# Specific transfer entropy and other state-dependent transfer entropies for continuous-state input-output systems

David Darmon*
*Department of Military and Emergency Medicine, Uniformed Services University of the Health Sciences, Bethesda, Maryland 20814, USA and The Henry M. Jackson Foundation for the Advancement of Military Medicine, Bethesda, Maryland 20817, USA*

Paul E. Rapp
*Department of Military and Emergency Medicine, Uniformed Services University of the Health Sciences, Bethesda, Maryland 20814, USA*
(Received 28 April 2017; published 9 August 2017)

Since its original formulation in 2000, transfer entropy has become an invaluable tool in the toolbox of nonlinear dynamicists working with empirical data. Transfer entropy and its generalizations provide a precise definition of uncertainty and information transfer that are central to the coupled systems studied in nonlinear science. However, a canonical definition of state-dependent transfer entropy has yet to be introduced. We introduce a candidate measure, the specific transfer entropy, and compare its properties to both total and local transfer entropy. Specific transfer entropy makes possible both state- and time-resolved analysis of the predictive impact of a candidate input system on a candidate output system. We also present principled methods for estimating total, local, and specific transfer entropies from empirical data. We demonstrate the utility of specific transfer entropy and our proposed estimation procedures with two model systems, and find that specific transfer entropy provides more, and more easily interpretable, information about an input-output system compared to currently existing methods.

DOI: 10.1103/PhysRevE.96.022121

## I. INTRODUCTION

One of the hallmarks of a complex system is that the interaction of relatively simple units gives rise to complex overall dynamics. Beyond the microscale behavior of individual units and the macroscale behavior of the overall system, an understanding of how the units interact and influence each other is also desired. In the absence of a model for a system, researchers turn to statistics of available observations from the system to quantify the relationship(s) between its components. For example, one of the earliest statistical measures for quantifying the impact of one system on another was Granger causality [1]. Granger causality quantifies the predictive impact of a candidate input system on a candidate output system accounting for the past of the candidate output system. In his original paper, Granger operationalized causality by considering if inclusion of the history of the input system reduces the residual variance of the optimal, in the minimum mean-squared error sense, predictor of the output future relative to the residual variance of the optimal predictor for the output future without the input system's history. He then further operationalized this definition in terms of optimal linear predictors, which he called "linear causality in mean." Thus, in the original formulation, Granger foreshadowed the so-called nonlinear or nonparametric Granger causalities [2,3], while restricting his main analysis to the linear case. A more modern formulation that subsumes and includes as special cases both linear and nonlinear Granger causality [4] is the transfer entropy [5,6] from the input system to the output system. Transfer entropy has the desirable property that it is zero precisely when the future of the output system is independent of the history of the input system conditional on the history of the output system. Thus, when transfer entropy is zero, the input system provides no predictive information about the output system beyond the predictive information already provided by its own past.[1] Transfer entropy has been applied across many disciplines, from the social sciences [10] and ecology [11] to genetics [12], neuroscience [13,14], and physiology [15]. It has become especially popular in neuroscience due to the availability of many open-source toolboxes [16,17].

Transfer entropy is defined via an ensemble average over all input-output pasts and output futures. As such, it quantifies the total predictive impact of the input system on the output system. However, for nonlinear systems, we expect the predictability, and thus the predictive impact of the input, to vary across the input-output state space. To this end, a local transfer entropy was proposed in Ref. [18] to quantify how predictive information varies across the input-output state space. However, the local transfer entropy has counterintuitive properties that make its interpretation difficult. In this paper, we develop a state-dependent transfer entropy, the specific transfer entropy,

*david.darmon.ctr@usuhs.edu



[1]It is in this restricted sense that transfer entropy and Granger causality can be considered as measures of causality or information flow. We will always consider transfer entropy as a measure of predictive information, since that is precisely what it measures. See Ref. [7] for additional discussion of interpretational pitfalls with transfer entropy. See Refs. [8,9] for discussions of alternative definitions of causality in the time series setting.

that does not share these counterintuitive properties and provides a direct information theoretic measure of the state-dependent impact of an input process on an output process.

Moreover, though the transfer entropy of a given stochastic input-output system is well defined,[2] the properties of estimators of transfer entropy in the absence of a model are not well studied. Since its formulation for continuous-valued input-output systems, it has been noted that the transfer entropy can be estimated using kernel density estimator-based [19] and $k$th-nearest-neighbor-based [20] estimators. From this perspective, the estimation of transfer entropy has two components: model selection, in the choice of the model order, and model estimation, in the choice of bandwidths for kernel density estimators and $k$ in $k$th-nearest-neighbor estimators. Heuristics are typically used for model selection, for example the Ragwitz criterion [21] in the neuroscience literature or embedology-based approaches adopted from attractor reconstruction [22].[3] After model selection, parameters for model estimation are also chosen in an *ad hoc* fashion by resorting to asymptotic results in, for example, choosing the nearest neighbor for $k$th-nearest-neighbor estimators. However, since we are always dealing with finite, and often quite small (relative to the dimension of the estimation problem) data sets, appeals to asymptotic results should be treated with skepticism. After model selection and estimation, the transfer entropy estimate is most often used for null hypothesis significance testing [24]. Thus, rather than attempting to estimate the transfer entropy with precision and accuracy, the easier problem of determining whether an input has any, no matter how small, predictive impact on an output is solved. This is typically done using a bootstrap method based on surrogates [25]. Because of this, little attention has been paid to how well proposed estimators perform. When the statistical properties of estimators of transfer entropy have been considered, it is typically in the context of linear vector autoregressive models for which transfer entropy has been computed analytically from Granger causality [4]. However, such systems are precisely those for which transfer entropy is least appropriate.

In the rest of this paper, we investigate the properties of total, local, and specific transfer entropies for an input-output process, and consider the statistical properties of estimators for these transfer entropies from finite samples. We present the total and local transfer entropies and develop the specific transfer entropy in Sec. II. In Sec. III, we present three methods for model selection and estimation of the transfer entropies from observations of an input-output system. We then consider how the transfer entropies and their estimators behave for an analytically tractable system in Sec. IV, and for a stochastic chaotic system in Sec. V. Finally, we conclude and consider potential directions for future research related to transfer entropy and its state-dependent variants in Sec. VI.

## II. TOTAL, LOCAL, AND SPECIFIC TRANSFER ENTROPIES

We begin by considering a nominal input-output system where we denote the time series for the input system by $\{Y_t\}_{t\in\mathbb{Z}}$ and the time series for the output system by $\{X_t\}_{t\in\mathbb{Z}}$. In this paper we consider the case where both the input and output time series are real valued. We will denote blocks of a time series from time $a$ to time $b > a$ by $Z_a^b = (Z_a, Z_{a+1}, \ldots, Z_{b-1}, Z_b)$. We denote the order-$p$ input-blind predictive density of the output process by $f_{X_t|X_{t-p}^{t-1}}(x_t \mid x_{t-p}^{t-1})$ and the order-$p$ input-conditioned predictive density of the output process by $f_{X_t|Y_{t-p}^{t-1},X_{t-p}^{t-1}}(x_t \mid y_{t-p}^{t-1}, x_{t-p}^{t-1})$. That is, $f_{X_t|X_{t-p}^{t-1}}(\cdot \mid x_{t-p}^{t-1})$ specifies the density over the output future $X_t$ conditional on the output past $x_{t-p}^{t-1}$ and $f_{X_t|Y_{t-p}^{t-1},X_{t-p}^{t-1}}(\cdot \mid y_{t-p}^{t-1}, x_{t-p}^{t-1})$ specifies the density over the output future $X_t$ conditional on the input-output past $(y_{t-p}^{t-1}, x_{t-p}^{t-1})$. We will assume that the output process is stationary conditional on its own past as well as on its own past and the past of the input process, that is the input-blind predictive density satisfies

$$f_{X_t|X_{-\infty}^{t-1}} = f_{X_{t+\tau}|X_{-\infty}^{t+\tau-1}} \tag{1}$$

and the input-conditioned predictive density satisfies

$$f_{X_t|Y_{-\infty}^{t-1},X_{-\infty}^{t-1}} = f_{X_{t+\tau}|Y_{-\infty}^{t+\tau-1},X_{-\infty}^{t+\tau-1}} \tag{2}$$

for all values of $t$ and $\tau$. In other words, the statistical properties of the future output of the process does not vary in time conditional on a sufficiently long output or input-output past. This assumption, called conditional stationarity [26], is weaker than the usual assumption of strong-sense stationarity. For example, a homogeneous Markov process of order $p$ with initial density not equal to a stationary density of the process is conditionally stationary after conditioning on $p$ steps into its past but is not strong-sense stationary. Because of this assumption, we will suppress the subscripts of the predictive densities where their arguments make their identities clear.

The transfer entropy from the input process to the output process is determined by the two predictive entropies for the output process. In the absence of information about the input process, the output process has an input-blind predictive entropy of order $p$ given by

$$h\left[X_{p+1}\middle|X_1^p\right] = -E\left[\ln f\left(X_{p+1}\middle|X_1^p\right)\right], \tag{3}$$

and all logarithms are taken base-$e$ where $E[\cdot]$ is the expectation operator. The input-blind predictive entropy quantifies the intrinsic uncertainty in the next output conditional on the previous $p$ outputs. Taking the limit as $p$ goes to infinity recovers the entropy rate of the process. With inclusion of the input process, the input-conditional predictive entropy is given by

$$h\left[X_{p+1}\middle|Y_1^p, X_1^p\right] = -E\left[\ln f\left(X_{p+1}\middle|Y_1^p, X_1^p\right)\right]. \tag{4}$$

Because conditioning reduces entropy [27], the inclusion of $Y_1^p$ can only decrease the predictive uncertainty relative to the input-blind predictive density. This motivates considering the difference between the input-blind and input-conditioned predictive entropies, which gives the transfer entropy of

[2] We emphasize that the system must be stochastic in a well-defined way for the requisite quantities to be well behaved. For example, if a purely deterministic relationship exists between the input and output, then infinities may occur among the quantities occurring in the definition of transfer entropy.

[3] See Ref. [23] for a discussion of how embedology techniques used for model selection in the predictive context can lead to suboptimal results.

order $p$ [5,6],

$$\mathrm{T}^{(p)}_{Y\to X} = h\left[X_{p+1}\middle|X_1^p\right] - h\left[X_{p+1}\middle|X_1^p, Y_1^p\right] \tag{5}$$

$$= E\left[\ln \frac{f\left(X_{p+1}\middle|Y_1^p, X_1^p\right)}{f\left(X_{p+1}\middle|X_1^p\right)}\right]. \tag{6}$$

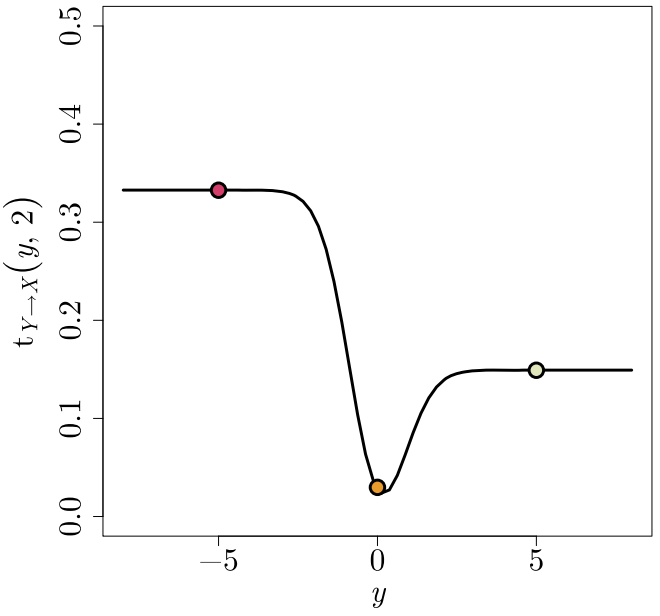
By manipulating (6), one can also show that the transfer entropy is equivalent to the mutual information between the next step future of the output process and the input past conditional on the output past, $I[X_{p+1} \wedge Y_1^p | X_1^p]$. In this form, we therefore immediately see that the transfer entropy is zero if and only if the next-step future of the output process is independent of the past of the input process conditional on the past of the output process.

The traditional transfer entropy as given by (6) averages over all input-output pasts and output futures. As such, it quantifies the total transfer entropy from the input process to the output process. However, one of the hallmarks of nonlinear dynamical systems is that their predictability can vary widely depending on where in state space the prediction occurs [28]. Thus, a state-dependent version of transfer entropy that quantifies the predictive impact of an input process on an output process as a function of the joint state space is desirable. One version of a state-dependent transfer entropy, the local transfer entropy, was developed in Refs. [18,29]. Local transfer entropy is derived from the expression (6) for the overall transfer entropy. The local transfer entropy of order $p$, denoted by $\widetilde{\mathrm{t}}_{Y\to X}(y_1^p, x_1^p, x_{p+1})$, is taken to be the expectand of (6):

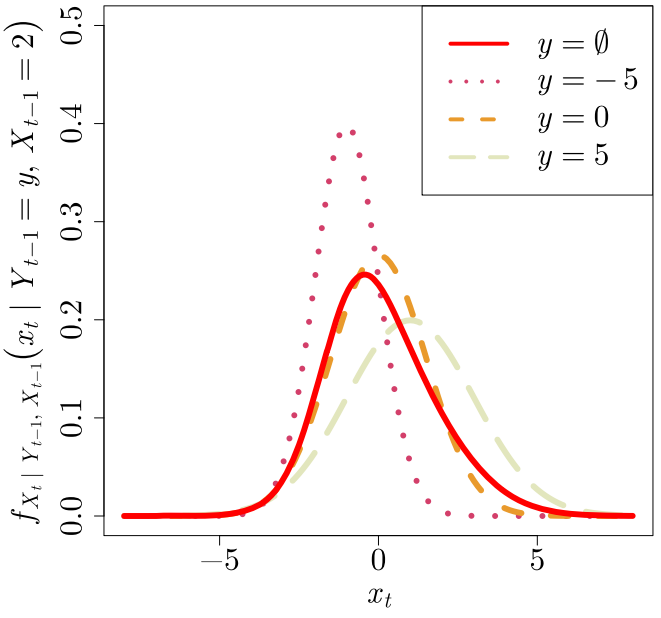
$$\widetilde{\mathrm{t}}_{Y\to X}\left(y_1^p, x_1^p, x_{p+1}\right) = \ln \frac{f\left(x_{p+1}\middle|y_1^p, x_1^p\right)}{f\left(x_{p+1}\middle|x_1^p\right)}. \tag{7}$$

The local transfer entropy has several desirable properties: it is identically zero precisely when total transfer entropy is zero, and taking its average over all input-output pasts and output futures gives total transfer entropy. However, its interpretation is made difficult by the fact that, though it averages to a non-negative value, for any given evaluation point it may be negative. In fact, it can take any real value. As an example, consider the right panel of Fig. 1, which shows the input-blind predictive density $f(x_t|x_{t-1})$ and several input-conditional predictive densities $f(x_t|y_{t-1}, x_{t-1})$ for various values of the input $y_{t-1}$ for the input-output system developed in Sec. IV. We see that the input-blind predictive density can be less than, equal to, or greater than the input-conditioned predictive density depending on the value of any one of $y_{t-1}, x_{t-1}$, or $x_t$. What, then, does a negative local transfer entropy mean in terms of the predictive impact of the input on the output? In Refs. [18,30], the authors state that in these cases the input is misleading and/or misinformative. This interpretation does not, however, agree with the fact that, conditional on that particular input-output past, the future will agree statistically with the input-conditioned predictive density and not the input-blind predictive density [31]. Thus, rather than being misinformative, the input is precisely correctly informative in these cases.

We now develop an alternative to local transfer entropy that shares its desirable properties while also having direct interpretability in terms of the state-dependent predictive impact of the input process on the output process. Like local transfer entropy, we begin from the definition of total transfer entropy given by (6). However, rather than taking the expectand as our definition, we first apply an iterated




(a)




(b)

FIG. 1. A demonstration of (a) specific transfer entropy and (b) input-blind $f(x_t|x_{t-1})$ and input-conditional predictive densities $f(x_t|y_{t-1}, x_{t-1})$ for the model system developed in Sec. IV at $x_{t-1} = 2$. The colors of the input-conditioned predictive densities correspond to colors of the points on specific transfer entropy curve, and the red density corresponds to the input-blind predictive density.

expectation,

$$\mathrm{T}^{(p)}_{Y\to X} = E\left[\ln \frac{f\left(X_{p+1}\middle|Y_1^p,X_1^p\right)}{f\left(X_{p+1}\middle|X_1^p\right)}\right] \tag{8}$$

$$= E\left[E\left(\ln \frac{f\left(X_{p+1}\middle|Y_1^p,X_1^p\right)}{f\left(X_{p+1}\middle|X_1^p\right)}\middle|Y_1^p,X_1^p\right)\right] \tag{9}$$

and consider the internal conditional expectation $E[\ln \frac{f(X_{p+1}|Y_1^p,X_1^p)}{f(X_{p+1}|X_1^p)}|Y_1^p = y_1^p, X_1^p = x_1^p]$. Unpacking this conditional expectation,

$$E\left[\ln \frac{f\left(X_{p+1}\middle|Y_1^p,X_1^p\right)}{f\left(X_{p+1}\middle|X_1^p\right)}\middle|Y_1^p = y_1^p, X_1^p = x_1^p\right] \tag{10}$$

$$= \int_{\mathbb{R}} f\left(x_{p+1}\middle|y_1^p,x_1^p\right) \ln \frac{f\left(x_{p+1}\middle|y_1^p,x_1^p\right)}{f\left(x_{p+1}\middle|x_1^p\right)}\, dx_{p+1} \tag{11}$$

$$= D_{KL}\left[f\left(\cdot\,\middle|y_1^p,x_1^p\right)\middle\|\, f\left(\cdot\,\middle|x_1^p\right)\right], \tag{12}$$

we see that it is equal to the Kullback-Leibler divergence from the input-blind predictive density to the input-conditioned predictive density conditional on the specific input-output past. We take this Kullback-Leibler divergence as our definition of specific transfer entropy $\mathrm{t}_{Y\to X}(y_1^p,x_1^p)$ of order $p$,

$$\mathrm{t}_{Y\to X}\left(y_1^p,x_1^p\right) = D_{KL}\left[f\left(\cdot\,\middle|y_1^p,x_1^p\right)\middle\|\, f\left(\cdot\,\middle|x_1^p\right)\right]. \tag{13}$$

This transfer entropy is specific in the sense that it depends on the specific history $(y_1^p,x_1^p)$ of the input-output system. This is in analogy to the specific entropy rate introduced in Ref. [32], which quantifies the intrinsic uncertainty associated with a specific past of a stochastic process. In fact, by expanding (13), we see that

$$\mathrm{t}_{Y\to X}\left(y_1^p,x_1^p\right) = \int_{\mathbb{R}} f\left(x_{p+1}\middle|y_1^p,x_1^p\right) \ln f\left(x_{p+1}\middle|y_1^p,x_1^p\right) \times dx_{p+1} - \int_{\mathbb{R}} f\left(x_{p+1}\middle|y_1^p,x_1^p\right) \times \ln f\left(x_{p+1}\middle|x_1^p\right) dx_{p+1} \tag{14}$$

$$= -h\left[X_{p+1}\middle|Y_1^p = y_1^p, X_1^p = x_1^p\right] + \mathrm{c}_{Y\to X}\left(y_1^p,x_1^p\right), \tag{15}$$

where $h[X_{p+1}|Y_1^p = y_1^p, X_1^p = x_1^p]$ is an input-conditioned specific entropy rate and $\mathrm{c}_{Y\to X}(y_1^p,x_1^p)$ is a specific cross entropy.

The specific transfer entropy shares the desirable properties of local transfer entropy. Taking its expectation with respect to the input-output past recovers the overall transfer entropy. Moreover, in the case that the overall transfer entropy $\mathrm{T}^{(p)}_{Y\to X}$ is zero, we immediately have that specific transfer entropy $\mathrm{t}_{Y\to X}(y_1^p,x_1^p)$ is identically zero, since in that case the input-conditioned predictive density is equal to the input-blind predictive density. Unlike local transfer entropy, specific transfer entropy, as a Kullback-Leibler divergence, is non-negative and zero precisely when the input-blind and input-conditional predictive densities are identical almost everywhere. Its deviation from 0 indicates how much the input-blind predictive density differs from the input-conditioned predictive density, giving a state-specific quantification of the predictive impact of the input-output past on the output future. Moreover, we can relate specific transfer entropy to local transfer entropy via a conditional expectation of the latter,

$$E\left[\widetilde{\mathrm{t}}_{Y\to X}\left(Y_1^p,X_1^p,X_{p+1}\right)\middle|Y_1^p = y_1^p, X_1^p = x_1^p\right] \tag{16}$$

$$= \int_{\mathbb{R}} f\left(x_{p+1}\middle|y_1^p,x_1^p\right) \ln \frac{f\left(x_{p+1}\middle|y_1^p,x_1^p\right)}{f\left(x_{p+1}\middle|x_1^p\right)}\, dx_{p+1} \tag{17}$$

$$= \mathrm{t}_{Y\to X}\left(y_1^p,x_1^p\right). \tag{18}$$

Thus, specific transfer entropy is equivalent to local transfer entropy averaged over future outputs. For an illustration of the interpretation of specific transfer entropy, again consider Fig. 1. We see that specific transfer entropy initially decreases as a function of the input $y$, since as $y$ increases the input-blind predictive density becomes an increasingly better approximation of the future relative to the input-conditioned predictive density. However, for values of $y$ greater than 0, specific transfer entropy again increases since the input-blind predictive density becomes an increasingly poor approximation.

## III. ESTIMATION OF TOTAL, LOCAL, AND SPECIFIC TRANSFER ENTROPIES FROM OBSERVATIONS

Thus far we have presented the definitions of the total and local transfer entropies, and a specific transfer entropy, which can be computed when a model input-output process is known. In practice, the model for a set of observations is unknown, and we must estimate the transfer entropies from the data in hand. We consider three approaches to estimating total, local, and specific transfer entropies from data: plug-in estimators via kernel density estimators with bandwidths based on a normal reference, plug-in estimators using kernel density estimators with bandwidths tuned by $l$-block cross validation, and plug-in estimators using $k$th-nearest-neighbor estimators.

We first consider the plug-in estimators using kernel density estimators. We present the estimator for the input-conditioned predictive density, for which the estimator for the input-blind predictive density immediately follows. Consider a time series $\{(Y_t,X_t)\}_{t=1}^T$ from a proposed input-output system. To estimate the total, local, and specific transfer entropies requires the predictive density $f(x_{p+1}|y_1^p,x_1^p)$. A plug-in estimator substitutes an estimator $\hat{f}(x_{p+1}|y_1^p,x_1^p)$ for the true predictive density in their definitions. Recalling that the predictive density is given by $f(x_{p+1}|y_1^p,x_1^p) = f(y_1^p,x_1^p,x_{p+1})/f(y_1^p,x_1^p)$ we can estimate the predictive density by estimating the joint density $f(y_1^p,x_1^p,x_{p+1})$ and its marginal density $f(y_1^p,x_1^p)$ and taking their ratio. We estimate the marginal and joint densities using kernel density estimators with product kernels and bandwidths $\mathbf{k}_y = (k_{y,1},\ldots,k_{y,p}), \mathbf{k}_x = (k_{x,1},\ldots,k_{x,p})$, and $k_{x,p+1}$. Note that the joint and marginal density estimators are coupled through the common bandwidths $\mathbf{k}_x$ and $\mathbf{k}_y$ used in both estimators. This coupling is necessary to ensure that $\hat{f}(x_{p+1}|y_1^p,x_1^p)$ is a probability density function, i.e., it integrates to 1 with respect to $x_{p+1}$.

The kernel density estimators require the specification of the kernel $K$, the model order $p$, and the bandwidths $\mathbf{k}_x,\mathbf{k}_y,k_{x,p+1}$. In practice, the kernel choice has little effect on

the estimator, and we use a multivariate product kernel with each univariate kernel given by $K(x) = \phi(x)$, the probability density function for a standard normal random variable. The bandwidths and model order will have a larger impact on the estimation of the transfer entropies. A rule of thumb for the bandwidths of a $d$-variate kernel density estimator using a normal reference density suggests bandwidths $k_j = T^{-1/(d+4)}\hat{\sigma}_j$ where $\hat{\sigma}_j$ is the sample standard deviation of the $j$th variate [33]. Thus, for the input-blind predictive density we take $k_{x,j} = T^{-1/(p+5)}\hat{\sigma}_X$ and for the input-conditioned predictive density we take $k_{x,j} = T^{-1/(2p+5)}\hat{\sigma}_X$ and $k_{y,j} = T^{-1/(2p+5)}\hat{\sigma}_Y$. This leaves the choice of the model order $p$. We choose the model order via $l$-block cross validation [34] of the negative log likelihood of the conditional density as in Ref. [32]. That is, we take the model order $p^* \in \{0,1,\ldots,p_{\max}\}$ for some prespecified $p_{\max}$ that minimizes

$$-\frac{1}{T - p_{\max}} \sum_{t=p_{\max}+1}^{T} \ln \hat{f}_{-t:l}\left(X_t \middle| Y_{t-p}^{t-1}, X_{t-p}^{t-1}\right), \qquad (19)$$

where $\hat{f}_{-t:l}$ is the kernel density estimator for the conditional density constructed using all of the data except the $2l+1$ observations about and including $t$. The half-window length $l$ is taken to be a fixed fraction of the time series length, typically $1/6$ or $1/8$ of the total length $T$. Thus, this objective function quantifies our average surprise at seeing a particular output future following a particular input-output past, without biasing the result by including the temporal dependencies around each evaluation point. For $p$ too small, we will have excess surprise about the future because we have not sufficiently modeled the input-output process. For $p$ too large, we will have excess surprise due to overfitting the predictive density. We therefore take a value of $p$ that balances these two sources of excess surprise.

The rule-of-thumb bandwidths are asymptotically optimal for the mean integrated squared error using a normal kernel with respect to a normal reference distribution. The rule-of-thumb bandwidths have the advantage that they are quick to compute, but they may result in suboptimal bandwidths for densities that deviate from joint normality or for small sample sizes. We thus also consider a kernel density estimator where the bandwidths are also chosen to minimize a cross-validation score, following Ref. [35]. In addition to the model order, we also choose the bandwidths to minimize the $l$-block cross-validation score given by (19). Both theoretical and empirical work has shown that choosing the bandwidth via cross validation can automatically remove irrelevant predictors by setting their bandwidths very large [35,36]. This is clearly desirable in the input-output time series case, since we desire to induce conditional independence between the future of the output process and the distant past of the input-output process, as well as to detect when the proposed input process is not relevant to the output process. Consider, for example, the extreme case where the past of a nominal input process is irrelevant to the future of the nominal output process for prediction. By this coupling, we can ignore the input past by setting the bandwidths $\mathbf{k}_y$ to large values. This has the effect of giving $\hat{f}(x_{p+1}|y_1^p, x_1^p) \approx \hat{f}(x_{p+1}|x_1^p)$ and recovering the appropriate conditional independence relationship. A similar advantage is gained if the most recent past of the input-output process screens off its distant past. We minimize the $l$-block cross-validation score with respect to the bandwidths using Nelder-Mead [37]. To accelerate the bandwidth selection process, we apply a warm start strategy in choosing the initial bandwidths for each new value of $p$. That is, after the optimal bandwidths for an order-$p$ model have been determined, we then take those bandwidths as the initial guess for the bandwidths for the output future and input-output past up to order $p$, and set the $(p+1)$st bandwidths according to the normal reference rule of thumb. In addition, if a bandwidth is set very large after optimization, we remove that lag from the model for all additional $p$.

We next consider estimators of the transfer entropies based on nearest-neighbor statistics. There are many $k$th-nearest-neighbor estimators for total transfer entropy. For a recent review, see Ref. [38]. We use the first estimator from Ref. [39] as implemented in JIDT [16]. This estimator applies the Kraskov-Stögbauer-Grassberger estimator for mutual information [40] to transfer entropy estimation. For a fixed nearest-neighbor value of $k$, the estimator is given by

$$\begin{aligned}\widehat{\mathrm{T}}_{Y\to X}^{(p)} = \frac{1}{T-p}\sum_{t=p+1}^{T} &\psi(k) + \psi[N_{\mathcal{X}^p}(t;\rho_{t,k}) + 1] \\ &-\psi[N_{\mathcal{X}^{p+1}}(t;\rho_{t,k}) + 1] - \psi[N_{\mathcal{Y}^p\times\mathcal{X}^p}(t;\rho_{t,k}) + 1],\end{aligned} \qquad (20)$$

where $\psi$ is the digamma function, $\rho_{t,k}$ is the distance to the $k$th nearest neighbor of $(Y_{t-p}^{t-1}, X_{t-p}^{t-1}, X_t)$ under the infinity norm, and $N_{\mathcal{S}}(t;\rho_{t,k})$ is the number of sample points within a distance $\rho_{t,k}$ in the space $\mathcal{S}$. The $k$th-nearest-neighbor estimator of local transfer entropy is then given by undoing the averaging in (20), giving

$$\begin{aligned}\widehat{\mathrm{t}}_{Y\to X}^{(p)}\left(y_{t-p}^{t-1}, x_{t-p}^{t-1}, x_t\right) = &\psi(k) + \psi[N_{\mathcal{X}^p}(t;\rho_{t,k}) + 1] \\ &-\psi[N_{\mathcal{X}^{p+1}}(t;\rho_{t,k}) + 1] \\ &-\psi[N_{\mathcal{Y}^p\times\mathcal{X}^p}(t;\rho_{t,k}) + 1].\end{aligned} \qquad (21)$$

The free parameters for the $k$th-nearest-neighbor estimators for total and local transfer entropy are the model order $p$ and the nearest-neighbor number $k$. For a fixed model order $p$, the nearest-neighbor number $k$ balances between the bias and variance of the estimator $\widehat{\mathrm{T}}_{Y\to X}^{(p)}$. The $k$th-nearest-neighbor estimator on which the Kraskov-Stögbauer-Grassberger estimator is based is known to be asymptotically unbiased for any fixed $k$. However, for finite sample sizes, the bias is nonzero, and will be smaller for smaller $k$. However, smaller $k$ will also result in a higher variance for the estimator. We follow the standard practice and fix $k = 4$ in all of our investigations. For model selection for the $k$th-nearest-neighbor-based estimator of the transfer entropies, we choose the model order $p$ to minimize the mean squared error between the $k^{\mathrm{reg}}$-nearest-neighbors prediction of the output future based on the output pasts of order $p$ and the true output future. This is the self-predictively optimal (SPO) formulation of [41] using the

model selection criterion from Ref. [21]. That is, we seek the $p$ that minimizes $\frac{1}{T-p_{\max}}\sum_{t=p_{\max}+1}^{T}(\hat{X}_t^{(p)}-X_t)^2$ where $\hat{X}_t^{(p)}=\sum_{t'=p_{\max}+1}^{T}W_{t',k^{\mathrm{reg}}}X_{t'}$ is the $k^{\mathrm{reg}}$-nearest-neighbors predictor of $X_t$ with $W_{t',k^{\mathrm{reg}}}=1/k^{\mathrm{reg}}$ when $X_{t'-p}^{t'-1}$ is one of the $k^{\mathrm{reg}}$ nearest neighbor to $X_{t-p}^{t-1}$ and zero otherwise. This requires the choice of $k^{\mathrm{reg}}$ for the $k^{\mathrm{reg}}$-nearest-neighbors predictor. While the number of nearest neighbors could be chosen in a data-driven manner, for computational efficiency we rely on a rule of thumb based on consistency results for $k^{\mathrm{reg}}$-nearest-neighbors regression. In particular, it has been shown that under relatively general conditions [42], a $k^{\mathrm{reg}}$-nearest-neighbors estimator is consistent if $k^{\mathrm{reg}}(T)$ is taken to grow faster than $\ln T$ but slower than $T$. We thus take $k^{\mathrm{reg}}(T)=\lfloor\sqrt{T}\rfloor$, which satisfies these bounds.

As noted in Sec. II, specific transfer entropy can be written as the expected value of local transfer entropy conditional on the input-output past. As such, an estimator of specific transfer entropy can be obtained from (21) by regressing local transfer entropy on the input-output past. Any nonparametric smoother may be used. We perform the smoothing using a $k^{\mathrm{reg}}$-nearest-neighbors regression. For a nearest-neighbor number $k^{\mathrm{reg}}$, the estimator of specific transfer entropy is then given by

$$\widehat{\mathrm{t}}_{Y\to X}^{(p)}\left(y_1^p,x_1^p\right)=\sum_{t=p+1}^{T}W_{t,k^{\mathrm{reg}}}\cdot\widehat{\widetilde{\mathrm{t}}}_{Y\to X}^{(p)}\left(Y_{t-p}^{t-1},X_{t-p}^{t-1},X_t\right),\tag{22}$$

where $W_{t,k^{\mathrm{reg}}}=1/k^{\mathrm{reg}}$ when $(Y_{t-p}^{t-1},X_{t-p}^{t-1})$ is one of the $k^{\mathrm{reg}}$ nearest neighbors to $(y_1^p,x_1^p)$ and zero otherwise. We take $k^{\mathrm{reg}}=\lfloor\sqrt{T}\rfloor$ as we did in model selection. Note that unlike the kernel density-based estimators of specific entropy rate, the $k$th-nearest-neighbor-based estimator can result in negative specific transfer entropies.

## IV. A SMOOTH THRESHOLD AUTOREGRESSIVE MODEL WITH EXOGENOUS DRIVER

As our first model system, we consider a smooth threshold autoregressive model with an exogenous driver (STARX) [43]. Threshold autoregressive (TAR) models were first systematically developed in Ref. [44] as simple nonlinear autoregressive models that captured many properties of nonlinear time series including subharmonics, amplitude-frequency dependence, and time irreversibility. In the simplest case, they achieve this goal by a piecewise linearization of the update equation according to thresholds in the state space of the system. Thus, such models are locally linear, and can parsimoniously approximate a nonlinear system.

For our purposes, we consider the STARX class of nonlinear input-output systems to allow for both nontrivial nonlinearity and analytical tractability. The STARX model incorporates an exogenous input whose value induces a smooth thresholding between two or more linear autoregressive models for the output. For our model system, we take the exogenous input time series $\{Y_t\}_{t\in\mathbb{Z}}$ to be a linear autoregressive process of order 1 [AR(1)], and the output time series $\{X_t\}_{t\in\mathbb{Z}}$ to switch smoothly between two AR(1) models depending on the previous value of the input time series. This model can be expressed as

$$Y_t=cY_{t-1}+d\eta_t\tag{23}$$

$$X_t=w(Y_{t-1})(b^{(1)}X_{t-1}+a^{(1)}\epsilon_t)+(1-w(Y_{t-1}))(b^{(0)}X_{t-1}+a^{(0)}\epsilon_t),\tag{24}$$

where $\{\eta_t\}_{t\in\mathbb{Z}}$ and $\{\epsilon_t\}_{t\in\mathbb{Z}}$ are mutually and serially independent, identically distributed standard normal random variables. The threshold function $w(y)$ controls the switching between the two AR(1) models, and we take $w(y)=\Phi(\frac{y}{s})$, a sigmoidal function given by the rescaling of the cumulative distribution function $\Phi(y)$ of a standard normal random variable.

To compute the specific transfer entropy of this system, we require the specific input-conditioned entropy rate and and specific cross entropy. The specific input-conditioned entropy rate $h[X_t|X_{t-1}=x,Y_{t-1}=y]$ of this model can be computed exactly. To see this, we begin by regrouping the state and dynamical noise terms in (24),

$$X_t=\{w(Y_{t-1})b^{(1)}+[1-w(Y_{t-1})]b^{(0)}\}X_{t-1}+\{w(Y_{t-1})a^{(1)}+[1-w(Y_{t-1})]a^{(0)}\}\epsilon_t,\tag{25}$$

which shows that conditional on $X_{t-1}$ and $Y_{t-1}$, $X_t$ is normally distributed with mean

$$E[X_t|X_{t-1}=x,Y_{t-1}=y]=\{w(y)b^{(1)}+[1-w(y)]b^{(0)}\}x\tag{26}$$

$$=[b^{(0)}+(b^{(1)}-b^{(0)})w(y)]x=m(x,y)\tag{27}$$

and variance

$$\mathrm{Var}[X_t|X_{t-1}=x,Y_{t-1}=y]=\{w(y)a^{(1)}+[1-w(y)]a^{(0)}\}^2=v(y).\tag{28}$$

Because $X_t$ is conditionally normal, we immediately have that the specific input-conditioned entropy rate of $X_t$ is

$$h[X_t|X_{t-1}=x,Y_{t-1}=y]=\tfrac{1}{2}\ln[2\pi e\cdot v(y)].\tag{29}$$

Thus, while our prediction of the future value of $X_t$ depends on both $x$ and $y$, our uncertainty about the future value of $X_t$ only depends on $y$. Depending on the conditional variance function $v(y)$, the output will be more or less predictable for different values of $y$. To compute the specific cross entropy, we also require the input-blind predictive density. Under the assumption of stationarity, the input-blind predictive density is computable from the joint transition density, and can be approximated numerically. See Appendix A for additional details.

We next consider the STARX system with parameters $c=0.8$, $d=1$, $a^{(1)}=2$, $a^{(0)}=1$, $b^{(1)}=1/2$, $b^{(0)}=-1/2$, and $s=1$. Figure 2 shows a particular realization from the input-output system. The output $X_t$ is shaded according to $Y_{t-1}$ to indicate how the dynamics vary according to the input, with red corresponding to negative $Y_{t-1}$ and blue corresponding to positive $Y_{t-1}$. When the previous inputs are negative, the output exhibits a negative autocorrelation (red), while when the previous inputs are positive, the output exhibits positive autocorrelation (blue). Thus, as the slowly varying input shifts from negative to positive values, the output shifts from exhibiting high-frequency dynamics to low-frequency dynamics, and

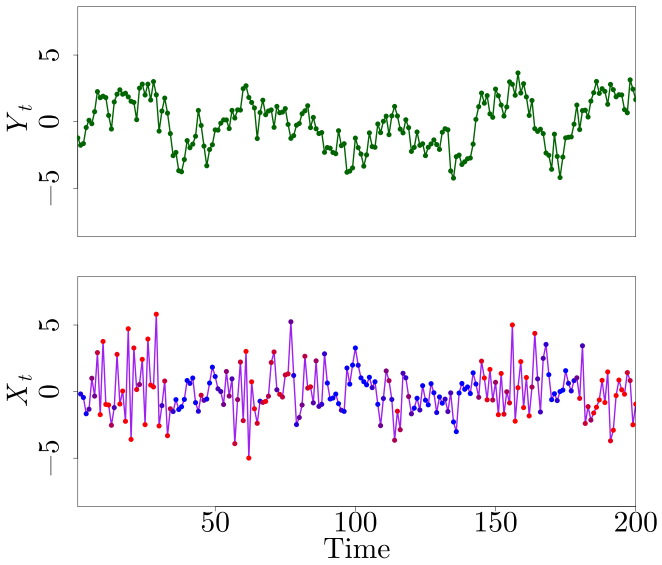



FIG. 2. An example realization of the input (top) and output (bottom) from the STARX input-output system with the parameters given in the text. The output $X_t$ is shaded according to $Y_{t-1}$ to indicate how the dynamics vary according to the input, with blue corresponding to negative $Y_{t-1}$ and red corresponding to positive $Y_{t-1}$.

the STARX system exhibits an amplitude-frequency coupling. Figure 3 shows the specific transfer entropy $\mathrm{t}_{Y\to X}(y,x)$ for the STARX system with these parameters. We see that if we fix the output past, the input past becomes more predictively informative away from 0, and becomes most informative as it becomes more negative. As the input past becomes more negative, $w(y_{t-1}) \to 0$, and the output dynamics are governed by $X_t = -1/2X_{t-1} + \epsilon_t$. In contrast, as the input past becomes more positive, $w(y_{t-1}) \to 1$, and the output dynamics are governed by $X_t = 1/2X_{t-1} + 2\epsilon_t$. Thus, for a negative input, the input-conditioned predictive density has a much smaller variance than for a positive input relative to

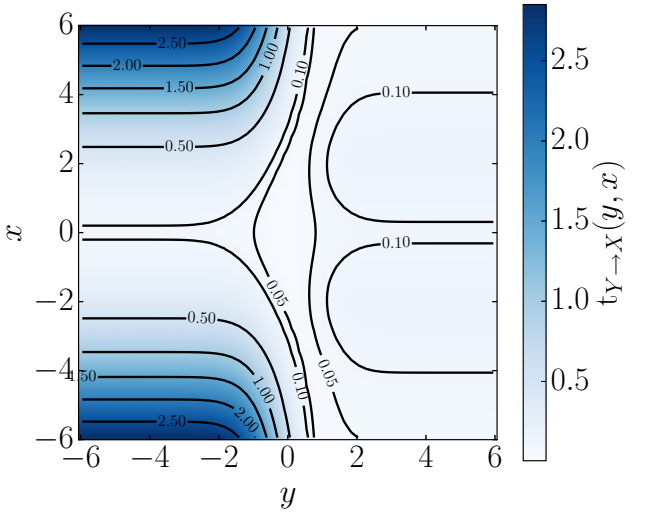



FIG. 3. A contour plot of the specific transfer entropy $\mathrm{t}_{Y\to X}(y,x)$ from the input process to the output process for the STARX model.

the input-blind predictive density, and thus provides a greater amount of predictive information. If we fix the input past, as the output past deviates from 0, the input past provides more predictive information because with the input it can distinguish between the negative lag-1 autocorrelation when $y_{t-1}$ is negative, and the positive lag-1 autocorrelation when $y_{t-1}$ is positive. Because the output process viewed marginally has a positive lag-1 autocorrelation, the input past is most informative for negative values.

To determine how well the three estimators for the total, local, and specific transfer entropies perform, we generated $B = 1000$ independent input-output time series of length $T = 1000$ and estimate the transfer entropies as described in Sec. III for the first $T_{\mathrm{sub}} = 100, 200, 500, 1000$ time points of the 1000-time-point series. For all estimators, we fix $p_{\max} = 10$, and use $l = \lfloor \frac{T_{\mathrm{sub}}}{8} \rfloor$ for the kernel density estimators and $k = 4$ and $k^{\mathrm{reg}} = \lfloor \sqrt{T_{\mathrm{sub}}} \rfloor$ for the $k$th-nearest-neighbor estimators. By considering the values of the estimates across the 1000 time series, we approximate the sampling distribution of the estimators. Figure 4 shows the local and specific transfer entropies as a function of time for a 100-time-step portion of one of the realizations. The top panels show the transfer entropies in the input-to-output direction where $\mathrm{T}_{Y\to X} > 0$ and the bottom panels show the transfer entropies in the output-to-input direction where $\mathrm{T}_{X\to Y} = 0$. The solid lines indicate the exact transfer entropies and the dotted lines indicate the transfer entropies estimated from the data. As expected, we see that both the exact local and specific transfer entropies vary with time for the input-to-output direction and are identically zero for the output-to-input direction. Moreover, we observe that while the local transfer entropy takes both positive and negative values, as discussed above, the specific transfer entropy is always non-negative. The estimates of local and specific transfer entropies track the true local and specific entropies with varying precision.

We summarize the sampling distributions of the different estimators for the total transfer entropy in Fig. 5 in the input-to-output ($Y \to X$) and output-to-input ($X \to Y$) directions. The points indicate the mean value of the estimates, the thick lines cover from the 0.25 sample quantiles to the 0.75 sample quantiles, and the thin lines cover from the 0.025 sample quantiles to the 0.975 sample quantiles. In the input-to-output direction where $\mathrm{T}_{Y\to X} > 0$, we see that the sampling distribution becomes more concentrated as $T_{\mathrm{sub}}$ increases. The dashed horizontal line indicates the order-1 total transfer entropy, $\mathrm{T}^{(1)}_{Y\to X} \approx 0.0939$. The mean value of the estimators approach a value smaller than $\mathrm{T}^{(1)}_{Y\to X}$. This is because the estimators are designed to estimate $\mathrm{T}^{(\infty)}_{Y\to X}$. While the overall STARX input-output system is a Markov process of order 1, the output process considered alone appears to have a Markov order larger than 1. This is typical for a subprocess of a vector autoregressive process [45], where the subprocess may have an infinite order despite the overall process having a finite order. Because the model selection process determines estimators for the best order-$p^*$ approximation to $\mathrm{T}^{(\infty)}_{Y\to X}$, the estimators will generally give values for the total transfer entropy less than $\mathrm{T}^{(1)}_{Y\to X}$. In the output-to-input direction, we have that $\mathrm{T}^{(\infty)}_{X\to Y} = \mathrm{T}^{(1)}_{X\to Y} = 0$. For this direction, the sampling distributions for the estimators based on the kernel density

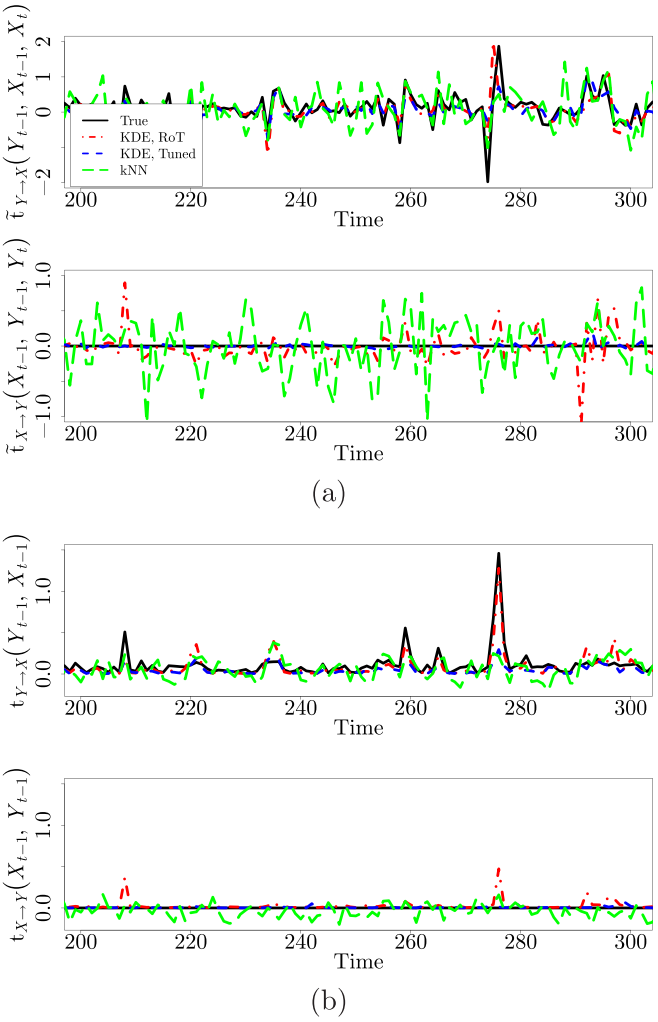



FIG. 4. The (a) local and (b) specific transfer entropies and their estimates in the input-to-output (top of panels) and output-to-input (bottom of panels) directions from a particular realization of the STARX system with $T_{\text{sub}} = 1000$. The exact transfer entropies are indicated by black (solid) lines, while the estimates based on kernel density estimation with rule of thumb (RoT) and tuned bandwidths and $k$th nearest neighbor (kNN) are indicated by red (dot-dashed), blue (short dashed), and green (long dashed) lines, respectively.

estimator with a tuned bandwidth and the $k$th nearest neighbors both concentrate at 0, with less sampling variability for the kernel density estimator with a tuned bandwidth. The reduced variability in this case occurs because with a long enough time series, the bandwidth tuning eliminates the irrelevant past of the output time series for predicting the input time series, and thus exactly recovers $\widehat{\mathrm{T}}^{(1)}_{Y\to X} = 0$. The estimator based on the kernel density estimators with rule-of-thumb bandwidths maintains a negative finite sample bias away from the true value of 0. Thus, we see the benefit of estimating the input-output predictive density directly via tuning the bandwidths rather than indirectly via attempting to estimate both the numerator and denominator of the input-conditioned predictive density.

We next consider the finite sample properties of the estimators of the local transfer entropy viewed as a time series.

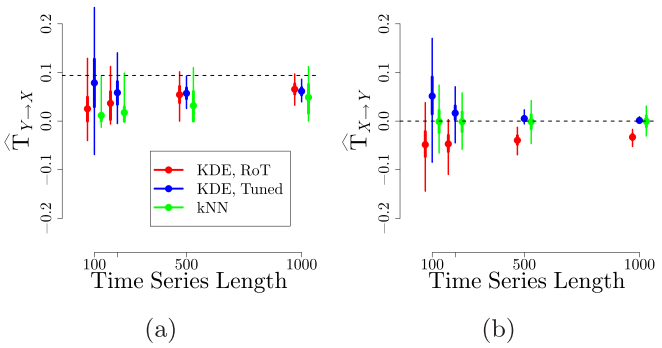



FIG. 5. A summary of the sampling distribution for the total transfer entropy estimates based on kernel density estimation with rule-of-thumb (RoT) and tuned bandwidths and $k$th nearest neighbor (kNN) in the (a) input-to-output and (b) output-to-input directions. Thick lines indicate the 0.25–0.75 quantiles of the 1000 realizations, and thin lines indicate the 0.025–0.975 quantiles. The dashed horizontal lines indicate $\mathrm{T}^{(1)}_{Y\to X} \approx 0.0939$ and $\mathrm{T}_{X\to Y} = 0$.

For each value of $T_{\text{sub}}$, we compute the mean absolute error between the true local transfer entropy and the estimate for the local transfer entropy over time. Figure 6 summarizes the performance of estimators under the mean absolute error. We see that in both the input-to-output and output-to-input directions, the estimators based on kernel density estimation have decreasing error with increasing time series length. The error in the input-to-output direction does not approach zero. This is again due to the discrepancy between considering an estimator for $\widetilde{\mathrm{t}}_{Y\to X}(y^{t-1}_{-\infty}, x^{t-1}_{-\infty}, x_t)$ and comparing to $\widetilde{\mathrm{t}}_{Y\to X}(y_{t-1}, x_{t-1}, x_t)$. In contrast, we see that the estimator based on the $k$th nearest neighbors has increasing error with increasing time series length. This counterintuitive result occurs because while the total transfer entropy estimator (20) is consistent for the total transfer entropy, each term in that estimator is not a consistent estimator for the local transfer entropy. This is due to the inconsistency of the $k$th-nearest-neighbor density estimator with $k$ fixed [46]. To recover consistency, we would need to take $k$ to grow with $T$. Thus, we see that if an estimator for the local

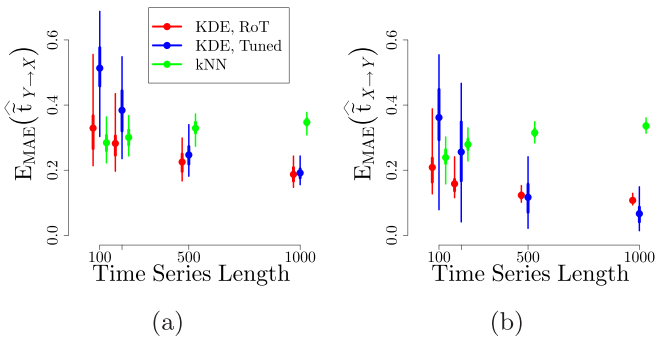



FIG. 6. A summary of the sampling distribution for the mean absolute error between the true local transfer entropy and the estimates based on kernel density estimation with rule-of-thumb (RoT) and tuned bandwidths and $k$th nearest neighbor (kNN) in the (a) input-to-output and (b) output-to-input directions. Thick lines indicate the 0.25–0.75 quantiles of the 1000 realizations, and thin lines indicate the 0.025–0.975 quantiles.

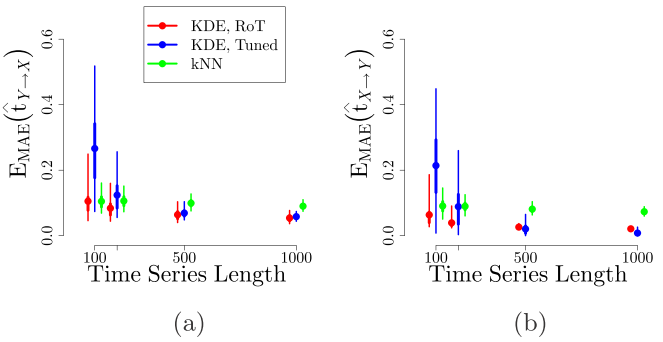



FIG. 7. A summary of the sampling distribution for the mean absolute error between the true specific transfer entropy and the estimates based on kernel density estimation with rule-of-thumb (RoT) and tuned bandwidths and $k$th nearest neighbor (kNN) in the (a) input-to-output and (b) output-to-input directions. Thick lines indicate the 0.25–0.75 quantiles of the 1000 realizations, and thin lines indicate the 0.025–0.975 quantiles.

transfer entropy is desired, the $k$th-nearest-neighbor estimator should use a $k$ that scales with the length of the time series.

Finally, we consider the finite sample properties of the estimators of the specific transfer entropy viewed as a time series. Now we compute the mean absolute error between the true specific transfer entropy and the estimate for the specific transfer entropy over time. Figure 7 summarizes the performance of estimators under the mean absolute error. Once again, all of the estimators exhibit decreasing error with increasing sample size. The estimator (22) based on $k$th nearest neighbors regains its consistency because the estimator incorporates an averaging over $k^{\text{reg}}$ neighbors with $k^{\text{reg}}$ growing with time series length. Thus, while the $k$th-nearest-neighbor estimator with $k$ fixed is inappropriate for local transfer entropy, it can be used for specific transfer entropy. However, we see that for moderate length time series, the estimators based on kernel density estimation outperform the $k$th-nearest-neighbor estimator. The estimator using kernel density estimation with tuned bandwidths performs especially well in the output-to-input direction, since the bandwidth tuning screens off the irrelevant output past for the input process, and thus recovers $\widehat{\mathrm{t}}_{X\to Y} \equiv 0$.

## V. COUPLED STOCHASTIC HÉNON MAPS

As our second example, we consider a stochastic version [47] of the coupled Hénon maps studied in [48]. We first consider the unidirectionally coupled case. The dynamics of the input system $Y$ and the output system $X$ are given by

$$Y_{1,t} = 1.4 - Y_{1,t-1}^2 + 0.3Y_{2,t-1} + \sigma_{\epsilon_1}\epsilon_{1,t} \tag{30}$$

$$Y_{2,t} = Y_{1,t-1} + \sigma_{\epsilon_2}\epsilon_{2,t} \tag{31}$$

$$X_{1,t} = 1.4 - [CY_{1,t-1} + (1-C)X_{1,t-1}] \times X_{1,t-1} + 0.3X_{2,t-1} + \sigma_{\eta_1}\eta_{1,t} \tag{32}$$

$$X_{2,t} = X_{1,t-1} + \sigma_{\eta_2}\eta_{2,t}, \tag{33}$$

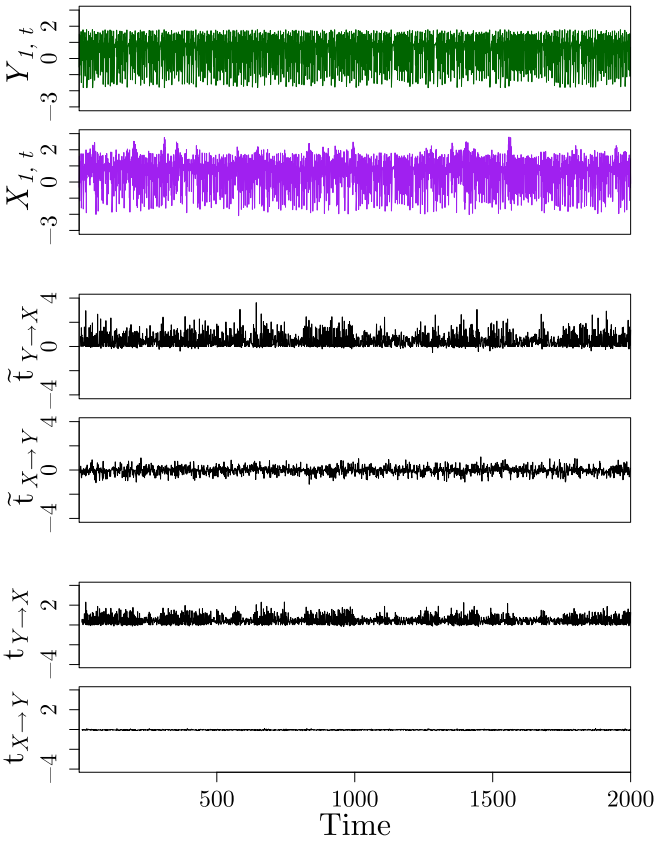



FIG. 8. Results for the unidirectionally coupled stochastic Hénon maps. (a) Input and output time series. (b) Local transfer entropy. (c) Specific transfer entropy.

where $\{\epsilon_{1,t}\}_{t\in\mathbb{Z}}, \{\epsilon_{2,t}\}_{t\in\mathbb{Z}}, \{\eta_{1,t}\}_{t\in\mathbb{Z}}, \{\eta_{2,t}\}_{t\in\mathbb{Z}}$ are mutually and serially independent, identically distributed standard normal random variables with associated dynamical noise amplitudes of $\sigma_{\epsilon_1}, \sigma_{\epsilon_2}, \sigma_{\eta_1}, \sigma_{\eta_2}$ and $C$ is the coupling strength from $Y$ to $X$.

To explore the local and specific transfer entropies for this system, we generate a length $T = 80000$ time series with $C = 0.6$ and $\sigma_{\epsilon_1} = \sigma_{\epsilon_2} = \sigma_{\eta_1} = \sigma_{\eta_2} = 0.004$ and estimate the total, local, and specific transfer entropies using the $k$th-nearest-neighbor estimators with the first state variables $\{(Y_{1,t}, X_{1,t})\}_{t=1}^{T}$ as the observations from the input-output system. Model selection chose $p^* = 4$ in the input-to-output direction and $p^* = 2$ in the output-to-input direction. The total transfer entropies were estimated to be $\widehat{\mathrm{T}}_{Y\to X} = 0.47$ in the input-to-output direction and $\widehat{\mathrm{T}}_{X\to Y} = -0.04$ in the output-to-input direction. Thus, the total transfer entropy correctly identifies that predictive information is positive from the input to the output and zero from the output to the input. Figure 8 shows the estimates for the [Fig. 8(a)] local and [Fig. 8(b)] specific transfer entropies for the first 2000 time steps of the time series. The estimate of the local transfer entropy in the input-to-output direction is generally positive, and varies dramatically depending on the state of the input-output system. The estimate of the local transfer entropy in the output-to-input direction varies around 0, and as with the STARX system, we see that the estimate has nonvanishing sampling variability due to fixing $k = 4$. The estimates of the specific transfer entropy more clearly show that predictive information is only

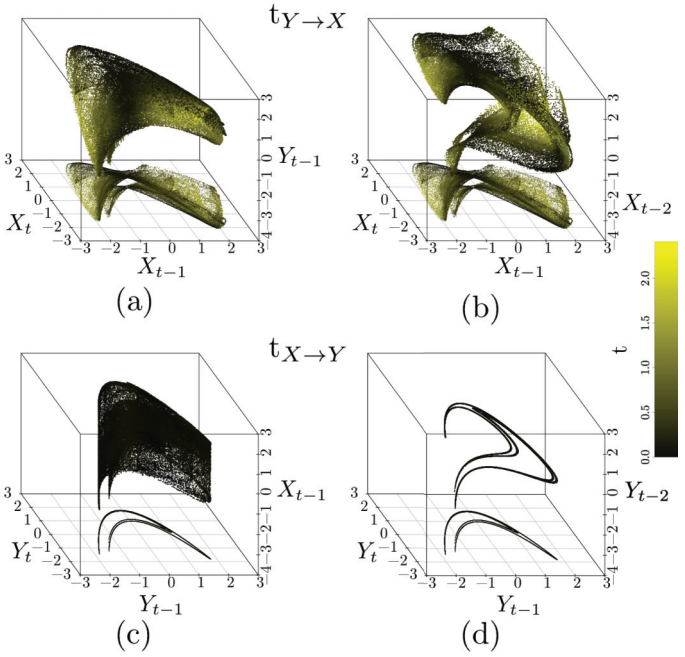



FIG. 9. The specific transfer entropies as a function of the observed state variables for coupled stochastic Hénon maps with unidirectional coupling. (a) and (c) show the nominal output future as a function of the nominal input-output past in the input-to-output and output-to-input direction while (b) and (d) show the nominal output future as a function of two lags of the nominal output past. The projections onto the planes show the nominal output future as a function of the most recent nominal output past. Each point is shaded according to the estimated specific transfer entropy associated with that nominal input-output state in the input-to-output direction [(a) and (b)] and output-to-input direction [(c) and (d)], with black (dark) indicating low specific transfer entropy and yellow (light) indicating high specific transfer entropy.

transferred from the input to the output: $\hat{t}_{Y\to X}$ is clearly positive for most values of the input-output past, while $\hat{t}_{X\to Y}$ is approximately zero for all values of the input-output past.

We have seen how the specific transfer entropy varies as a function of time. Next we consider how the specific transfer entropy varies as a function of the input-output pasts. Because the model orders $p^*$ are greater than 1, we cannot directly visualize the output future as a function of the $p^*$ lags of the input and output pasts. Instead, we consider two projections of the input-output pasts: the previous input-output pair and the previous two output pairs. We show the nominal output future as a function of these two projections in Fig. 9. For reference, we also show the nominal output future as a function of the most recent nominal output past projected onto the plane. Each point is shaded according to the order-$p^*$ specific transfer entropy estimated for that input-output past. Consider Figs. 9(a) and 9(b). These correspond to the case with a positive coupling. We see that the regions of the input-output past that provide predictive information correspond to those regions where inclusion of the input past unfolds the attractor more than inclusion of an additional lag of the output past. For example, for large magnitude $X_{1,t-1}$, $Y_{1,t-1}$ provides more information about $X_{1,t}$, while for $X_{1,t-1}$ close to 0, this is not the case. This agrees with the update equation for $X_{1,t}$, since the impact of $Y_{1,t-1}$ is modulated by the magnitude of $X_{1,t-1}$. Contrast this with Figs. 9(c) and 9(d), which correspond to the output-to-input direction. Here, $X_{1,t-1}$ provides no predictive information for $Y_{1,t}$, and thus the attractor does not unfold upon inclusion of $X_{1,t-1}$, while it unfolds completely after inclusion of $Y_{1,t-2}$. Again, this agrees with the update equation for $Y_{1,t}$ since knowledge of $Y_{1,t-1}$ and $Y_{1,t-2}$ completely specifies $Y_{1,t}$ up to dynamical noise.

As our final example, we consider stochastic Hénon maps where the system switches from no coupling to bidirectional coupling. In the case of time-dependent coupling, the dynamics of the two systems are given by

$$Y_{1,t} = 1.4 - (C_{X\to Y}(t)X_{1,t-1} + (1 - C_{X\to Y}(t))Y_{1,t-1}) \times Y_{1,t-1} + 0.3Y_{2,t-1} + \sigma_{\epsilon_1}\epsilon_{1,t} \quad (34)$$

$$Y_{2,t} = Y_{1,t-1} + \sigma_{\epsilon_2}\epsilon_{2,t} \quad (35)$$

$$X_{1,t} = 1.4 - (C_{Y\to X}(t)Y_{1,t-1} + (1 - C_{Y\to X}(t))X_{1,t-1}) \times X_{1,t-1} + 0.3X_{2,t-1} + \sigma_{\eta_1}\eta_{1,t} \quad (36)$$

$$X_{2,t} = X_{1,t-1} + \sigma_{\eta_2}\eta_{2,t}, \quad (37)$$

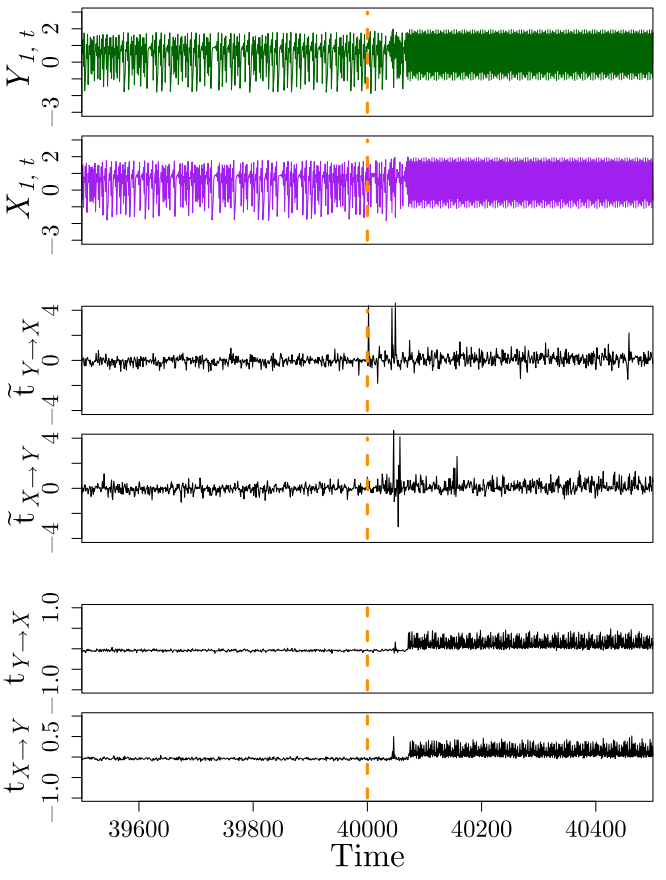



FIG. 10. Results for the stochastic Hénon maps with switched coupling. (a) Input and output time series. (b) Local transfer entropy. (c) Specific transfer entropy. The dashed orange line indicates when the coupling switches on.

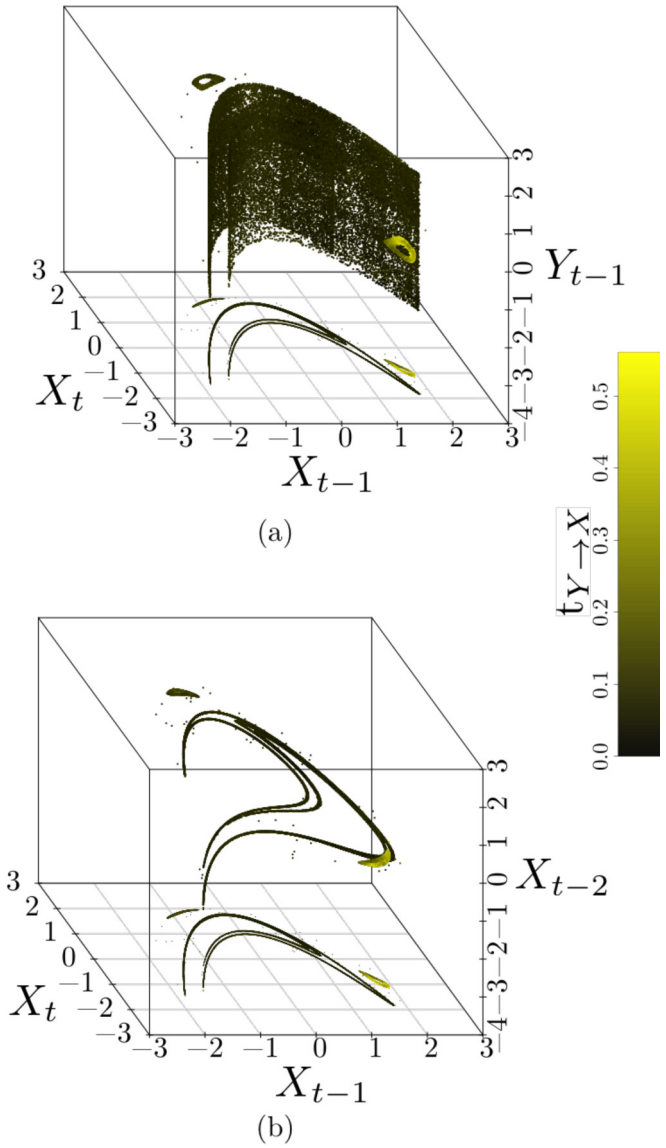



FIG. 11. The specific transfer entropies as a function of the observed state variables for coupled stochastic Hénon maps with switching coupling. (a) shows the nominal output future as a function of the nominal input-output past and (b) shows the nominal output future as a function of two lags of the nominal output past. The projections onto the planes show the nominal output future as a function of the most recent nominal output past. Each point is shaded according to the specific transfer entropy associated with that nominal input-output state, with black (dark) indicating low specific transfer entropy and yellow (light) indicating high specific transfer entropy.

where the dynamical noise terms are as before and the coupling terms $C_{X\to Y}(t)$ and $C_{Y\to X}(t)$ can vary as a function of time.

We again generate a length $T = 80000$ time series from this system with the dynamical noise amplitudes as before and with the coupling terms in both directions given by $C_{X\to Y}(t) = C_{Y\to X}(t) = 0$ for $t \leqslant 40000$ and $C_{X\to Y}(t) = C_{Y\to X}(t) = 0.27$ for $t > 40000$. Note that in this case, the input-output system is not conditionally stationary: to induce conditional stationarity, we would also have to condition on $C_{X\to Y}(t)$ and $C_{Y\to X}(t)$. However, we proceed to see how the estimators for total, local, and specific transfer entropies behave under this violation. Model selection chose $p^* = 3$ in both directions. The overall transfer entropy estimates are $\widehat{\mathrm{T}}_{Y\to X} = 0.06$ and $\widehat{\mathrm{T}}_{X\to Y} = 0.06$. This is a mixing of the estimates from the first half of the time series, where $\widehat{\mathrm{T}}_{Y\to X} = -0.04$ and $\widehat{\mathrm{T}}_{X\to Y} = -0.04$, and the second half, where $\widehat{\mathrm{T}}_{Y\to X} = 0.15$ and $\widehat{\mathrm{T}}_{X\to Y} = 0.15$. Figure 10 shows the estimates for the (a) local and (b) specific transfer entropies from 500 time points before to 500 time points after the switch from $C_{X\to Y}(t) = C_{Y\to X}(t) = 0$ to $C_{X\to Y}(t) = C_{Y\to X}(t) = 0.27$. After an initial transient, we see that the coupled system transitions from asynchronous dynamics to synchronous dynamics. This is especially apparent from the estimated specific transfer entropy, which transitions from being nearly identically zero before the coupling to positive after the coupling. The transition is less obvious from the local transfer entropy.

As with the unidirectionally coupled system, we also consider how the specific transfer entropy varies across the input-output state space. Figure 11 shows the nominal output future as a function of [Fig. 11(a)] the nominal input-output past and [Fig. 11(b)] two time steps of the nominal output past again colored according to the estimated specific transfer entropy. We only consider a single direction, $Y \to X$, due to the symmetric nature of the system. We see that during the period without coupling the specific transfer entropy is nearly identically zero because, as expected, the nominal input past provides no predictive information relative to the nominal output pasts. However, once coupling begins and after an initial transient, the nominal input past now does provide predictive information but only for positive values of the nominal output past. In the coupled state, the system exhibits quasiperiodic dynamics, and the nominal input past only helps resolve the nominal output future during a portion of the quasiperiod.

## VI. CONCLUSION

In this paper, we have developed specific transfer entropy, and compared its theoretical properties to both total and local transfer entropies. Specific transfer entropy shares the favorable properties of total and local specific entropies, and does not share their undesirable properties. We have seen that specific transfer entropy provides a directly interpretable measure of the predictive impact of a nominal input on a nominal output as a function of the input-output state. As such, specific transfer entropy provides both time-dependent and state-dependent information about input-output systems.

We have seen that total, local, and specific transfer entropies can be reliably estimated from observations of input-output systems. We found that for short to moderate length time series, plug-in estimators based on kernel density estimators outperformed plug-in estimators based on $k$th nearest neighbors. Moreover, we found that $k$th-nearest-neighbor-based estimators of local transfer entropy with $k$ fixed do not consistently estimate the local transfer entropy, and should not be used if a state- or time-resolved analysis of the system is desired. One topic we have not addressed is the computational complexity of these estimators. The kernel density estimators require $O(T^2)$ operations to evaluate, compared to $O(T \ln T)$ for $k$th-nearest-neighbor estimators. In addition, the optimization

of the bandwidths for the tuned kernel density estimator may require many evaluations of the kernel density estimator. Thus, for long enough time series, to gain computational tractability, the $k$th-nearest-neighbor-based estimator may be preferred despite worse statistical performance. In addition, for an input-output system where the specific transfer entropy varies greatly over the input-output state space, the $k$th-nearest-neighbor estimator may perform better since it adapts locally to the density at any given point.

In this paper, we have focused on transfer entropies for discrete-time systems. Extensions to continuous-time systems are clearly of universal interest. Extensions along these lines have been made for Granger causality for stochastic delay differential equations [49] and transfer entropy for jump and point processes [50]. A related literature on multiscale systems has also been developed [51–53]. We leave extensions of specific transfer entropy to continuous time for future work.

Specific transfer entropy may find applicability anywhere that Granger causality, total transfer entropy, or local transfer entropy have been used. As one example, local transfer entropy was found to perform well as a spatiotemporal filter for discrete-state systems [18,54]. The extension of this approach to continuous-state systems via the specific transfer entropy is straightforward. This approach might then be used, for example, to filter a network of coupled oscillators in order to determine those regions of the network and durations of its dynamic when predictive information is maximized or minimized.

## ACKNOWLEDGMENTS

We thank Chao Wang, Amy Trongnetrpunya, David Keyser, and Chris Cellucci for valuable discussions. We acknowledge support from the Uniformed Services University and the Defense Medical Research and Development Program. The opinions and assertions contained herein are the private opinions of the authors and are not to be construed as official or reflecting the views of the United States Department of Defense.

## APPENDIX: APPROXIMATION OF THE LOCAL AND SPECIFIC TRANSFER ENTROPY FOR THE STARX SYSTEM

To compute the first-order local and specific transfer entropies $\widetilde{\mathrm{t}}_{Y\to X}(y_{t-1},x_{t-1},x_t)$ and $\mathrm{t}_{Y\to X}(y_{t-1},x_{t-1})$, we require the conditional density of $X_t$ given $X_{t-1}$. By the law of total probability, we have that

$$f(x_t|x_{t-1}) = \int_{\mathbb{R}} f(x_t|x_{t-1},y_{t-1}) f(y_{t-1}|x_{t-1})\, dy_{t-1}. \quad \text{(A1)}$$

The first density in the integral is known from (24), thus we only require the conditional density of $Y_{t-1}$ given $X_{t-1}$. By assuming stationarity, we can compute this conditional density from the stationary density of the joint process $\{(X_t,Y_t)\}_{t\in\mathbb{Z}}$. The stationary density of the joint process is given as the solution to the eigenproblem

$$f(x_t,y_t) = \int_{\mathbb{R}^2} K(x_{t-1},y_{t-1},x_t,y_t) f(x_{t-1},y_{t-1})\, dx_{t-1} dy_{t-1}, \quad \text{(A2)}$$

where $K$ is the transition kernel given by

$$K(x_{t-1},y_{t-1},x_t,y_t) = f(x_t,y_t|x_{t-1},y_{t-1}), \quad \text{(A3)}$$

i.e., the transition density from $(x_{t-1},y_{t-1})$ to $(x_t,y_t)$. Note that because of the conditional independence relationships implicit in the STARX model, we have that the transition density factors as $f(x_t,y_t|x_{t-1},y_{t-1}) = f(x_t|x_{t-1},y_{t-1}) f(y_t|y_{t-1})$, where both of these conditional densities are normal by (23)–(24). We thus approximate the solution to (A2) using Nyström's method with Gauss-Legendre quadrature [55,56], and then compute the conditional densities $f(y_{t-1}|x_{t-1})$ and $f(x_t|x_{t-1})$ via Gauss-Legendre quadrature.

[1] C. W. J. Granger, Investigating causal relations by econometric models and cross-spectral methods, Econometrica **37**, 424 (1969).
[2] D. Marinazzo, M. Pellicoro, and S. Stramaglia, Kernel Method for Nonlinear Granger Causality, Phys. Rev. Lett. **100**, 144103 (2008).
[3] L. Faes, G. Nollo, and A. Porta, Information-based detection of nonlinear granger causality in multivariate processes via a nonuniform embedding technique, Phys. Rev. E **83**, 051112 (2011).
[4] L. Barnett, A. B. Barrett, and A. K. Seth, Granger Causality and Transfer Entropy are Equivalent for Gaussian Variables, Phys. Rev. Lett. **103**, 238701 (2009).
[5] T. Schreiber, Measuring Information Transfer, Phys. Rev. Lett. **85**, 461 (2000).
[6] A. Kaiser and T. Schreiber, Information transfer in continuous processes, Physica D: Nonlinear Phenomena **166**, 43 (2002).
[7] R. G. James, N. Barnett, and J. P. Crutchfield, Information Flows? A Critique of Transfer Entropies, Phys. Rev. Lett. **116**, 238701 (2016).
[8] D. Chicharro and A. Ledberg, When two become one: The limits of causality analysis of brain dynamics, PLoS One **7**, e32466 (2012).
[9] S. Li, J. Ernest, and P. Bühlmann, Econometrics and Statistics **2**, 81 (2017).
[10] T. Bossomaier, L. Barnett, and M. Harré, Information and phase transitions in socio-economic systems, Complex Adapt. Syst. Model. **1**, 9 (2013).
[11] L. J. Moniz, J. D. Nichols, and J. M. Nichols, Mapping the information landscape: Discerning peaks and valleys for ecological monitoring, J. Biol. Phys. **33**, 171 (2007).
[12] D. M. Budden and E. J. Crampin, Information theoretic approaches for inference of biological networks from continuous-valued data, BMC Sys. Bio. **10**, 89 (2016).
[13] M. Wibral, B. Rahm, M. Rieder, M. Lindner, R. Vicente, and J. Kaiser, Prog. Biophys. Mol. Biol. **105**, 80 (2011).
[14] C. Gómez, J. T. Lizier, M. Schaum, P. Wollstadt, C. Grützner, P. Uhlhaas, C. M. Freitag, S. Schlitt, S. Bölte, R. Hornero, and M. Wibral, Reduced predictable information in brain signals in autism spectrum disorder, Front. Neuroinform. **8** (2014).

[15] G. Valenza, L. Faes, L. Citi, M. Orini, and R. Barbieri, Instantaneous transfer entropy for the study of cardio-respiratory dynamics, In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE* (IEEE, Piscataway, 2015), pp. 7885–7888.

[16] J. T. Lizier, JIDT: An information-theoretic toolkit for studying the dynamics of complex systems, Front. Robot. AI **1**, 1085 (2014).

[17] M. Lindner, R. Vicente, V. Priesemann, and M. Wibral, TRENTOOL: A Matlab open source toolbox to analyze information flow in time series data with transfer entropy, BMC Neurosci. **12**, 1 (2011).

[18] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, Local information transfer as a spatiotemporal filter for complex systems, Phys. Rev. E **77**, 026110 (2008).

[19] M. P. Wand and M. C. Jones, *Kernel Smoothing* (CRC Press, Boca Raton, 1994).

[20] L. F. Kozachenko and N. N. Leonenko, Sample estimate of the entropy of a random vector, Problemy Peredachi Informatsii **23**, 95 (1987).

[21] M. Ragwitz and H. Kantz, Markov models from data by simple nonlinear time series predictors in delay embedding spaces, Phys. Rev. E **65**, 026209 (2002).

[22] T. Sauer, J. A. Yorke, and M. Casdagli, Embedology, J. Stat. Phys. **65**, 579 (1991).

[23] J. Garland, R. James, and E. Bradley, Model-free quantification of time-series predictability, Phys. Rev. E **90**, 052910 (2014).

[24] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, Transfer entropy - A model-free measure of effective connectivity for the neurosciences, J. Comput. Neurosci. **30**, 45 (2011).

[25] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Testing for nonlinearity in time series: The method of surrogate data, Physica D **58**, 77 (1992).

[26] S. Caires and J. A. Ferreira, On the non-parametric prediction of conditionally stationary sequences, Statistical inference for stochastic processes **8**, 151 (2005).

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, 2012).

[28] J. Fan and Q. Yao, *Nonlinear Time Series: Nonparametric and Parametric Methods* (Springer-Verlag, Berlin, 2003).

[29] J. T. Lizier and M. Prokopenko, Differentiating information transfer and causal effect, Eur. Phys. J. B **73**, 605 (2010).

[30] M. Prokopenko, J. Lizier, and D. Price, On thermodynamic interpretation of transfer entropy, Entropy **15**, 524 (2013).

[31] T. Gneiting, F. Balabdaoui, and A. E. Raftery, Probabilistic forecasts, calibration and sharpness, J. Roy. Stat. Soc. B **69**, 243 (2007).

[32] D. Darmon, Specific differential entropy rate estimation for continuous-valued time series, Entropy **18**, 190 (2016).

[33] J. S. Simonoff, *Smoothing Methods in Statistics* (Springer Science & Business Media, Berlin, 2012).

[34] P. Burman, E. Chow, and D. Nolan, A cross-validatory method for dependent data, Biometrika **81**, 351 (1994).

[35] P. Hall, J. Racine, and Q. Li, Cross-validation and the estimation of conditional probability densities, J. Am. Stat. Assoc. **99**, 1015 (2004).

[36] S. Efromovich, Dimension reduction and adaptation in conditional density estimation, J. Am. Stat. Assoc. **105**, 761 (2010).

[37] C. T. Kelley, *Iterative Methods for Optimization* (SIAM, Philadelphia, 1999).

[38] J. Zhu, J. J. Bellanger, H. Shu, and R. Le Bouquin Jeannès, Contribution to transfer entropy estimation via the k-nearest-neighbors approach, Entropy **16**, 5263 (2015).

[39] M. Wibral, R. Vicente, and M. Lindner, Transfer entropy in neuroscience, in *Directed Information Measures in Neuroscience*, pages 3–36 (Springer, Berlin, 2014).

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

[41] Michael Wibral, Raul Vicente, and Joseph T. Lizier, *Directed Information Measures in Neuroscience* (Springer, 2014), pp. 13–14.

[42] L. Devroye, L. Gyorfi, A. Krzyzak, and G. Lugosi, On the strong universal consistency of nearest neighbor regression function estimates, Ann. Stat. **22**, 1371 (1994).

[43] H. Tong, Threshold models in time series analysis - 30 years on, Stat. Interface **4**, 107 (2011).

[44] H. Tong and K. S. Lim, Threshold autoregression, limit cycles and cyclical data, J. Roy. Stat. Soc. B **42**, 245 (1980).

[45] L. Barnett and A. K. Seth, The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference, J. Neurosci. Methods **223**, 50 (2014).

[46] G. Biau and L. Devroye, *Lectures on the Nearest Neighbor Method* (Springer, Berlin, 2015), pp. 30–31.

[47] I. Bashkirtseva and L. Ryashko, Attainability analysis in the problem of stochastic equilibria synthesis for nonlinear discrete systems, Int. J. Appl. Math. Comput. Sci. **23** (2013).

[48] R. Quian Quiroga, J. Arnhold, and P. Grassberger, Learning driver-response relationships from synchronization patterns, Phys. Rev. E **61**, 5142 (2000).

[49] L. Barnett and A. K. Seth, Detectability of granger causality for subsampled continuous-time neurophysiological processes, J. Neurosci. Methods **275**, 93 (2017).

[50] R. E. Spinney, M. Prokopenko, and J. T. Lizier, Transfer entropy in continuous time, with applications to jump and neural spiking processes, Phys. Rev. E **95**, 032319 (2017).

[51] M. Costa, A. L. Goldberger, and C-K. Peng, Multiscale Entropy Analysis of Complex Physiologic Time Series, Phys. Rev. Lett. **89**, 131601 (2002).

[52] M. Lungarella, A. Pitti, and Y. Kuniyoshi, Information transfer at multiple scales, Phys. Rev. E **76**, 056117 (2007).

[53] M. Paluš, Cross-scale interactions and information transfer, Entropy **16**, 5263 (2014).

[54] M. Wibral, J. T. Lizier, S. Vögler, V. Priesemann, and R. Galuske, Local active information storage as a tool to understand distributed neural information processing, Front. Neuroinform. **8** (2014).

[55] J. Anděl, I. Netuka, and K. Zvára, On threshold autoregressive processes, Kybernetika **20**, 89 (1984).

[56] William H. Press, *Numerical Recipes 3rd Edition: The Art of Scientific Computing* (Cambridge University Press, Cambridge, 2007).