Transfer entropy computation using the Perron-Frobenius operator

David Diego,[*] Kristian Agasøster Haaga,[†] and Bjarte Hannisdal

Department of Earth Science, University of Bergen, PO Box 7803, NO-5020 Bergen, Norway

We propose a method for computing the transfer entropy between time series using Ulam's approximation of the Perron-Frobenius (transfer) operator associated with the map generating the dynamics. Our method differs from standard transfer entropy estimators in that the invariant measure is estimated not directly from the data points, but from the invariant distribution of the transfer operator approximated from the data points. For sparse time series and low embedding dimension, the transfer operator is approximated using a triangulation of the attractor, whereas for data-rich time series or higher embedding dimension, we use a faster grid approach. We compare the performance of our methods with existing estimators such as the *k* nearest neighbors method and kernel density estimation method, using coupled instances of well known chaotic systems: coupled logistic maps and a coupled Rössler-Lorenz system. We find that our estimators are robust against moderate levels of noise. For sparse time series with less than 100 observations and low embedding dimension, our triangulation estimator shows improved ability to detect coupling directionality, relative to standard transfer entropy estimators.

I. INTRODUCTION

Time series analysis is used to study the dynamics of complex systems across many disciplines, including macroscale activity of the brain [1] and interactions in the global climate system [2]. A long-standing problem in time series analysis is the detection of causal connections between different components of a system from observed time series. Several theoretical frameworks have been proposed to address this problem [3], including information-theoretic approaches [4,5]. A popular information-theoretic method is the so-called transfer entropy (TE) [6], or conditional mutual information [7], which quantifies whether knowledge of changes in one variable reduces uncertainty about changes in another variable. For deterministic systems, the concept of information entropy relies on the existence of invariant densities associated to attractors [8]. Standard methods for computing mutual information (and from it, TE), estimate the invariant distribution directly from the embedding of the data. For instance, the *k* nearest neighbors method (kNN) [9], which uses counting of nearest neighbors, or approaches using visitation frequency, either directly [6] or through kernel density estimation (KDE) [10]. Other TE estimators use the concept of permutation entropy [11]. In our work, we propose to compute TE (or any standard information-theoretic measure [12]) between time series, based on a numerical approximation of the Perron-Frobenius (transfer) operator of the underlying dynamics. The transfer operator [13] dictates how densities in phase space are transformed under the action of maps and its approximation has been used to identify long-term emergent behavior in dynamical systems. Applications include computation of the

stretching rate of chaotic maps, and identification of attracting regions in ocean circulation [14,15]. The transfer operator enables estimation of invariant distributions from which information entropies may be computed [16–18]. Transfer operator approximation may also be used to generate dynamically informed surrogates for null-hypothesis testing, and to interpolate and forecast time series. These possibilities are explored in a forthcoming study.

In the following, we describe our approach starting with the general notion of TE, and a brief overview of entropies in the context of deterministic systems (for a more in-depth review, see [3,8]).

II. TRANSFER ENTROPY

Suppose that for two variables X_1 and X_2, one is given the probability density $P(X_1 = x_1, X_2 = x_2)$ with support S. From this density one may compute their mutual information [12,19] as

$$I(X_1, X_2) = \int_S P(x_1, x_2) \log_2 \frac{P(x_1, x_2)}{P(x_1)P(x_2)}. \tag{1}$$

For the case of three variables X_1, X_2, and X_3, a related quantity is the TE

$$T_{X_1, X_2|X_3} = \int_S P(x_1, x_2, x_3) \log_2 \frac{P(x_1|x_2, x_3)}{P(x_1|x_3)}. \tag{2}$$

One can easily check the identity

$$T_{X_1, X_2|X_3} = I(X_1, X_{23}) - I(X_1, X_2), \tag{3}$$

with $X_{23} = (X_2, X_3)$. $T_{X_1, X_2|X_3}$ thus quantifies the amount of information shared between X_1 and (X_2, X_3) beyond the information already shared between X_1 and X_2. The TE was originally introduced in the context of time series analysis [6] as a way of estimating the information transfer from one time series to another. Given two time series X and Y, the

---
[*]david.castro@uib.no; diegocastro.david@gmail.com
[†]https://www.earthsystemevolution.com;
https://github.com/kahaaga/

TE measures how much information is lost by assuming that the variables X and Y are independent, i.e., by assuming that $P(x(t + \tau)|x(t), y(t)) = P(x(t + \tau)|x(t))$. There is, however, no absolute scale associated with information entropy [12,19], hence, the TE only determines whether the information transfer from Y to X is greater than in the opposite direction.

III. MEASURE THEORETIC ENTROPIES FOR DETERMINISTIC SYSTEMS

In the context of dynamical systems, mutual information (or any information-theoretic measure) between variables may be computed from the density distribution of invariant measures associated with the attractor of the dynamical system. In the following, we briefly review these concepts.

A. Attracting sets and attractors

Assume the dynamics is generated by a diffeomorphism [20] $\psi : \mathbb{R}^n \to \mathbb{R}^n$. A set $\mathbb{A} \subset \mathbb{R}^n$ is said to be an attracting set for ψ if the following conditions are met [21]:

(1) There is an open set $U \supset \mathbb{A}$ and a natural number N such that for any open set $V \supset \mathbb{A}$, $\psi^m(U) \subset V$ for all $m \geqslant N$.

(2) $\psi(\mathbb{A}) \subset \mathbb{A}$.

Here, ψ^m denotes the m-fold iterate of ψ. The open set U is called a fundamental neighborhood of $\mathbb{A}$. This definition implies [21] that $\mathbb{A} = \cap_{m \geqslant 1} \psi^m(U)$ and $\psi(\mathbb{A}) = \mathbb{A}$. Moreover, if there is an open set $U \subset R^n$ such that for all m big enough $\psi^m(U)$ has compact closure contained in U, then $\mathbb{A} = \cap_{r \geqslant 1} \psi^r(U)$ is a compact attracting set with fundamental neighborhood U. Because the open set V can be arbitrarily small around $\mathbb{A}$, all the trajectories entering U asymptotically approach $\mathbb{A}$. In addition, $B = \cup_{m \geqslant 0} \psi^{-m}(U)$, where ψ^{-m} denotes the preimage of the m-fold iterate of ψ, is such that for any $p \in B$ there is m with $\psi^m(p) \in U$. Thus, the corresponding orbit approaches $\mathbb{A}$ asymptotically. B is called the basin of attraction of $\mathbb{A}$ and if $B = \mathbb{R}^n$, $\mathbb{A}$ is called a global attracting set. An attractor is, however, a somewhat more restrictive concept than an attracting set. Roughly speaking, an attractor is what is left of an attracting set after removing the wandering points (see [21] for a precise definition). Trivial examples of attractors are asymptotically stable fixed points while less trivial ones are stable limit cycles or quasiperiodic limiting orbits [22]. More complicated ("strange") attractors contain unstable orbits (i.e., sensitivity to initial conditions) and usually fractal geometries. Most dynamical systems associated with natural processes, even simple processes involving very few variables, give rise to highly complex dynamics in the form of strange attractors [23–25].

B. Invariant measures and ergodicity

The trajectory of a typical orbit of a dynamical system having an attractor generates a distribution of points in the phase space with a certain density which seems to be intrinsic to the system. Different portions of the attractor are visited by the orbit with different frequency, and this frequency of visitations naturally defines a density on the attractor, clearly invariant under the dynamics. The notion of invariance leads to the notion of *ergodicity*. Intuitively, a dynamical system is said to be ergodic if a generic trajectory fills in the

attractor (according to the above notion of invariant density). A crucial result pertaining to ergodic systems is the celebrated Birkhoff's ergodic theorem:

Given a space M and a map $h : M \to M$, let μ be a measure on M such that μ is invariant under h. Then for any $\phi : M \to \mathbb{R}$ measurable, it holds that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} \phi(h^k(x)) = \int_M \phi \, d\mu \,, \tag{4}$$

for μ almost every $x \in M$.

Applied to the characteristic function [26] of any measurable set K, the theorem implies that the measure of K, $\mu(K)$, equals the frequency of visits to K, in concordance with the above notion of invariant measure. In Appendix B, we show the equality between time and spatial averages obtained from the estimates of invariant measures considered in this work.

Assuming that a density δ of μ is well defined for μ almost all points [27] [$\mu(K) = \int_K \delta \, dm$, with dm the Euclidean (Riemannian) volume element induced on the attractor], any standard information-theoretic entropy [12] can be computed from δ.

Axiom-A systems are known to possess an (unique) invariant measure of physical significance [28] (compatible with the volume measure on the attractor). The property of being axiom A refers to the existence of a continuous and invariant splitting of the tangent space, at each point of the attractor, into stable and unstable directions [29]. This property is difficult (if not impossible) to check from an observed times series. However, a system with a sufficiently large number of degrees of freedom, and in a stationary state, can be regarded, for the purpose of computing macroscopic properties, as a smooth dynamical system with a transitive axiom-A global attractor [30].

C. Estimation of the transfer operator and invariant measures

Let $\psi : \mathbb{R}^n \to \mathbb{R}^n$ be differentiably invertible and $\mathbb{A} \subset \mathbb{R}^n$ a compact attractor with m denoting the induced volume measure on $\mathbb{A}$. Suppose that μ is a measure compatible with m, with support contained in $\mathbb{A}$ and having density δ with respect to m. The map ψ acts on the measure as $(\psi_* \mu)(K) := \mu(\psi^{-1}(K))$, for any measurable set $K \subset \mathbb{A}$. Thus, its density is modified as $(\psi_* \delta)(x) := |d_x \psi^{-1}| \delta \circ \psi^{-1}(x)$, $| \cdot |$ denoting the absolute value of the determinant. The linear map between functions

$$\mathcal{P}(f)(x) := |d_x \psi^{-1}| f \circ \psi^{-1}(x) \tag{5}$$

is known as the Perron-Frobenius (or transfer) operator associated to the map ψ. If μ is a ψ-invariant measure, then

$$\delta(x) = |d_x \psi^{-1}| \delta \circ \psi^{-1}(x) \,, \tag{6}$$

and thus ψ-invariant densities correspond to fixed points of $\mathcal{P}$. There is a rich literature on the approximation of the transfer operator and the estimation of invariant measures [14,15]. Ulam's method [31] approximates the transfer operator by a row stochastic Markov matrix acting on distributions defined over a given partition of $\mathbb{A}$. More specifically, let $\{B_1, \ldots, B_N\}$ be a partition of $\mathbb{A}$ into measurable sets and for each $1 \leqslant a \leqslant N$, let $\chi_a : \mathbb{R}^n \to \mathbb{R}$ be defined as $\chi_a(x) = 1$ if $x \in B_a$ and 0,

otherwise. Any (measurable) function $\rho : \mathbb{R}^n \to \mathbb{R}_+$ can be approximated as constant on each partition element, that is

$$\rho = \sum_{i=1}^{N} \frac{\int_{B_i} dm \, \rho}{m(B_i)} \chi_i \,, \tag{7}$$

where $m(B_i)$ denotes the volume of the partition element B_i and dm is the (Lebesgue) volume element. From this piecewise constant approximation and from Eq. (5) it follows that

$$\begin{aligned} \mathcal{P}(\chi_i) &= \sum_{j=1}^{N} \frac{1}{m(B_j)} \int_{B_j} dm \, |d_x \psi^{-1}| \, \chi_i \circ \psi^{-1} \, \chi_j \\ &= \sum_{j=1}^{N} \frac{\int_{\psi^{-1}(B_j)} dm \, \chi_i}{m(B_j)} \chi_j \\ &= \sum_{j=1}^{N} \frac{m(B_i \cap \psi^{-1}(B_j))}{m(B_j)} \chi_j \,. \end{aligned} \tag{8}$$

The linearity of $\mathcal{P}$ implies

$$\tilde{\rho} := \mathcal{P}(\rho) = \sum_{i,j} \frac{1}{m(B_j)} \rho_i \frac{m(B_i \cap \psi^{-1}(B_j))}{m(B_i)} \chi_j \,, \tag{9}$$

where $\rho_i := \int_{B_i} dm \, \rho$ is the measure of B_i according to the density ρ. Taking $\tilde{\rho}_i := \int_{B_i} dm \, \tilde{\rho}$ as the updated measure of B_i, one finds

$$\tilde{\rho}_j = \sum_{i=1}^{N} \rho_i \frac{m(B_i \cap \psi^{-1}(B_j))}{m(B_i)} \,. \tag{10}$$

The row stochastic matrix with entries

$$P_{ij}^{(N)} = \frac{m(B_i \cap \psi^{-1}(B_j))}{m(B_i)} \tag{11}$$

constitutes the Ulam's approximation to the transfer operator and it approximates how distributions defined over a given partition do change under the map generating the dynamics. Accordingly, its left invariant distribution ($\rho^{(N)} P^{(N)} = \rho^{(N)}$) corresponds to the approximation of the invariant density of the system subject to the partition. From the left invariant distribution of $P^{(N)}$, a measure on $\mathbb{A}$ can be defined as

$$\mu_N(K) = \sum_{i=1}^{N} \rho_i^{(N)} \frac{m(K \cap B_i)}{m(B_i)} \,. \tag{12}$$

In [15] it is shown that using a piecewise linear approximation of the map ψ, the above sequence of measures $\{\mu_N\}$ approaches a ψ-invariant measure as the partition gets infinitely refined (the maximum size of the sets in the partition approaches 0 as $N \to \infty$). In the following sections, we give a detailed description of the implementation of these approximations.

D. Computation of TE

Suppose X and Y are time series of two variables of a dynamical system with attractor $\mathbb{A}$. The attractor may be reconstructed using a generalized delay embedding from both time series as $(x^{(l)}(t), y^{(k)}(t))$ [32].

Where $x^{(l)}(t) = (x(t), \dots, x[t - (l-1)\tau])$ and $y^{(k)}(t) = (y(t), \dots, y[t - (k-1)\tau])$, for appropriate delay τ and embedding dimension $l + k$. Denote the resulting embedded attractor by $\tilde{\mathbb{A}}$. The transfer operator and the invariant measure may be approximated using Eqs. (11) and (12) in the $(x^{(l)}, y^{(k)})$ embedding space. Suppose $\tilde{\mu}$ is the invariant measure on $\tilde{\mathbb{A}}$ and $P(x^{(l)}, y^{(k)})$ is the corresponding density of $\tilde{\mu}$. The TE from $Y \to X$ is then

$$T_{Y \to X} = \int_{\tilde{\mathbb{A}}} dm \, P(x^{(l)}, y^{(k)}) \log_2 \frac{P(x^{(j)} | x^{(l-j)}, y^{(k)})}{P(x^{(j)} | x^{(l-j)})} \,. \tag{13}$$

Notice that this procedure can be easily extended to compute any of the standard information-theoretic measures by using the appropriate embedding. For instance, for the conditional TE, $T_{Y \to X | Z}$, one may use a generalized embedding of the form $(x^{(l)}(t), y^{(m)}(t), z^{(k)}(t))$ [33].

During the revision of this paper, we were made aware of the work by Bollt [17], who proposed to use Ulam's approximation to the transfer operator to estimate the transfer entropy between coupled systems in order to identify synchronization. In his work, Bollt interprets the transfer matrix, that constitutes the Ulam's approximation, as a conditional probability between states in the phase space and computes the TE using Bayes's rule. In our work, we follow a different strategy: we use a generalized embedding to approximate the transfer operator and the invariant distribution of this transfer operator is then interpreted as a joint probability on the phase space, from which TE is computed. Our method also differs from Bollt's in the use of a triangulation estimator for sparse time series, as described in the following section.

E. Numerical implementation

Suppose that $X = \{x_1, \dots, x_N\}$ and $Y = \{y_1, \dots, y_N\}$ are time series of two variables of some dynamical system generated by the map ψ, and that $T_{Y \to X}$ is to be computed. The collection of points $E = \{(x_{n+k}^{(j)}, x_n^{(l)}, y_n^{(r)})\}_{n=1}^{N}$ for $x_n^{(l)} = (x_n, \dots, x_{n-l+1})$ (analogously for $y_n^{(r)}$) is a reconstruction of the attractor for suitable time delay k and embedding dimension $j + l + r$ [32]. Several methods for estimating both parameters can be found in the literature [3,34]. Suppose for the moment that the transfer operator has already been approximated using Eq. (11) and an estimate for an invariant measure for E, μ has been obtained from Eq. (12). For convenience, we relabel the axes corresponding to $x_{n+k}^{(j)}, x_n^{(l)}, y_n^{(r)}$ as 1,2,3, respectively. We use a regular grid into (hyper)rectangular bins, say $\{C_i\}_{i=1,\dots,J}$, and uniquely decompose each bin index i into the triplet (i_1, i_2, i_3) (Appendix A for details). The integral expression for the TE, Eq. (2), can then be approximated as

$$T_{Y \to X} \simeq \sum_{i_1, i_2, i_3} m(C_i) P(i_1, i_2, i_3) \log_2 \frac{P(i_1 | i_2, i_3)}{P(i_1 | i_2)} \,, \tag{14}$$

where $m(C_i)$ is the Euclidean volume of the ith bin and $P(i_1, i_2, i_3) = \frac{\mu(C_i)}{m(C_i)}$, i.e., the density of μ over the bin C_i. Defining $\mu_{i_1 i_2 i_3} = \mu(C_i)$, one easily checks that

$$\sum_{i_1, i_2, i_3} m(C_i) P(i_1, i_2, i_3) \log_2 \frac{P(i_1 | i_2, i_3)}{P(i_1 | i_2)}$$

$$= -H(i_1, i_2, i_3) - H(i_2) + H(i_1, i_2) + H(i_2, i_3) \,, \tag{15}$$

where $H(i_1, i_2, i_3)$ denotes the Shannon entropy of the distribution $\mu_{i_1 i_2 i_3}$, $H(i_1, i_2)$ is the entropy of the marginal distribution $\sum_k \mu_{i_1 i_2 k}$, $H(i_2)$ corresponds to the Shannon entropy of the marginal distribution $\sum_{l,k} \mu_{l i_2 k}$, and so on.

To estimate the transfer operator and the invariant measure, we propose two different approaches depending on the length of the time series (the number of observations). Denote the set of points in the reconstructed attractor E by $\{p_n\}_{n=1}^N$.

1. Grid estimator

If the time series X and Y contain a sufficient number of observations [35], the transition matrix in Eq. (11) can be approximated by a coarse-grained estimation [36] as

$$P_{ij} \simeq \frac{\sharp\{p_n \mid \psi(p_n) \in C_j \cap p_n \in C_i\}}{\sharp\{p_m \mid p_m \in C_i\}}, \tag{16}$$

with $\sharp$ denoting the cardinal. In this case, the left invariant distribution ρ from Eq. (12) coincides with the measure of the bins, that is, $\rho_i = \mu(C_i) = \mu_{i_1 i_2 i_3}$. We clarify that the name grid estimator refers to the fact that the transfer operator is approximated using a partition into rectangular bins, as opposed to using a partition consisting of simplices, which we consider in the next section. The word grid does not imply that we use a visitation frequency estimator. For time series with a sufficient number of observations, however, the visitation frequency and the grid estimators converge to the same invariant distribution (Appendix C). The motivation for obtaining an estimate of the transfer operator is that it provides an approximation to the underlying map that has applications beyond the computation of TE. For the purpose of this study, and as we show in the following sections, computing TE from the transfer operator is advantageous for sparse and noisy time series.

2. Triangulation estimator

For time series with fewer observations, the estimation of the transfer operator using Eq. (16) might become inaccurate. Actually, the transfer matrix obtained with the grid method for time series with few observations might fail to be Markovian. This is because the bin containing the last point in the embedding might not contain any other point. In that case, such a bin is a sink of information. We thus adopt the method developed in [15], by which the reconstructed attractor E is triangulated into simplices [37] (the vertices of each simplex being points from the embedding E). Suppose $\{S_1, \ldots, S_N\}$ is such a triangulation. The map ψ is then approximated by a linear map $\tilde{\psi}$ on each simplex such that if $\{p_{a_0}, \ldots, p_{a_d}\}$ are the vertices of the simplex S_a, then $\{p_{a_0+1}, \ldots, p_{a_d+1}\}$ are the vertices of its image under the map $\tilde{\psi}(S_a)$. The transfer matrix is obtained as

$$P_{ab} = \frac{m(S_b \cap \tilde{\psi}(S_a))}{m(\tilde{\psi}(S_a))}, \tag{17}$$

and according to Eq. (12), the measure of the simplex a is ρ_a. To compute the simplex intersection volume we follow a direct approach outlined in Appendix D, although several methods for polytope volume computation can be found in the literature [38]. Once we have obtained the invariant distribution over the simplices, finding $\mu(C_i)$ exactly is computationally rather demanding. Instead, we estimate the measure of each bin by evenly sampling each simplex of the triangulation with M_s points, and then assuming that each sampling point carries a fraction $1/M_s$ of the measure of the simplex they belong to. Thus, if the set of sampling points belonging to the bin C_i is formed by N_a points from simplex a, for $a = 1, \ldots, N$ (possibly with some N_a being zero), its measure is estimated as

$$\mu(C_i) \simeq \sum_{a=1}^N \frac{N_a}{M_s} \rho_a. \tag{18}$$

Using this sampling enables a virtually unlimited number of points to estimate the density of the measure. Notice that we do not introduce any bias by doing so (provided the sampling is even over each simplex) because the sampling points do not contain any information beyond that encoded in ρ.

Using embedding dimension 3 and time series with a few hundred observations, the number of simplices with positive measure out of the triangulation is on the order of hundreds (Appendix F). For the examples we study here, TE becomes independent of the (total) number of sampling points beyond ~ 5000 [39]. Therefore, M_s will be on the order of tens. We subsample the simplices using a shape preserving simplex splitting routine developed in [40]. This algebraic procedure uses an edgewise splitting factor r and splits a simplex in dimension d into r^d subsimplices, all with the same volume. We use the centroids of the subsimplices resulting from the splitting as the sampling points.

IV. EXAMPLE DYNAMICAL SYSTEMS

We apply our TE estimators to time series generated by coupled instances of well known dynamical systems. Several realizations of the time series are generated from randomly chosen initial values. In all cases, TE in each direction is computed using a bin size adapted to the number of points available [41] and to the size of the reconstructed attractor (Appendix A). We will generically denote the TE computed from time series X and Y, as $T_{X \to Y}$ and $T_{Y \to X}$, where $X \to Y$ corresponds to the direction of the coupling (in the case of unidirectional coupling) or to the direction of the strongest coupling (in the case of bidirectional coupling). In both cases, one expects $T_{X \to Y} - T_{Y \to X} > 0$. To check the ability of our methods to detect the direction of the coupling between time series, we study the dependence of the average values of $T_{X \to Y}$ and $T_{Y \to X}$ across realizations, on the number of observations in the time series. We also study the response of our estimators to the strength of the coupling and to observational and dynamical noise. We compare our results with those obtained with the kNN [9] and the KDE [10] estimators. Although these methods, strictly speaking, estimate mutual information, TE can be computed from the identity in Eq. (3). Because we are primarily interested in the sensitivity of our new estimators to noise and time series length, and not the absolute value of the TE, we do not apply any bias correction to the estimators [42–44]. In Appendix A we detail the embedding used for computing TE for each dynamical system example.

A. Dynamical and measurement noise

Measurement noise is independently added to each time series as follows: given a time series X, measurement noise of intensity ϵ is simulated by adding independently to each observation of X a random value drawn from a Gaussian distribution with zero mean and standard deviation $\epsilon\sigma$, where σ is the standard deviation corresponding to the values in X, and $0 \leqslant \epsilon \leqslant 1$. For instance, a measurement noise of intensity $\epsilon = 0.1$ will be referred to as 10% measurement noise. Dynamical noise is simulated for each system as explained in the following sections [Eqs. (21), (24), and (31)].

B. Unidirectionally coupled logistic maps (UCLM)

The logistic map is one of the hallmarks of chaotic behavior in dynamical systems [24]. It was originally proposed by May as a model for population growth. Here, we consider two logistic maps unidirectionally coupled, given by

$$x(n+1) = 3.78\, x(n)[1 - x(n)], \tag{19}$$

$$y(n+1) = 3.66\, f_n[1 - f_n],$$
$$f_n = \frac{y(n) + c\, x(n)}{1 + c}. \tag{20}$$

We also allow for the presence of a moderate level of dynamical noise by modifying the sequence f_n as

$$\tilde{f}_n = \frac{y(n) + c\,[x(n) + \epsilon\,\xi]}{1 + c(1 + \epsilon)}, \tag{21}$$

where ξ is a random number drawn from [0,1] with a flat distribution and $0 \leqslant \epsilon \leqslant 0.5$.

C. Bidirectionally coupled logistic maps (BCLM)

In this case we consider the system generated by the map

$$x(n+1) = \frac{3.78\, x(n)[1 - x(n)] + 0.03\, y(n)^2}{1.03}, \tag{22}$$

$$y(n+1) = \frac{3.66\, g_n[1 - g_n] + c\, x(n)^2}{1 + c},$$
$$g_n = \frac{y(n) + 0.06\, x(n)}{1.06}. \tag{23}$$

As in the UCLM case, we also allow for the presence of dynamical noise by modifying the sequence g_n as

$$\tilde{g}_n = \frac{y(n) + 0.06\,[x(n) + \epsilon\,\xi]}{1 + 0.06(1 + \epsilon)}, \tag{24}$$

where ξ is a random number drawn from [0,1] with a flat distribution and $0 \leqslant \epsilon \leqslant 0.5$.

D. Coupled Rössler-Lorenz system

Historically relevant in the study of chaos are also the Lorenz system [23] and the Rössler system [45]. The former was developed by Lorenz in 1963 as a simplified model of viscous fluid flow. Rössler proposed his system in 1976 as a simpler version of the Lorenz attractor in order to more easily study its chaotic properties. Here, we study a coupled
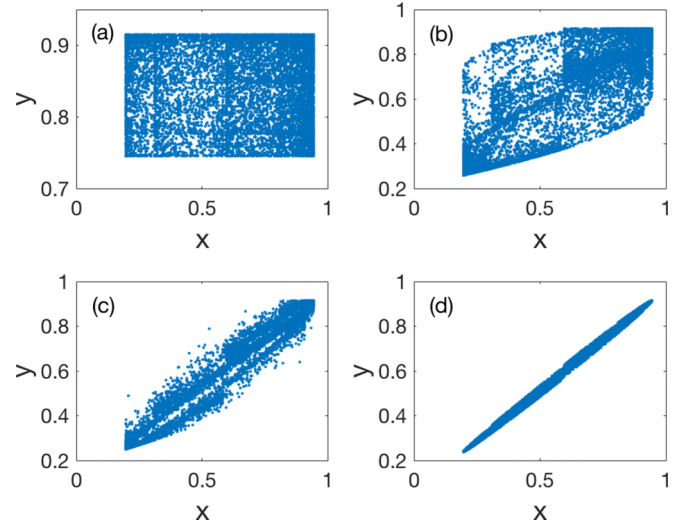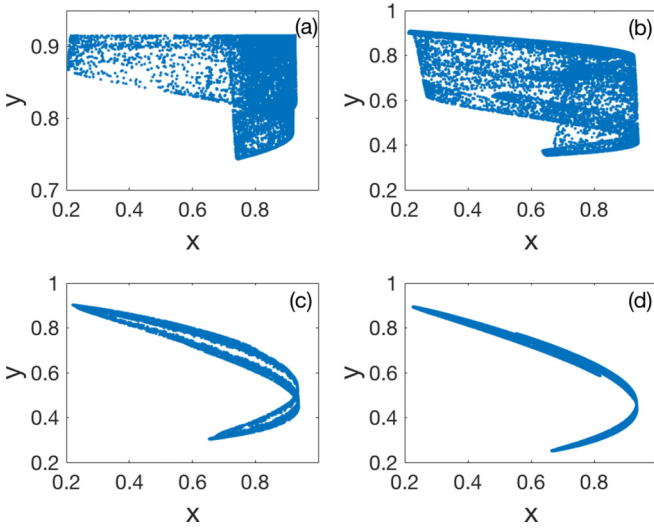

FIG. 1. x-y phase space generated from 10^4-point long orbits of the UCLM without noise and for different values of the coupling constant: $c = 0$ (a), $c = 0.4$ (b), $c = 0.6$ (c), and $c = 1$ (d).

version of both systems also studied in [41,46], generated by the vector field

$$\dot{x}_1 = -6\,(x_2 + x_3), \tag{25}$$

$$\dot{x}_2 = 6\,(x_1 + 0.2x_2), \tag{26}$$

$$\dot{x}_3 = 6\,[0.2 + x_3\,(x_1 - 5.7)], \tag{27}$$

$$\dot{y}_1 = 10\,(y_2 - y_1), \tag{28}$$

$$\dot{y}_2 = y_1\,(28 - y_3) - y_2 + c\,(x_2)^2, \tag{29}$$

$$\dot{y}_3 = y_1\,y_2 - (8/3)\,y_3. \tag{30}$$

In this case, the dynamical noise is introduced by modifying the coupling term in Eq. (29) above as

$$c\,(x_2)^2(1 + \epsilon\,\xi)^2, \tag{31}$$

where ξ is a random number drawn from $[-1, 1]$ with a flat distribution and $0 \leqslant \epsilon \leqslant 0.5$.

E. Synchronization

The synchronization threshold for the coupled instances of the logistic maps can be easily estimated from the shape of the attractor itself. It turns out that for both cases the synchronization seems to take place around $c \sim 1$. In the UCLM case, the attractor clearly shrinks to the diagonal for $c \sim 1$ (Fig. 1). For the BCLM, however, one has a generalized synchronization and hence the attractor does not collapse to the diagonal when synchronization sets in (Fig. 2). In fact, by taking the limit when $c \to \infty$, the BCLM system reduces to the new dynamical system

$$x(n+1) = 3.67\, x(n)[1 - x(n)] + 0.029\, y(n)^2, \tag{32}$$

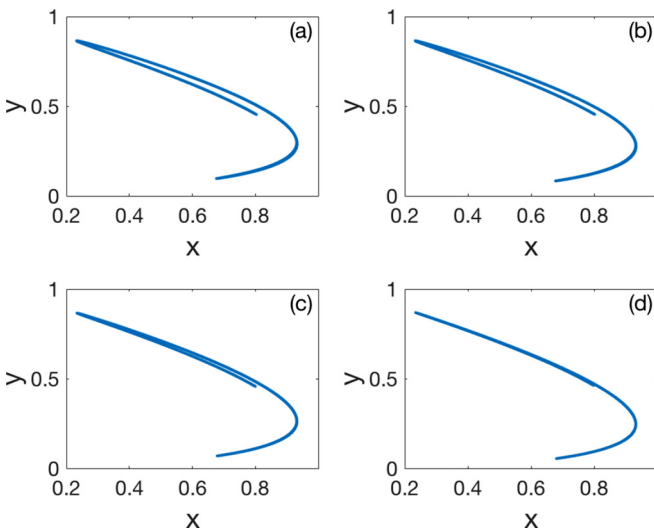$$y(n+1) = x(n)^2. \tag{33}$$

FIG. 2. x-y phase space generated from 10^4-point long orbits of the BCLM without noise and for different values of the coupling constant: $c = 0$ (a), $c = 0.2$ (b), $c = 0.5$ (c), and $c = 1$ (d).
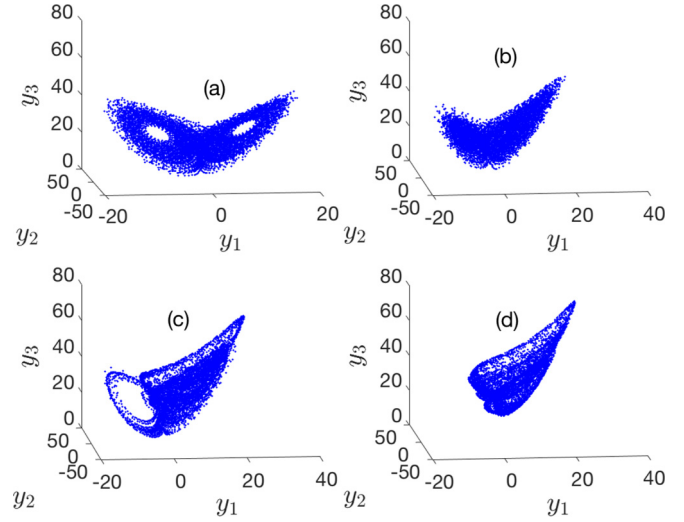

FIG. 4. 3D sections of the Rössler-Lorenz phase space generated from 10^4-point long orbits without noise and for different values of the coupling: $c = 0$ (a), $c = 1$ (b), $c = 2.5$ (c) and $c = 3.5$ (d).

As c increases, the attractor generated by the BCLM approaches the attractor obtained with the system (32) and (33) (Fig. 3).

For the case of the Rössler-Lorenz system, the generalized synchronization seems to take place around $c \gtrsim 2$ [46]. Indeed, one can observe a great distortion of the usual butterfly shape Lorenz attractor for $c > 2.5$ (Fig. 4).

V. SENSITIVITY OF TE TO TIME SERIES LENGTH

In this section we study the dependence of the TE computed using our estimators on the number of observations in the time series.

FIG. 3. x-y phase space generated from 10^4-point long orbits of the BCLM without noise and for high values of the coupling constant: $c = 10$ (a), $c = 15$ (b), $c = 30$ (c), and $c = \infty$ [the limiting system given by Eqs. (32) and (33)] (d).

A. Coupled logistic maps

1. Data-rich time series

In both UCLM and BCLM systems, we compute the TE as a function of the time series length in the range 1000 to 5000 observations and with a low level (10%) of measurement noise. For the UCLM case the coupling constant is set to $c = 0.4$, while for the BCLM instance we set $c = 0.2$.

In both cases of coupled logistic maps, the grid and the kNN estimators seem to be the least sensitive to the number of observations in the time series [Figs. 5(a) and 5(c) and 6(a) and 6(c)], while the KDE estimator shows a mildly stronger dependence on the time series length. For the grid estimator,
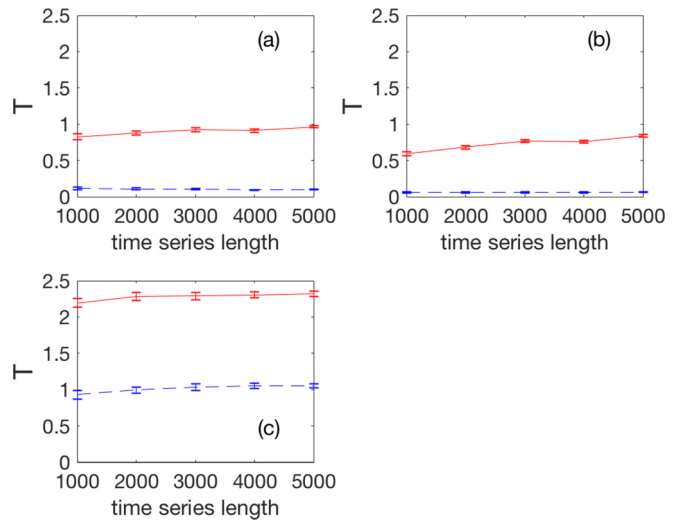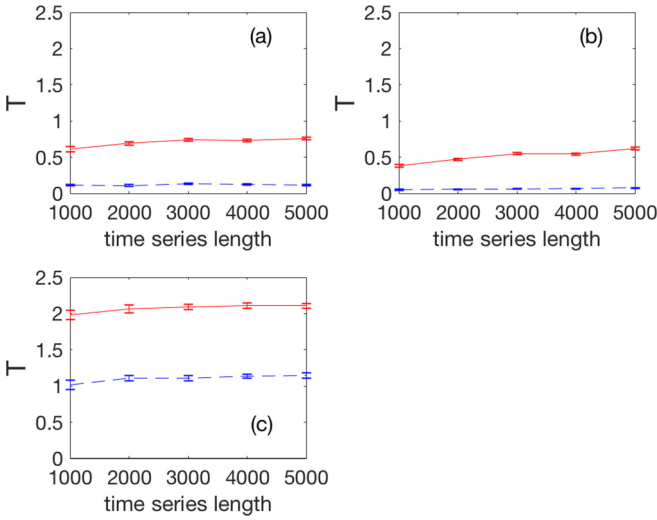

FIG. 5. Sensitivity of TE to time series length for the UCLM with coupling constant $c = 0.4$ and 10% measurement noise, using data rich time series. Values are the mean and standard deviation of $T_{x \to y}$ (red line) and $T_{y \to x}$ (dashed blue line) over 50 realizations, computed with the grid estimator (a), the KDE estimator (b) and the kNN estimator (c).

FIG. 6. Sensitivity of TE to time series length for the BCLM with coupling constant $c = 0.2$ and 10% measurement noise, using data-rich time series. Values are the mean and standard deviation of of $T_{x \to y}$ (red line) and $T_{y \to x}$ (dashed blue line) over 50 realizations, computed with the grid estimator (a), the KDE estimator (b), and the kNN estimator (c).
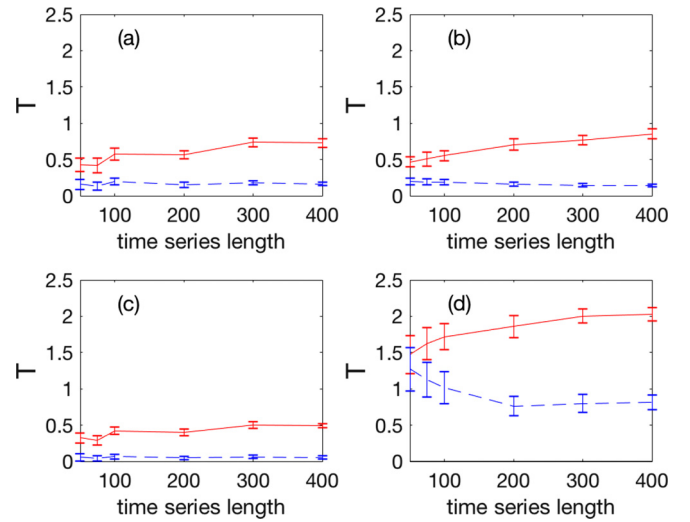

FIG. 7. Sensitivity of TE to time series length for the UCLM with coupling constant $c = 0.4$ and 10% measurement noise, using sparse time series. Values are the mean and standard deviation of of $T_{x \to y}$ (red line) and $T_{y \to x}$ (dashed blue line) over 50 realizations, computed with the grid estimator (a), the triangulation estimator (b), the KDE estimator (c), and the kNN estimator (d).

and using the adapted bin size described in Appendix A, TE for both UCLM and BCLM saturates to a fixed value for time series with more than ∼5000 observations.

2. Sparse time series

In this section we check the ability of our estimators to yield directional asymmetry in the TE for sparse data in the range 50 to 400 observations and adding 10% measurement noise. We also test the stability of the TE against the time series length. For both UCLM and BCLM systems, our estimators yield the correct TE asymmetry ($T_{x \to y} > T_{y \to x}$) even for time series sparsely sampled with 50 values [Figs. 7(a) and 7(b) and 8(a) and 8(b)]. In the case of the BCLM system, the triangulation estimator outcompetes the rest of the estimators at detecting asymmetry in the TE (in the expected direction) for very sparse time series, with less than 100 observations (Fig. 8). As for the sensitivity of the TE on the number of observations, our estimators yield relatively stable TE in the range 100–300 observations [Figs. 7(a) and 7(b) and 8(a) and 8(b)]. The KDE estimator seems to be the least sensitive to the number of observations while the kNN estimator shows the highest sensitivity [Figs. 7(c) and 7(d) and 8(c) and 8(d)].

B. Coupled Rössler-Lorenz system

Due to the dimensionality of this system, relatively data-rich time series are required to obtain reliable TE estimates, hence, we use time series with 2000 to 10 000 observations. Again we compare the TE computed using our grid transfer operator estimator with that of the kNN and the KDE estimators. The triangulation approach becomes prohibitively time demanding for high embedding dimension $d \geqslant 5$ (Appendix F). The grid estimator is the least sensitive to the number of observations in the time series [Fig. 9(a)] while the kNN estimator arguably is the most sensitive [Fig. 9(c)]. For the

grid estimator, and using the adapted bin size described in Appendix A, TE for the Rössler-Lorenz system saturates to a fixed value for time series with more than ∼12 000 observations. For completeness, we also applied the triangulation estimator to the Rössler-Lorenz system using 3D embeddings $[(x_2(i + 1), x_2(i), y_2(i))$, for computing $T_{y_2 \to x_2}$ and $(y_2(i + 1), y_2(i), x_2(i))$, for computing $T_{x_2 \to y_2}]$ and time series with 50–400 observations. The estimator detects a marginal causal
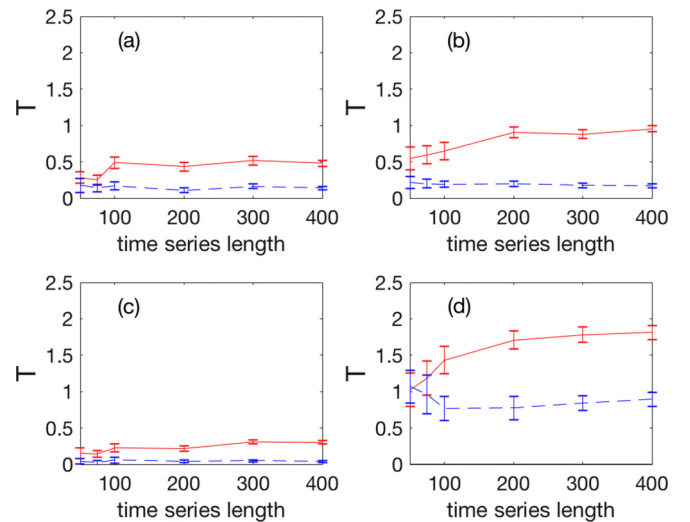

FIG. 8. Sensitivity of TE to time series length for the BCLM with coupling constant $c = 0.2$ and 10% measurement noise, using sparse time series. Values are the mean and standard deviation of of $T_{x \to y}$ (red line) and $T_{y \to x}$ (dashed blue line) over 50 realizations, computed with the grid estimator (a), the triangulation estimator (b), the KDE estimator (c), and the kNN estimator (d).
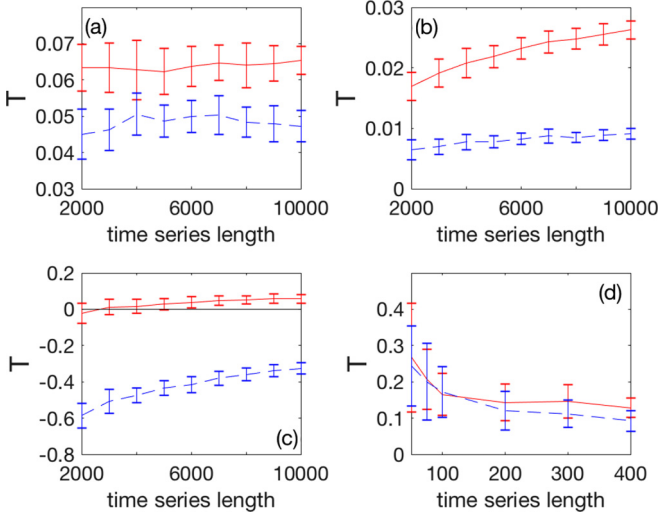
FIG. 9. Sensitivity of TE to time series length for the Rössler-Lorenz system with coupling constant $c = 1.5$ and 10% measurement noise. Values are the mean and standard deviation of $T_{x_2 \to y_2}$ (red line) and $T_{y_2 \to x_2}$ (dashed blue line) over 50 realizations, computed with the grid estimator (a), the KDE estimator (b), and the kNN estimator (c). We also include the TE computed using the triangulation estimator with 3D embeddings and using sparse time series (d).

signal for time series with more than ~ 300 observations [Fig. 9(d)].

We note that the computational requirements of the triangulation estimator become prohibitive for embedding dimensions beyond ~ 5. The current implementation of this estimator is based on computing exact simplex volume intersections (Appendix D) which suffers from the curse of dimensionality. There is room for optimization, however, through more efficient (approximate) polytope intersection algorithms. Alternatively, a coarse-grained sampling of the simplices might also be used to estimate the volume intersection. We have implemented both exact and approximate volume intersection routines in our CAUSALITYTOOLS.JL [47] JULIA[48] package, which also provides an implementation of the grid estimator. We also suggest a more efficient alternative in Appendix E. We leave a comprehensive investigation of these optimizations for future work.

VI. DEPENDENCE OF TE ON THE COUPLING CONSTANT AND ITS RESPONSE TO NOISE

In this section we study the dependence of the TE computed using our estimators on the coupling constant and its response to observational and dynamical noise. We compare the results with the above standard estimators. Because our main interest in this study is the estimation of TE from sparse time series, we fix the time series length to 100 observations for the coupled logistic maps (both UCLM and BCLM) and to 1000 observations for the Rössler-Lorenz system.

A. Coupled logistic maps

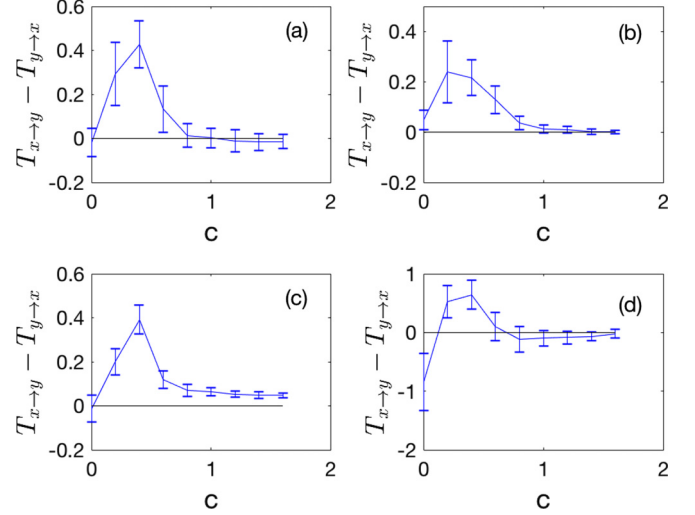For the coupled logistic maps, we compute the TE for values of the coupling constant in the range 0 to 1.6 in steps of

FIG. 10. Dependence of $T_{x \to y} - T_{y \to x}$ on the coupling constant for the noise free UCLM. Values are the mean and standard deviation of $T_{x \to y} - T_{y \to x}$ over 50 realizations, computed with the grid estimator (a), the triangulation estimator (b), the KDE estimator (c), and the kNN estimator (d).

0.2. We also include dynamical noise as well as measurement noise with intensities ranging from 0 to 0.5 in steps of 0.1. For the case of UCLM without noise, the asymmetry $T_{x \to y} - T_{y \to x}$, computed with our estimators, starts at zero (or a very small value) for $c = 0$, it then increases up to a maximum value around $c \sim 0.2$–0.4 and decreases back to zero as the synchronization triggers for $c \gtrsim 1$ [Figs. 10(a) and 10(b)]. In the UCLM system, synchronization causes the evolution of y to closely follow the evolution of x (Fig. 1). Accordingly, one expects $T_{x \to y} \to T_{y \to x}$ as c increases beyond 1. However, when dynamical noise is added, the effect of the synchronization is attenuated [Figs. 11(a) and 11(b)]. Dynamical noise may be interpreted as a hidden process affecting the system. It
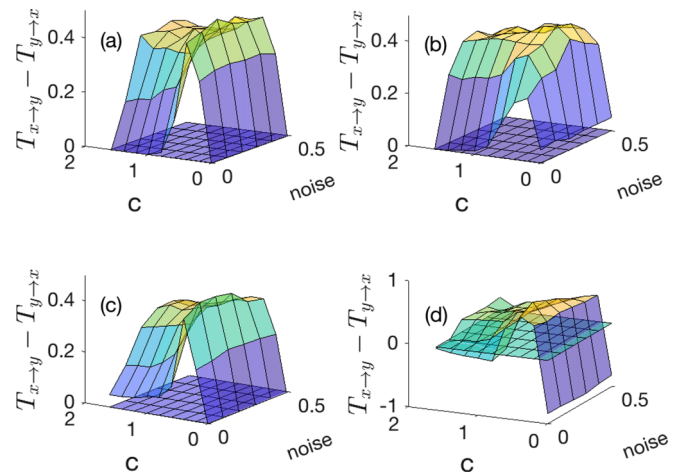

FIG. 11. Dependence of $T_{x \to y} - T_{y \to x}$ on the coupling constant and dynamical noise level for the UCLM. Values are the mean and standard deviation of $T_{x \to y} - T_{y \to x}$ over 50 realizations, computed with the grid estimator (a), the triangulation estimator (b), the KDE estimator (c) and the kNN estimator (d).
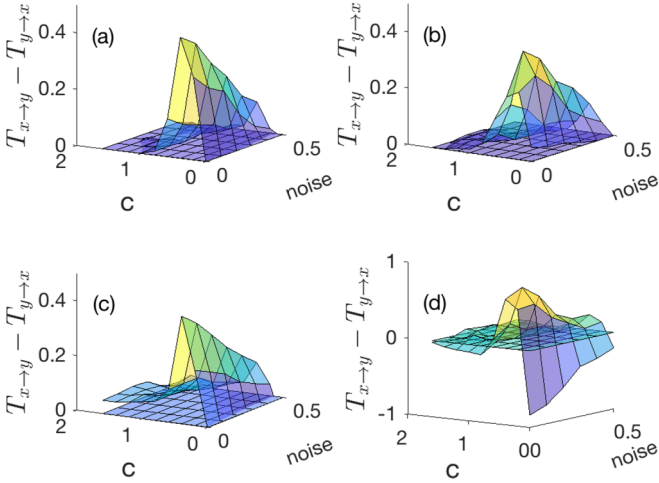
FIG. 12. Dependence of $T_{x \to y} - T_{y \to x}$ on the coupling constant and measurement noise level for the UCLM. Values are the mean and standard deviation of $T_{x \to y} - T_{y \to x}$ over 50 realizations, computed with the grid estimator (a), the triangulation estimator (b), the KDE estimator (c) and the kNN estimator (d).

is then expected that the synchronization effect breaks down for sufficiently intense dynamical noise. On the other hand, when measurement noise is added, the asymmetry $T_{x \to y} - T_{y \to x}$ decreases with increasing noise intensity (Fig. 12), as the effect of the coupling is masked by the noise. Remarkably, the ability of the triangulation estimator to detect the correct directionality of the coupling enhances for low to moderate levels of measurement noise and weak coupling ($c \sim 0.2$–0.4) [Fig. 12(b)]. This finding is congruent with the concept of random perturbation approximation to the map, on which the triangulation estimator is based [15]. In the case of BCLM, synchronization does not decrease the asymmetry between $T_{x \to y}$ and $T_{y \to x}$. Synchronization reduces the BCLM system to the map in Eqs. (32) and (33). In that limiting case, the coupling in the $x \to y$ direction is much stronger than in the opposite direction. Both the grid and the triangulation estimators do capture this saturation of the asymmetry in the TE for high values of the coupling constant [Figs. 13(a) and 13(b)]. In contrast, the KDE estimator yields a monotonically decreasing asymmetry $c \gtrsim 1$ [Fig. 13(c)] whereas the kNN estimator shows a less obvious decrease [Fig. 13(d)]. When dynamical noise is included into the BCLM system, the asymmetry $T_{x \to y} - T_{y \to x}$ responds similarly to that of the UCLM system (Fig. 14). Observational noise, however, causes the TE asymmetry to decreases with increasing levels of noise, as expected. (Fig. 15). Although less obvious than for the UCLM system, the triangulation estimator also shows a local maximum in TE asymmetry for weak coupling and moderate observational noise [Fig. 15(b)].

B. Rössler-Lorenz system

For the Rössler-Lorenz system, we compute the TE for values of the coupling constant in the range 0 to 4 in steps of 0.2 also including observational and dynamical noise with intensities ranging from 0 to 0.5 in steps of 0.1. The asymmetry $T_{x_2 \to y_2} - T_{y_2 \to x_2}$ computed with the grid estimator
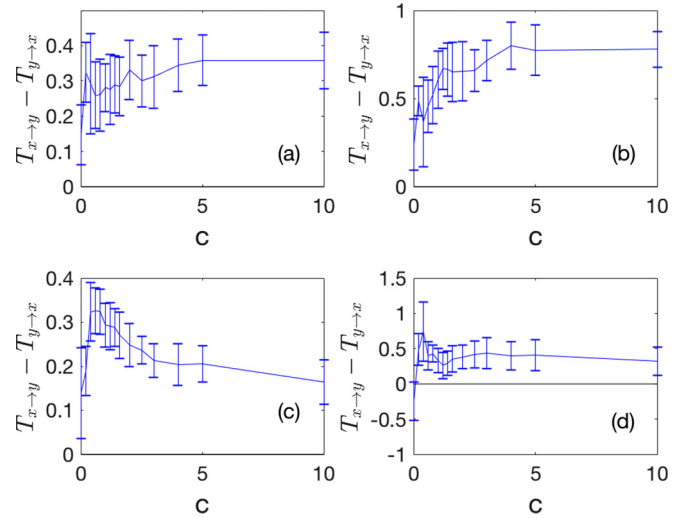

FIG. 13. Dependence of $T_{x \to y} - T_{y \to x}$ on the coupling constant for the noise free BCLM. Values are the mean and standard deviation of $T_{x \to y} - T_{y \to x}$ over 50 realizations, computed with the grid estimator (a), the triangulation estimator (b), the KDE estimator (c) and the kNN estimator (d).

saturates for high values of dynamical noise and coupling constant [Fig. 16(d)], a trait also seen for the kNN estimator [Fig. 16(c)]. The TE asymmetry computed using the grid estimator increases for strong coupling and high levels of dynamical noise [Fig. 16(a)]. However, the asymmetry saturates for higher levels of dynamical noise [Fig. 16(d)]. With increasing levels of measurement noise, all estimators yield a decreasing value for the TE asymmetry (Fig. 17). Remarkably, nonetheless, the TE asymmetry computed with the grid estimator only starts to decrease substantially once the level of measurement noise goes beyond 50% [Fig. 17(d)]. These findings suggest that our grid estimator is robust to both observational and dynamical noise.
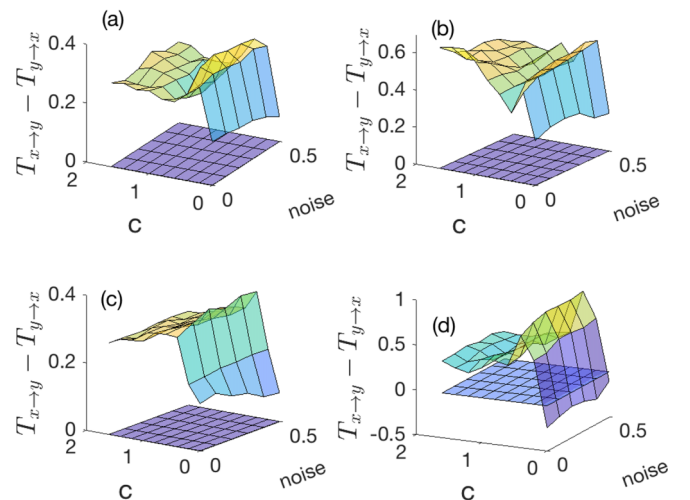

FIG. 14. Dependence of $T_{x \to y} - T_{y \to x}$ on the coupling constant and dynamical noise level for the BCLM. Values are the mean and standard deviation of $T_{x \to y} - T_{y \to x}$ over 50 realizations, computed with the grid estimator (a), the triangulation estimator (b), the KDE estimator (c), and the kNN estimator (d).
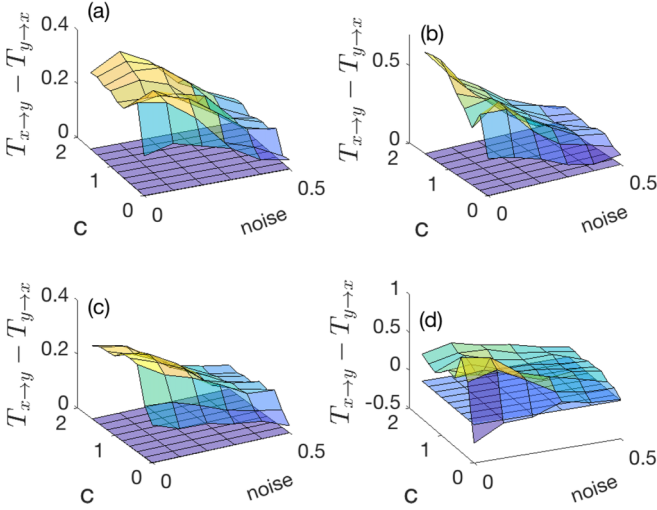
FIG. 15. Dependence of $T_{x \to y} - T_{y \to x}$ on the coupling constant and measurement noise level for the BCLM. Values are the mean and standard deviation of $T_{x \to y} - T_{y \to x}$ over 50 realizations, computed with the grid estimator (a), the triangulation estimator (b), the KDE estimator (c), and the kNN estimator (d).

VII. DIRECT VS INDIRECT COUPLING

At the end of Sec. III D, we mentioned that our approach to estimate TE can be easily extended to compute conditional transfer entropy between three time series. As an example of this, we apply the grid estimator to a chain of three coupled Lorenz systems, studied in [49]. The flow is generated by the vector field

$$\dot{x}_1 = 10(y_1 - x_1), \tag{34}$$

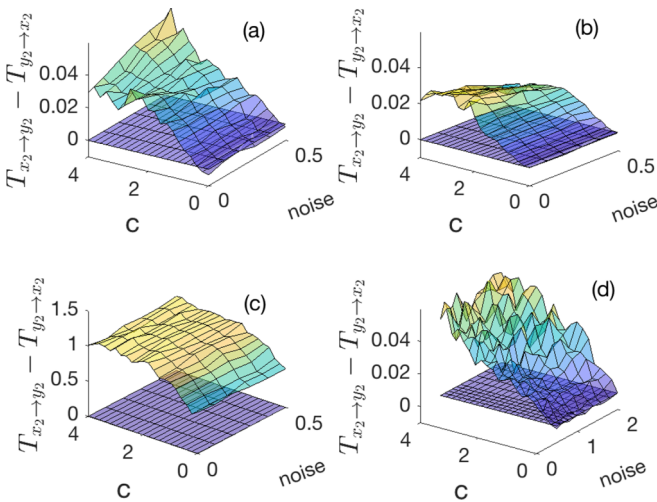$$\dot{y}_1 = x_1(28 - z_1) - y_1, \tag{35}$$

FIG. 16. Dependence of $T_{x_2 \to y_2} - T_{y_2 \to x_2}$ on the coupling constant and dynamical noise level for the Rössler-Lorenz system. (a) Mean value of $T_{x_2 \to y_2} - T_{y_2 \to x_2}$ over 50 realizations computed with the grid estimator; (b) the same for the KDE estimator; (c) using the kNN estimator; (d) using the grid estimator and extending level of dynamical noise up to 2.
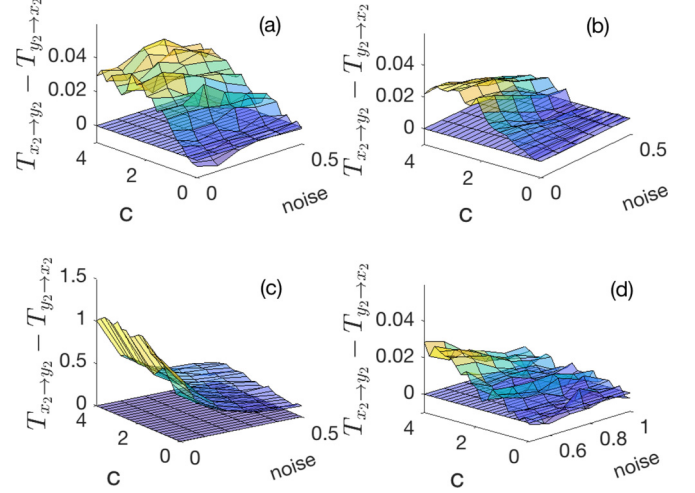

FIG. 17. Dependence of $T_{x_2 \to y_2} - T_{y_2 \to x_2}$ on the coupling constant and measurement noise level for the Rössler-Lorenz system. (a) Mean value of $T_{x_2 \to y_2} - T_{y_2 \to x_2}$ over 50 realizations computed with the grid estimator; (b) the same for the KDE estimator; (c) using the kNN estimator; (d) using the grid estimator and extending the level of measurement noise up to 100%.

$$\dot{z}_1 = x_1 y_1 - 8/3 \, z_1, \tag{36}$$

$$\dot{x}_i = 10(y_i - x_i) + c \, (x_{i-1} - x_i), \tag{37}$$

$$\dot{y}_i = x_i(28 - z_i) - y_i, \tag{38}$$

$$\dot{z}_i = x_i y_i - 8/3 \, z_i, \tag{39}$$

with $i = 2, 3$. The direct coupling chain is $x_1 \to x_2 \to x_3$. We use coupling constant values in the range 0 to 8 in steps of 0.4 (according to [49], the full synchronization takes place for $c > 8$). For each instance of the coupling constant, we generate 50 orbits starting at randomly chosen initial conditions and consisting of 10^4 observations. The data generation for this system and the delay embeddings used to compute TE are specified in Appendix A. Our grid estimator detects the direct coupling $x_1 \to x_2$ for $c \gtrsim 2$ [Fig. 18(a)] and the coupling $x_2 \to x_3$ is detected for $c \gtrsim 4$ [Fig. 18(b)]. In addition, the grid estimator detects the indirect coupling $x_1 \to x_3$ for $c \gtrsim 4$ [Fig. 18(c)]. On the other hand, when the transfer entropy is conditioned on the mediating variable x_2, the TE for the indirect coupling $x_1 \to x_3$ vanishes [Fig. 18(d)], which indicates that our method holds some promise for detecting indirect coupling.

VIII. CONCLUSIONS

In this work we propose the computation of transfer entropy (TE) between time series corresponding to variables of some dynamical system, based on a numerical approximation of the Perron-Frobenius operator (transfer operator) associated to the map (or vector field) giving rise to the dynamics. More specifically, the TE is computed using the invariant distribution of the transfer operator. Depending on the number of observations in the time series and the embedding
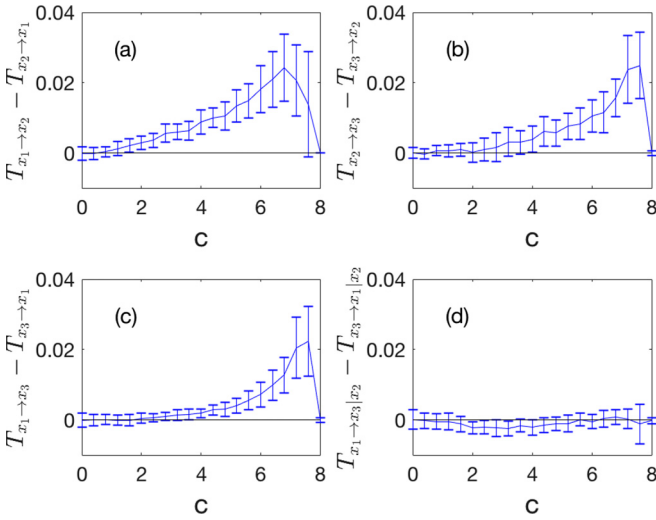
FIG. 18. Distinguishing direct from indirect coupling for the chain of coupled Lorenz systems. Values are the mean and standard deviation of TE asymmetry over 50 realizations of the chain of coupled Lorenz systems [Eqs. (34)–(39)]. TE asymmetry for the direct coupling $x_1 \rightarrow x_2$ (a), direct coupling $x_2 \rightarrow x_3$ (b), indirect coupling $x_1 \rightarrow x_3$ (c), and conditional TE for the indirect coupling given the mediating variable (d).

dimension, we propose two methods to estimate the transfer operator. For sparse time series (less than a few hundred points) and low embedding dimension, we use a triangulation of the delay reconstructed attractor to estimate the transfer operator, whereas for data-rich time series (thousands of points) or high embedding dimension we estimate TE using a faster rectangular grid approach. The TE computed using our estimators shows robustness to both observational and dynamical noise, even for a high dimensional system such as the Rössler-Lorenz system. Our results suggest that our estimators of TE are relevant for the detection of causal directionality between sparse and noisy time series, which are commonly encountered in many disciplines.

Computer code required for reproducing the numerical results presented in this work is available in our CAUSALITY-TOOLS.JL JULIA package [47].

APPENDIX A: NUMERICAL IMPLEMENTATION DETAILS

1. Generating time series and embedding

a. Logistic maps

To generate the time series for the UCLM and BCLM systems, the variables x and y are sampled every second iterate, after a lapse of 10^3 iterations. For the computation of $T_{x \rightarrow y}$, we used the embedding $(y(i+1), y(i), x(i))$ while the computation of $T_{y \rightarrow x}$ was done with the embedding $(x(i+1), x(i), y(i))$.

b. Rössler-Lorenz system

The system of equations (25)–(30) is solved using a fourth order Runge-Kutta routine with time step $dt = 0.005$. The time series are generated by recording the variables every 6 time steps of integration and after an initial lapse of 500 steps, to avoid transients. The embeddings we used in this case were $(y_2(i+3), y_2(i+2), y_2(i+1), x_2(i+2), x_2(i+1), x_2(i))$ for $T_{x_2 \rightarrow y_2}$ and $(x_2(i+3), x_2(i+2), x_2(i+1), y_2(i+2), y_2(i+1), y_2(i))$ for $T_{y_2 \rightarrow x_2}$.

c. Chain of coupled Lorenz systems

The system of equations (34)–(39) is also solved using a fourth order Runge-Kutta method with the same integration step as for the Rössler-Lorenz system. The variables x_1, x_2, and x_3 are sampled using the same sampling time and initial lapse. To compute $T_{x_i \rightarrow x_j}$, with $i \neq j \in \{1, 2, 3\}$, we use the embedding $(x_j(t+3), x_j(t+2), x_j(t+1), x_i(t+2), x_i(t+1), x_i(t))$ while for the estimation of the conditional transfer entropy $T_{x_1 \rightarrow x_3 | x_2}$ we used the embedding $(x_3(t+3), x_3(t+2), x_1(t+2), x_2(t+2), x_2(t+1), x_2(t))$. For the case $T_{x_3 \rightarrow x_1 | x_2}$, the embedding is obtained as in the case of $T_{x_1 \rightarrow x_3 | x_2}$ by simply interchanging the roles of x_3 and x_1.

2. Bin sizes

The choice of the size of the intervals along each axis is adapted to the size of the attractor and the number of points available. Following [41], if N is the number of points furnishing the (embedded) attractor and d is the embedding dimension, the number of intervals along each axis is taken as $N_{\text{int}} = \min\{\text{ceil}(N^{1/(d+1)}), n_{\max}\}$, where ceil($\cdot$) denotes the ceiling and $n_{\max}$ is taken to be 9 for $d = 3$ and 4 for $d = 6$. If $\{p_i\}$ is the set of points furnishing the attractor (it could either be the actual set of points in the embedding or the result of the sampling of the simplices in the triangulation), the size of the intervals along the ath axis is chosen as follows: Let $O_a = (1 - \frac{1}{10 N_{\text{int}}}) \min\{(p_i)_a | 1 \leqslant i \leqslant N\}$ and $T_a = (1 + \frac{1}{10 N_{\text{int}}}) \max\{(p_i)_a | 1 \leqslant i \leqslant N\}$. O will be referred to as the origin of the attractor. The interval size along the ath axis is determined as $\epsilon_a = (T - O)_a / N_{\text{int}}$.

The results for the KDE estimator are obtained using the minimum interval size for each case, that is, $\min\{\epsilon_a\}$.

3. Constructing the binning

Suppose $E = \{p_1, \ldots, p_N\} \subset \mathbb{R}^d$ is the set of points furnishing the reconstructed attractor (in the case of the grid estimator) or the set of final sampling points (in the case of the triangulation estimator). Let $\epsilon = (\epsilon_a)$ and $O = (O_a)$ be the bin size and the origin of the attractor (see the previous section). Call $(x_1, \ldots, x_d)$ the coordinate axes on the embedding space, and generically denote as $A_{n+1} := (x_1, \ldots, x_{n_1})$, $A_n := (x_{n_1+1}, \ldots, x_{n_2})$, and $B_n := (x_{n_2+1}, \ldots, x_d)$, the variables on which the transfer entropy $T_{B \rightarrow A}$ is computed. Each point $p_l \in E$ is assigned a unique triplet of integer tuples $I_l = (i_1^l, i_2^l, i_3^l)$, with $i_1^l = (j_1^l, \ldots, j_{n_1}^l)$, $i_2^l = (k_{n_1+1}^l, \ldots, k_{n_2}^l)$, and $i_3^l = (m_{n_2+1}^l, \ldots, m_d^l)$, and such that $O_a + (j_a^l - 1)\epsilon_a < (p_l)_a \leqslant O_a + j_a^l \epsilon_a$, for all $1 \leqslant a \leqslant n_1$ (analogously for i_2^l and i_3^l). The unique elements in the set $\{I_1, \ldots, I_N\}$, say

$\{I_1, \ldots, I_M\}$, identify the bins that contain at least one point from the set E and constitute the binning used to compute the transfer operator and the TE (in the case of the grid estimator) and just the TE in the case of the triangulation estimator.

4. Grouping of variables for TE computation

For the case of the coupled logistic maps (both UCLM and BCLM), the TE corresponding to $x \to y$ uses the grouping of variables $A_{n+1} = (x(n+1))$, $A_n = (x(n))$, and $B_n = (y(n))$, while the TE corresponding to $y \to x$ is computed with the variable grouping $A_{n+1} = (y(n+1))$, $A_n = (y(n))$, and $B_n = (x(n))$.

For the case of the Rössler-Lorenz system, the gathering of variables used to compute $T_{x_2 \to y_2}$ is $A_{n+1} = (y_2(i+3))$, $A_n = (y_2(i+2), y_2(i+1))$, and $B_n = (x_2(i+2), x_2(i+1), x_2(i))$. For computing $T_{y_2 \to x_2}$ we use the same gathering of variables but interchanging the symbols x_2 and y_2.

The TE corresponding to the coupling $x_i \to x_j$ (both direct and indirect) in Sec. VII is computed using the gathering of variables $A_{n+1} = (x_j(t+3))$, $A_n = (x_j(t+2))$, and $B_n = (x_i(t+2), x_i(t+1), x_i(t))$.

5. Computation of conditional TE

Given the variables A_{n+1}, A_n, B_n, and C_n, the conditional transfer entropy $T_{B \to A|C}$ is computed as

$$\int P(A_{n+1}, A_n, B_n, C_n) \log_2 \frac{P(A_{n+1}|A_n, B_n, C_n)}{P(A_{n+1}|A_n, C_n)}.$$

The conditional $T_{x_1 \to x_3|x_2}$ in Sec. VII is computed using the gathering of variables $A_{n+1} = (x_3(t+3))$, $A_n = (x_3(t+2))$, $B_n = (x_1(t+2))$ and $C_n = (x_2(t+2), x_2(t+1), x_2(t))$. The conditional $T_{x_3 \to x_1|x_2}$ is computed using the same gathering of variables but interchanging x_1 and x_3.

6. k nearest neighbors counting

For the case of kNN estimator of mutual information for dimension 3 or less, we used 5 nearest neighbors to compute $I(a, (b, c))$ while 10 nearest neighbors were used to compute $I(a, b)$, where a, b, and c denote generic variables. For higher dimensions ($\geqslant 4$), the same number of nearest neighbors may be taken for both mutual informations (Fig. 16 in Ref. [9]).

APPENDIX B: ERGODICITY CROSS CHECK

As a way of testing the ergodicity of the invariant measure estimated with our method, we compare the temporal and spatial averages of several functions for the UCLM and for the Rössler-Lorenz system. In particular, we consider the functions $h_a := \text{sech}(\sqrt{x^2 + y^2})$, $h_b := \beta(1 + x^2, 1 + y^2)$, and $h_c = \psi(\sqrt{x^2 + y^2})$, for the coupled logistic maps, and $h_d := \text{sech}(\sqrt{(x_2)^2 + (y_2)^2})$, $h_e := \beta(1 + (x_2)^2, 1 + (y_2)^2)$ and $h_f := \psi(\sqrt{(x_2)^2 + (y_2)^2})$, for the Rössler-Lorenz system, where $\beta(x, y)$ is the Euler β function and $\psi(x)$ is the digamma function. There is no particular reason behind the choice of these functions, other than being complicated functions having no obvious connection with the systems.
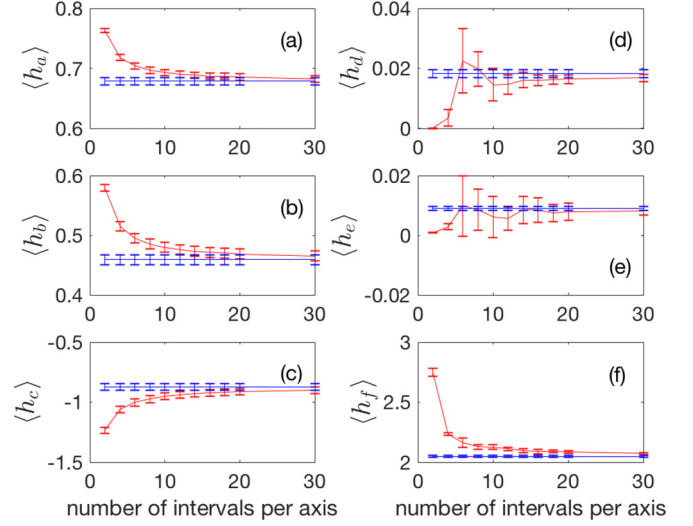

FIG. 19. Ergodicity test for 1000-point long orbits from the UCLM with $c = 0.4$ and no noise ((a), (b), and (c)) and for 10 000-point long orbits from the Rössler-Lorenz system with $c = 1.5$ and no noise ((d), (e), and (f)). Blue lines show the temporal average and red lines the spatial average of the different functions (see text). The x axis indicates the number of intervals that are taken along each axis in the embedding space for defining the grid. Error bars indicate the standard deviation over 50 realizations.

The spatial averages rapidly converge to the temporal averages as the bin size decreases (Fig. 19). Also, we point out that the rate of convergence seems to be fairly independent of the functions chosen to be averaged. This is maybe not so evident in the case of the Rössler-Lorenz system [Figs. 19(d)–19(f)] but all the spatial averages seem to saturate beyond $\sim$15 intervals per axis. This rate of convergence is rather dependent on the system and, likely more strongly, on the embedding dimension, suggesting that such a saturation could be used as a criterion for choosing a suitable bin size.

APPENDIX C: GRID ESTIMATOR VS VISITATION FREQUENCY ESTIMATOR

Here, we compare the invariant density obtained using the grid estimator with the density that a direct visitation frequency estimation yields, as a function of the number of observations in the time series. For each instance of time series length, say n, we set a bin size (as explained in Appendix A) and consider a partition into rectangular bins. We then apply the grid estimator to 50 realizations of time series with n observations and generated from randomly chosen initial values. Hence, we obtain 50 estimates for the invariant distribution, say $\rho_{\text{grid}}(r, n)$, for $1 \leqslant r \leqslant 50$. Using the same time series, we also compute the visitation frequency to each bin, obtaining thus $\rho_{vf}(r, n)$. We consider the discrepancy measure

$$\delta(n) = \frac{1}{50} \sum_{r=1}^{50} \frac{\| \rho_{\text{grid}}(r, n) - \rho_{vf}(r, n) \|}{\max\{\| \rho_{\text{grid}}(r, n) \|, \| \rho_{vf}(r, n) \|\}}$$
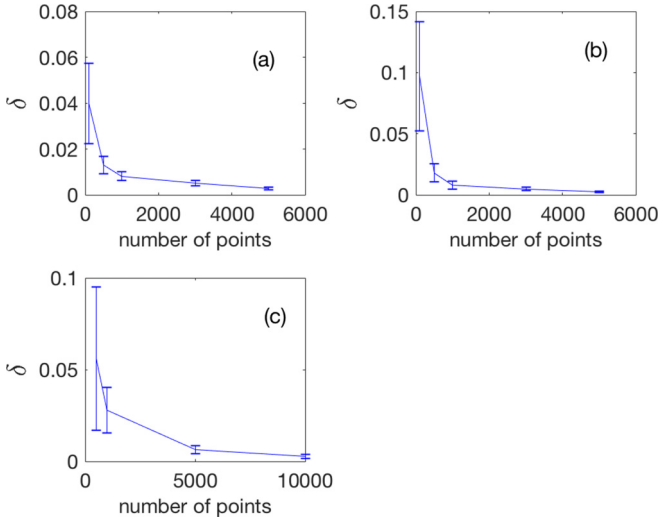
with $\| v \| := \max\{|v_a|\}$.

FIG. 20. Mean value and standard deviation over 50 realizations of the discrepancy between the invariant densities computed with the grid estimator and via computation of the visitation frequency (Appendix C), as a function of the number of observations in the time series for (a) UCLM with $c = 0.4$ and no noise; (b) BCLM with $c = 0.2$ and no noise; (c) Rössler-Lorenz system with $c = 1.5$ and no noise.

We apply this procedure to our example systems and find that both methods for estimating invariant densities produce the same outcomes (within very small discrepancies) for long enough time series (Fig. 20). By virtue of the ergodic theorem, the invariant density of the transfer operator and the invariant density yielded by the frequency of visitations must coincide. Hence, the convergence of the density estimates is expected for long enough time series.

APPENDIX D: COMPUTING THE INTERSECTING VOLUME BETWEEN SIMPLICES IN DIMENSION d

We outline the method used in this paper for computing the volume of the intersection between two simplices. Let $V = \{p_0, \dots, p_d\}$ be $d + 1$ affinely independent points in $\mathbb{R}^d$. The simplex S with vertices V is the convex hull of them, usually denoted as $S = CH(V)$, and defined as all the points in R^d constructed as $x = \alpha_0 p_0 + \cdots + \alpha_d p_d$, with $\alpha_i \geqslant 0$ and $\alpha_0 + \cdots + \alpha_d = 1$. In addition, x lies in the interior of S, denoted as $x \in \mathring{S}$, if and only if all $\alpha_i > 0$. A generic boundary of S is the simplex with vertices $\{p_{\sigma_0}, \dots, p_{\sigma_k}\}$, where $\{\sigma_0 < \cdots < \sigma_k\}$ is a (nonempty) selection of $\{0, \dots, d\}$, for $k = 0, \dots, d$. The proper faces of the simplex correspond to $k = d - 1$. Suppose $S_1 = CH(V_1)$ and $S_2 = CH(V_2)$, with $V_1 = \{p_0, \dots, p_d\}$ and $V_2 = \{q_0, \dots, q_d\}$. The method for computing the volume of the intersection $S_1 \cap S_2$ used in this work is based on the following result (the proof of which is given at the end of this Appendix):

Theorem 1. Let $S_1, S_2 \subset \mathbb{R}^n$ be two simplices of dimensions $n \geqslant m \geqslant 1$, respectively, and with $S_1 \cap S_2 \neq \emptyset$. Let $\mathcal{I}$ be the set of points in $\mathbb{R}^n$ constructed as follows: $p \in \mathcal{I}$ if $p \in \mathring{B}_1 \cap \mathring{B}_2$, where B_1 and B_2 are boundaries of S_1 and S_2, respectively, and not supporting any common direction. Then, it holds that $S_1 \cap S_2 = CH(\mathcal{I})$.

In other words, if $\mathcal{I} = \{x_1, \dots, x_N\}$ is such a set, then $S_1 \cap S_2$ consists of all the points of the form $\beta_1 x_1 + \cdots + \beta_N x_N$, for $\beta_n \geqslant 0$ and $\beta_1 + \cdots + \beta_N = 1$. The set $\mathcal{I}$ may be found as follows: let $B_1 = CH(\{p_{\rho_1}, \dots, p_{\rho_r}\})$ and $B_2 = CH(\{q_{\sigma_1}, \dots, q_{\sigma_s}\})$ and, with no loss of generality, assume that $r \geqslant s$. Reorder the vertices of S_1 as $\{p_{\rho_1}, \dots, p_{\rho_r}, p_{\rho_{r+1}}, \dots, p_{\rho_{d+1}}\}$. Every point in $\mathbb{R}^d$ can be expressed as a unique affine linear combination of these vertices, possibly with negative coefficients. In particular, $q_{\sigma_i} = \sum_{j=1}^{d-r+1} \gamma_{ji} p_{\rho_{r+j}} + \cdots$, where the coefficients on the rest of the vertices of S_1 are omitted. Denote the least dimensional affine space containing the boundary B_1 (respectively B_2) as Π_1 (respectively Π_2). By definition, the affine spaces generated by the sets of (affinely independent) points $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_m\}$, intersect uniquely if the equations

$$\sum_{i=1}^{n} \alpha_i x_i = \sum_{j=1}^{m} \beta_j y_j , \tag{D1}$$

$$\sum_{i=1}^{n} \alpha_i = \sum_{j=1}^{m} \beta_j = 1 \tag{D2}$$

have unique solution. In our case, this translates into conditions

$$\text{rank } \gamma = s - 1, \tag{D3}$$

$$\text{rank}\begin{pmatrix} \gamma \\ \mathbf{1} \end{pmatrix} = s, \tag{D4}$$

where γ is the $(d - r + 1) \times s$ matrix with entries γ_{ab} and $\mathbf{1}$ denotes a row of ones. If x is the unique intersecting point, then

$$x = \sum_{i=1}^{r} \alpha_i p_{\rho_i} = \sum_{j=1}^{s} \beta_j q_{\sigma_j} , \tag{D5}$$

with $\sum_i \alpha_i = \sum_j \beta_j = 1$, and only if $\alpha_i, \beta_j > 0$, the point x is in the interior of both boundaries and thus an element of $\mathcal{I}$.

Once the set $\mathcal{I}$ has been found it can be further decomposed as $\mathcal{I} = \cup_{a=1,2; j=0,\dots,d} \mathcal{I}_j^{(a)}$ where $\mathcal{I}_j^{(a)} := \mathcal{I} \cap F_j^{(a)}$ and $F_j^{(a)}$ denotes the face of the simplex S_a that lies opposite to the jth vertex of that simplex. Some of these subsets might be empty and there might be repetitions. Call $\mathcal{F}$ the set of the unique elements in the collection $\{\mathcal{I}_j^{(a)}\}_{a=1,2; j=0,\dots,d}$ that appear at most once for $a = 1$ or once for $a = 2$. If the element $\mathcal{I}_i^{(1)}$ appears only for $a = 1$, then it is easy to check that the intersection $F_i^{(1)} \cap S_2$ produces a convex set of dimension $d - 1$, i.e., a proper face of $S_1 \cap S_2$. On the other hand, Theorem 1 applied to the simplices $F_i^{(1)}$ and S_2 precisely yields the set of vertices $\mathcal{I}_i^{(1)}$. The same holds if the set of vertices $\mathcal{I}_i^{(2)}$ appear only for $a = 2$. In the case where any such set $\mathcal{I}_j^{(a)}$ appears for both $a = 1$ and 2, its convex hull is a proper face of $S_1 \cap S_2$ only when the corresponding faces containing $\mathcal{I}_j^{(a)}$ are parallel. Each face of $S_1 \cap S_2$ may be triangulated (some of them might be already simplices) and the triangulation of these faces, together with any point in the interior of $S_1 \cap S_2$, for instance its centroid, form a triangulation of the intersection between the simplices. Its volume is then computed as the sum of the volumes of the simplices in this last triangulation.

1. Proof of Theorem 1

The statement in Theorem 1 is actually a corollary of a well known result on convex geometry:

Theorem 2 (Minkowski). Every convex and compact set in $\mathbb{R}^n$ is the convex hull of its extreme points.

The proof of Theorem 2 can be found in Ref. [50]. A point x in a convex set P is said to be extreme if the equality $x = \lambda y + (1 - \lambda) z$, for $y, z \in P$ and $0 < \lambda < 1$, requires $x = y = z$. In other words, x is not found in the interior of any segment contained in P. For our purposes, it is convenient to use an (easily shown to be) equivalent definition: x is extreme if for every unit vector u and for every $\epsilon > 0$, there is $|\lambda| < \epsilon$, such that $x + \lambda u \notin P$. The set of extreme points of P is denoted as ext(P).

Proof of Theorem 1. We claim that the set $\mathcal{I}$, as defined in Theorem 1, equals ext($S_1 \cap S_2$). To see this, let $\{p_0, \ldots, p_n\}$ be the vertices of S_1 and $\{q_0, \ldots, q_m\}$ be the vertices of S_2 and let $x \in \mathring{B}_1 \cap \mathring{B}_2$ with B_1 and B_2, boundaries verifying the properties required in Theorem 1. Next, let u be an arbitrary unit vector and assume, without loss of generality, that the direction u is not supported by Π_1 (the least dimensional affine space supporting B_1). Further assume, without loss of generality, that $\{p_0, \ldots, p_r\}$ are the vertices of the boundary B_1. Then, $u = \sum_{i=1}^{r} \gamma_i (p_i - p_0) + \sum_{j=r+1}^{n} \mu_j (p_j - p_0)$ with not all μ_j vanishing. Assume, without loss of generality, that $\mu_{r+1} \neq 0$. Therefore, $x + \lambda u = \lambda \mu_{r+1} p_{r+1} + \sum_{i=0}^{r} \alpha_i(\lambda) p_i + \sum_{j>r+1} \beta_j(\lambda) p_j$. Given $\epsilon > 0$ arbitrary, take $\lambda = -\text{sign}(\mu_{r+1}) \frac{\epsilon}{2}$, it then holds that $x + \lambda u = -|\mu_{r+1}| \frac{\epsilon}{2} p_{r+1} + \cdots$ and therefore $x + \lambda u \notin S_1$. This shows that x is an extreme point of $S_1 \cap S_2$ and since $x \in \mathcal{I}$ was arbitrary, it follows that $\mathcal{I} \subset \text{ext}(S_1 \cap S_2)$.

To see the reverse inclusion, let $x \in \text{ext}(S_1 \cap S_2)$ arbitrary. Without loss of generality, assume that $x = \sum_{i=0}^{r} \alpha_i p_i = \sum_{j=0}^{s} \beta_i q_i$ with all $\alpha_i > 0$ and all $\beta_j > 0$. Call Π_r the least dimensional affine space containing the vertices $\{p_0, \ldots, p_r\}$ and Π_s the affine space with the same property with respect to the vertices $\{q_0, \ldots, q_s\}$. It holds that Π_r and Π_s do not support any common direction. Indeed, suppose u is a unit vector along a direction supported by both Π_r and Π_s then, having that all the coefficients α_i and β_j are strictly positive, it follows that for some $\epsilon > 0$ small enough, $x + \lambda u \in S_1 \cap S_2$, for all $|\lambda| < \epsilon$, contradicting that x is an extreme point. This shows that $\mathcal{I} = \text{ext}(S_1 \cap S_2)$. Given that both S_1 and S_2 are compact and convex sets, so it is $S_1 \cap S_2$ and the proof is completed by using Theorem 2. ∎

APPENDIX E: SAMPLE-AND-FILTER APPROACH

Here, we provide a tentative modification of the triangulation estimator in order to reduce its high computational demands. As described in Sec. III E 2, the triangulation estimator is based on an initial partition of the embedded attractor into simplices and the map generating the dynamics is approximated linearly onto each simplex [15], say $\tilde{\psi}$. Given the initial triangulation of the attractor, we make use of the piecewise linear approximation of the map to generate sampling points as an input to the grid estimator [Eq. (16)]:
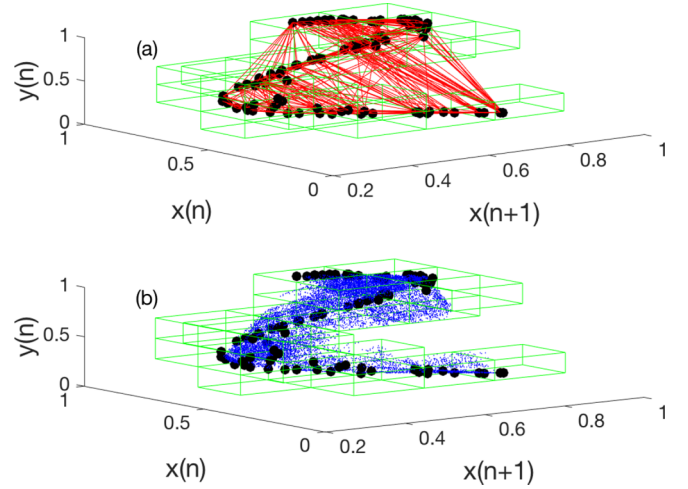

FIG. 21. Example of the sampling and filtration of 100 embedding points for the UCLM system with $c = 0.4$, no noise, and using the embedding $(x(n + 1), x(n), y(n))$. Solid black points represent the original embedding points and the grid is shown in green. (a) Initial triangulation of the reconstructed attractor. (b) Each simplex is sampled with 80 points (small blue points), where those not lying in the grid have been filtered out. The number of points after filtering (small blue points) is about 23 000.

(1) Given an embedded attractor in a d dimensional space (we assume it consists of few hundred points), let $\{S_1, \ldots, S_N\}$ be its triangulation into d simplices and $\{B_1, \ldots, B_M\}$ be the set of bins in a regular grid that are visited by the embedded points [Fig. 21(a)]. The size of the bins in the regular grid is adapted to the number of points in the reconstructed attractor (Appendix A).

(2) Each simplex is sampled with N_s points using a predefined matrix of convex coefficients. More specifically, let C be a $N_s \times (d + 1)$ matrix such that $C_{ai} \geqslant 0$, no two rows are equal and $\sum_{j=1}^{d+1} C_{aj} = 1$, for all $1 \leqslant a \leqslant N_s$ and $1 \leqslant i \leqslant d + 1$. Let $\{v_1, \ldots, v_{d+1}\}$ be the vertices of the simplex S_n in the triangulation of the attractor and $\{\tilde{\psi}(v_1), \ldots, \tilde{\psi}(v_{d+1})\}$ be the vertices of the simplex being the image of S_n under the map (see Sec. III E 2 for details). The ath sampling point of the simplex S_n is given by $p_{n,a} := \sum_{j=1}^{d+1} C_{aj} v_j$ and its image under $\tilde{\psi}$ is given by $\tilde{\psi}(p_{n,a}) = \sum_{j=1}^{d+1} C_{aj} \tilde{\psi}(v_j)$ (the map $\tilde{\psi}$ is linear on each simplex). Call $P := \{p_{n,a}\}_{1 \leqslant n \leqslant N, \, 1 \leqslant a \leqslant N_s}$, the set of all the sampling points of the simplices in the triangulation and $\tilde{\psi}(P) = \{\tilde{\psi}(p_{n,a})\}$, the set of image points.

(3) From P, discard all those points not lying in $\cup_{i=1}^{M} B_i$. The resulting set $\bar{P}$ contains (possibly thousands of) points that are distributed more tightly to the volume occupied by the reconstructed attractor [Fig. 21(b)].

(4) Finally, a new bin size is adapted to the number of points in $\bar{P}$. Their images are found in the set $\tilde{\psi}(\bar{P}) = \{\tilde{\psi}(p_{n,a}) \mid p_{n,a} \in \cup_i B_i\}$. Therefore, the transfer operator may be approximated using Eq. (16), which using the above notation reads as

$$P_{ij} \simeq \frac{\sharp\{p_{n,a} \mid \tilde{\psi}(p_{n,a}) \in B_j \, \cap \, p_{n,a} \in B_i\}}{\sharp\{p_{m,b} \mid p_{m,b} \in B_i\}}.$$
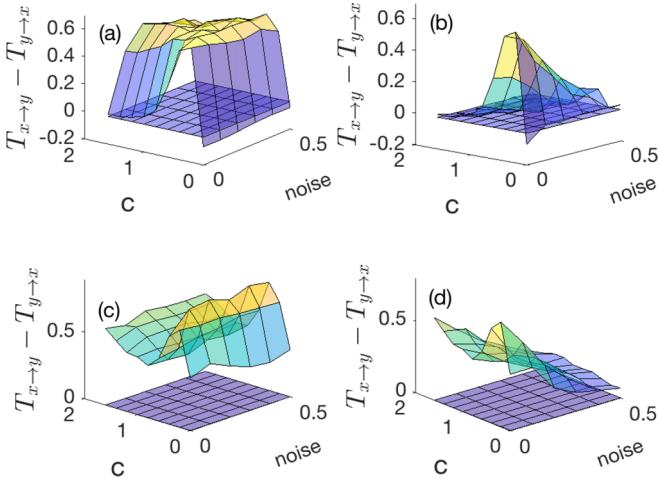
FIG. 22. Mean value of $T_{x \to y} - T_{y \to x}$ over 50 realizations as a function of coupling constant and noise level obtained with the sample-and-filter approach. (a) UCLM with dynamical noise; (b) UCLM with measurement noise; (c) BCLM with dynamical noise; (d) BCLM with measurement noise.

Note that generating sampling points in this manner only assumes piecewise linearity of the map and does not introduce any further bias. We applied this approach to both UCLM and BCLM, using time series with 100 observations and 50 realizations from randomly chosen initial conditions. The dependence of $\Delta T := T_{x \to y} - T_{y \to x}$ on the coupling constant and noise is comparable to that obtained from the grid estimator [Figs. 22(a) and 22(b)]. Interestingly, the values for the ΔT obtained with the sample-and-filter approach are, in general, higher than those obtained with the grid estimator (Fig. 23). This result suggests that from a sparse time series one may generate thousands of points from a piecewise linear approximation and obtain reliable results for the TE. We also
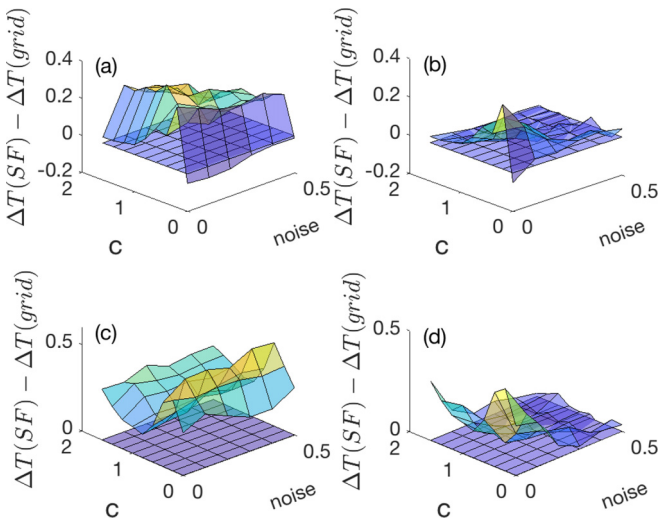

FIG. 23. $\Delta T(\text{SF}) - \Delta T(grid)$ (SF indicates sample-and-filter method) as a function of the coupling constant and noise level for (a) UCLM with dynamical noise; (b) UCLM with measurement noise; (c) BCLM with dynamical noise; (d) BCLM with measurement noise.
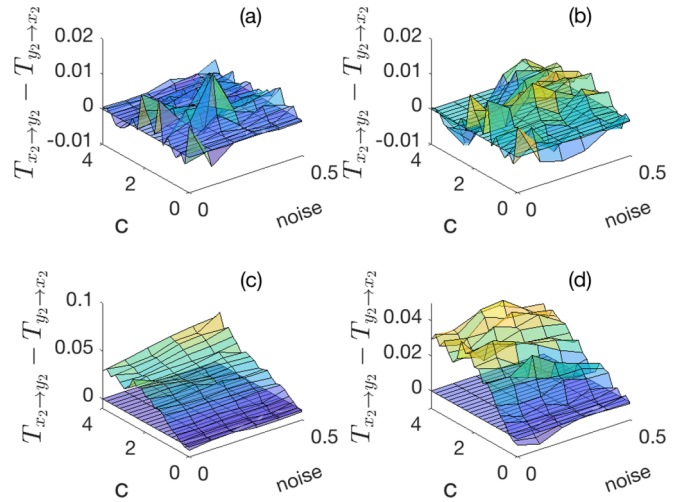

FIG. 24. Mean value of $T_{x_2 \to y_2} - T_{y_2 \to x_2}$ over 50 realizations as a function of the coupling constant and noise level for the Rössler-Lorenz system and (a) with dynamical noise obtained with the sample-and-filter approach; (b) the same as in (a) but with measurement noise instead; (c) obtained with the grid estimator and adding dynamical noise; (d) the same as in (c) but with measurement noise instead.

apply the sample-and-filter approach to the Rössler-Lorenz system, in which case the results are less impressive (Fig. 24). Note, however, that these analyses used time series with only 200 observations in embedding dimension 6 (Appendix A).
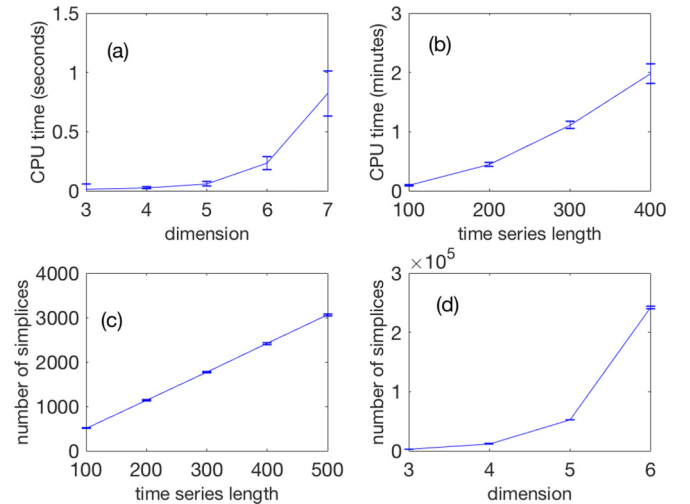

FIG. 25. (a) Mean and standard deviation over 50 realizations of the CPU time needed to obtain the volume of a nontrivial simplex intersection; (b) mean and standard deviation over 10 realizations of the CPU time required to obtain an estimate of the transfer operator for to the UCLM system with $c = 0.4$ and no noise, in embedding dimension 3. The timing values were obtained using the *cputime* function in MATLAB; (c) mean and standard deviation over 50 realizations of the number of simplices in a triangulation in 3D; (d) the same as in (c), fixing the number of observations in the time series to 500.

APPENDIX F: COMPUTATIONAL TIMES FOR THE TRIANGULATION ESTIMATOR

In this Appendix we provide CPU times required to obtain the transfer operator using the triangulation estimator [Fig. 25(b)] as well as the CPU time required for obtaining the volume of a (nontrivial) simplex intersection [Fig. 25(a)]. The simulations were run in MATLAB using a MacBook Pro with a 2.8 GHz Intel Core i7 processor. Computation times increase for higher dimensions because the CPU time to obtain the volume for a simplex intersection and the number of simplices in the triangulation both scale exponentially [Figs. 25(a) and 25(d)]. In particular, for dimension 5, the CPU time required to obtain the volume of a nontrivial simplex intersection is $t_s \sim 5 \times 10^{-2}$ s and the typical number of simplices in a triangulation generated from 500 points is $n_s \sim 5 \times 10^4$. If each simplex intersects nontrivially with just 10 of the simplices in the triangulation (which is a quite optimistic estimate) we are left with a computation time for obtaining the transfer operator in the order $t_{\mathrm{TO}} \sim 10\, n_s\, t_s \sim 2.5 \times 10^4$ s $\sim 7\,h$. For comparison, our grid estimator applied to 500-point long time series in dimension 5 requires a CPU time of $\sim 5 \times 10^{-3}$ s to estimate the transfer operator.

[1] M. Breakspear, Nat. Neurosci. **20**, 340 (2017); S. L. Bressler and A. K. Seth, NeuroImage **58**, 323 (2011); T. Kreuz, D. Chicharro, M. Greschner, and R. G. Andrzejak, J. Neurosci. Methods **195**, 92 (2011); K. Lehnertz and C. E. Elger, Phys. Rev. Lett. **80**, 5019 (1998); C. E. Elger and K. Lehnertz, Eur. J. Neurosci. **10**, 786 (1998).

[2] J. Runge, V. Petoukhov, J. F. Donges, J. Hlinka, N. Jajcay, M. Vejmelka, D. Hartman, N. Marwan, M. Paluš, and J. Kurths, Nat. Commun. **6**, 8502 (2015); M. Paluš, D. Hartman, J. Hlinka, and M. Vejmelka, Nonlin. Proc. Geophys. **18**, 751 (2011); M. Ghil, M. Allen, M. Dettinger, K. Ide, D. Kondrashov, M. Mann, A. W. Robertson, A. Saunders, Y. Tian, F. Varadi *et al.*, Rev. Geophys. **40**, 3 (2002); M. E. Schlesinger and N. Ramankutty, Nature (London) **367**, 723 (1994).

[3] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, 2nd ed. (Cambridge University Press, Cambridge, UK, 2003).

[4] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya (unpublished).

[5] P. O. Amblard and O. J. Michel, Entropy **15**, 113 (2013).

[6] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).

[7] M. Paluš, V. Komárek, Z. HrnčíǍŹ, and K. Štěrbová, Phys. Rev. E **63**, 046211 (2001).

[8] J.-P. Eckmann and D. Ruelle, Rev. Mod. Phys. **57**, 1115 (1985).

[9] A. Kraskov, H. Stögbauer, and P. Grassberger, Phys. Rev. E **69**, 066138 (2004).

[10] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, Bioinformatics **18**, S231 (2002).

[11] M. Staniek and K. Lehnertz, Phys. Rev. Lett. **100**, 158101 (2008); H. Dickten and K. Lehnertz, Phys. Rev. E **90**, 062706 (2014); C. Bandt and B. Pompe, Phys. Rev. Lett. **88**, 174102 (2002).

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley, Hoboken, NJ, 2006), p. 748.

[13] A. Lasota and M. C. Mackey, *Applied Mathematical Sciences*, 2nd ed. (Springer, New York, 1994), p. 472; L. Vinet and A. Zhedanov, J. Phys. A: Math. Theor. **44** 085201 (2011); A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences* (Academic, New York, 1979), p. 337; C. Beck and F. SchloǏgl, *Thermodynamics of Chaotic Systems: An Introduction* (Cambridge University Press, Cambridge, UK, 1993), p. 286.

[14] G. Froyland and K. Padberg-Gehle, Phys. D (Amsterdam) **241**, 1612 (2012); P. Miron, F. J. Beron-Vera, M. J. Olascoaga, J. Sheinbaum, P. Pérez-Brunius, and G. Froyland, Sci. Rep.

7, 1 (2017); G. Froyland, C. Horenkamp, V. Rossi, and E. van Sebille, Chaos **25**, 083119 (2015); G. Froyland, R. M. Stuart, and E. van Sebille, *ibid.* **24**, 033126 (2014); E. Ser-Giacomi, V. Rodríguez-Méndez, C. López, and E. Hernández-García, Eur. Phys. J.: Spec. Top. **226**, 2057 (2017); C. Maes, N. Grima, B. Blanke, E. Martinez, T. Paviet-Salomon, and T. Huck, Geophys. Res. Lett. **45**, 1915 (2018); R. McAdam and E. van Sebille, J. Geophys. Res., (Oceans) **123**, 514 (2018).

[15] G. Froyland, Bull. Austral. Math. Soc. **56**, 157 (1997).

[16] G. Froyland, Nonlinearity **12**, 79 (1999).

[17] E. M. Bollt, Int. J. Bifurc. Chaos **22**, 1250261 (2012).

[18] X. S. Liang and R. Kleeman, Phys. Rev. Lett. **95**, 244101 (2005).

[19] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).

[20] The same would be true for a dynamical system generated by a vector field.

[21] D. Ruelle, Commun. Math. Phys. **82**, 137 (1981).

[22] A typical example of a quasiperiodic orbit is given by the map on the torus defined as $T_t(x, y) = (x + t, y + rt)$ (each component taken modulo 1) for $r > 0$ nonrational and $0 \leqslant x, y < 1$. It can be shown that such a trajectory returns arbitrarily close and infinitely often to itself, but it never closes up exactly.

[23] E. N. Lorenz, J. Atmos. Sci. **20**, 130 (1963).

[24] R. M. May, Nature (London) **261**, 459 (1976).

[25] D. Ruelle and F. Takens, Commun. Math. Phys. **20**, 167 (1971).

[26] The characteristic function of a set A returns 1 if x belongs to A and 0, otherwise.

[27] If μ assigns positive measure to sets when, and only when, the sets have positive volume (and then μ is said to be compatible with Lebesgue), a unique density δ of μ is guaranteed to exist by the Radon-Nikodym theorem [51]. In the case of dynamical systems, axiom-A maps are known to possess a unique invariant measure compatible with Lebesgue [28].

[28] R. Bowen, *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, Lecture Notes in Mathematics, Vol. 470 (Springer, Berlin, 1975); R. Bowen and D. Ruelle, Invent. Math. **29**, 181 (1975); D. Ruelle, Am. J. Math. **98**, 619 (1976).

[29] S. Smale, Bull. Am. Math. Soc. **73**, 747 (1967).

[30] G. Gallavotti, J. Stat. Phys. **84**, 899 (1996); G. Gallavotti and E. G. D. Cohen, *ibid.* **80**, 931 (1995).

[31] S. M. Ulam, *Problems in Modern Mathematics* (Interscience, New York, 1964), p. 150.

[32] F. Takens, Lect. Notes Math. **898**, 366 (1981); T. Sauer, J. A. Yorke, and M. Casdagli, J. Stat. Phys. **65**, 579 (1991); E. R. Deyle and G. Sugihara, PLoS ONE **6**, e18295 (2011).

[33] With mild modifications in the derivation of the main result of Chap. 4 in [15], it can be shown that the estimate of the invariant measure is independent of the embedding.

[34] A. M. Fraser and H. L. Swinney, Phys. Rev. A **33**, 1134 (1986); G. Mishne, R. Talmon, R. Meir, J. Schiller, M. Lavzin, U. Dubin, and R. R. Coifman, IEEE Journal of Selected Topics in Signal Processing **10**, 1238 (2016); W. Liebert and H. G. Schuster, Phys. Lett. A **142**, 107 (1989); H. Kim, R. Eykholt, and J. Salas, Phys. D (Amsterdam) **127**, 48 (1999).

[35] This number is likely system dependent.

[36] M. Dellnitz, G. Froyland, and O. Junge, *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems* (Springer, Berlin, 2001), pp. 145–174.

[37] Built-in routines for Delaunay triangulation in arbitrary dimensions can be found in standard numerical softwares, for instance the QHULL library.

[38] B. Büeler, A. Enge, and K. Fukuda, in *Polytopes Combinatorics and Computation*, edited by G. Kalai *et al.* (Birkhäuser, Basel, 2000), Chap. 6, pp. 131–154.

[39] We expect this number to be dependent on the system and the embedding dimension.

[40] H. Edelsbrunner and D. R. Grayson, Discrete Comput. Geom. **24**, 707 (2000).

[41] A. Krakovská, J. Jakubík, M. Chvosteková, D. Coufal, N. Jajcay, and M. Paluš, Phys. Rev. E **97**, 042207 (2018).

[42] R. Marschinski and H. Kantz, Eur. Phys. J. B **30**, 275 (2002).

[43] B. Gourévitch and J. J. Eggermont, J. Neurophysiol. **97**, 2533 (2007).

[44] T. Bossomaier, L. Barnett, M. Harré, and J. T. Lizier, *An Introduction to Transfer Entropy* (Springer, Berlin, 2016).

[45] O. E. Rössler, Phys. Lett. A **57**, 397 (1976).

[46] R. Quian Quiroga, J. Arnhold, and P. Grassberger, Phys. Rev. E **61**, 5142 (2000).

[47] CausalityTools.jl is a registered Julia package and the source code is found at https://github.com/kahaaga/CausalityTools.jl.

[48] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, SIAM Rev. **59**, 65 (2017).

[49] D. Kugiumtzis, Phys. Rev. E **87**, 062918 (2013).

[50] B. Grünbaum, *Convex Polytopes*, Volume 221 of Graduate Texts in Mathematics (Springer, Berlin, 2003).

[51] D. L. Cohn, *Measure Theory*, 2nd ed. (Springer, New York, 2013), p. 466.