generalized from pairs of binary events to pairs of numerical variables, and then again to pairs of vectors, such as time series (Appendix 1.1). Importantly, time series are often autocorrelated (e.g. what occurs today influences what occurs tomorrow), so that a single pair of time series contains information from only a single trial. Thus, attempting to discern whether two time series are statistically dependent is like attempting to divine whether diabetes and vision loss are dependent with only a single patient (i.e. we have an "n-of-one problem").

In practice, this n-of-one problem is often addressed by a technique called surrogate data testing. Specifically, one begins with two time series X and Y and computes some measure of correlation between them. Next, one uses a computer to simulate how independent replicates of Y might have looked. These simulated Y time series are called "surrogate" Ys. Finally, one computes the correlation between X and each surrogate. The probability of observing the original correlation (or a stronger one) under the null hypothesis that X and Y are independent (i.e. the p-value) is then determined by counting how many of the surrogate Ys produce a stronger correlation than the real Y. For example, if we produced 19 surrogates and found the real correlation to be stronger than all 19 surrogate correlations, then we would write down a p-value of 0.05. Ideally, if two time series are not causally related, then we should register a a p-value of 0.05 (or less) in only 5% of cases.

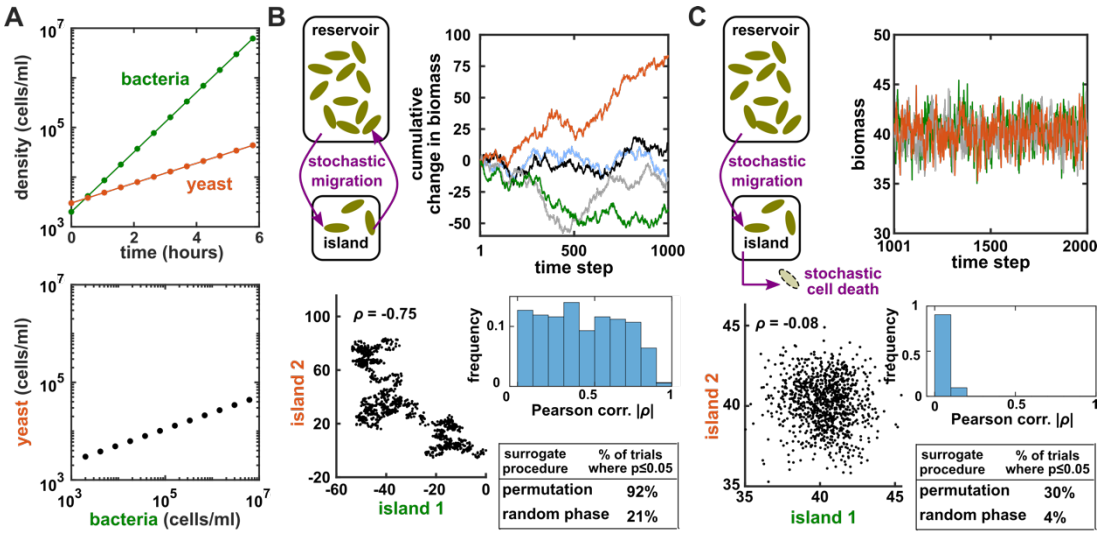

Figure 2: Two independent temporal processes can appear significantly correlated. (**A**) Densities of yeast and bacteria cultures growing exponentially in two different tubes are correlated. (**B**, **C**) Time series data of two independent island populations can appear casually related. Island population is controlled by migration (B) or migration and death (C). In both cases, we simulated 1000 pairs of independent islands, and correlated the biomass of one island with that of the other island. Example time series (upper right), a correlation plot between two islands (lower left), and the histogram of Pearson correlation strength (blue shading) are shown. We tabulated the fraction of simulations in which the correlation was deemed significant ($p \leq 0.05$; suggestive of a causal relation) by surrogate data tests using either permutation or phase randomization (see main text). (**B**) Consider an island population whose size is driven by individuals stochastically migrating to and from a "reservoir" (a random walk). At each time step, the net change in island biomass is drawn from a standard normal distribution (mean = 0; standard deviation = 1 biomass unit). (**C**) Consider a population which receives cells through migration and loses cells to death and whose size has reached an equilibrium. Specifically, an island starts with an initial biomass of 40, and at each time step 25% of the biomass is lost to death, and the biomass gained through migration is randomly chosen from a normal distribution with a mean of 10 biomass units and standard deviation of 1 unit. We correlated the final 1000 steps of one island with the final 1000 steps of the other island. The strength of correlation is weaker in this case compared to (B), yet still often causal according to the permutation test.

Several procedures are commonly used for generating surrogate data, each corresponding to an assumption about the Y time series. One popular procedure for generating surrogate Y time series is to shuffle the values of Y in time [38, 62, 63, 64]. This procedure, often called permutation, assumes that all time points of Y are independent of each other. A rejection of this null hypothesis (i.e. a low p-value) often provides little evidence of a causal link because rarely is this assumption met. For example, for independent time series in Fig. 1B-C, this test returns $p < 0.05$ at rates of 30~92%, much higher than 5%. Nevertheless, the permutation test sneaks into many experimental research articles, perhaps because it is the default option in some popular software packages. Another procedure for generating surrogate Y data has been called phase randomization. It first uses the Fourier transform to represent a time series as a sum of sine waves, then randomly shifts each of the component sine waves in time, and finally sums the phase-shifted components [26, 66] (Fig 12). This

procedure assumes that Y is Gaussian, linear, and stationary [66, 65], where "Gaussian" roughly means that Y's values follow a Gaussian distribution, "stationary" roughly means that this distribution does not change over time, and "linear" means that future Y values depend linearly on past Y values and past random events ("process noise"). Indeed, this test performed well (assigning dependence to two independent time series with a probability of 4%) when Y satisfied the three requirements (Fig 2C), and relatively poorly otherwise (Fig 2B, $21\% > 5\%$). Other surrogate data procedures include time shifting [66] and the twin method [70].

In sum, the mere observation of an apparent correlation between two time series is usually not compelling evidence for a causal relation. Surrogate data allow a researcher to test for dependence under some assumption about the data-generating process. It is not always feasible to verify these assumptions in a data-driven way. Nevertheless, an appropriately chosen surrogate data null model can provide some suggestive evidence of dependence, and thus evidence of a causal relation. Tests for dependence indicate only the presence and not the direction of causation. As we will see below, Granger causality and convergent cross mapping attempt to go further by inferring causal direction.

Granger causality: intuition, pitfalls, and implementation

In simple language, X Granger-causes Y if a collection of time series containing all historical measurements predicts Y's future behavior better than a similar collection that excludes the history of X. An immediate and important consequence of this definition is that Granger causality excludes indirect causes. This is illustrated in Fig 3A, where X causes Z, through a mechanism mediated entirely by a third variable Y. Here, although X causes Z, X does not Granger-cause Z. This is because once the history of Y is known, the history of X does not improve a prediction of future Z behavior. Thus, Granger causality is designed to reveal direct causal relationships. In practice, whether a causal relationship is direct or indirect depends on which variables are observed. If Y were not observed in (Fig 3A), then X would "directly" cause (and Granger-cause) Z.

How might Granger causality fail as an indicator of direct causality? Fig 3B lists four conceptually or practically important ways. First, if a system has deterministic dynamics (see Appendix 1.6 for a discussion of deterministic versus stochastic systems), then Granger causality may fail to detect causal relations (Fig 3Bi). We do not expect population dynamics to be completely deterministic, so the real lesson here is that when dynamics have a low degree of randomness (i.e. are almost deterministic), Granger causality may suffer low sensitivity [11] and other methods have been developed to address exactly this situation [5]. Second, Granger causality may erroneously assign a direct causal relation between a pair of variables that have an unobserved common cause (Fig 3Bii). Third, when sampling frequency is low, the causal signal may

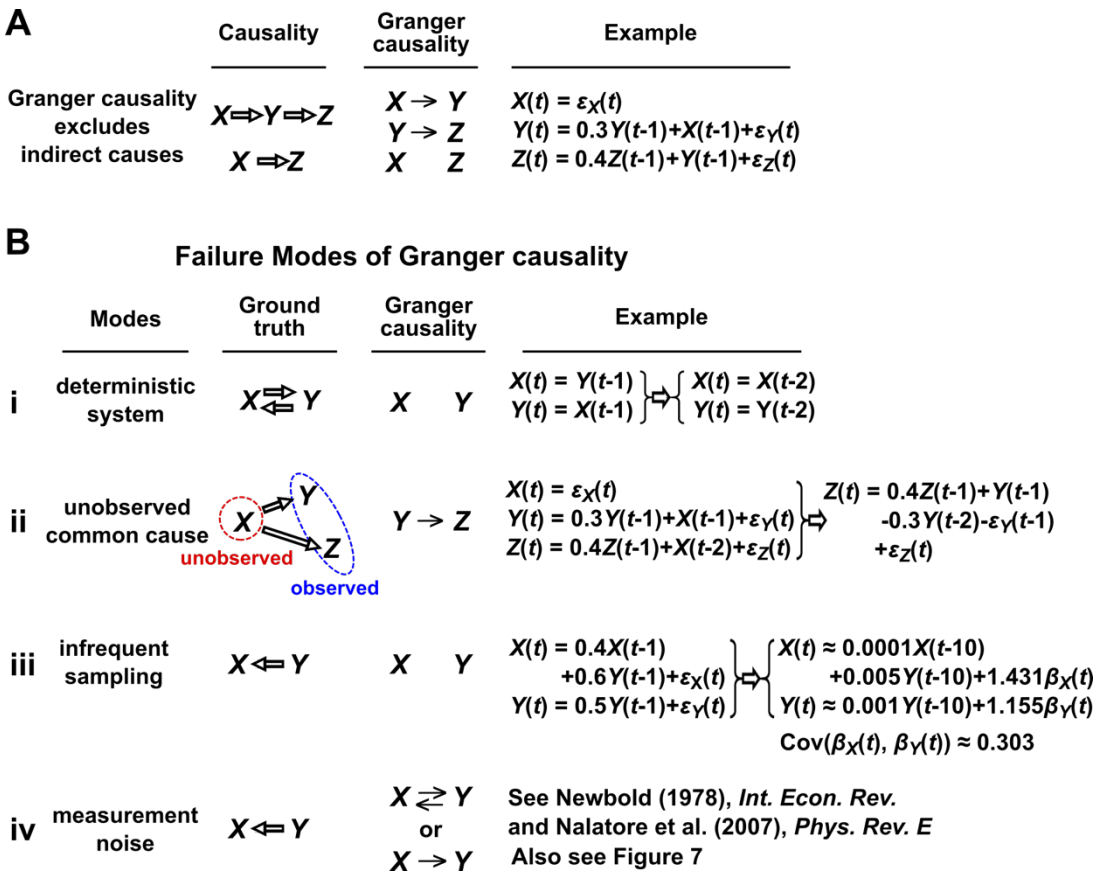

Figure 3: Causality versus Granger causality. (**A**) Granger causality is not designed to uncover indirect causes. Although X causes Z, X does not Granger-cause Z because with the history of Y available, the history of X no longer adds value for predicting Z. In this example, we can also see that Granger causality is not transitive: X Granger-causes Y and Y Granger-causes Z, but X does not Granger-cause Z. (**B**) Failure modes of Granger causality when inferring direct causality. ϵ_X, ϵ_Y, ϵ_Z, β_X, and β_Y represent process noise and are normal random variables with mean of 0 and variance of 1. ϵ_X, ϵ_Y, and ϵ_Z are uncorrelated with one another. Coefficients are chosen to ensure stationarity, except in (**ii**). (**i**) False negative due to lack of stochasticity. X and Y mutually and deterministically cause one another through a copy operation [11]: $X(t)$ copies $Y(t-1)$ and vice versa. In this case, since $X(t-2)$ already contains sufficient information to know $X(t)$ exactly, the history of Y cannot improve prediction of X, and so Y does not Granger-cause X. By symmetry, X does not Granger-cause Y. (**ii**) False positive due to unobserved common cause. X causes Y with a delay of 1, and causes Z with a delay of 2. We only observe Y and Z. Since Y receives the same "information" before Z, the history of Y helps to predict Z, and thus Y Granger-causes Z, resulting in a false positive. (**iii**) Infrequent sampling induces false negatives by reducing the correlation between current value of causee and previous value(s) of causer. The quantitative relationship is derived in Appendix 1.8. Note that in the equation of X, the ratio of signal (the coefficient of causer Y) to noise (the coefficient of process noise β or ε) is much smaller when sampling is every ten time steps (right) compared to every one time step (left). (**iv**) Measurement noise can lead Granger causality to return both false positives and false negatives.

become too low for Granger causality to detect (Fig 3Biii). Lastly, Granger causality can run into trouble when measurements are noisy, for instance due to instrument uncertainty (Fig 3Biv). In the presence of measurement noise, Granger causality can assign false interactions and also fail to detect true causality [51, 52].

Granger causality assumes many related but nonequivalent quantitative incarnations in the literature, including several that were proposed by Granger himself as he repeatedly grappled with the idea [20, 30]. In Box 1, we write two definitions: one based on a linear regression which we call "linear Granger causality" [4, 9, 40, 33] and a second, more general, definition which we call "general Granger causality" [43, 44, 41, 42, 39]. Linear Granger causality is a special case of general Granger causality.

Box 1: Granger causality

1. Linear Granger causality: Suppose one measures several variables over time index t, represented as X_t, Y_t, and perhaps others (e.g. Z_t). Under linear Granger causality, X Granger-causes Y if including the history of X in a linear autoregressive model (Eq 1) allows for a better prediction of future Y than not including the history of X (i.e. setting all α_k coefficients to zero). By "linear autoregressive model", we mean that the future value of variable Y is modeled as a linear combination of historical values of X and Y and all other included variables "$\cdots$":

$$Y_{t+1} = c + \sum_{k=0}^{n} (\alpha_k X_{t-k} + \beta_k Y_{t-k} + \cdots) + \varepsilon_t \tag{1}$$

Here, c is a constant, $k = 0, 1, \cdots, n$ are the time lags which index history, coefficients such as α_k and β_k represent the strength of contributions from the respective terms, and ε_t represents independent and identically-distributed (IID, 1.2) process noise (Fig 7A). To test whether X Granger-causes Y, we fit our data to the above equation (without the process noise term ε_t which now shows up in residuals), and test whether the α_k coefficients differ significantly from 0.

2. General Granger causality [30]: Let X_t Y_t, and Z_t be series of random variables indexed by time t. X Granger-causes Y with respect to the information set $\{X_t, Y_t, Z_t\}$ if:

$$f(Y_t|\{X_k, Y_k, Z_k \text{ for all } k < t\}) \neq f(Y_t|\{Y_k, Z_k \text{ for all } k < t\}) \tag{2}$$

where $f(Y_t|\mathcal{S})$ is the probability distribution of Y_t conditional on $\mathcal{S}$. Note that Z_k in Eq. 2 may include multiple variables and thus plays the same role as "$\cdots$" in Eq. 1.

Linear Granger causality, popular in microbiome studies [4, 9, 40, 33], uses standard parametric hypothesis tests to ask whether any coefficients are nonzero: if any of the α_k terms in Eq.1 is nonzero, then X linear

Granger-causes Y. Parametric tests are computationally inexpensive and available in multiple free and well-documented software packages [48, 31]. These tests assume that time series are "covariance-stationary", which means that certain statistical properties of the time series are time-independent [31] (see Appendix 1.5), and can fail when this assumption is violated [46, 45, 47]. Additionally, when linear Granger causality is applied to nonlinear systems, it can produce wrong causal inferences, even if none of the hazards from Fig 3B are at play [71]. One can assess whether the linear model (Eq. 1) is a reasonable approximation, for instance by checking whether the model residuals ε_t are uncorrelated across time [53] (as is assumed by Eq. 1).

Tests for general Granger causality often use surrogate data null models, with the same procedures and caveats discussed previously. In these tests, it is common to compute a statistic known as transfer entropy. In theory, X Granger-causes Y if and only if the true transfer entropy from X to Y is nonzero (Appendix 1.7). In practice, surrogate data tests find statistically significant evidence of X Granger-causing Y if the estimated transfer entropy from X to Y is greater than some proportion (e.g. 95%) of the transfer entropy values obtained when one of the time series is replaced with a surrogate (e.g. [39]). Tests for general Granger causality using transfer entropy are available for two-variable systems (e.g. the package RTransferEntropy [37]) and for multivariable systems (IDTxl [36]).

State space reconstruction (SSR): intuition, pitfalls, and implementations

The term "state space reconstruction" (SSR) refers to a broad swath of techniques for prediction, inference, and estimation in time series analysis [74, 72, 73, 5, 18]. In this article, when we use the term SSR, we refer only to SSR methods for causality detection. SSR enjoys widespread use in empirical ecology [54, 55, 69, 68, 67] and is often suggested as an alternative to model-based inference for microbial ecosystems [8, 7, 32, 58, 40, 57]. SSR methods are intended to complement Granger causality: Whereas Granger causality has trouble with deterministic dynamics (Fig 3B), the SSR approach is explicitly designed for systems that are primarily deterministic [5]. Since SSR is less intuitive than correlation or Granger causality, we introduce it with an example rather than a definition.

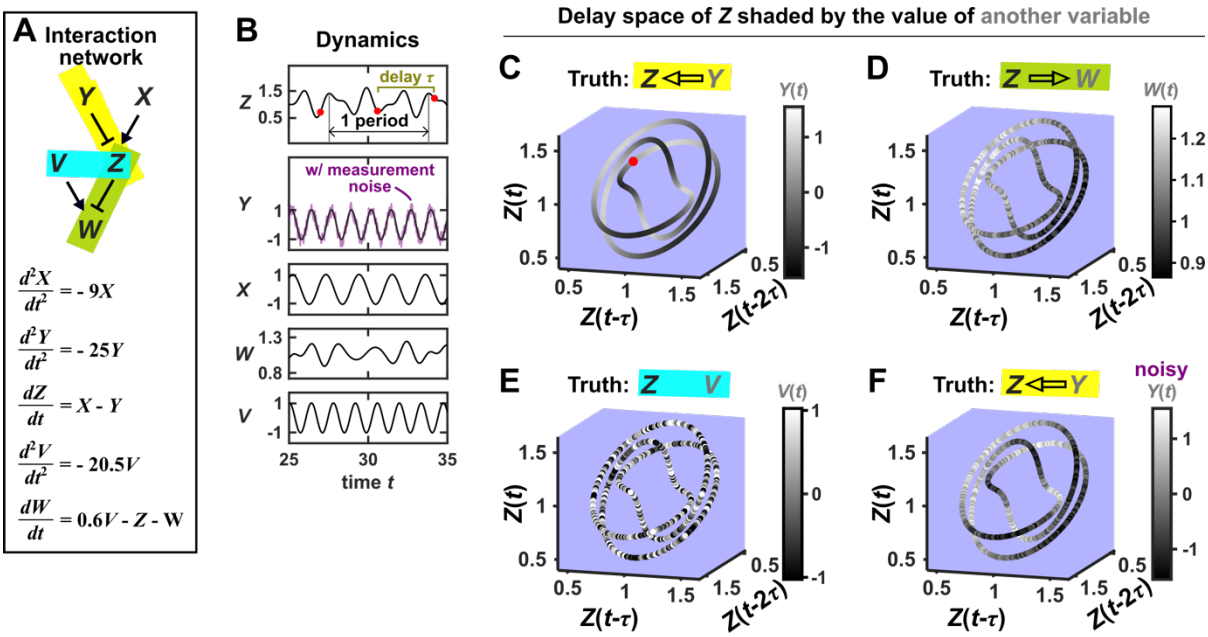

Figure 4: SSR causal inference looks for a continuous map from the delay space of a causee to the causer, and becomes more difficult in the presence of noise. (**A**) A 5-variable toy system. Filled arrow and blunt head arrow represent activation and inhibition, respectively. (**B**) Time series. The delay vector $[Z(t), Z(t-\tau), Z(t-2\tau)]$ can be represented as a single point in the 3-dimensional Z delay space (**C**, red dot). (**C**) We then shade each point in the Z delay space by its corresponding contemporaneous value of $Y(t)$ (without measurement noise). The shading is continuous (i.e. gradual transitions in shade), and note that Y causes Z. (**D**) When we repeat this procedure, but now shade the Z delay space by $W(t)$, the shading is bumpy, and note that W does not cause Z (even though Z causes W). (**E**) Shading the delay space of Z by the causally unrelated V is also bumpy. (**F**) Dynamics as in (**C**), but now with noisy measurements of Y (purple in **B**). The shading is no longer gradual. Thus with noisy data, causal inference is more difficult.

Visualizing SSR causal inference

Consider the deterministic dynamical system described by the equations in Fig 4A (time series in Fig 4B). In this system, Z is causally driven by Y (and X), but not by W or V. We can make a vector out of the current value $Z(t)$ and two past values $Z(t-\tau)$ and $Z(t-2\tau)$ (Fig 4B, red dots), where τ is the time delay and $[Z(t), Z(t-\tau), Z(t-2\tau)]$ is called a "delay vector". The delay vector can be represented as a single point in the 3-dimensional Z "delay space" (Fig 4C, red dot). We then shade each point of the Z delay space according to the contemporaneous value of its causer $Y(t)$. Since in this example each point of the Z delay space corresponds to one and only one $Y(t)$ value, we call this a "delay map" from Z to Y. Notice that the $Y(t)$ gradient in this plot looks gradual in the sense that if two time points are nearby in the delay space of Z, then their corresponding $Y(t)$ shades are also similar. This property is called "continuity" (Fig 17). We provide additional details on continuity, and the more stringent criterion of smoothness, in Appendix 1.9. Overall, there is a continuous map from the Z delay space to Y, or more concisely, a "continuous delay

map" from Z to Y. A similar continuous delay map also exists from Z to causer X. On the other hand, if we shade the delay space of Z by W (which does not cause Z), we do not get a continuous delay map (Fig 4D). We also do not get a continuous delay map from Z to the causally unrelated variable V (Fig 4E).

At least in this example, there is a continuous delay map from a causee to a causer, but not the other way around, and also no continuous delay map between causally unrelated variables. If this behavior is not just a coincidence, but rather an instance of a broader principle, then perhaps continuous delay maps can be used to infer the presence and direction of causation. Is there in fact a broader principle?

In fact, there is a sort of broader principle, but it may not be fully satisfying for causal inference. The principle is a classic theorem due to Floris Takens [22]. Very roughly, Takens's theorem can be interpreted for our situation as follows: If X and Y are variables in a deterministic dynamical system, and if X influences Y, and if several other technical requirements are met, then we are almost guaranteed to find a continuous delay map from Y to X. We walk through these details with visual examples in Appendix 1.11. In principle, this phenomenon is mostly insensitive to the choice of delay τ and the length of delay vector (see Appendix 1.10 for practical considerations in delay vector construction). Takens's theorem and later results along the same lines [25] form the theoretical basis of SSR techniques. However, we must stress that Takens's theorem does *not* state that a continuous delay map implies a causal relationship.

In sum, SSR techniques attempt to detect a smooth delay map (or a related feature) between two variables and use this to make inferences about the presence and direction of causation [5, 50, 17]: A smooth delay map from Y to X is taken as an indication that X causes Y. However, Takens's theorem leaves open the possibility of a smooth delay map without the corresponding causation. Indeed, we now illustrate some scenarios in which the SSR approach can be misleading.

SSR failure modes

Fig 5 illustrates some failure modes where a causal relationship and a continuous delay map do not coincide. In the first failure mode which we call "nonreverting continuous dynamics" (top row of Fig 5; Appendix 1.12), a continuous map arises from the delay space of X to Z because a continuous map can be found from the delay space of X to time ("nonreverting") and from time to Z ("continuous"). This pathology leads to false causal inference and may explain apparently causal results in some early works that applied SSR methods to time series with a clear trend. We are not aware of statistical tests for this problem, but Clark et al. [21] recommend shading points in the delay space with their corresponding time to visually check for any time trend. In the second failure mode, called "strong forcing" or "synchrony" [5] (Fig 5, second row), one variable drives another variable very strongly so that the dynamics of the two variables are synchronized.

Consequently, although the true causal relationship is unidirectional, bidirectional causality is inferred. A countermeasure (the "prediction lag test"; Fig 6B right panel) has partially alleviated the problem associated with synchrony [14, 15], but we find that this countermeasure can be misled by various circumstances, including when a common cause is present (Appendix 1.13). In the third failure mode (Fig 5 third row; [15]), X and Z both oscillate and X's oscillation period is an integer multiple of Z's oscillation period. In this case, Z is inferred to cause X even though they are causally unrelated. In the fourth failure mode (Fig 5, bottom row), SSR gives a false negative error due to "pathological symmetry", although this may be rare in practice.

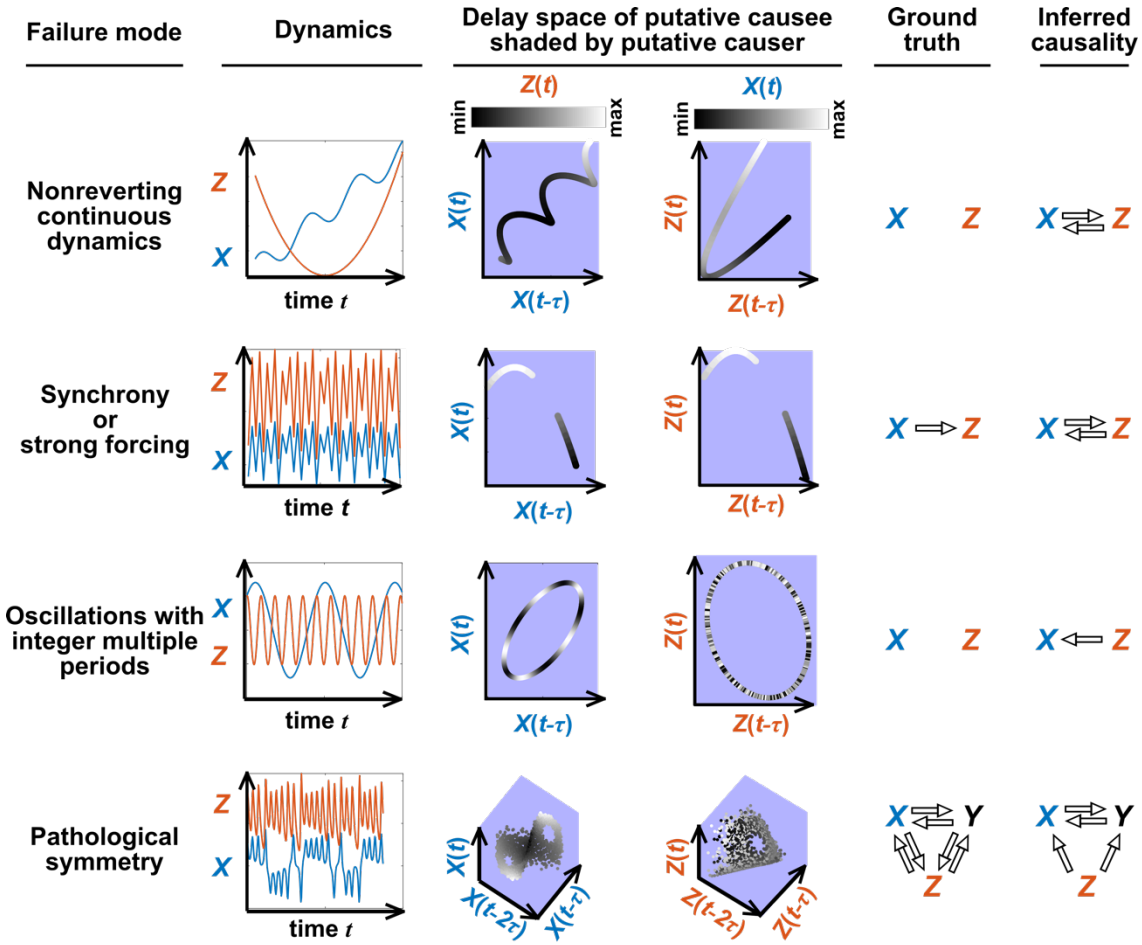

Figure 5: Failure modes associated with state space reconstruction. **Top row**: A smooth delay map can occur between two causally unrelated time series when we have nonreverting continuous dynamics. This example consists of two causally unrelated time series: a wavy linear increase and a parabolic trajectory. Despite being unrelated, we can find continuous delay maps between them. This is because there is (i) a continuous map from the delay vector $[X(t), X(t-\tau)]$ to t (X is "nonreverting"), and (ii) a continuous map from t to Z (Z is "continuous"), and thus there is a continuous delay map from X to Z ("nonreverting continuous dynamics"). Nonreverting continuous dynamics may lead one to infer causality where there is none. **Second row**: X drives Z in a manner such that their dynamics are "synchronzied", and consequently, we find a continuous delay map also from X to Z even though Z does not drive X. Note that the extent of synchronization is not always apparent from inspecting equations (e.g. Fig 12 of [16]) or dynamics (Fig 21). See rows 1 and 5 in Fig 21 for examples of false positives due to strong forcing. **Third row**: X oscillates at a period that is 5 times the oscillatory period of Y. There is a continuous delay map from X to Z even through X and Z are causally unrelated. Note that sometimes true causality also induces oscillations with integer multiples periods (e.g. in Fig 4, the period of Z is 3 times the period of X). **Bottom row**: In the classic chaotic Lorenz attractor, the Z variable is caused by X and Y, but we do not see a continuous map from the delay space of Z to either X (shown) or Y (not shown). This is because, as mentioned earlier, satisfying the conditions in Takens's theorem makes a continuous mapping likely but not guaranteed (Appendix 1.11). Here, Z is an example of this lack of guarantee [24] due to a symmetry in the system that causes the dynamics of X and Y to fold over on themselves when viewed from the Z axis ("Background definitions for causation in dynamic systems" in Supplementary Information of [5]).

Convergent cross mapping: Detecting SSR causal signals from real data

Despite the caveats above, one might still attempt to use SSR to hypothesize causal relations. In this case, one must test for continuity. However, while continuity is well defined in theory (Figure 17), it is not clear what continuity means when applied to real time series, which are noisy and sampled at discrete times. Thus, if we observe a delay map from a putative causee to a putative causer that is not continuous, we need to decide whether our delay map is discontinuous only because of noise and discrete sampling, or instead because the putative causal relationship is not there (Fig 4, compare D and F).

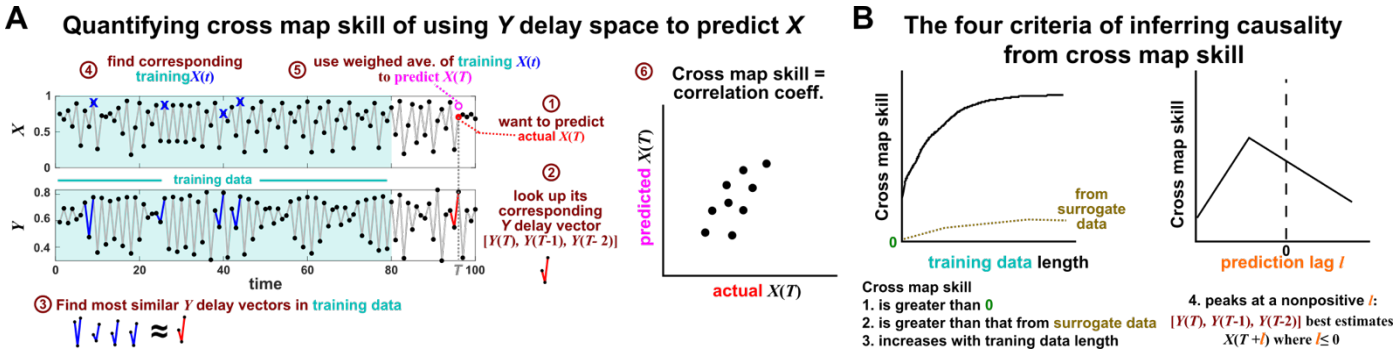

Figure 6: Illustration of the convergent cross mapping (CCM) procedure for testing whether X causes Y. (**A**) Quantifying cross map skill. Consider $X(T)$ denoted by the red dot ("actual $X(T)$" in **1**), which we want to predict from Y delay vectors. We first look up the contemporaneous Y delay vector ($[Y(T), Y(T-1), Y(T-2)]$) (**2**, red dynamics), and identify times within our training data when delay vectors of Y were the most similar to our red delay vector (**3**, blue segments). We then look up their contemporaneous values of X (**4**, blue crosses), and use their weighted average to predict $X(T)$ (**5**, open magenta circle). The weights are given as equations S2 and S3 in the supplement of [5]. We repeat this procedure for many choices of $X(T)$ and calculate the Pearson correlation between the actual values of $X(T)$ and our predicted values of $X(T)$ (**6**). This correlation is called the "cross map skill". While other measures of cross map skill, such as mean squared error, may also be used [5], our choice follows the convention of [5]. (**B**) Four criteria for inferring causality from the cross map skill. Data points in **A** are marked by dots and connecting lines serve as visual aids.

Several methods have been used to detect SSR causal signals by detecting approximate continuity [18] or related properties [5, 50, 17]. The most popular is convergent cross mapping (CCM) [5], which is based on a statistic called "cross map skill". Cross map skill quantifies how well a causer can be predicted from delay vectors of its causee (Fig 6A), conceptually similar to checking for gradual transitions when shading the causee delay space by causer values (Fig 4). Four criteria have been proposed to infer causality [5, 14, 15] (Fig 6B): First, the cross map skill must be positive. That is, when using the causee delay space to predict the causer, the Pearson correlation coefficient between true values and predictions must be positive. Second, the cross map skill must be significant according to some surrogate data test. Third, the cross map skill must increase with an increasing amount of training data. Lastly, cross map skill must be greater when predicting past values of causer than when predicting future values of causer (the previously mentioned prediction lag

test [14, 15], but see Appendix 1.13). In practice, many if not most CCM analyses use only a subset of these four criteria [5, 54, 55, 56]. Other approaches to detect various aspects of continuous delay maps have also been proposed [50, 18, 17, 82]. We do not know of a systematic comparison of these alternatives.

External drivers and noise jointly influence causal inference performance

In this section we examine how environmental drivers, process noise, and measurement noise can influence the performance of Granger causality and CCM, using computer simulations. We constructed a toy ecological system with a known causal structure, obtained its dynamics (with noise) through simulations, and applied a linear Granger causality test (using the MVGC package) and CCM (using the rEDM package) to test how well we can infer causal relationships.

We simulated a two-species community in which one species (S_1) causally influences the other species (S_2) but S_2 has no influence on S_1 (Fig 7B). Additionally, S_1 is causally influenced by an unobserved periodic external driver and S_2 either is (Fig 7D) or is not (Fig 7E) causally influenced by its own (also unobserved) periodic external driver. We added process noise to model the stochastic nature of natural ecosystems and added measurement noise to model measurement uncertainty. Process noise propagates to future time steps and can result from, for instance, stochastic migration and death (Fig 7A). In contrast, measurement noise does not propagate over time, and includes instrument noise as well as ecological processes during sampling (e.g. cell births within a stool sample as the stool passes through the intestine). Unlike in linear Granger causality, there is no default test procedure for CCM causality criteria [15, 40]. We therefore tested for CCM criteria using two different procedures (Fig 7 legend and Methods).

Granger causality and CCM can perform well when their respective requirements are met, but both are fairly sensitive to the levels of process and measurement noise (Fig 7D and E, correct inferences colored as green in pie charts). In reality, where a system lies in the spectrum of process versus measurement noise is often unknown, and we are not aware of any test that reliably distinguishes between process noise and measurement noise. Both Granger causality and CCM are also sensitive to details of the ecosystem (whether or not S_2 has its own external driver; compare Fig 7D with E). CCM is additionally sensitive to test procedure details (Fig 7 D and E, olive brackets).

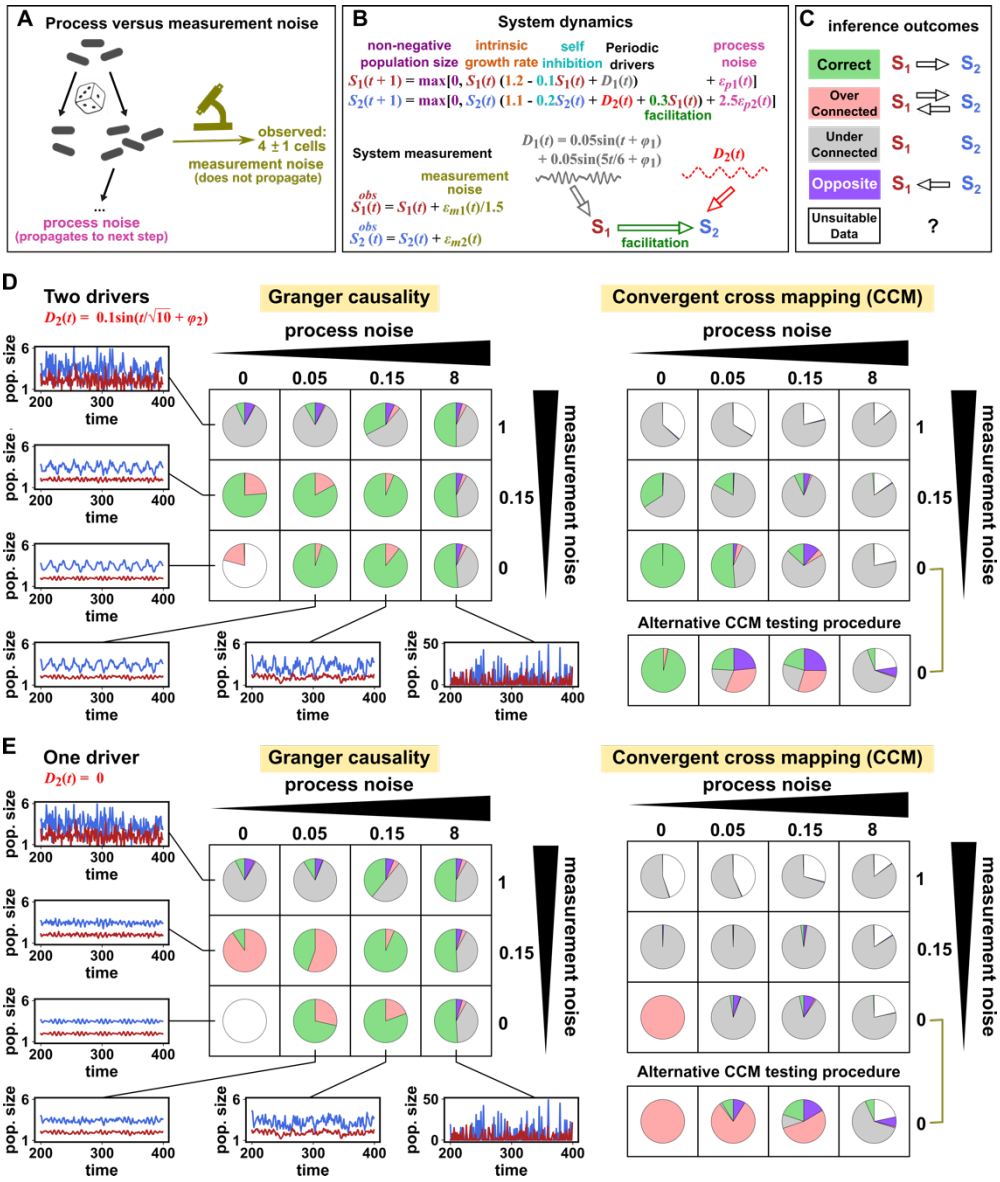

Figure 7: Performance of Granger causality and convergent cross mapping in a toy model with noise. (**A**) The effect of process noise but not measurement noise propagates to samples taken at subsequent time points. (**B**) We simulated a two-species community where S_1 promotes the growth of S_2. S_1 has an external driver, while S_2 may (**D**) or may not (**E**) have its own external driver. The process noise terms $\epsilon_{p1}(t)$ and $\epsilon_{p2}(t)$, as well as the measurement noise terms $\epsilon_{m1}(t)$ and $\epsilon_{m2}(t)$, are IID normal random variables with a mean of zero and a standard deviation whose value we vary. (**C**) Five possible outcomes of causal inference. (**D**, **E**) Community dynamics and causal inference outcomes. We varied the level (i.e. standard deviation) of process noise and measurement noise. Each pie chart shows the distribution of inference outcomes from 1000 independent replicates. Granger causality: In the deterministic setting (lower left corner), the MVGC tool correctly rejects the data as inappropriate. With modest (but not high) process noise and no measurement noise, Granger causality frequently infers the correct causal structure. High measurement noise prevents the detection of true causality, as expected. When process noise is low (i.e. ≤ 0.05), adding an intermediate level of measurement noise can increase the false positive rate. CCM: The performance of CCM is sensitive to test procedure (olive brackets) and ecological details (one versus two drivers), and quickly deteriorates with measurement or process noise. When S_2 does not have its own external driver (**E**), CCM cannot infer causality in the deterministic case due to the (visually apparent) synchrony. In both the main and alternative CCM procedures, criterion 1 (positive ρ) was checked directly and random phase surrogate data were used to test criterion 2 (significance of ρ). Because the dynamics are periodic, criterion 4 (prediction lag test) was not used (Fig 21). The two procedures differ only in how they test criterion 3 (ρ increases with more training data): the main procedure uses bootstrap testing following [15] while the alternative procedure uses a Kendall's τ as suggested by [59].

Handling high-dimensional data

To keep things simple, we have so far considered scenarios with only a few variables. However, some studies identify hundreds of species or more [79, 10]. Such high-dimensional data poses a serious challenge for causal inference. When confronted with such data, pairwise correlation or SSR techniques may return large "hairball-like" networks that can be difficult to interpret [18, 12], whereas Granger causality can become overwhelmed by many possible common causes and thus struggle to detect interactions [76, 77]. A variety of methods, whose details are beyond the scope of this article, have been developed to address this problem. A technique called regularization can be used with Granger causality to impose the assumption that only relatively few causal links exist [4]. Additionally, the PC algorithm and its several cousins handle high-dimensional data by constructing the causal network in smaller and more manageable steps, rather than all at once as in Granger causality [78, 76, 77].

Model-free causality tests are not assumption-free

Returning to the three questions we initially posed, we can summarize answers in Figure 8. Although the techniques explored in this article are model-free and do not depend on prior mechanistic knowledge, they are by no means free from assumptions. For example, if one wishes to draw causal inferences using random phase surrogate data, a linear, Gaussian, stationary process must be assumed [66, 70]. Such assumptions are not necessarily evils in and of themselves. Rather, the severe danger in replacing knowledge-based modeling with model-free inference is that we as scientists can be tempted or fooled into replacing explicitly stated assumptions with unstated and unscrutinized assumptions. Too often both methodological and applied works fall into this trap. Nevertheless, our view is that when assumptions and caveats are clearly articulated, model-free causal inference techniques have the potential to jump-start the discovery process when little mechanistic information is known.

	What can it tell us if the null hypothesis is rejected?	Implied causal statement	What are some possible failure modes?
Correlation (surrogate data)	The apparent correlation between X and Y cannot be explained by the surrogate null model.	X causes Y, Y causes X, or Z causes both	Assumptions of the surrogate null model may be violated.
Granger Causality	The history of X contains unique information that is useful for predicting the future of Y.	X directly causes Y	hidden common cause; deterministic system; infrequent sampling; measurement noise
Convergent Cross Mapping	X contains topological information about Y.	Y causes X	nonreverting continuous dynamics; strong forcing/synchrony; interger multiple periods; patholoigcal symmetry; measurement or process noise

Figure 8: A comparison of three statistical causal inference approaches.

Methods

Methodological details for Fig 2

For panel B, we simulated the random walk system

$$X(t+1) = X(t) + \epsilon(t)$$

where $\epsilon(t)$ terms were drawn independently from a normal distribution with mean of 0 and standard deviation of 1. We simulated this system from the initial condition of $X(1) = 0$ through 999 subsequent steps. For panel C, we simulated the autoregressive system

$$X(t+1) = 0.75X(t) + 10 + \epsilon(t)$$

where $\epsilon(t)$ terms were again drawn independently from a normal distribution with mean of 0 and standard deviation of 1. We simulated this system from the initial condition of $X(1) = 40$ for 1999 subsequent steps. We only used the final 1000 steps for computing the correlation between two time series.

To compute the significance of the Pearson correlation between two time series, we used surrogate data generated by either permutation or the random phase procedure. Permutation surrogate time series were generated by randomly shuffling data. Random phase surrogate time series were generated by Ebisuzaki's random phase method [26] as implemented in the rEDM (version 1.5) function make_surrogate_data. For a pair of time series $[X_1(1), X_1(2), ..., X_1(1000)], [X_2(1), X_2(2), ..., X_2(1000)]$, we first computed the Pearson

correlation $\hat{\rho}$ between the two time series. We then replaced the X_2 values with surrogate time series and recomputed the Pearson correlation as $\tilde{\rho}$. We computed this shuffled correlation 9999 times (permutation) or 499 times (random phase) to get a null distribution $[\tilde{\rho}_1, \tilde{\rho}_2, ..., \tilde{\rho}_n]$. Following [86], we computed the p value as

$$p = (N_{stronger} + 1)/(N_{surr} + 1) \tag{3}$$

where N_{surr} is the number of surrogates, $N_{stronger}$ is the number of surrogate correlations $\tilde{\rho}$ whose magnitude was greater than or equal to the magnitude of the original correlation $\hat{\rho}$, and the "+1" terms account for $\hat{\rho}$.

Methodological details for Fig 4

The system of equations was numerically integrated using the ode45 method in Matlab from $t = 0$ to $t = 200$ in time steps of 0.03, and plotted in the delay space Z with $\tau = 3.6$. The initial condition for all state variables (V, W, X, Y, Z, $\frac{dX}{dt}$, $\frac{dY}{dt}$, and $\frac{dV}{dt}$) was 1. For panel F, measurement noise was added to $Y(t)$. Specifically, noisy data were generated as:

$$Y^{obs}(t) \sim \text{Unif}\left(Y(t) - 3^{1/2}\left(0.15\Delta_Y\right), Y(t) + 3^{1/2}\left(0.15\Delta_Y\right)\right)$$

where $\text{Unif}(a, b)$ is a uniform random variable bounded by a and b, and Δ_Y is the difference between the maximum and minimum values of $Y(t)$ between $t = 0$ and $t = 200$. These noise parameters are chosen so that $Y^{obs}(t)$ is centered at $Y(t)$ and has a standard deviation of $0.15\Delta_Y$.

Methodological details for Fig 5

The dynamics in the top row of Fig 5 were generated from the equations:

$$X(t) = \sin(t) + 0.5t$$

$$Z(t) = 0.1(t - 10)^2$$

This continuous-time system was discretized from $t = 1$ to $t = 20$ on an evenly spaced grid of 400 data points for visualizing delay spaces where the time delay is 50 time points (i.e. $\tau = 50(20 - 1)/(400 - 1)$).

The dynamics in the second row of Fig 5 were generated from the equations:

$$X(t+1) = X(t)(3.61 - 3.61X(t))$$
$$Z(t+1) = Z(t)(3.61 - 3.61X(t))$$

with initial conditions of $X(1) = 0.4$ and $Z(1) = 0.7$. For this system, $\tau = 1$ and $t = 1, 2, ..., 2000$ were used to make the plots of delay spaces.

The dynamics in the third row of Fig 5 were generated from the equations:

$$\frac{dX^2}{dt} = -X(t)$$
$$\frac{dZ^2}{dt} = -25Z(t)$$

with initial conditions of $X(1) = X'(1) = Z(1) = Z'(1) = 1$. For this system, $\tau = 0.9$ was used for delay spaces. This continuous-time system was numerically integrated using the ode45 method in Matlab from $t = 0$ to $t = 13.998$ on a grid of 4667 evenly-spaced time points for plotting dynamics, and time points $t = 0.003$ through $t = 7.698$ were used for visualizing delay spaces.

The dynamics in the bottom row of Fig 5 were generated from the classic Lorenz attractor equations:

$$\frac{dX}{dt} = -10X(t) + 10Y(t)$$
$$\frac{dY}{dt} = 28X(t) - Y(t) - X(t)Z(t)$$
$$\frac{dZ}{dt} = -\frac{8}{3}Z(t) + X(t)Y(t)$$

with initial conditions of $X(0) = Y(0) = Z(0) = 1$. A delay of $\tau = 0.14$ was used to make delay spaces. This continuous-time system was numerically integrated using the ode45 method in Matlab from $t = 0$ to $t = 399.98$ on an evenly spaced grid of 5715 data points for visualizing delay spaces.

Methodological details for Fig 7

Ground truth model and data generation

We used the ground truth model:

$$S_1(t+1) = \max\left(0, S_1(t)\left(1.2 - 0.1S_1(t) + D_1(t)\right) + \epsilon_{p1}(t)\right)$$

$$S_2(t+1) = \max\left(0, S_2(t)\left(1.1 - 0.2S_2(t) + D_2(t) + 0.3S_1(t)\right) + 2.5\epsilon_{p2}(t)\right)$$

$S_1(t)$ and $S_2(t)$ represent the population sizes of species 1 and 2 at time t. $D_1(t)$ and $D_2(t)$ are the values of periodic drivers at time t. Specifically, in both the two-driver and one-driver cases:

$$D_1(t) = 0.05\sin\left(t + \phi_1\right) + 0.05\sin\left(\frac{5t}{6} + \phi_1\right)$$

In the two-driver case:

$$D_2(t) = 0.1\sin\left(\frac{t}{\sqrt{10}} + \phi_2\right)$$

Conversely, in the one-driver case $D_2(t) = 0$. The process noise terms $\epsilon_{p1}(t)$ and $\epsilon_{p2}(t)$ are both IID normal random variables with mean of 0 and with shared standard deviation σ_p. Specifically, for any pair of times $t_1 \neq t_2$, $\epsilon_{p1}(t_1)$ and $\epsilon_{p1}(t_2)$ are independent, and similarly for ϵ_{p2}. Also, all values $\epsilon_{p1}(1), \epsilon_{p1}(2), \cdots$ are independent of all values $\epsilon_{p2}(1), \epsilon_{p2}(2), \cdots$. At the beginning of each replicate simulation, the phases ϕ_1 and ϕ_2 are independently assigned a random number from a uniform distribution between 0 and 2π, and do not change with time.

To generate data without measurement noise, we simulated this system for $t = 1, 2, ..., 400$ with the initial conditions $S_1(1) = 2$; $S_2(1) = 4.5$. We used the final 200 time points for inference to help ensure that the system had reached equilibrium behavior.

We also introduced additive measurement noise to simulate instrument uncertainty:

$$S_1^{obs}(t) = S_1(t) + \epsilon_{m1}(t)/1.5$$

$$S_2^{obs}(t) = S_2(t) + \epsilon_{m2}(t)$$

where S_1^{obs} and S_2^{obs} represent the observed values (i.e. noisy measurements) of S_1 and S_2. $\epsilon_{m1}(t)$ and $\epsilon_{m2}(t)$ are also IID normal random variables with mean of 0 and standard deviation σ_m. The tables in Fig 7D are generated by varying σ_p from 0 to 8 and varying σ_m from 0 to 1.

Causal inference using Granger causality and CCM

For each combination of σ_m and σ_p (the standard deviation of measurement noise and process noise, respectively), we generated 1000 time series for S_1 and S_2 as described above. For each replicate pair of time series, we used Granger causality and CCM to infer whether S_1 causes S_2 (it does) and whether S_2 causes S_1 (it does not).

Granger causality inference

We used the multivariate Granger causality Matlab package (MVGC, [31]). We used the following settings:

- regmode = 'OLS' (We fit the autoregressive model by the ordinary least squares method).
- icregmode = 'LWR' (We determined the information criterion using the LWR algorithm. This is the default setting).
- morder = 'AIC' (We used Akaike information criterion to determine the number of lags in the autoregressive model).
- momax = 50 (We used a maximum of 50 lags in the autoregressive model).
- tstat = '' (We used Granger's F-test for statistical significance. This is the default setting).

We inferred the presence of a causal link if the p-value was less than or equal to 0.05. We inferred no causal link otherwise. When σ_m and σ_p were both 0, the MVGC package (correctly) exited with an error on most trials. We reported this as "unsuitable data" in Fig 7D & E.

When σ_m and σ_p are both 0, the inferred spectral radius of the stochastic process is close to 1, and the MVGC routines can be prohibitively slow (i.e. when running 1000 trials, the program would hang at an early stage for hours). In this case, the authors note that switching from the package's default single-regression mode to an alternative dual-regression mode may improve runtime [31]. We thus switched to the dual-regression mode when the spectral radius was between 0.9999 and 1 (a spectral radius of 1 or more causes an error). This fix had no effect on benchmark results as long as at least one of σ_m and σ_p was not 0.

Convergent cross mapping

Convergent cross mapping looks for a delay map from X to Y. That is, CCM looks for a map from $[X(t), X(t-\tau), X(t-2\tau), ..., X(t-(E-1)\tau)]$ to $Y(t)$. Thus in order to apply CCM one needs to choose the delay τ and the vector length (dimension of the delay space) E. E and τ should ideally be "generic" in

the sense of Takens's theorem: we want to avoid line-crossing (such as the symbol "∞") in the delay space, because otherwise, Φ^{-1} in Fig 18 does not exist. There are different ways to do this, but no method is obviously the best ([15, 17]).

Following [15] and [5] we chose τ and E to maximize univariate one-step-ahead forecast of the putative causee X. That is, for $X(n)$, we try to predict $X(n+1)$ using the simplex projection method by finding delay vectors in the training data of X that are most similar to $[X(n), X(n-\tau), X(n-2\tau), ..., X(n-(E-1)\tau)]$, and take weighed average of their X values 1 step in the future (i.e. Fig 6A where $X = Y$ and the prediction lag is 1). If the delay space has a line crossing, then at the cross-point, a one-step-ahead forecast may have more than one possible outcome and thus perform poorly. In more detail, we made one-step-ahead forecasts within the time range 201-400 (we did not use time range 1-200 to avoid transient dynamics). As per the field standard, we used leave-one-out cross-validation to do simplex projection. That is, when making a forecast for a time t, we used all times within 201-400 other than t as training data (200 time points). We performed a grid search, varying τ from 1 to 6 and varying E from 1 to 6. We then used the combination of τ and E that maximized the forecast skill (the Pearson correlation between forecasts and true values) for subsequent CCM analysis. Additionally, following [5], if the optimal combination of τ and E failed to give a significantly positive forecast skill, we did not report CCM results for that trial and reported the trial as "unsuitable data". To test whether forecast skill is "significantly positive", we ask whether it is robust to small changes in the training dataset. To do so, we used a naive bootstrap approach to create different versions of training libraries composed of randomly chosen delay vectors (sampling with replacement: some vectors may not be sampled and others may be sampled more than once) from the original training data using the 'random_libs' setting in the rEDM (version 1.5) ccm method. The training library size (the number of delay vectors in the library) was chosen to be 200. We then calculated forecast skills with 300 such libraries and considered the forecast skill "significant" if at least least 95% gave a forecast skill greater than 0.

Having chosen τ and E, we used three CCM criteria to infer causality (criteria 1-3 in Fig 6). We did not use the fourth criterion (the prediction lag test) since its interpretation is unclear for periodic systems (Fig 21). For all three criteria, we used the same cross-validation setting that we used to choose τ and E. The first CCM criterion is that cross map skill is greater than 0. Thus, we computed cross map skill using the maximum possible number of distinct delay vectors $(200 - (E-1)\tau)$ and compared this value to 0.

The second CCM criterion is that the cross map skill from causee to causer with real data must be greater than the cross map skill when the putative causer is replaced with surrogate data. To test this criterion, we first computed cross map skill using the same training and testing time points as before to obtain a single cross map skill value. We then repeatedly (1000 times) computed cross map skill in the same way, but now with the putative causer time series replaced with random phase surrogate data. Random phase surrogate

data were generated by Ebisuzaki's method as implemented in the rEDM function make_surrogate_data. We then computed the p-value according to Eq. 3. A putative causal link would pass this criterion if the p-value was less than or equal to 0.05.

The third CCM criterion is that cross map skill increases with more training data. Following [15], we again used a naive bootstrap approach to test for this criterion. Specifically, we computed the cross map skill with a training library composed of randomly chosen delay vectors sampled with replacement from the original training data time points. We used either a large library with $200 - (E-1)\tau$ available training vectors as used previously, or a small library with 15 training vectors. For each of 1000 bootstrap trials, we compared the cross map skill from a randomly chosen small library to the cross map skill from a randomly chosen large library. We said that the cross map skill increased with training data if the cross map skill of the large library was greater than that of the small library in at least 95% of the 1000 bootstrap trials.

For "alternative" CCM testing, we only changed how the third CCM criterion (cross map skill increases with more training data) were tested. Here, instead of using the bootstrap test of [15], we tested the third CCM criterion using Kendall's τ test as suggested in [59]. To do this, we varied the library size from a minimum of 15 vectors to the the maximum library size $(200 - (E-1)\tau)$, in increments of 3 vectors. For each library size, we computed cross map skill using 50 libraries randomly sampled without replacement (e.g. the 50 libraries would be identical at the maximal library size). We then computed the median cross map skill for each library size. Finally we ran a 1-tailed Kendall's τ test for a positive association between library size and median cross map skill. We used the function stats.kendalltau from the Python package SciPy to compute a 2-tailed p-value, and then divided this p-value by 2 to get a 1-tailed p-value. We said that cross map skill increased with training data if the τ statistic was positive and the 1-tailed p-value was ≤ 0.05.

Methodological details for Fig. 11

The original subpopulation distributions are normal distributions with standard deviation of 10 and mean of 100 (male) or 130 (female). Each sampling plot shows 300 samples randomly drawn from the appropriate mixture distribution.

Methodological details for Fig 18

To generate data for panels C-J, the system of panel A was numerically integrated using the ode45 method in Matlab with a time step of 0.005 and with the initial condition that X, Y, Z, $\frac{dX}{dt}$, $\frac{dY}{dt}$ were all set to 1 at $t = 0$. Panels C , E, F, G, I, and J (non-inset) show data from a single period. For panel H the system was integrated for about 5 periods to more clearly visualize the lack of a continuous delay map. For the panel

J inset the system was integrated for about 12 periods to better see the separated legs of the curve upon zooming in. Panels C, D, F, H, and J were colored $\mathrm{mod}(t, T)$. That is, they were colored by the remainder of t (time) after dividing by T (here $T = 2\pi$). $\tau = 3.6$ was used for all delay spaces.Methodological details for Fig 19

All systems were discretized from $t = 1$ to $t = 20$ on an evenly spaced grid of 200 points for visualizing delay spaces.

The dynamics in the top row were generated from the equations:

$$X(t) = \sin(t) + 0.5t$$
$$Y(t) = \sin(1.3t)$$

A delay time of 12 time indices (i.e. $\tau = 12(20-1)/(200-1)$) was used for constructing delay spaces.

The dynamics in the second row were generated from the equations:

$$X(t) = \sin(1.3t)$$
$$Y(t) = Y(t-\delta)(3.77 - 3.77 * Y(t-\delta))$$

with $\delta = (20-1)/(200-1)$ and the initial condition $Y(1) = 0.3$. A delay time of 25 time indices (i.e. $\tau = 25(20-1)/(200-1)$) was used for constructing delay spaces.

In the third row the dynamics are identical to the first row, except that X and Y are switched, and $\tau = 25$ time indices was used for constructing delay spaces.

Methodological details for Fig 20

Top row: For this system, we used the initial conditions $W(0) = \dot{W}(0) = X(0) = Y(0) = \dot{Y}(0) = 1$. We numerically integrated this system using ode45 in Matlab with a time step of 0.03. We composed delay vectors of length $E = 3$ with a delay of $\tau = 3.6$. We visualized the delay space using data from $t = 0$ through $t = 29.97$ (time indices 1-1000). For CCM with temporally separate training and testing sets, we used data from $t = 0$ through $t = 14.97$ (time indices 1-500) for training data and data from $t = 15$ through $t = 29.97$ (time indices 501-1000) for testing. Specifically, in the rEDM (version 0.7.2) ccm method we set the lib argument to "(1, 500)" and set the pred argument to "(501, 1000)". For CCM with temporally interspersed training and testing sets, we set both the lib and pred arguments to "(1,1000)". This setting instructs rEDM

to use leave-one-out cross-validation.

Second row: For this system, we used the initial conditions $W(0) = \dot{W}(0) = X(0) = \dot{X}(0) = Y(0) = 1$. We numerically integrated this system using ode45 in Matlab with a time step of 0.03. We visualized the delay space using data from $t = 0$ through $t = 29.97$ (time indices 1-1000). We used the delay vector parameters ($E = 3, \tau = 3.6$). For CCM with temporally separate training and testing sets, we used data from $t = 0$ through $t = 14.97$ (time indices 1-500) for training data and data from $t = 15$ through $t = 29.97$ (time indices 501-1000) for testing. For CCM with temporally interspersed training and testing sets, we used cross-validation over the entire range $t = 0$ through $t = 29.97$.

Third row: For this system, we used the initial conditions $W(0) = 0, \dot{W}(0) = 1/1.5, X(0) = 1.3, Y(0) = 1.5$. We numerically integrated this system using ode45 in Matlab with a time step of 0.1. We visualized the delay space using data from $t = 0$ through $t = 40$ (time indices 1-401). We used the delay vector parameters ($E = 3, \tau = 3.0$). For CCM with temporally separate training and testing sets, we used data from $t = 0$ through $t = 19.9$ (time indices 1-200) for training data and data from $t = 20$ through $t = 39.9$ (time indices 201-400) for testing. For CCM with temporally interspersed training and testing sets, we used cross-validation over the entire range $t = 0$ through $t = 39.9$.

Bottom row: We discretized this system with a time step of 0.05. We visualized the delay space using data from $t = 0$ through $t = 20$ (time indices 1-401). We used the delay vector parameters ($E = 2, \tau = 2.5$). For CCM with temporally separate training and testing sets, we used data from $t = 0$ through $t = 9.95$ (time indices 1-200) for training data and data from $t = 10$ through $t = 19.95$ (time indices 201-400) for testing. For CCM with temporally interspersed training and testing sets, we used cross-validation over the entire range $t = 0$ through $t = 19.95$.

For convergent cross mapping, we used the same τ and E as for visualizing delay spaces (see above). "Training data size" on the horizontal axis is the number of delay vectors in the training library. Each dot in these CCM plots represents the average forecast skill over 300 randomly chosen libraries. Error bars represent the 95% confidence interval as calculated by the bias-corrected and accelerated bootstrap (1000 bootstraps) as implemented in Matlab's bootci function. Error bars are the same color as the dots and so are not visible when they fit inside the dots.

Methodological details for Fig 21

Top row: For this system, we used the initial conditions $X(1) = 0.2, Y(1) = 0.4$ and composed delay vectors of length $E = 2$ with a delay of $\tau = 2$. We visualized the delay space using data from time points 501-2000. We used points 801-1000 for training data and points 1001-2000 for testing cross map predictions.

Second row: For this system, we used the initial conditions $X(1) = 0.4, Y(1) = 0.2$ and the delay vector parameters $(E = 2, \tau = 1)$. We visualized the delay space using data from time points 501-2000. We used points 801-1000 for training data and points 1001-2000 for testing cross map predictions.

Third row: For this system, we used the initial conditions $X(1) = 0.2, Y(1) = 0.4$ and the delay vector parameters $(E = 3, \tau = 2)$. We visualized the delay space using data from time points 501-2000 (time points 1-6 $\times 10^5$ for the zoomed-in inset). We used points 801-1000 for training data and points 1001-2000 for testing cross map predictions.

Fourth row: For this system, we used the initial conditions $W(0) = Y(0) = 0$ and $X(0) = \dot{W}(0) = \dot{Y}(0) = 1$. We numerically integrated this system using ode45 in Matlab with a time step of 0.1. We visualized the delay space using data from $t = 50.1$ through $t = 200$ (time indices 501-2000). We used the delay vector parameters $(E = 3, \tau = 7.2)$. We used data from $t = 70.1$ through $t = 100$ (time indices 701-1000) for training data and data from $t = 100.1$ through $t = 200$ (time indices 1001-2000) for testing cross map predictions.

Fifth row: For this system, we used the initial conditions $X(1) = 0.2, Y(1) = 0$ and composed delay vectors of length $E = 2$ with a delay of $\tau = 2$. We visualized the delay space using data from time points 501-2000. We used points 801-1000 for training data and points 1001-2000 for testing cross map predictions.

Sixth row: For this system, the "initial" conditions specified the first 3 timepoints since we included a lag of 3. Thus, for $k = 1, 2, 3$, $W(k) = 0.2$, $X(k) = 0.4$, and $Y(k) = 0.3$. We composed delay vectors of length $E = 3$ with a delay of $\tau = 1$. We visualized the delay space using data from time points 501-2000. We used points 801-1000 for training data and points 1001-2000 for testing cross map predictions.

For convergent cross mapping, we used the same τ and E as for visualizing delay spaces. The training data size is the number of delay vectors in the training library. For the plots in the fourth column, we chose 300 random libraries of training delay vectors with variable training data size, and used the standard prediction lag of 0. Delay vectors were chosen without replacement. Note that at large training data size, some or all of the 300 random libraries can be identical. Each dot in these CCM plots represents the average forecast skill over all 300 randomly-chosen libraries. Error bars represent the 95% confidence interval as calculated by the bias-corrected and accelerated bootstrap (1000 bootstraps) as implemented in Matlab's bootci function. Error bars are the same color as the dots and so are not visible when they fit inside the dots.

In all rows, the cross map skill for the putative causer Y was greater than for at least 95% of random phase surrogate time series (purple dot). The 5% cutoff value was computed for the maximum library size (156 for row 4 and ~ 200 for all other rows) by running the CCM procedure after replacing the putative causer Y with 500 random phase surrogate time series generated using the rEDM function make_surrogate_data.

For the plots in the fifth column we used the full library contained within the training data window (156 delay vectors for row 4 and ~ 200 for all other rows) and varied the prediction lag. We did not use random libraries for these plots.

Methodological details for Fig 22

To generate randomized parameter sets, we randomly selected R_X, R_Y, A_{XY} and A_{YX} from uniform distributions. We also randomly selected the initial conditions $X(1)$ and $Y(1)$ from uniform distributions. To make systems in the "friendly" parameter regime, we drew R_X and R_Y independently from the range $3.7 - 3.9$, we drew A_{XY} and A_{YX} independently from the range $0.05 - 0.1$, and we drew $X(1)$ and $Y(1)$ independently from the range $0.01 - 0.99$. These are the same parameters used in the randomized numerical simulations of [14]. Next, to make systems in the "pathological" parameter regime, we drew R_X from the range $3.7 - 3.9$, we drew R_Y from the range $3.1 - 3.3$, we drew A_{XY} and A_{YX} independently from the range $0.15 - 0.2$, and we drew $X(1)$ and $Y(1)$ independently from the range $0.01 - 0.99$. For both parameter regimes we randomly chose 500 sets of parameters and ran the system for 3000 time points. Occasionally a randomly chosen system would leave the basin of attraction and reach large values, represented on the computer as positive or negative infinity, or "not a number". When this occurred, we discarded the data and resampled parameters.

To apply CCM on each system, we generated a training library of delay vectors of X by randomly selecting 200 vectors from among time points 100-2000. We then evaluated cross map skill from delay vectors of X to values of Y at points 2001-3000. Following [14], we used delay vectors of length $E = 2$ and a delay duration of $\tau = 1$. We evaluated cross map skill with a prediction horizon of -8 through 8.

Acknowledgments

Ideally, writers of this essay should be Renaissance scholars with mastery of divergent fields. However, human knowledge has expanded since the Renaissance and we are not modern da Vincis. We therefore sought feedback and advice from domain experts: Bree Cummins (Montana State University) and Tim Sauer (George Mason University) discussed topology with us; Sean Gibbons (Institute for Systems Biology), Nathan Kutz (University of Washington), and Chris Barnes and Bianca De Stavola (UCL) critiqued our manuscript. We thank David Fredricks and Sujatha Srinivasan (Fred Hutch) for discussions that inspired this effort, and members of the Shou group for helpful comments.

References

[1] R. R. Stein, V. Bucci, N. C. Toussaint, C. G. Buffie, G. Ratsch, E. G. Pamer, C. Sander, and J. B. Xavier, "Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota," *PLoS computational biology*, vol. 9, no. 12, p. e1003388, 2013.

[2] V. Bucci, B. Tzen, N. Li, M. Simmons, T. Tanoue, E. Bogart, L. Deng, V. Yeliseyev, M. L. Delaney, Q. Liu, B. Olle, R. R. Stein, K. Honda, L. Bry, and G. K. Gerber, "Mdsine: Microbial dynamical systems inference engine for microbiome time-series analyses," *Genome Biology*, vol. 17, p. 121, Jun 2016.

[3] C. K. Fisher and P. Mehta, "Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression," *PLOS ONE*, vol. 9, pp. 1–10, 07 2014.

[4] 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.

[5] 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.

[6] B. Momeni, L. Xie, and W. Shou, "Lotka-volterra pairwise modeling fails to capture diverse pairwise microbial interactions," *Elife*, vol. 6, p. e25051, 2017.

[7] K. Faust, L. Lahti, D. Gonze, W. M. De Vos, and J. Raes, "Metagenomics meets time series analysis: unraveling microbial community dynamics," *Current opinion in microbiology*, vol. 25, pp. 56–66, 2015.

[8] G. K. Gerber, "The dynamic microbiome," *FEBS letters*, vol. 588, no. 22, pp. 4131–4139, 2014.

[9] D. Ai, X. Li, G. Liu, X. Liang, and L. C. Xia, "Constructing the microbial association network from large-scale time series data using granger causality," *Genes*, vol. 10, no. 3, p. 216, 2019.

[10] L. A. David, A. C. Materna, J. Friedman, M. I. Campos-Baptista, M. C. Blackburn, A. Perrotta, S. E. Erdman, and E. J. Alm, "Host lifestyle affects human microbiota on daily timescales," *Genome biology*, vol. 15, no. 7, p. R89, 2014.

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

[12] A. Carr, C. Diener, N. S. Baliga, and S. M. Gibbons, "Use and abuse of correlation analyses in microbial ecology," *The ISME journal*, p. 1, 2019.

[13] 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.

[14] 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.

[15] 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.

[16] D. Mønster, R. Fusaroli, K. Tylén, A. Roepstorff, and J. F. Sherson, "Causal inference from noisy time-series data-testing the convergent cross-mapping algorithm in the presence of noise and external influence," *Future Generation Computer Systems*, vol. 73, pp. 52–62, 2017.

[17] 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.

[18] B. Cummins, T. Gedeon, and K. Spendlove, "On the efficacy of state space reconstruction methods in determining causality," *SIAM Journal on Applied Dynamical Systems*, vol. 14, no. 1, pp. 335–381, 2015.

[19] J. Pearl, *Causality*. Cambridge university press, 2000.

[20] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[21] A. T. Clark, H. Ye, F. Isbell, E. R. Deyle, J. Cowles, G. D. Tilman, and G. Sugihara, "Spatial convergent cross mapping to detect causal relationships from short time series," *Ecology*, vol. 96, no. 5, pp. 1174–1181, 2015.

[22] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical systems and turbulence, Warwick 1980*, pp. 366–381, Springer, 1981.

[23] M. Luo, H. Kantz, N.-C. Lau, W. Huang, and Y. Zhou, "Questionable dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature," *Proceedings of the National Academy of Sciences*, vol. 112, no. 34, pp. E4638–E4639, 2015.

[24] E. R. Deyle and G. Sugihara, "Generalized theorems for nonlinear state space reconstruction," *PLOS ONE*, vol. 6, pp. 1–8, 03 2011.

[25] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of statistical physics*, vol. 65, no. 3-4, pp. 579–616, 1991.

[26] 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.

[27] H. Alrasheed, R. Jin, and J. S. Weitz, "Caution in inferring viral strategies from abundance correlations in marine metagenomes," *Nature communications*, vol. 10, no. 1, p. 501, 2019.

[28] E. B. Baskerville and S. Cobey, "Does influenza drive absolute humidity?," *Proceedings of the National Academy of Sciences*, vol. 114, no. 12, pp. E2270–E2271, 2017.

[29] L. Barnett, A. B. Barrett, and A. K. Seth, "Misunderstandings regarding the application of granger causality in neuroscience," *Proceedings of the National Academy of Sciences*, vol. 115, no. 29, pp. E6676–E6677, 2018.

[30] C. W. Granger, "Testing for causality: a personal viewpoint," *Journal of Economic Dynamics and control*, vol. 2, pp. 329–352, 1980.

[31] L. Barnett and A. K. Seth, "The mvgc multivariate granger causality toolbox: a new approach to granger-causal inference," *Journal of neuroscience methods*, vol. 223, pp. 50–68, 2014.

[32] E. Margolis, A. Oot, and D. N. Fredricks, "Human Microbiome Dynamics: Causality Detection With Convergent Cross Mapping," *Open Forum Infectious Diseases*, vol. 3, 10 2016. 2224.

[33] K. Mainali, S. Bewick, B. Vecchio-Pagan, D. Karig, and W. F. Fagan, "Detecting interaction networks in the human microbiome with conditional granger causality," *PLoS computational biology*, vol. 15, no. 5, p. e1007037, 2019.

[34] D. M. Hausman and J. Woodward, "Independence, invariance and the causal markov condition," *The British journal for the philosophy of science*, vol. 50, no. 4, pp. 521–583, 1999.

[35] J. Woodward, "Causation and manipulability," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, winter 2016 ed., 2016.

[36] P. Wollstadt, J. Lizier, R. Vicente, C. Finn, M. Martinez-Zarzuela, P. Mediano, L. Novelli, and M. Wibral, "Idtxl: The information dynamics toolkit xl: a python package for the efficient analysis of multivariate information dynamics in networks," *Journal of Open Source Software*, vol. 4, no. 34, p. 1081, 2019.

[37] S. Behrendt, T. Dimpfl, F. J. Peter, and D. J. Zimmermann, "Rtransferentropy—quantifying information flow between different time series using effective transfer entropy," *SoftwareX*, vol. 10, p. 100265, 2019.

[38] 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.

[39] 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.

[40] F. Barraquand, C. Picoche, M. Detto, and F. Hartig, "Inferring species interactions using granger causality and convergent cross mapping," *arXiv preprint arXiv:1909.00731*, 2019.

[41] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, "Transfer entropy—a model-free measure of effective connectivity for the neurosciences," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 45–67, 2011.

[42] F. Roux, M. Wibral, W. Singer, J. Aru, and P. J. Uhlhaas, "The phase of thalamic alpha activity modulates cortical gamma-band activity: evidence from resting-state meg recordings," *Journal of Neuroscience*, vol. 33, no. 45, pp. 17827–17835, 2013.

[43] C. Diks and V. Panchenko, "A new statistic and practical guidelines for nonparametric granger causality testing," *Journal of Economic Dynamics and Control*, vol. 30, no. 9-10, pp. 1647–1669, 2006.

[44] S. D. Bekiros and C. G. Diks, "The nonlinear dynamic relationship of exchange rates: Parametric and nonparametric causality testing," *Journal of macroeconomics*, vol. 30, no. 4, pp. 1641–1650, 2008.

[45] L. E. Ohanian, "The spurious effects of unit roots on vector autoregressions: A monte carlo study," *Journal of Econometrics*, vol. 39, no. 3, pp. 251–266, 1988.

[46] H. Y. Toda and P. C. Phillips, "The spurious effect of unit roots on vector autoregressions: an analytical study," *Journal of Econometrics*, vol. 59, no. 3, pp. 229–255, 1993.

[47] Z. He and K. Maekawa, "On spurious granger causality," *Economics Letters*, vol. 73, no. 3, pp. 307–313, 2001.

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

[49] C. Hitchcock and M. Rédei, "Reichenbach's common cause principle," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, spring 2020 ed., 2020.

[50] H. Ma, K. Aihara, and L. Chen, "Detecting causality from nonlinear dynamics with short-term time series," *Scientific reports*, vol. 4, p. 7464, 2014.

[51] P. Newbold, "Feedback induced by measurement errors," *International Economic Review*, pp. 787–791, 1978.

[52] H. Nalatore, M. Ding, and G. Rangarajan, "Mitigating the effects of measurement noise on granger causality," *Physical Review E*, vol. 75, no. 3, p. 031123, 2007.

[53] E. L. Feige and D. K. Pearce, "The casual causal relationship between money and income: Some caveats for time series analysis," *The Review of Economics and Statistics*, pp. 521–533, 1979.

[54] 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.

[55] K. L. Cramer, A. O'Dea, T. R. Clark, J.-x. Zhao, and R. D. Norris, "Prehistorical and historical declines in caribbean coral reef accretion rates driven by loss of parrotfish," *Nature communications*, vol. 8, no. 1, pp. 1–8, 2017.

[56] Y. Wang, J. Yang, Y. Chen, P. De Maeyer, Z. Li, and W. Duan, "Detecting the causal effect of soil moisture on precipitation using convergent cross mapping," *Scientific reports*, vol. 8, no. 1, pp. 1–8, 2018.

[57] A. R. Coenen, S. K. Hu, E. Luo, D. Muratore, and J. S. Weitz, "A primer for microbiome time-series analysis," *Frontiers in Genetics*, vol. 11, 2020.

[58] J. S. Biteen, P. C. Blainey, Z. G. Cardon, M. Chun, G. M. Church, P. C. Dorrestein, S. E. Fraser, J. A. Gilbert, J. K. Jansson, R. Knight, *et al.*, "Tools for the microbiome: nano and beyond," 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] J. T. Wootton, "Indirect effects in complex ecosystems: recent progress and future challenges," *Journal of Sea Research*, vol. 48, no. 2, pp. 157–172, 2002.

[61] S. H. Levine, "Competitive interactions in ecosystems," *The American Naturalist*, vol. 110, no. 976, pp. 903–910, 1976.

[62] 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.

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

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] S. Li, Y. Xiao, D. Zhou, and D. Cai, "Causal inference in nonlinear systems: Granger causality versus time-delayed mutual information," *Physical Review E*, vol. 97, no. 5, p. 052216, 2018.

[72] D. Kugiumtzis, B. Lillekjendlie, and N. D. Christophersen, "Chaotic time series: Part 1: estimation of some invariant properties in state space," *Modeling, identification and control*, vol. 15, no. 4, pp. 205–224, 1994.

[73] T. Asefa, M. Kemblowski, U. Lall, and G. Urroz, "Support vector machines for nonlinear state space reconstruction: Application to the great salt lake time series," *Water resources research*, vol. 41, no. 12, 2005.

[74] M. Casdagli, S. Eubank, J. D. Farmer, and J. Gibson, "State space reconstruction in the presence of noise," *Physica D: Nonlinear Phenomena*, vol. 51, no. 1-3, pp. 52–98, 1991.

[75] 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.

[76] 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.

[77] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science Advances*, vol. 5, no. 11, p. eaau4996, 2019.

[78] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in genetics*, vol. 10, p. 524, 2019.

[79] J. G. Caporaso, K. Paszkiewicz, D. Field, R. Knight, and J. A. Gilbert, "The western english channel contains a persistent microbial seed bank," *The ISME journal*, vol. 6, no. 6, pp. 1089–1093, 2012.

[80] A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. N. Balasubramanian, "Neural network attributions: A causal perspective," in *International Conference on Machine Learning*, pp. 981–990, PMLR, 2019.

[81] B. C. Daniels and I. Nemenman, "Automated adaptive inference of phenomenological dynamical models," *Nature communications*, vol. 6, no. 1, pp. 1–8, 2015.

[82] S. Leng, H. Ma, J. Kurths, Y.-C. Lai, W. Lin, K. Aihara, and L. Chen, "Partial cross mapping eliminates indirect causal influences," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

[83] N. M. Mangan, S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Inferring biological networks by sparse identification of nonlinear dynamics," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 2, no. 1, pp. 52–63, 2016.

[84] R. Sanchez-Romero, J. D. Ramsey, K. Zhang, M. R. Glymour, B. Huang, and C. Glymour, "Estimating feedforward and feedback effective connections from fmri time series: Assessments of statistical methods," *Network Neuroscience*, vol. 3, no. 2, pp. 274–306, 2019.

[85] C. Hitchcock, "Causal Models," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, summer 2020 ed., 2020.

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