Comparison of six methods for the detection of causality in a bivariate time series

Anna Krakovská,* Jozef Jakubík, and Martina Chvosteková
Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9, 842 19 Bratislava, Slovak Republic

David Coufal, Nikola Jajcay, and Milan Paluš
Institute of Computer Science, Czech Academy of Sciences, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic

(Received 6 October 2017; published 10 April 2018)

In this comparative study, six causality detection methods were compared, namely, the Granger vector autoregressive test, the extended Granger test, the kernel version of the Granger test, the conditional mutual information (transfer entropy), the evaluation of cross mappings between state spaces, and an assessment of predictability improvement due to the use of mixed predictions. Seven test data sets were analyzed: linear coupling of autoregressive models, a unidirectional connection of two Hénon systems, a unidirectional connection of chaotic systems of Rössler and Lorenz type and of two different Rössler systems, an example of bidirectionally connected two-species systems, a fishery model as an example of two correlated observables without a causal relationship, and an example of mediated causality. We tested not only 20 000 points long clean time series but also noisy and short variants of the data. The standard and the extended Granger tests worked only for the autoregressive models. The remaining methods were more successful with the more complex test examples, although they differed considerably in their capability to reveal the presence and the direction of coupling and to distinguish causality from mere correlation.

DOI: 10.1103/PhysRevE.97.042207

I. INTRODUCTION

Detection of causality is not a closed issue, even for a bivariate case. Applying different causal methods to real-world data very often leads to contradictory results. New methods that seem to offer superior solutions at first will later prove to be effective only for suitably selected examples. This experience forced us to test multiple currently used causality methods on artificially generated data with clearly defined causal relationships. In this study, we focused on detecting uni- and bidirectional causal links between just two systems, each of which was represented by a single time series. Unlike other comparative studies, a wide variety of methodologies have been tested here on different types of time series. This has enabled us to demonstrate clearly that use of a particular method for inappropriate type of data can have critical consequences. Moreover, as shown in the following, in some cases, even methods designed specifically for a particular type of data, produced conflicting results. This calls into question the reliability of some relatively popular methods and shows that it is important to keep a critical eye on the limits of the individual methods.

Six frequently used methods were tested. The first one concerns the mathematical approach proposed by Granger in 1969 [1]. In Granger's sense, a time series x is said to cause y if it can be shown, through statistical hypothesis tests, that past x values provide significant information about the future values of y. Using Granger's test is problematic in the case of dynamically linked variables. Moreover, the initially linear Granger concept also requires generalizations to enable the investigation of complex nonlinear processes. Therefore, new approaches have been proposed, including the extended Granger causality test (EG) [2], which is the second method examined in this study. EG operates in state spaces, but with direct reference to the initial Granger's idea. The method starts with a standard delay embedding reconstruction [3]. Then, locally, the dynamics is approximated using an autoregressive (AR) model. Granger's causality is assessed on short pieces of the trajectory, and the results are averaged over the entire state portrait.

The third method represents another way of generalizing the original Granger test for nonlinear cases. It is a kernel version of the Granger test (KG), introduced in Ref. [4]. The method performs linear Granger causality in the feature space of suitable kernel functions, assuming arbitrary degree of nonlinearity.

The fourth method is based on transfer entropy or conditional mutual information [5,6]. This information-theoretic measure assesses the change in uncertainty of the future of a signal when estimates with and without additional knowledge of another system are compared.

The fifth method belongs to approaches that rely on evaluating the distances of conditioned neighbors in reconstructed state spaces. It works with the assumption that the investigated time series x and y are manifestations of dynamical systems X and Y, respectively. Therefore, as a first step, a d-dimensional manifold M_X is reconstructed from lags of observable x so that the state of the system in time t is

$$[x(t),x(t-\tau),x(t-2\tau),\ldots,x(t-(d-1)\tau)].$$

The manifold M_Y is reconstructed analogously. Given certain conditions, the reconstructed manifold is diffeomorphic to the

*krakovska@savba.sk

 ©2018 American Physical Society

original one [3]. Consequently, they share the same features in many ways. Most importantly for us, points that are close in the original manifold are also close in the reconstruction. Several methods have been proposed to infer causality from the asymmetry of cross mappings between close neighbors in M_X and M_Y [7–13]. Different authors have defined different criteria to quantify the identification capability. As shown later, the idea can be applied through measures such as M or $\mathcal{L}$ introduced in Refs. [11] and [12] or through a method called convergent cross mapping [13].

The state space reconstruction, which is essential for some of the tested methods, is determined by two parameters, namely, the dimension d of the reconstructed space and the delay τ. The choice of parameters is usually based on the method of nearest neighbors and the first minimum of mutual information, respectively. The optimal value of d must be at least as large as the dimensionality of the original attractor. It is also true that the length of the embedding time window ($\tau.d$) is more important than the delay time and the embedding dimension separately [14].

The last reviewed method evaluates predictability improvement (PI). It uses mixed predictions [15] and again belongs to the state-space approaches. It determines whether the prediction of a time series made in a reconstructed space improves when data from another system are included in the space reconstruction. The potential of this idea has been identified in Refs. [16,17].

In the following section, we introduce the data sets selected for the comparison of the methods. Then, the six methods for detecting causal relationships are explained. Three of the methods were originally designed to offer their results in graphical form only. Here, these methods are complemented by surrogate analysis to allow a more objective comparison of the performance of the methods. The fourth section presents the results.

It is confirmed that the traditional Granger's test is a clear choice, fast and reliable when applied to appropriate data. But the new methods designed for complex data are far from reliable. It turns out, for example, that some of the popular methods have extremely low specificity—they produce a large number of false detections of causality. Finally, the findings are summarized and discussed.

II. DATA

To test the methods described below, we used the next seven examples of coupled systems.

A. Coupled autoregressive models

In the first experiment, we studied the next system consisting of two bidirectionally coupled autoregressive processes of first order,

$$\begin{aligned} x(t) &= 0.5x(t-1) + 0.2y(t-1) + \epsilon_x(t) \\ y(t) &= Cx(t-1) + 0.7y(t-1) + \epsilon_y(t), \end{aligned} \tag{1}$$

where $\epsilon_x(t)$ and $\epsilon_y(t)$ are independent Gaussian random processes with zero mean and common variance $\sigma_x^2 = \sigma_y^2 = 0.1$. The parameter C regulates the coupling strength; for $C = 0$, the coupling is unidirectional: Only system Y drives X. In our experiment, we considered $C \in \{0,0.2,0.4,0.6\}$. For each coupling, we generated a bivariate time series $\{x(t), y(t)\}_t$ of length $N = 120\,000$ (random initial conditions extracted from the normal distribution with zero mean and unit variance) and the initial 10^5 iterations were discarded.

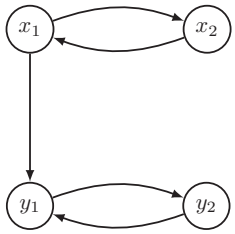

FIG. 1. Interaction graph for the coupling of two Hénon systems.

B. Hénon → Hénon

In the second example, we used two unidirectionally coupled identical Hénon maps. The first two lines correspond to the driving system X, and the last two equations describe the response system Y:

$$\begin{aligned} x_1(t+1) &= 1.4 - x_1^2(t) + 0.3x_2(t) \\ x_2(t+1) &= x_1(t) \\ y_1(t+1) &= 1.4 - \left[Cx_1(t)y_1(t) + (1-C)y_1^2(t)\right] + 0.3y_2(t) \\ y_2(t+1) &= y_1(t). \end{aligned} \tag{2}$$

C controls the strength of the coupling. The so-called interaction graph (see Fig. 1), which is easy to interpret from Eqs. (2), shows that the systems are coupled through a one-way driving relationship between variables x_1 and y_1. In the interaction graph, the nodes representing the variables are connected by directed edges whenever one variable directly drives another. As emphasized by the authors in Ref. [18], both direct and indirect interactions can be recovered, although they cannot be reliably distinguished using the comparisons of state space reconstructions.

We generated pairs of time series x_1, y_1 for 21 coupling strengths, chosen from 0 to 0.8 with a step of 0.04. Coupling strengths C above 0.7 give rise to synchronization [8]. In each case, the starting point was [0.7,0,0.91,0.7].

The first 1000 data points were discarded, whereas the next 20 000 ones were saved. The same Hénon-Hénon system was studied in Refs. [6,8,19,20].

C. Rössler → Lorenz

The next example was a unidirectional coupling of two different chaotic systems. The Rössler system X drives the Lorenz system Y:

$$\begin{aligned} \dot{x}_1 &= -6(x_2 + x_3) \\ \dot{x}_2 &= 6(x_1 + 0.2x_2) \\ \dot{x}_3 &= 6[0.2 + x_3(x_1 - 5.7)] \\ \dot{y}_1 &= 10(-y_1 + y_2) \\ \dot{y}_2 &= 28y_1 - y_2 - y_1y_3 + Cx_2^2 \\ \dot{y}_3 &= y_1y_2 - \tfrac{8}{3}y_3. \end{aligned} \tag{3}$$

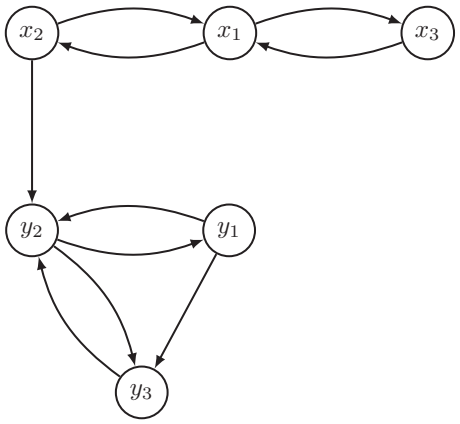

FIG. 2. Interaction graph for the Rössler system driving the Lorenz system.

Generalized synchronization seems to take place just before coupling $C = 3$ [19].

For testing purposes, we will assume that we know variable x_2 of the driving Rössler system and variable y_2 of the driven Lorenz system and we would like to recover the corresponding causal link (see Fig. 2). The solutions were obtained numerically by the fourth-order Runge-Kutta method. They were computed for 21 coupling strengths chosen from 0 to 4 with a step of 0.1. The starting point was [0,0,0.4,0.3,0.3,0.3]. The first 1000 data points were discarded, whereas the next 20 000 data points were saved. As regards the sampling of the resulting trajectory, one typical run around the attractor took about 10–15 points. The same example of the Rössler → Lorenz system was studied, among others, in Refs. [6,11,14,19–21].

D. Rössler 0.5 → Rössler 2.5

In the next testing example, we used unidirectionally coupled nonidentical Rössler systems at different coupling strengths C:

$$\begin{aligned}
\dot{x}_1 &= -\omega_1 x_2 - x_3 \\
\dot{x}_2 &= \omega_1 x_1 + 0.15 x_2 \\
\dot{x}_3 &= 0.2 + x_3(x_1 - 10) \\
\dot{y}_1 &= -\omega_2 y_2 - y_3 + C(x_1 - y_1) \\
\dot{y}_2 &= \omega_2 y_1 + 0.72 y_2 \\
\dot{y}_3 &= 0.2 + y_3(y_1 - 10),
\end{aligned} \tag{4}$$

where

$$\omega_1 = 0.5, \quad \omega_2 = 2.515.$$

The two Rössler systems represent distinct dynamical subsystems X and Y coupled through a one-way driving relationship between variables x_1 and y_1. This causal link is what we would like to recover. Mutual interconnections of the variables may again be visualized by an interaction graph (Fig. 3).

In total, 21 000 data points of x_1 and y_1 were generated by a MATLAB solver of ordinary differential equations (ode45) at an integration step of 0.3. The starting point was [0, 0, 0.4, 0, 0, 0.4]. The first 1000 data points were discarded. The coupling strength C was chosen from 0 to 1.1 with a step size of 0.1. The synchronization takes place at a coupling of about 1. In addition, these systems also differ in that the first is chaotic, whereas the other is quasiperiodic [6]. The frequency ratio of about 1:5 used in this example is reminiscent of cardiorespiratory interactions, where the influence of the (slower) respiratory dynamics on the (faster) cardiac dynamics is larger than in the opposite direction. The system was first used in Ref. [6] to demonstrate that, in this case, the problem of detecting directionality is much more challenging than in cases of two systems with more closely related natural frequencies.

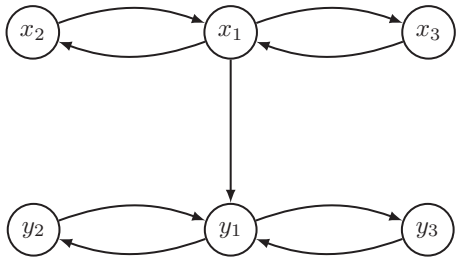

FIG. 3. Interaction graph for the coupling of two Rössler systems.

E. Bidirectional two-species model

The next example represents a bidirectional causality between two coupled logistic difference equations:

$$\begin{aligned}
x(t+1) &= x(t)[3.78 - 3.78x(t) - 0.07y(t)] \\
y(t+1) &= y(t)[3.77 - 3.77y(t) - 0.08x(t)].
\end{aligned} \tag{5}$$

Like in Ref. [22], the system was initialized at $x(1) = 0.2$ and $y(1) = 0.4$ and was run for 21 000 time steps. The last 20 000 data points were saved.

F. Fishery model

The fishery model, which was also used in Refs. [13] and [15], represents the situation of a correlation between systems that can be falsely declared as causality. We analyzed a standard logistic model of two noninteracting fish populations that share common environmental (e.g., weather) forcing:

$$\begin{aligned}
R_x(t+1) &= x(t)\{3.1\,[1 - x(t)]\}\,\exp(0.5\,Z) \\
R_y(t+1) &= y(t)\{2.9\,[1 - y(t)]\}\,\exp(0.6\,Z) \\
x(t+1) &= 0.4\,x(t) + \max[R_x(t-3),0] \\
y(t+1) &= 0.35\,y(t) + \max[R_y(t-3),0].
\end{aligned} \tag{6}$$

The variables x and y denote the sizes of the fish populations, R_x and R_y are the recruitments of the populations, and Z is an environmental influence represented here by a red noise. Figure 4 shows the corresponding interaction graph. We generated data with starting points $R_x(i) = R_y(i) = x(i) = y(i) = 0.5$ for $i \in \{1,2,3,4\}$. Z was defined as follows:

$$Z(i) = p \sum_{j=i-14}^{i} Z'(j),$$

$$Z' \sim N(0,1).$$

The variable p was chosen so that var(Z) was equal to 1.

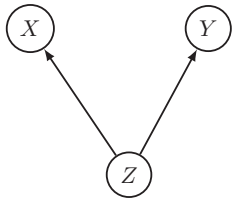

FIG. 4. Interaction graph for the fishery model (6) describing two populations, X and Y, driven by an environmental variable Z.

G. Mediated link

In the last example, like in the fishery case, two uncoupled subsystems X and Y are driven by the third one, Z. However, we only tested the case with a coupling of $C = 0.06$ for which Z and X were already synchronized, which was reflected as a mediated causal effect from X to Y. The equations are as follows:

$$\begin{aligned}
\dot{z}_1(t) &= -\omega_1 z_2(t) - z_3(t)\\
\dot{z}_2(t) &= \omega_1 z_1(t) + 0.15 z_2(t)\\
\dot{z}_3(t) &= 0.2 + z_3(t)(z_1(t) - 10)\\
\dot{x}_1(t) &= -\omega_2 x_2(t) - x_3(t) + C[z_1(t) - x_1(t)]\\
\dot{x}_2(t) &= \omega_2 x_1(t) + 0.15 x_2(t)\\
\dot{x}_3(t) &= 0.2 + x_3(t)[x_1(t) - 10]\\
\dot{y}_1(t) &= -\omega_3 y_2(t) - y_3(t) + C[z_1(t) - y_1(t)]\\
\dot{y}_2(t) &= \omega_3 y_1(t) + 0.15 y_2(t)\\
\dot{y}_3(t) &= 0.2 + y_3(t)[y_1(t) - 10]. \qquad (7)
\end{aligned}$$

The frequencies ω_1, ω_2, and ω_3 were selected to have the following, approximately unit, ratios: $\omega_1 = 0.985$, $\omega_2 = 1$, and $\omega_3 = 1.015$, respectively.

The data were generated by numerical integration based on the adaptive Bulirsch-Stoer method [23] using a sampling interval of 0.314. The initial 5000 samples were discarded, whereas the next 20 000 samples were saved. We assumed that the only two known variables were x_1 and y_1.

Faes *et al.* suggested that, sometimes, indicators of causal relations can be enhanced in the presence of noise compared to clean data [24]. To see if this is true for our examples, we also prepared noisy versions of the test time series. Specifically, we added white Gaussian noise, which is considered a suitable model of noise typically present in real-world data. The signal-to-noise ratio per sample was 20 dB, which corresponded to an amplitude ratio of 10 and a power ratio of 100.

III. METHODS

A. Granger's vector autoregressive test

Assume two jointly vector-valued stochastic variables x and y. One approach to evaluate the causal relations between two time series is to examine if the prediction of one series based on the information about the past of the series itself could be improved by incorporating the information about the past of the other. Granger (1969) [1] formalized this idea of Wiener (1956) [25] in terms of the vector autoregressive (VAR) modeling of stochastic processes. Specifically, variable x is said to Granger-cause variable y if the inclusion of past observations of x reduces the variance of the prediction error of y in a VAR model. In its simplest form, the Granger causality from x to y is motivated as follows: suppose that time series x and y can be described by a bivariate autoregressive model

$$\begin{aligned}
x(t) &= \sum_{j=1}^{p} a_{xx,j} x(t-j) + \sum_{j=1}^{p} a_{xy,j} y(t-j) + \epsilon_{x/y}(t)\\
y(t) &= \sum_{j=1}^{p} a_{yx,j} x(t-j) + \sum_{j=1}^{p} a_{yy,j} y(t-j) + \epsilon_{y/x}(t),
\end{aligned}$$

where p is the model order; $a_{xx,j}$, $a_{xy,j}$, $a_{yx,j}$, and $a_{yy,j}$ are coefficients of the model; and $(\epsilon_{x/y}(t), \epsilon_{y/x}(t))'$ is a (2×1) unobservable zero mean white noise vector process with time-invariant covariance matrix Σ. In a bivariate VAR(p) model for $(x, y)'$, x fails to Granger-cause y if $\mathrm{var}(\epsilon_{y/x}) > \mathrm{var}(\epsilon_y)$, where $\epsilon_y(t)$ is defined as

$$y(t) = \sum_{j=1}^{p} a_{y,j} y(t-j) + \epsilon_y(t).$$

The likelihood ratio test statistic for the null hypothesis of non-Granger causality x to y, i.e., $H_0 : \forall j : a_{yx,j} = 0$, is defined as

$$\frac{\left(R^2_{H_0} - R^2\right)/p}{R^2/(T-3p)} \sim_{as.} F_{p,T-3p}$$

and under the null hypothesis has an asymptotic F distribution with p and $T - 3p$ degrees of freedom, where $R^2_{H_0} = \sum_t \hat{\epsilon}^2_y(t)$ is the sum of the squared residuals of the restricted model given by H_0, $R^2 = \sum_t \hat{\epsilon}^2_{y/x}(t)$ is the sum of the squared residuals of the full model, and T is the number of observations. Rejection of the null hypothesis means that the coefficients corresponding to the data from variable x are statistically significantly different from zero, and it is concluded that the process x Granger-causes y. The roles of the two time series can be reversed to investigate the causal influence in the opposite direction. Note that the model order p for the VAR model is determined using a model selection criterion, with the most common information criteria being the Akaike information criterion and the Schwartz-Bayesian information criterion. Unknown coefficients of the models may be estimated by ordinary least squares.

B. Extended Granger test

Although a rather general class of covariance stationary multivariate processes can be modeled as VARs, and not just stochastic processes generated by a linear autoregressive scheme, it will be shown under Results that Granger's VAR test often fails to identify the correct causal influence for nonlinear time series. Chen *et al.* proposed an extension of Granger's approach to nonlinear multivariate time series called *extended Granger causality* [2]. The authors started with standard delay-coordinate state space reconstruction and assumption that one can approximate the dynamics linearly. Then, the original Granger's idea about causality detection was

applied to δ neighborhoods and the results were averaged over the entire state portrait of the dynamics.

In the first step of the procedure, the dynamics is reconstructed using delay vector $Z(t) = [X(t), Y(t))]$, where

$$X(t) = [x(t), x(t-\tau), \ldots, x(t-(d-1)\tau)],$$
$$Y(t) = [y(t), y(t-\tau), \ldots, y(t-(d-1)\tau)],$$

where τ is the time delay and d_i is the embedding dimension. In the next step, a VAR model

$$x(t) = \sum_{j=1}^{d} a_{xx,j} x(t-j\tau) + \sum_{j=1}^{d} a_{xy,j} y(t-j\tau) + \epsilon_{x/y}(t),$$
$$y(t) = \sum_{j=1}^{d} a_{yx,j} x(t-j\tau) + \sum_{j=1}^{d} a_{yy,j} y(t-j\tau) + \epsilon_{y/x}(t)$$

and an autoregressive model for y time series

$$y(t) = \sum_{j=1}^{d} a_{y,j} y(t-j\tau) + \epsilon_y(t)$$

are fitted for all points in the neighborhood Θ of a reference point Z_0, where $\Theta = \{Z : |Z - Z_0| \leqslant \delta\}$. Then the error ratio is calculated,

$$1 - \frac{\mathrm{var}(\epsilon_{y/x})}{\mathrm{var}(\epsilon_y)}, \tag{8}$$

which can be interpreted as the ratio of the variation in y, which is not explained by y itself and can be explained by x. The fitting process is repeated for a randomly selected set of reference points Z_0's. The causal relationship from x to y is evaluated by the extended Granger causality index (EGCI), which is estimated as the average of the values (8) from all selected local neighborhoods. Analogously, for the evaluation of the causal relationship from y to x, the autoregressive model on the individual x is fitted in the second step of the procedure and EGCI is determined as the average of $1 - \mathrm{var}(\epsilon_{x/y})/\mathrm{var}(\epsilon_x)$ from all selected local neighborhoods. The critical issue of this method is how to determine the optimal neighborhood size. Therefore, EGCI is finally displayed as a function of δ and the conclusion should result from a visual assessment of the function. There has not been any statistical test proposed to identify the causal relationship between the series.

However, due to more objective comparison of results, we had to add a testing method. We adopted the procedure that draws on employing Fourier transform surrogates (FT surrogates) [26]. For a suitable δ, the surrogates were used to compute the EGCI in the same way as was done for the original data. We used 50 surrogates, and the obtained set of EGCI values was characterized by its empirical mean and standard deviation (SD). Comparing the EGCI obtained for the original data with the empirical distribution of the surrogates, we inferred the significance of the EGCI value if it was greater than the surrogates' mean+2SD.

C. Kernel Granger test

Another attempt to extend Granger causality to nonlinear cases by using the theory of reproducing kernel Hilbert spaces was proposed in Ref. [4]. The method investigates Granger causality in the feature space of suitable kernel functions.

Let $X(t) = \{x(t), x(t-1), \ldots, x[t-(m-1)]\}$, $Y(t) = \{y(t), y(t-1), \ldots, y[t-(m-1)]\}$, and $Z(t) = [X(t), Y(t)]$.

Denote $u_1, \ldots, u_m$ the eigenvectors of matrix K with elements $K_{ij} = k(Y(i), Y(j))$, where k is a kernel function.

Let $H \subseteq R^{T-m}$ be the range of the matrix K. Then, the estimate of $y(t)$ based on $y(t-1), \ldots, y(t-m)$, denoted as $\tilde{y}(t)$, is the projection of $y(t)$ on the space H. In other words, $\tilde{y}(t) = Py(t)$, where $P = \sum u_i u_i^T$ is the projector on H.

Analogously for the estimate of $y(t)$ based on $y(t-1), \ldots, y(t-m), x(t-1), \ldots, x(t-m)$, denoted as $\tilde{y}'(t)$, we have $\tilde{y}'(t) = P'y(t)$, where P' is the projector on H' and $H' \subseteq R^{T-m}$ is the range of matrix K' with elements $K'_{ij} = k(Z(i), Z(j))$. The space H' ($H' \subseteq H$) may be decomposed as $H' = H \otimes H^{\perp}$, where $H^{\perp}$ is the space of the vectors of H' orthogonal to the vectors of H. The space $H^{\perp}$ is the range of the matrix $\tilde{K} = K' - PK' - K'P + PK'P$.

Calling $P^{\perp}$ the projector on $H^{\perp}$, the causality index to detect the causal relation $X \to Y$ can be written as $\sum_t (P^{\perp} y(t))^2 / \sum_t [y(t) - \tilde{y}(t)]^2$. To avoid problems with overfitting, the value $\sum_t [P^{\perp} y(t)]^2$ is calculated as a sum of squared statistically significant Pearson's correlation coefficients $\{r_{i'}\}$ of eigenvectors of $\tilde{K}$ and observations of y, i.e., $\sum_t [P^{\perp} y(t)]^2 = \sum_{i'} r_{i'}^2$. Since only significant correlations, which pass the Bonferroni test with expected fraction of false positive equal to 0.05, are summed, the causality index, denoted δ_F, is called filtered linear Granger causality index.

The roles of the two time series can be reversed to evaluate the causality index in the opposite direction.

Authors in Ref. [4] considered two choices of kernel functions: the inhomogeneous polynomial of the appropriate order and the Gaussian kernel. In this study, we tested both of these variants using the MATLAB code freely available on Ref. [27].

This method, again, does not include statistical testing. The presence of the causal link should be estimated visually from the plotted δ_F values. Therefore, in this case, we also used surrogate testing, analogously to the previous method.

D. Conditional mutual information (transfer entropy)

The next tested method is rooted in information theory [28,29]. The method computes the conditional mutual information (CMI) between two variables of interest for an appropriately selected condition. A significantly high value of the CMI indicates the presence of a causal link between the investigated variables. The direction of the link is determined by the formulation of the condition. More details follow.

Let $\{x(t)\}_{t=1}^{L}$ and $\{y(t)\}_{t=1}^{L}$, $L \in \mathbb{N}$ be two finite time series that represent the coordinates of two (possibly) coupled dynamical systems, namely X and Y, respectively. In the CMI,

the measure for detection of coupling direction reads as

$$I(x(t), y(t+\tau)|y(t), y(t-\eta), \ldots, y(t-D\eta)). \quad (9)$$

Given that (9) reaches a significantly high value, we indicate an information flow between $x(t)$ and $y(t+\tau)$ conditioned on $y(t)$ and its history. If this holds for various forward lags $\tau \in \mathbb{N}$ after averaging, then we conclude that there exists a causal link from X to Y. The inference of a link in the opposite direction is provided by the CMI (9) in which the variables x and y are swapped. Computationally, obtaining the CMI (9) relies on the binning method [30]. Paluš proposed that the number of bins $Q \in \mathbb{N}$ in each marginal should not exceed the bound $\sqrt[d+1]{N}$, i.e., $Q \leqslant \sqrt[d+1]{N}$, where d is the dimension of the data and N is the sample size.

Although the univariate component $x(t)$ is sufficient for a proper application of the CMI method, the dimension $D \in \mathbb{N}$ of condition (the embedding dimension) is important as it should provide the access to the full information about the present state of system Y.

Taken's theorem [3] provides a theoretical justification for mapping a $D+1$-dimensional dynamical system by the set of time-lagged coordinates $y(t), y(t-\eta), \ldots, y(t-D\eta)$, where $\eta \in \mathbb{N}$ is the backward time lag used in the embedding reconstruction.

Given the original data $\{y(t)\}_{t=1}^{L}$ and the construction of their shifted versions $\{y(t+\eta)\}_{t=1}^{L-\eta}$, $\eta \in \mathbb{N}$, the value of η appearing in the condition of (9) may be selected as the locus of the first minimum on a graph showing the mutual information between the original data and their shifted versions for various $\eta \in \mathbb{N}$ [31].

As the forward time lag τ is not usually known *a priori*, the CMI is computed as the function of τ. In our application, τ ranges from 1 to 60. The resulting value of the CMI is then computed as the average over the individual forward lags.

The crucial question is how to test if the obtained value of the CMI is significant. In analogy to the previous two methods, we adopted the procedure that draws on employing Fourier transform surrogates [26]. Fifty surrogates were used to compute the CMI in the same way as how it was done for the original data. We inferred the significance of the CMI value if it was greater than the surrogates' mean+2SD.

E. Cross mappings

1. Measure $\mathcal{L}$

Cross-mapping methods operate in state spaces of the dynamical systems. Again, we only consider the knowledge of the observables x and y. Therefore, as a first step, d-dimensional manifolds M_X and M_Y are reconstructed from the lags of the observables x and y as explained above.

As stated in the Introduction, close points remain close after the reconstruction and, therefore, the neighborhoods are preserved. This is important since the detection of causality between systems X and Y is based on assessing whether the time indices of nearby points in the historical data of manifold M_Y can be used to identify the nearby points in the reconstructed manifold M_X. If so, then we are dealing with a causal link from X to Y.

Considering the direction $X \to Y$, for points x_i and its neighbors, we can evaluate the rank-based measure $\mathcal{L}$, which was introduced by Chicharro and Andrzejak in 2009 [12]:

$$\mathcal{L}(X|Y) = \frac{1}{N}\sum_{i=1}^{N}\frac{G_i(X) - G_i^k(X|Y)}{G_i(X) - G_i^k(X)},$$

where $G_i(X)$ is the mean rank of all neighbors ($\frac{N}{2}$), $G_i^k(X)$ is the minimal mean rank of k nearest neighbors ($\frac{k+1}{2}$), and $G_i^k(X|Y)$ is the Y-conditioned mean rank of k neighborhoods.

2. Convergent cross mapping

Convergent cross mapping (CCM), which was introduced by Sugihara *et al.* in 2012 [13], builds on earlier cross-mapping techniques and results in essentially the same conclusions as if, for example, the above-mentioned $\mathcal{L}$ measure was used. However, in this case, the evaluation of the success of the cross mapping is done as follows.

Given $k = d+1$ nearest neighbors, $y(t)$ (based on the information from M_X) is approximated by

$$\hat{y}(t)|M_X = \sum_{i=1}^{d+1} w_i y(t_i),$$

where

$$w_i = u_i \Big/ \sum_{j=1}^{d+1} u_j,$$

$$u_i = e^{-\frac{d(X(t),X(t_i))}{d(X(t),X(t_1))}},$$

and $d(X,Y)$ is the Euclidean distance between vectors X and Y. The cross mapping from Y to X is defined analogously. The skill of cross mapping is evaluated by the correlation coefficient ρ between Y and $\hat{Y}|M_X$. A high value of the correlation indicates that system X drives system Y. By analogy, a high correlation between X and $\hat{X}|M_Y$ indicates a causal link from Y to X.

Sugihara *et al.* emphasize that a key feature of the CCM method is the capability to distinguish causation from mere correlation. Specifically, they monitored whether the cross-mapped estimates converged to the correct values for an increasing number of used data points. For causally coupled systems, as opposed to merely correlated ones, the estimates improve with the length of the time series. Lack of convergence should indicate the absence of actual causality.

For the CCM and $\mathcal{L}$, we used the same MATLAB implementation as in Ref. [32]. However, the method as proposed in Ref. [13] requires obtaining the results through a visual assessment of the images. For our purposes, extending the method through some form of a statistical test or surrogate analysis is necessary, as has been done, for example, in Ref. [33] and also in this study.

F. Predictability improvement

The last method, with special attention attributed to the optimization of the reconstructed space based on the weighting

of each of its coordinates, was proposed in Ref. [15]. Because of its computational demands, we decided to work with a slightly simpler variant in this study.

The basic idea lies in the so-called mixed-state analysis presented in Refs. [16,17]. Consider two systems X and Y, represented by time series x and y, respectively. Similarly, as in the case of Granger's meaning, we say that variable x causes variable y if better prediction of y can be obtained by using the information from both x and y rather than by using only the information from y. However, unlike the Granger method, this one operates in multidimensional state spaces.

To find the optimal choice of the reconstruction parameters, rather than trust the false nearest-neighbor method without reservation, we evaluated the prediction error for various combinations of possible embedding parameters. Consequently, the lowest errors led us to the proper choices of d and τ.

Regarding the actual method of prediction, we used the method of analogs [34] that finds historical data similar to the current situation and that assumes that the system will continue just as it did in the past. There are several ways to predict the follower of point $Y(t)$, the simplest one being finding the time index i of its nearest neighbor from past states on the reconstructed trajectory and declaring $\hat{Y}(t+1) = Y(i+1)$. A modification, which was used here, improves the simplest version by averaging the followers of several neighbors while considering exponential weighting based on the distances of the neighbors from $Y(t)$.

The specific steps of the method to detect the causality from X to Y are as follows:

(i) The manifold M_Y is performed. This provides space to get the predictions of Y without using additional information from X. The one-point predictions $\hat{Y}$ of a large-enough statistical sample of N points over the reconstructed trajectory are computed. The resulting errors are given by $e_Y(t) = Y(t) - \hat{Y}(t)$.

(ii) To get the predictions of Y using information from both X and Y, we reconstruct the manifold M_{X+Y} (mixed-state space). M_{X+Y} contains some of the coordinates from M_Y and some from M_X. If we used the full number of coordinates, then the state corresponding to time t would be $\{y(t), y(t-\tau_Y), \ldots, y[t-(d_Y-1)\tau_Y], w.x(t), w.x(t-\tau_X), \ldots, w.x[t-(d_X-1)\tau_X]\}$, where the weight w represents the impact of system X. Analogously as in step (i), the predictions of system Y, denoted as $\hat{Y}_{X+Y}(t)$ in time t, are computed. The corresponding errors are $e_{X+Y}(t) = Y(t) - \hat{Y}_{X+Y}(t)$.

(iii) To decide whether the addition of information from X improves the prediction of Y (X drives Y), we use the Welch test to test the null hypothesis H_0 that the errors come from independent random samples from normal distributions with equal means and equal but unknown variances against the alternative hypothesis that the mean of errors e_{X+Y} is less than the mean of e_Y. If H_0 is rejected on a 5 % significance level, then we accept that $e_{X+Y} < e_Y$ or, equivalently, that the inclusion of the knowledge of X significantly improves the prediction of Y.

Causality in the opposite direction, i.e., from Y to X, is investigated analogously—after exchanging the roles of X and Y in the above instructions.

IV. RESULTS

In the following, we demonstrate selected examples on how each method mastered the causality detection. The percentages of false-positive and -negative results are then presented in Table I.

In case the reader is interested, more detailed results for all methods and each individual test example can be found in the Supplemental Material [35].

A. Granger's test result

The Schwartz-Bayesian information criterion was used to determine the model order p. The maximal number of time lags of 15 was considered for all examples. The regression coefficients were fitted by ordinary least squares, and the residual covariance matrix was obtained as the sample covariance of the residual errors $\hat{\Sigma} = \sum_{t=p+1}^{T} \hat{\epsilon}(t)\hat{\epsilon}(t)'/(T-p)$. Before using Granger's VAR analysis, we checked the VAR's stability and the autocorrelation of the error term. A VAR system is stable if the eigenvalues of the companion matrix are less than 1. Although the maximal eigenvalue was quite close to unity for some examples, we assumed all fitted models as stable. Since the autocorrelation of the error term has influence on Granger's VAR test, we tested by the multivariate Ljung-Box portmanteau test (Q test) if the error term was not serially correlated. In the case where the Q test rejects the null hypothesis that the error term is not correlated with the past values, it can be concluded that Granger's VAR test is not appropriate for analyzing causality and its result is not valid. We report that only for the data of coupled AR models did the Q test not reject the hypothesis that the residuals in fitted VAR are a white-noise sequence (p value ≈ 0.42); for the rest of the examples, the null hypothesis of the Q test was rejected (p value $< 10^{-15}$).

In the case of the AR processes, the Granger causality was correctly detected in the direction $Y \to X$ (p value $< 10^{-4}$) for $C = 0$ and in both the directions $Y \to X$ (p value $< 10^{-15}$) and $X \to Y$ (p value $< 10^{-15}$) for $C = \{0.2, 0.4, 0.6\}$.

Although without the support of the Q test, the correct causal links were detected significantly for the bidirectional two-species model and for few cases of week couplings between the chaotic dynamical systems. For the rest of the examples, causality was incorrectly detected as bidirectional (p value $< 10^{-10}$).

As expected, the application of Granger's VAR test turned out to be fast and accurate for the AR models but inappropriate for most of the considered examples of coupled nonlinear systems. Table I shows that the method too often detected causality in cases where no causal relationship existed (high false-positive rate).

B. Extended Granger's test results

The set of reference points in the reconstructed phase space consisted of 200 randomly chosen points in the long data set and 100 points in the short data sets. The time delay for all examples was set to $\tau = 1$. The embedding dimension was $d = 2$ for the coupled AR models, unidirectionally coupled Henón systems, and bidirectional two-species model; $d = 3$ for the unidirectionally coupled Rössler systems and the example of a mediated causal link; and $d = 5$ for the fishery model.

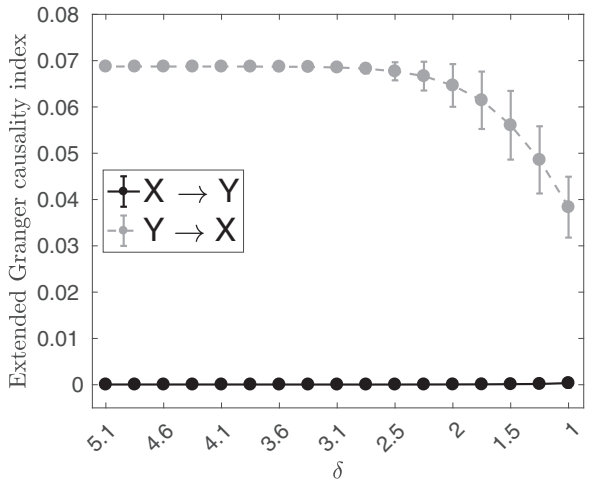

FIG. 5. Extended Granger causality index for unidirectionally coupled autoregressive processes [Eqs. (1) for $C = 0$] for 20 000 points and decreasing δ. The index correctly suggests the causal link in the direction $Y \rightarrow X$. Zero index values in the opposite direction mean that there was no link from X to Y.

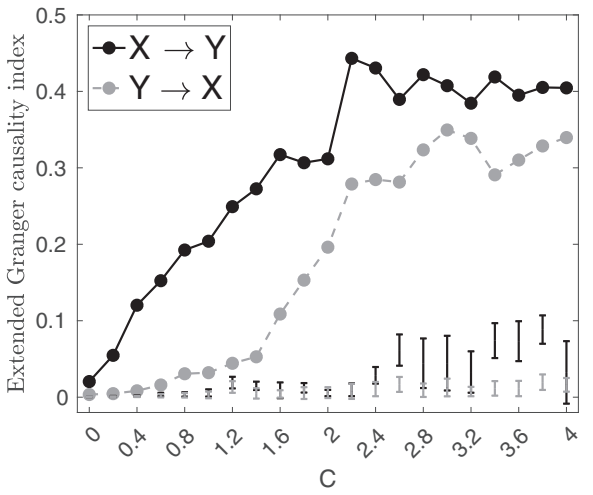

FIG. 7. Extended Granger causality index for unidirectionally coupled Rössler and Lorenz systems [Eqs. (3)] for clean, long time series, fixed δ, and increasing couplings C. Bars at the bottom show the results for the surrogates. The figure indicates bidirectional causality even in cases where only $X \rightarrow Y$ is true.

A difficult issue in the EG method is how to determine the optimal neighborhood size δ. The number of points in the neighborhood should be large enough because of statistics, but the size of the neighborhood must be small due to the usability of the linearization. Chen *et al.* have not provided instructions for identifying the optimal interval; they just suggest to seek a compromise by examining the index as a function of δ [2]. Figures 5 and 6 show the estimated EGCI as a function of size of neighborhood δ averaged over five experimental runs (the error bars denote the standard deviation). The values of δ mean percentages of the diameter of the reconstructed state portrait.

As expected, the size of δ was not critical when analyzing interconnected autoregressive processes. The result for Eqs. (1), $C = 0$, is shown in Fig. 5.

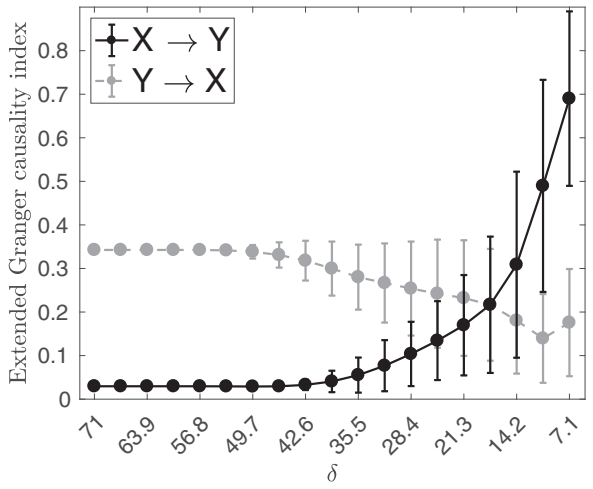

FIG. 6. Extended Granger causality index for unidirectionally coupled Rössler and Lorenz systems [Eqs. (3) for $C = 2$] for clean, long time series and decreasing δ. In this example, $X \rightarrow Y$. However, the figure wrongly indicates causality in both directions for low δ.

However, in more complex situations, finding the relevant δ neighborhoods and reading the resulting graph were problematic or even impossible.

An example can be seen in Fig. 6, which shows the causality index for decreasing δ in coupled Rössler and Lorenz systems. Surrogate testing (see Fig. 7), resulting in the detection of bidirectional causality, confirmed the failure of the EG method in the Rössler $\rightarrow$ Lorenz example.

We report that the results obtained by employing the surrogate analysis to the EG method were invalid for most of the nonlinear examples. In the neighborhoods selected according to the above instructions, the use of VAR models did not seem appropriate.

C. Kernel Granger's test results

The Kernel Granger's test was employed using both inhomogeneous polynomial functions and Gaussian kernel functions [4]. Both variants led mostly to the same conclusions regarding causality, but the polynomial kernel version was considerably faster than using the Gaussian kernel. In testing, we used different combinations of parameters of the method. The resulting images for all tested examples can be found in the Supplemental Material [35].

The KG method detected causal links between AR models without problems.

KG was also successful with linked Hénon systems. Figure 8 shows the estimated causality index δ_F for different coupling values. The nonzero values of the index correctly indicate causality in the $X \rightarrow Y$ direction, while δ_F remains zero for unlinked direction and also after synchronization.

However, in more complex situations, the KG method tends to give false-positive results. Take an example of the unidirectionally coupled Rössler and Lorenz systems for which the KG test (Fig. 9) incorrectly indicates bidirectional causality.

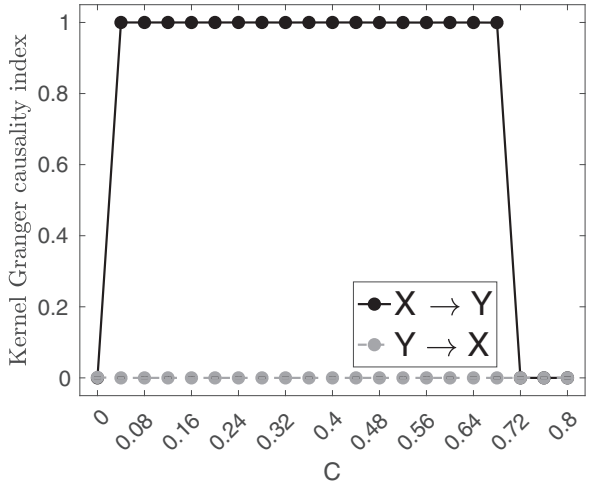

FIG. 8. Kernel Granger causality index δ_F for unidirectionally coupled Hénon systems [Eqs. (2)] for increasing couplings C. Inhomogeneous polynomial kernel, $m = 2$, $p = 2$; 20 000 clean data points used for computation. Results for surrogate data are so close to zero that they cannot be seen. The picture indicates correctly the $X \to Y$ link.

D. Conditional mutual information results

In the case of the VAR model, the CMI method led to correct results for the 20 000-point long time series. Detection was less successful when using only 1000 data points. In cases of weak couplings, the values of the examined index were indistinguishable from the values computed for the surrogates; however, for strong couplings, the causal link was still determined correctly.

Causal analysis of unidirectionally coupled Hénon systems was also managed very well by the CMI method.

The test example of Rössler → Lorenz turned out to be much more difficult for the CMI method. Figure 10 shows the results for all investigated coupling values when using short, noisy time series. In fact, when using long and clean data, the resulting figures indicated the same thing. For weak couplings, correctly, only the link $X \to Y$ was detected; for strong couplings, however, the CMI method came to the wrong conclusion that $Y \to X$ also.

Analysis of the Rössler 0.5 → Rössler 2.5 example was slightly better than in the previous case. Figure 11 shows that, for long time series and coupling values well below the threshold of synchronization, the method correctly detected that X

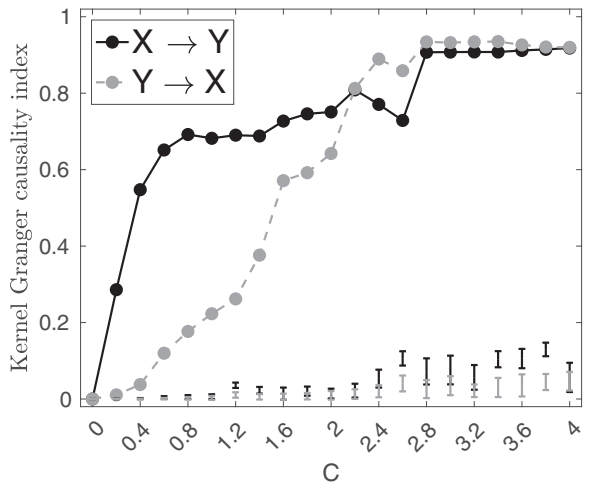

FIG. 9. Kernel Granger causality index δ_F for unidirectionally coupled Rössler and Lorenz systems [Eqs. (3)] for increasing couplings C. Inhomogeneous polynomial kernel, $m = 3$, $p = 3$; 20 000 clean data points used for computation. Bars at the bottom show the results for the surrogates. The figure indicates bidirectional causality even in cases where only $X \to Y$ is true.

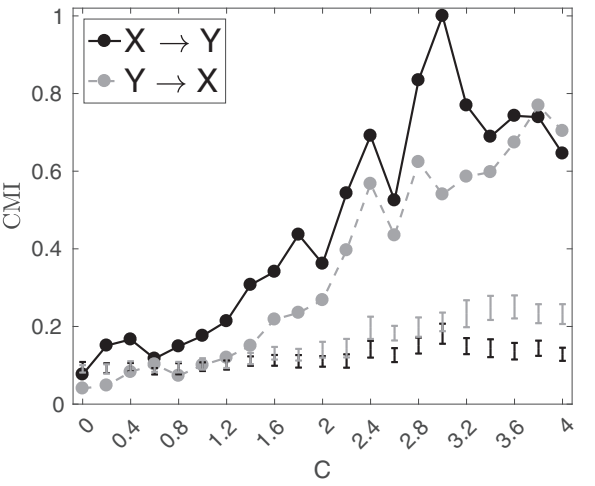

FIG. 10. CMI for unidirectionally coupled Rössler and Lorenz systems [Eqs. (3)] for noisy, 1000-point long time series and increasing coupling values C. Embedding dimension $d = 3$ and delay $\tau = 3$ for the direction $X \to Y$ and $\tau = 6$ for the opposite direction were used for the reconstruction. Bars at the bottom show the results for the surrogates. Although the correct answer was $X \to Y$, except for very weak couplings, the CMI method indicated causal links in both directions.

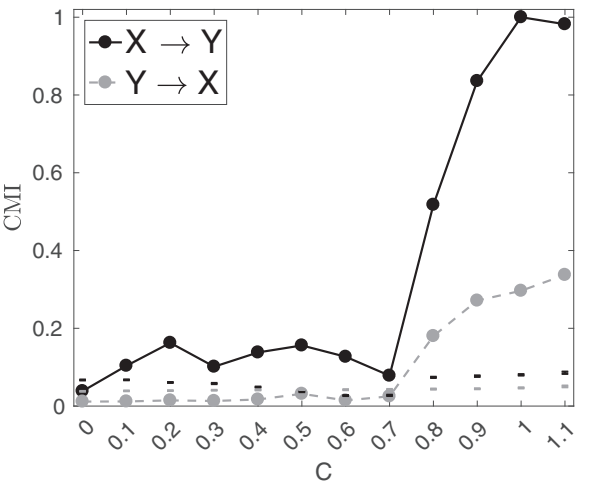

FIG. 11. CMI for unidirectionally coupled Rössler systems [Eqs. (4)] for clean, long time series and increasing coupling values C. Embedding dimension $d = 3$ and delay $\tau = 13$ for the direction $X \to Y$ and $\tau = 2$ for the opposite direction were used for the reconstruction. Bars at the bottom show the results for the surrogates.

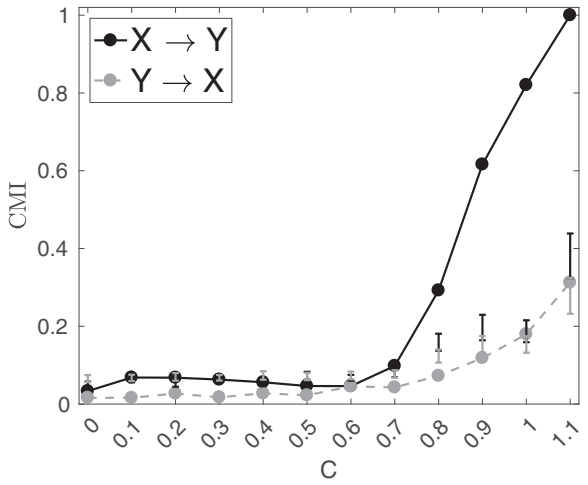

FIG. 12. CMI for unidirectionally coupled Rössler systems [Eqs. (4)]. Same as in Fig. 11 but for the 1000-point long time series.

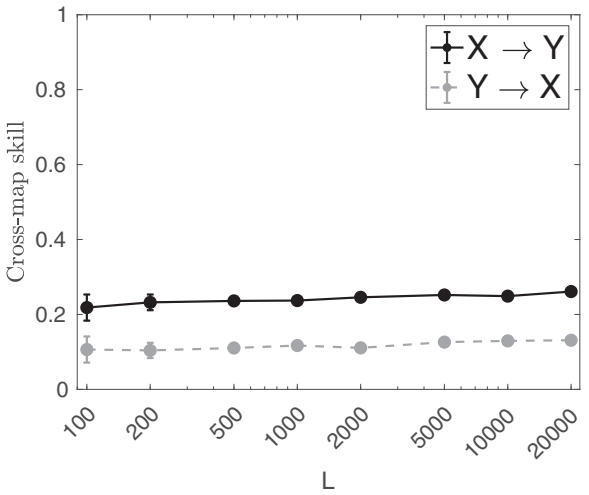

FIG. 13. Cross-map skill ρ for unidirectionally coupled autoregressive processes [Eqs. (1) for $C = 0$] for an increasing number of used data points. Embedding dimension $d = 2$ and delay $\tau = 1$ were used for the reconstruction. The CCM method led to false-positive results. Cross-map skill in the direction $X \to Y$ being higher than in the opposite direction suggests a causal link from X to Y and a very weak or no link from Y to X. In fact, in this case, the correct answer would be that Y drives X.

drives Y. However, with short time series, the detection was no longer possible, as the surrogate analysis shows in Fig. 12.

In the two-species model, the bidirectional causal link was clearly identified for both clean and noisy data. In the case of this example, the CMI method was the only method that was able to find the causal link even after the addition of noise.

In the fishery model of two correlated but uncoupled time series, the surrogate analysis incorrectly indicated bidirectional causality.

The mediated link from the last test example [Eqs. (7)] was clearly detected by the CMI method when using 20 000 data points. For the short time series, the detection was no longer possible.

We used the straightforward application of the CMI test with three-dimensional condition in the case of three-dimensional systems. This setting requires the estimation of five-dimensional probability distribution for which 1000 data samples are insufficient especially for the binning estimator. We have repeated the tests using the k-nearest-neighbor estimator which is considered more effective by many authors. Testing for the presence of causality, however, the k-NN algorithm did not bring significant improvement (see Supplemental Material [35]).

In some cases, deeper understanding of the analyzed data can bring improvement. For instance, in the case of coupled Rössler systems, the CMI method can be applied to the instantaneous phases, which effectively require just one-dimensional condition in the CMI. Then, the sensitivity and specificity are increased even when using the binning estimator, and a series length of order 1000 would be sufficient for a successful detection of causality. See Ref. [6] for details.

E. Cross-mappings results

In the case of linked AR processes, the results were mixed. For zero coupling, the CCM method failed to detect that Y drives X unidirectionally (see Fig. 13). For nonzero couplings, the method led to a significant cross-map skill in both directions, correctly indicating a bidirectional link.

Causal analysis of the second data sets, the unidirectionally coupled Hénon systems, was not well managed by the CCM method. The unidirectional links were in 88% falsely detected as bidirectional.

The test example of Rössler → Lorenz has also proved to be unmanageable for the CCM method. Figure 14 shows the results for the 21 coupling values when 20 000 clean data points were used for the analysis. The CCM method incorrectly indicated bidirectional causality.

Since the authors in Ref. [13] pointed out that, with limited or noisy data, causality should be demonstrated by estimation

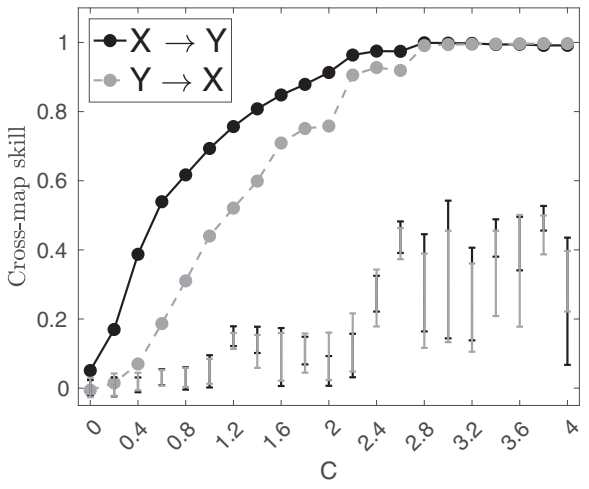

FIG. 14. Cross-map skill ρ for unidirectionally coupled Rössler and Lorenz systems [Eqs. (3)] for increasing coupling values C; 20 000 clean data points used for computation. Embedding dimension $d = 8$ and delay $\tau = 1$ were used for the reconstruction. For each coupling, the boxplots include the results for 50 surrogate time series. In cases where the correct answer was $X \to Y$, the CCM method mostly indicated causal links in both directions.

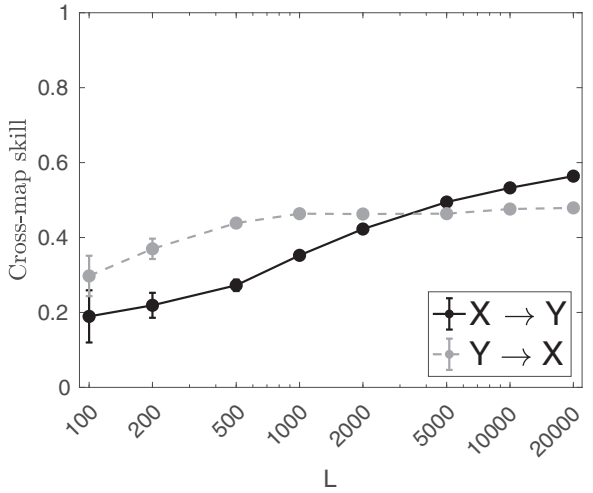

FIG. 15. Cross-map skill ρ for unidirectionally coupled Rössler and Lorenz systems [Eqs. (3), $C = 1$] for an increasing number of used data. Although the correct answer was $X \to Y$, the CCM method wrongly indicated a bidirectional link.

precision increasing with the number of used data points, we were also looking for this phenomenon. This is illustrated by Fig. 15, which shows the cross-map skill under coupling $C = 1$ for an increasing number of used data points. However, this figure incorrectly implies bidirectional causality. If we used fewer than about 4000 data points, then the nonexistent link in the direction $Y \to X$ would even appear to be stronger than the $X \to Y$ link, which was the only one present here.

Analysis of the interconnected slower and faster Rössler systems came in the same way as in the previous example. The unidirectional links were in 82% falsely detected as bidirectional.

Figures 16 and 17 illustrate how the use of different measures affects the evaluation of the cross-map skill. We can see that the use of the rank-based measure $\mathcal{L}$ suggested by the

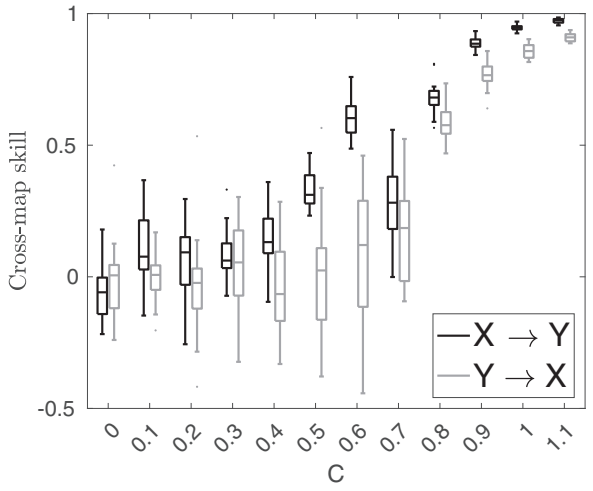

FIG. 16. Cross-map skill ρ for unidirectionally connected Rössler systems [Eqs. (4)] for increasing coupling values C. Embedding dimension $d = 8$ and delay $\tau = 1$ were used for the reconstruction. Each boxplot includes the results for 20 time series of 1000 clean data points.

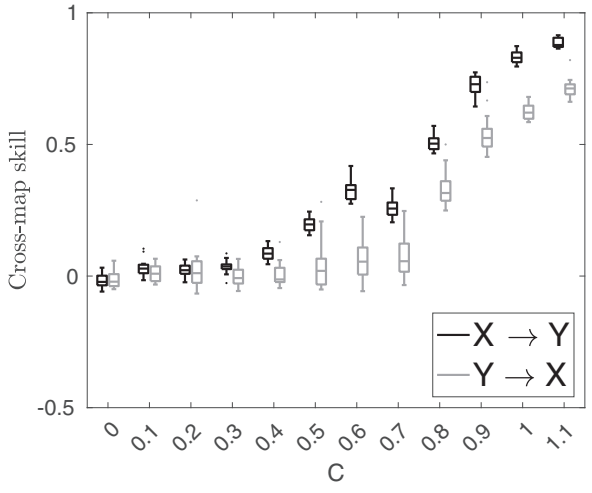

FIG. 17. Cross-map skill ρ for unidirectionally connected Rössler systems [Eqs. (4)] for increasing coupling values C. The only difference compared to the previous figure is that here the measure $\mathcal{L}$ [12] was used for the evaluation of the cross mapping, which obviously led to more stable results.

authors in Ref. [12] led to more stable results than the measure used in the CCM method [13].

In the two-species model, the bidirectional causal link was clearly identified for the clean data. After the addition of noise, on the other hand, detection of the causality was no longer possible. The reason could be that the coupling used in this example was relatively weak and that the noise made it impossible to detect.

In the fishery model, the significant cross-correlation between species suggests that they might be coupled. On the other hand, the cross-mapping skill did not improve for an increasing number of data points in any of the two directions, indicating that we might be not dealing with causality here. Unfortunately, the surrogate analysis closed it as a bidirectional causality, which belongs to false-positive results.

For the 20 000-point long, clean data, the mediated link in Eqs. (7) was clearly detected by the CCM. For the short or noisy data, it was no longer possible.

As regards the CCM method, Table I shows that it belonged to the least successful. CCM produced false results even for long clean time series. However, this failure is not entirely against the theory, which claims that the direction of the causal link implies good cross mapping but not that the direction of no causality produces a poor mapping. As a result, the criterion based on the improvement in the cross-map quality with the number of observations can lead to false-positive results, especially in the case of strongly coupled observables. Like the authors in Ref. [33], we must warn against using this method for real-world data.

F. Predictability improvement results

Investigation of the predictability improvement was done in the reconstructed state spaces. In the case of long time series, 1000 predictions were made on the basis of the previous 19 000 data points. For the short time series, 200 points were predicted on the basis of the previous 800 points.

In the case of coupled AR models, the results were mixed. For $C = 0$, the null hypothesis of no predictability improvement was not rejected ($p = 0.27$) in the direction $X \to Y$, which was right. For nonzero couplings, there was a causal link in the direction $X \to Y$. The examined predictabilities improved nonsignificantly for the weakest coupling ($p = 0.09$) but were highly significant for the two stronger couplings ($p < 0.0008$). However, there was a link in the opposite direction of $Y \to X$ in all four cases, but it could not be confirmed because the significance level was above the cut-off value ($0.12 < p < 0.18$). For 20 short time series (1000 data points each), the results were the same, only even less significant.

In the Hénon $\to$ Hénon example, for the 20 000-point long time series, the causal links were identified correctly with high significance. This was true for both clean and noisy data and for any of the 21 coupling values. For the 1000-point long time series, the conclusions were the same. Only in the case of the short noisy series was the predictability improvement sometimes not significant for very small and very large (around the synchronization threshold) coupling values. However, neither here nor in the other examples did we ever receive any false-positive results.

In the third example (Rössler $\to$ Lorenz), M_X was reconstructed with delay $\tau = 2$ and embedding dimension $d = 3$, whereas for M_Y the delay was 1 and the embedding dimension was up to 9 (the stronger the coupling, the higher the optimal embedding window $\tau_Y.d_Y$). The conclusions were similar as in the previous case. The predictions of y_2 without considering x_2 [Eqs. (3)] were clearly poorer than the predictions made with the help of x_2. For the long time series, the link $X \to Y$ was detected with high statistical significance (p value $< 10^{-15}$ for the clean data and p value $< 10^{-7}$ for the noisy data). When we focused on the opposite direction, we found that adding information from system Y did not improve the prediction of the representative of X, confirming that Y does not drive X. For the 1000-point long time series, the conclusions were the same. Only in the case of the short noisy series (see Fig. 18) was the predictability improvement sometimes not significant for very small and very large coupling values.

In the fourth test example of a slower Rössler system X driving a faster Rössler system Y [Eqs. (4)], the observable from X helps to predict Y with high significance ($p < 0.0001$), except for the two weakest coupling values and states after the synchronization. Testing the opposite direction showed no predictability improvement and rejected the causal link of $Y \to X$.

For the clean data of the two-species system, bidirectional causality was detected with high statistical significance (p value $< 10^{-50}$ for the long time series and p value $< 10^{-10}$ for the short data). In the noisy data, the presence of bidirectional causality could not be confirmed because the significance level of the predictability improvement was above the cut-off value.

In the fishery model, no causal link was found between the two time series. Regardless of whether we were testing long, short, clean, or noisy data, we failed to reject the null hypothesis, which was the correct outcome in this example.

The last example concerned a mediated link. The PI method correctly failed to find any link in the $Y \to X$ direction. However, the expected predictability improvement in the direction $X \to Y$ was not confirmed either, being insignificant

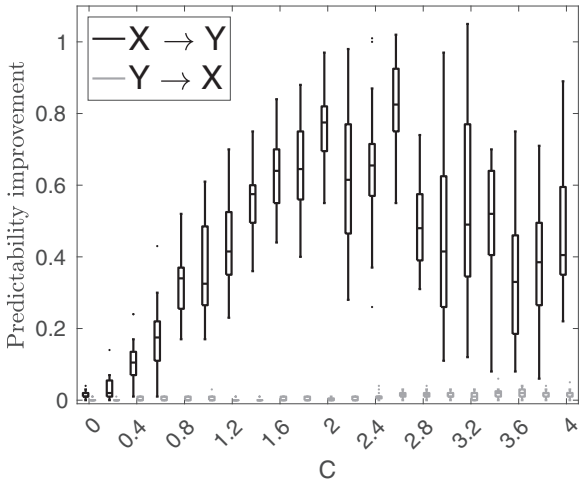

FIG. 18. Predictability improvement for unidirectionally coupled Rössler and Lorenz systems [Eqs. (3)] for increasing coupling values C. Each boxplot includes the results for 20 time series of 1000 noisy data points (the figure for the clean data looks very similar). For most values of C, the method revealed that $X \to Y$.

($p = 0.11$) for long clean data and at the limit of significance ($p = 0.048$) for long noisy data. With the short time series, the conclusion was the same.

We also touched on the impact of the noise level and the number of data on the performance of individual methods, although it was not systematically studied here. As regards the impact of noise on the results, we tested only cases with 20-dB additive noise. We found that this level of noise did not affect considerably the performance of the methods.

As far as the data length is concerned, we used 20 000 or 1000 data points. It happened quite often that 20 000 points allowed for the correct detection of causality, but 1000 points were no longer enough.

The results for both long and short, as well as results for clean and noisy data can be found in the Supplemental Material [35].

Synchronized states would also deserve more attention. Once the systems were synchronized, causality should not be able to detect [6]. This was mostly true when we used the PI method with long and clean data sets. On the other hand, noisy and short data might behave somewhat differently, as shown in Fig. 18, where the synchronized signals were able to help a little with each other's prediction. However, some methods led to confusing results when analyzing synchronized time series. For example, the cross-map skills evaluated in the CCM method were close to the maximum in both directions when analyzing synchronized signals. Then, it was impossible to guess from the resulting images whether we were dealing with synchronized systems, highly correlated systems, or strongly bidirectionally interconnected systems. Another of the methods, the CMI method, as expected, did not see any causality in the Hénon $\to$ Hénon example after the synchronization. However, in some other test examples, the method incorrectly indicated strong causal links between synchronized time series (see Figs. 10, 11, and 12).

To compare the six methods to each other, the results for clean, long versions of each individual test data were

TABLE I. Rates (in %) of false-negative and false-positive results of the six methods (Granger, extended Granger, kernel Granger, cross mapping, predictability improvement, and conditional mutual information) for the seven test examples. Clean, 20 000-point-long time series were used for the evaluation. The results of the detection after the synchronization were not included. The false-negative rates represent the cases when the method failed to find an existing causal link. The false-positive rate is the proportion of all cases without causal links, where the causal effect was incorrectly detected.

		G	EG	KG	CMI	CCM	PI
AR models	False negative	0	0	0	0	0	71
	False positive	0	0	0	0	100	0
Hénon → Hénon	False negative	0	0	0	0	0	0
	False positive	65	100	0	0	88	0
Rössler → Lorenz	False negative	0	0	0	0	0	0
	False positive	87	100	87	60	87	0
Rössler 0.5	False negative	0	0	0	0	0	22
→ Rössler 2.5	False positive	45	100	64	18	82	0
Two-species	False negative	0	0	0	0	0	0
Fishery model	False positive	100	0	100	100	100	0
Mediated link	False negative	0	0	0	0	0	100
	False positive	100	100	100	0	0	0
All together	False negative	0	0	0	0	0	16
	False positive	68	94	49	28	85	0

summarized in Table I. Because of the problems described above, we decided to base the ratings in the table only on the results obtained for the pre-synchronization states.

V. CONCLUSION

In this study, the performance of six methods for the detection of causality was assessed. The methods included the Granger VAR test, the extended Granger test, the kernel Granger test, cross-mapping techniques, conditional mutual information, and assessment of the predictability improvement. Comparisons were made on test examples of systems with different types of interconnections.

Unfortunately, the results of different methods often contradicted each other. Take the test data of Rössler → Lorenz, for example. Although we had a causal link here in just one direction, the Granger causality was detected highly significantly in both directions. False causality from the Lorenz system to the Rössler system was also detected by the EG, KG, CMI, and CCM methods. PI was the only method that did not produce the false-positive Lorenz → Rössler result here.

In the case of unidirectionally coupled AR processes [Eqs. (1) for $C = 0$], for a change, the PI method did not find any causality and the Granger tests and the CMI method correctly detected that $Y \to X$, whereas the CCM method wrongly suggested the opposite link of $X \to Y$.

The outputs of the methods were difficult to compare. The G, PI, and CMI method were based on statistical hypothesis testing and surrogate analysis, respectively, whereas the EG, KG, and CCM methods were designed to rely only on a visual assessment of the images. In order to compare the sensitivity and specificity of all tested methods to each other, we extended the later three methods in the sense that the statistical significance of the results was evaluated by the surrogate test technique.

This allowed us to compare the results for all methods applied to each individual test data within Table I.

The table, among other things, shows that it is important to choose the right method for a particular type of data. We found that detecting causality in coupled autoregressive processes was mastered quickly and flawlessly by the Granger test and by the extended and kernel Granger tests. The CMI method was also fairly successful, while the task was much more difficult for the CCM and PI methods.

As expected, the Granger test failed to correctly detect causality in more complex test cases. Surprisingly, the extended Granger test did not do better either. In 94% of noncausal examples, the EG method falsely detected causality. The kernel method seemed to be more promising, although even this, too, often led to false-positive results.

In fact, low specificity was the problem of most methods, as evidenced by numbers in the bottom row of Table I. The only method without false-positive results was PI. On the other hand, PI had some difficulties with detecting existing causal links in AR processes and for very weak couplings.

To summarize, this study showed that blind application of any causality test easily leads to incorrect conclusions. Causality analysis of our example of coupled AR processes was simple and correct with the Granger tests but was tedious and mostly incorrect when using other approaches. On the other hand, time series produced by more complex dynamical systems need carefully designed information-theoretic or state space-based methods such as CMI and PI.

ACKNOWLEDGMENTS

This work was supported by the Slovak Grant Agency for Science (Grants No. 2/0011/16 and No. 2/0047/15) and by Slovak Grant Agency: Agentúra na Podporu Výskumu a Vývoja (Grant No. APVV-15-0295). M. Paluš was supported by the Czech Health Research Council (Project No. NV15-33250A).

[1] C. W. Granger, Econometrica **37**, 424 (1969).
[2] Y. Chen, G. Rangarajan, J. Feng, and M. Ding, Phys. Lett. A **324**, 26 (2004).
[3] F. Takens, in *Dynamical Systems and Turbulence*, edited by D. A. Rand and L. S. Young (Springer-Verlag, Berlin, 1981), pp. 366–381.
[4] D. Marinazzo, M. Pellicoro, and S. Stramaglia, Phys. Rev. Lett. **100**, 144103 (2008).
[5] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).
[6] M. Paluš and M. Vejmelka, Phys. Rev. E **75**, 056211 (2007).
[7] N. F. Rulkov, M. M. Sushchik, L. S. Tsimring, and H. D. I. Abarbanel, Phys. Rev. E **51**, 980 (1995).

[8] S. J. Schiff, P. So, T. Chang, R. E. Burke, and T. Sauer, Phys. Rev. E **54**, 6708 (1996).
[9] J. Arnhold, P. Grassberger, K. Lehnertz, and C. E. Elger, Physica D **134**, 419 (1999).
[10] R. Q. Quiroga, A. Kraskov, T. Kreuz, and P. Grassberger, Phys. Rev. E **65**, 041903 (2002).
[11] R. G. Andrzejak, A. Kraskov, H. Stögbauer, F. Mormann, and T. Kreuz, Phys. Rev. E **68**, 066202 (2003).
[12] D. Chicharro and R. G. Andrzejak, Phys. Rev. E **80**, 026217 (2009).
[13] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, Science **338**, 496 (2012).
[14] A. Krakovská, K. Mezeiová, and H. Budáčová, J. Complex Syst. **2015**, 932750 (2015).
[15] A. Krakovská and F. Hanzely, Phys. Rev. E **94**, 052203 (2016).
[16] M. Wiesenfeldt, U. Parlitz, and W. Lauterborn, Int. J. Bifurcat. Chaos **11**, 2217 (2001).
[17] U. Feldmann and J. Bhattacharya, Int. J. Bifurcat. Chaos **14**, 505 (2004).
[18] B. Cummins, T. Gedeon, and K. Spendlove, SIAM J. Appl. Dynam. Syst. **14**, 335 (2015).
[19] R. Q. Quiroga, J. Arnhold, and P. Grassberger, Phys. Rev. E **61**, 5142 (2000).
[20] M. Paluš, V. Komárek, Z. Hrnčíř, and K. Štěrbová, Phys. Rev. E **63**, 046211 (2001).
[21] M. Le Van Quyen, J. Martinerie, C. Adam, and F. J. Varela, Physica D **127**, 250 (1999).
[22] H. Ye, E. R. Deyle, L. J. Gilarranz, and G. Sugihara, Sci. Rep. **5**, 14750 (2015).
[23] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, Cambridge, MA, 1986).
[24] L. Faes, A. Porta, and G. Nollo, Phys. Rev. E **78**, 026201 (2008).
[25] N. Wiener, in *Modern Mathematics for Engineers*, edited by E. F. Backenbach (McGraw Hill, New York, 1956), pp. 165–190.
[26] M. Paluš, Contemp. Phys. **48**, 307 (2007).
[27] D. Marinazzo, Kernel granger causality test—matlab function, https://github.com/danielemarinazzo/KernelGrangerCausality.
[28] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 1991).
[29] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[30] M. Paluš, Physica D **80**, 186 (1995).
[31] A. M. Fraser and H. L. Swinney, Phys. Rev. A **33**, 1134 (1986).
[32] A. Krakovská, J. Jakubík, H. Budáčová, and M. Holecyová, arXiv:1511.00505v2 (2016).
[33] S. Cobey and E. B. Baskerville, PLoS ONE **11**, e0169050 (2016).
[34] E. N. Lorenz, J. Atmos. Sci. **26**, 636 (1969).
[35] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.97.042207 for results of all causal methods and each individual test example.