Causality indices for bivariate time series data:
a comparative review of performance

Tom Edinburgh,[1, a)] Stephen J. Eglen,[1] and Ari Ercole[2]

[1)]*Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, UK*

[2)]*Cambridge Centre for Artificial Intelligence in Medicine and Division of Anaesthesia, Department of Medicine, University of Cambridge, Cambridge CB2 0QQ, UK*

Inferring nonlinear and asymmetric causal relationships between multivariate longitudinal data is a challenging task with wide-ranging application areas including clinical medicine, mathematical biology, economics and environmental research. A number of methods for inferring causal relationships within complex dynamic and stochastic systems have been proposed but there is not a unified consistent definition of causality in the context of time series data. We evaluate the performance of ten prominent causality indices for bivariate time series, across four simulated model systems that have different coupling schemes and characteristics. Pairwise correlations between different methods, averaged across all simulations, show there is generally strong agreement between methods, with minimum, median and maximum Pearson correlations between any pair (excluding two similarity indices) of 0.298, 0.719 and 0.955 respectively. In further experiments, we show that these methods are not always be invariant to real-world relevant transformations (data availability, standardisation and scaling, rounding error, missing data and noisy data). We recommend transfer entropy and nonlinear Granger causality as particularly strong approaches for estimating bivariate causal relationships in real-world applications. Both successfully identify causal relationships and a lack thereof across multiple simulations, whilst remaining robust to rounding error, at least 20% missing data and small variance Gaussian noise. Finally, we provide flexible open-access Python code for computation of these methods and for the model simulations.

Quantifying causal relationships between longitudinal observations of a complex system is essential to an understanding of the interactions between sub-components of the system and is subsequently key to building better and more parsimonious models[1,2]. In many real-world applications, we are rarely able to access or describe an underlying graphical network of these interactions *a priori*, and we are typically limited to observing simultaneously recorded variables from each subsystem as a multivariate time series. Two key properties that are widely regarded as crucial in defining causal relationships are: that the effect is temporally preceded by the cause, and that external changes to values of the causal variable propagate to values of the effect variable and do not break the causal structure[3]. Correlation or synchronisation in these multivariate time series does not necessarily imply a causal relationship between variables, and counter-examples are easy to find[4]. Further, a lack of correlation does not imply a lack of causality, and a reliance on correlation-based measures may result in nonlinear causal relationships being obscured, e.g. Ref 5. In recent decades, various mathematical frameworks[1,6,7] have been described to allow identification of nonlinear (and asymmetric) causal structure within complex systems, primarily driven by domain-specific applications, from diverse application areas including as statistical economics[8,9], climate science[10–12] and computational neuroscience[13,14].

I. INTRODUCTION

No general method exists to identify causal structure within complex systems, and there is no single consistent and unifying notion of quantitative causality estimation for time series data. Published methods can be broadly categorised into the following groups:

1. regression-based indices that use 'recent history' vectors as predictors in a model (e.g. Granger causality),

2. information-theoretic indices that build upon ideas of conditional mutual information (e.g. transfer entropy),

3. indices based on state space dynamics, such as local neighbourhoods and trajectories (e.g. convergent cross mapping),

4. graphical models that scale causal inference estimation to high-dimensional multivariate systems for causal identification.

There exist common themes between these methods, and membership within these groups is sometimes somewhat blurred. Figure 1 sets out key properties and similarities between methods from groups 1-3. Previous reviews of the literature[15–18] typically focus on a subset of methods from one of these groups. The suitability, interchangeability and performance of published methods, particularly where they span more than one of these groups, has received relatively little attention. In this work, we identified and assessed a widely used subset of indices for directed bivariate causality inference, concentrating on methods involving univariate embeddings to describe the recent history of the system (Section **II.**). A review of such methods has been published previously by Lungarella *et al.*[19]. We reproduce these results for
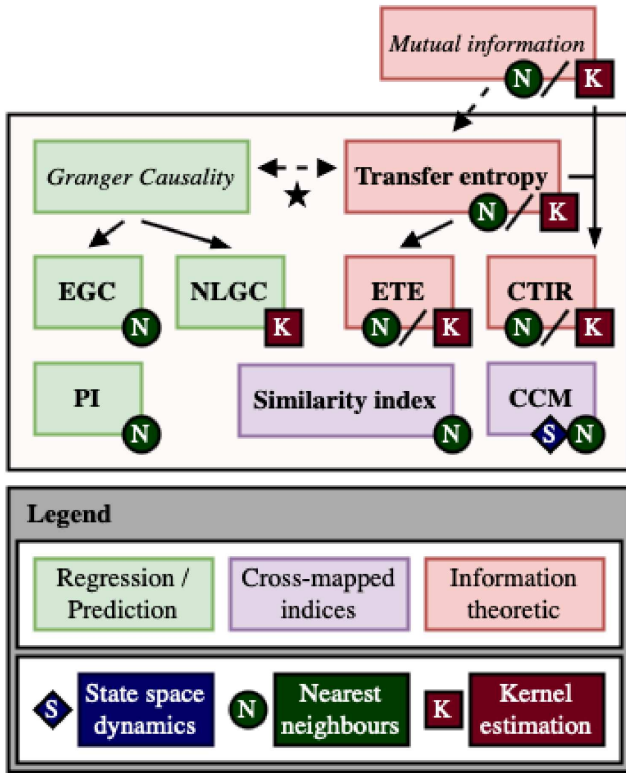
[a)]Electronic mail: te269@cam.ac.uk

FIG. 1. Causality indices described in this paper, which represent a widely-used but non-exhaustive subset of this field of research. The indices are as follows (where GC is Granger causality): extended GC (EGC), nonlinear GC (NLGC), predictability improvement (PI), transfer entropy (TE), effective transfer entropy (ETE), coarse-grained transinformation rate (CTIR), similarity indices (SI), convergent cross mapping (CCM). We classify these indices into three categories, and highlight commonalities between the approaches and their estimation (state space dynamics, nearest neighbour computation, kernel estimation).

the original and newer methods. We also extend this work by proposing a set of modifications that can be made to simulated data prior to causal estimation, in order to investigate sensitivity of each method to data availability, scaling, missing data, rounding and Gaussian noise (Section **III.**). Each of these reproduces phenomena that often occur in real-world data, such as when instruments have a fixed measurement precision and data is reported with rounding error. We believe these tests should provide in-depth benchmarking criteria for new proposed methodologies. Finally, we summarise the strengths and weaknesses of these approaches, and identify key areas of further research (Section **IV.**).

II. METHODS

We observe a complex system as a set of variables within a multivariate time series. The time series $X = (x_1, \ldots, x_T)$ and $Y = (y_1, \ldots, y_T)$ describe a bivariate system with state $\mathbf{s}_t = (x_t, y_t)$ at time t. A critical implicit assumption is that the

series has unit time, or equivalently that the data is observed at a fixed constant frequency. Underpinning all methods is the key assumption that cause directly precedes effect. As a result, a preliminary step is the construction of time-delay embedding vectors $\mathbf{x}_t^{m,\tau}$ in m-dimensional state space $\mathscr{X} \cong \mathbb{R}^m$. An important distinction in the defined methods is whether their theoretical basis is stochastic or deterministic. Both employ the same time-delay vectors, though only in deterministic systems are the vectors $\mathbf{x}_t^{m,\tau}$ considered elements within an m-dimensional state space $\mathscr{X} \cong \mathbb{R}^m$. We construct equivalent embedding vectors $\mathbf{y}_t$ (and state space $\mathscr{Y} \cong \mathbb{R}^m$) for Y, and a joint embedding vectors $\mathbf{z}_t$ (and state space $\mathscr{Z} \cong \mathbb{R}^{2m}$):

$$\mathbf{x}_t = \mathbf{x}_t^{m,\tau} = \left(x_{t-(m-1)\tau}, x_{t-(m-2)\tau}, \ldots, x_{t-\tau}, x_t\right)' \in \mathscr{X}$$

$$\mathbf{z}_t = \begin{pmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{pmatrix} \in \mathscr{Z}, \quad t = (m-1)\tau + 1, \ldots, T$$

In practice, many real-world systems are stochastic, with some level of noise or randomness in at least part of the system. A further assumption for stochastic causality estimation is that of separability, which states that there is unique information about the effect variable that is contained only within the causal variable. The standard approach here is to describe or model the current value x_t of X as conditional upon upon the 'recent history' joint embedding vector $\mathbf{z}_{t-1}$ (full model). Separability means that removing the causal variable Y eliminates the information it contains about the effect X, which we observe either by identifying non-zero coefficients in the full model or constructing a reduced model, conditioned only upon $\mathbf{x}_{t-1}$. These methods are generally described with time index shifted $t \mapsto t+1$, though the interpretation ('current' and 'recent history') remains the same.

Granger causality[1] (GC), a notable and popular method for causality estimation in time series, fits autoregressive models on the time series to this end. Extensions of GC to nonlinear systems include a locally linear version called **extended Granger causality**[20] (EGC) and **nonlinear Granger causality**[21] (NLGC), which performs a 'global' nonlinear autoregression using radial basis functions (RBFs). **Predictability improvement**[22] (PI) is another locally constant linear regression of 'recent history' embeddings, which measures a reduction in mean squared error when $\mathbf{z}_t$ is used for predicting a 'horizon' value x_{t+h} instead of $\mathbf{x}_t$ alone.

Information theory is a natural framework for describing causal relationships. **Transfer entropy**[7] (TE) measures deviation from the generalised Markov property $p(x_{t+1}|\mathbf{x_t}) = p(x_{t+1}|\mathbf{x_t}, \mathbf{y_t})$ as a conditional mutual information. With weak coupling and limited data, transfer entropy can suffer from finite sample effects and **effective transfer entropy**[23] (ETE) corrects for this using shuffled versions of the causal variable. TE reduces to vanilla GC under the assumption of Gaussian variables[24] ($\star$, Figure 1), and non-zero GC implies violation of the generalised Markov property and non-zero TE[25]. **Coarse-grained transinformation rate**[26] (CTIR) is based upon 'coarse-grained entropy rates', and measures the rate of net information flow, averaged over different lags τ. Often the difficulty in information theoretic methods (described in depth in Ref 15) is the robust estimation of joint probabilities or entropy values, which in turn form building blocks

for these methods. We use a histogram binning partition (H) and the (hypercube) Kraskov-Stögbauer-Grassberger (KSG) estimate[27], which is a technique involving k-nearest neighbour statistics. All information theoretic computation here is in 'nats' (logarithm base e).

Fully deterministic dynamical systems, which evolve according to a differential equation or difference equation, do not necessarily satisfy the separability condition. In these systems, x_t can often be reformulated as a function of only past values of X, which makes the potential causal role of Y in the coupled system less clear, as highlighted by Granger[1]. Causal relationships in a coupled deterministic system are instead observed via the event that each variable belongs to a shared attractor manifold $A \subset \mathscr{Z}$. A consequence of Takens' embedding theorem[28] is that the 'library of historical behaviour' of X preserves the topology of A and, by transitivity, local neighbourhoods in $\mathscr{X}$ those in $\mathscr{Y}$ and vice versa[5]. It is possible to detect unidirectional causal influence, where only the dynamics of a causal variable propagate to the response variable in this way. Sugihara et $al.$[5] argue that the inferred direction of unidirectional causal influence is counter-intuitively reversed (i.e. cross mapping from $\mathscr{X}$ to $\mathscr{Y}$ reveals causal influence from Y to X).

The key assumption of cross mapped indices is that causal relationships are observed in the similarity between sets of (subscript) indices denoting the nearest neighbours for each set of embedding vectors, which can be mapped from one variable to the other to reveal interdependency. This is the idea behind the **similarity indices**: two similarity indices we test here, denoted $\text{SI}_{Y \to X}^{(1)}$ and $\text{SI}_{Y \to X}^{(2)}$, are $H(\mathscr{X}|\mathscr{Y})$ in Ref 29 and $E(\mathscr{X}|\mathscr{Y})$ in Ref 30 respectively. **Convergent cross mapping**[5] (CCM) computes the correlation ρ between the cross mapped estimate and the true value, with convergence in ρ as T increases "a key property that distinguishes causation from simple correlation"[5].

III. RESULTS

Our results are split into two parts. First, we reproduce the results from Ref 19, evaluating the performance of all methods including the additional CTIR and CCM, plus ETE using histogram binning and TE using KSG. In these simulations, we choose the same simulation model parameters and causality index parameters as in Ref 19 (Table I). In the second part, we investigate sensitivity to common issues relevant to real-world data, using the Ulam lattice system to illustrate these.

A. Numerical simulations

We investigate performance on four simulated model systems (Table II). In each simulation, we assess the causality estimates of each method by varying the coupling strength λ. These simulated systems are widely studied in chaos theory, e.g. Ref 38, and also appear elsewhere in the literature, e.g. Ulam lattice in Ref 7.

Linear process:

$$x_{t+1} = b_x x_t + \lambda y_t + \varepsilon_{x,t}, \quad y_{t+1} = b_y y_t + \varepsilon_{y,t} \quad (1)$$
$$\varepsilon_{x,t} \sim N(0, \sigma_x^2), \quad \varepsilon_{y,t} \sim N(0, \sigma_y^2)$$

Ulam lattice:

$$s_{t+1,l+1} = f(\lambda s_{t,l} + (1-\lambda)s_{t,l+1}), \quad l = 1, \ldots, N_L - 1 \quad (2)$$
$$s_{t+1,1} = f(\lambda s_{t,N_L} + (1-\lambda)s_{t,1}), \quad f(s) = 2 - s^2$$
$$x_t = s_{t,1}, \quad y_t = s_{t,2}$$

Hénon unidirectional map:

$$x_{t+2} = a - x_{t+1}^2 + b_x x_t \quad (3)$$
$$y_{t+2} = a - (\lambda x_{t+1} + (1-\lambda)y_{t+1})y_{t+1} + b_y y_t$$

Hénon bidirectional map:

$$x_{t+2} = a - x_{t+1}^2 + \lambda_{yx}(x_{t+1}^2 - y_{t+1}^2) + b_x x_t \quad (4)$$
$$y_{t+2} = a - y_{t+1}^2 + \lambda_{xy}(y_{t+1}^2 - x_{t+1}^2) + b_y y_t$$

We reproduce figures for all simulations and methods in Figures 3-5, and summarise our results in Figure 2, which shows correlations between each pair of indices. For linear process and Ulam lattice simulations, we report causality estimates in both directions, i.e. $i_{X \to Y}$ and $i_{Y \to X}$ (where i denotes any of the causality indices). For Hénon maps, we instead use the directed index $D_{X \to Y} = i_{X \to Y} - i_{Y \to X}$, following Ref 19. In general, all measures exhibit a small standard deviation relative to the absolute value of the index, indicating that random initial conditions during data simulation has at most a very small influence on the causal structure, when initial transients are discarded. Though we are able to replicate the findings in Ref 19 in most cases, we occasionally find minor differences between their results and ours. In particular, we sometimes find results of a similar profile but different magnitude, as λ varies. We observe this for: EGC and linear processes; PI and all simulations, $\text{SI}^{(1)}$ and Ulam lattice; TE and Hénon unidirectional maps. Though we handle numerical outliers differently in our visualisation of results for Hénon bidirectional maps, our results for these simulations appear largely comparable in magnitude and profile. There is no mention of a data standardisation step in Ref 19 and the results we report do not involve any pre-processing, though this did not appear to rectify these differences. In one notable difference between identical Hénon bidirectional map results, Lungeralla et $al.$[19] find a region in λ-space (namely $\{(\lambda_{xy}, \lambda_{yx}): \lambda_{xy} > 0.1, \lambda_{yx} > 0.1, \lambda_{xy} + \lambda_{yx} < 0.35\}$) in which they identify general synchronisation between X and Y and have difficulty estimating indices due to numerical instabilities, yet we do not observe this. We have followed the implementation in Ref 19 as closely as possible and it is unclear why these differences exist.

We knowingly deviated from the implementations in Ref 19 only in the case of NLGC, in which we preferred to use k-means rather than fuzzy c-means clustering to determine RBF centers, after finding similar or improved results but with a much reduced computational cost. Lungarella et $al.$[19] note that NLGC is numerically unstable for 'small' T and computationally expensive for 'large' T, which we suggest may be

TABLE I. Causality indices in this review and their parameters. The indices are as follows (where GC is Granger causality): extended GC (EGC), nonlinear GC (NLGC), predictability improvement (PI), transfer entropy (TE), effective transfer entropy (ETE), coarse-grained transinformation rate (CTIR), similarity indices (SI) and convergent cross mapping (CCM). Table S.I (**Supplementary materials**) provides more detail on the parameter choices for individual simulation results.

	Method	Parameters / other choices		Notes / suggestions	Values used here
Embedding	All	T	Time series length	Depends on data availability	10^p, $p = 3,4,5$
		h	Time horizon value	Normally $h = 1$, generalised to $h \geq 1$ (in PI)	$h = 1$
	All \ CTIR	m	Embedding dimension	'Optimal'[31] vs 'empirical' ($m = 1,...,5$)	$m = 1$ or 2
		τ	Time-delay lag	'Optimal'[32] vs 'empirical' ($\tau = 1,2,3$)	$\tau = 1$
Regression error	**EGC**[20]	Nearest neighbour metric		ℓ_p, may depend on state space / distribution	ℓ_1
		L	No. of neighbourhoods	$L = 100$ in Refs. 19 and 20 (depends on T)	$L = 20$ or 100
		δ	Neighbourhood size	Compute EGC for $\delta \downarrow 0$ (Ref 20)	Various (Table S.I)
	NLGC[21]	Radial basis function (RBF)		Gaussian RBFs in Refs. 19 and 21	Gaussian
		P	No. of RBFs	e.g. gap statistics[33]	Various (Table S.I)
		$\mathbf{x}_\rho$	Gaussian RBF centers	via k-means or fuzzy c-means clustering	via k-means
		σ^2	Gaussian RBF variance	*A priori* fixed e.g. $\sigma^2 = 0.05$ in Refs 19 and 21	$\sigma^2 = 0.05$
	PI[22]	Nearest neighbour (NN) metric		ℓ_p, may depend on state space / distribution	ℓ_2
		R	No. of NNs	*A priori* unclear, e.g. $R = 1, 10$ in Refs. 19 and 22	$R = 1$ or 10
		h	Time horizon value	As above, e.g. $h = 1$ in Refs. 19 and 22	$h = 1$
Information theory	Estimation	Estimation method		e.g. KSG, histogram partition	Both (H / KSG)
		Nearest neighbour metric (KSG)		ℓ_∞ (for hypercube dimensions)[27]	ℓ_∞
		k	No. of NNs (KSG)	Small values e.g. $k = 2,3,4$[27]	$k = 4$
		N	No. of bins (histogram)	e.g. via minimum description length[34,35]	$N = 8$
	TE[7]	n/a		No parameters besides estimation (above)	n/a
	ETE[23]	N_{shuffle}	No. of shuffled X or Y	*A priori* unclear, single shuffle in Ref 23	$N_{\text{shuffle}} = 10$
	CTIR[26]	$\tau_{\max}$	Max time-delay lag	$\tau_{\max} : I(x_t, x_{t+\tau'}) \approx 0, \forall \tau' \geq \tau_{\max}$[26]	$\tau_{\max} = 5$ or 20
		τ_I, ε_I	For estimation of $\tau_{\max}$	$\tau_{\max} = \min_{\tau'}\{\tau' \leq \tau_I : I(x_t, x_{t+\tau'}) < \varepsilon_I\}$	Unused, fixed $\tau_{\max}$
Cross mapped	**SI**[29,30]	Nearest neighbour (NN) metric		ℓ_p, may depend on state space / distribution	ℓ_2
		R	No. of NNs	*A priori* unclear, e.g. $R = 10$ in Refs. 29 and 30	Various (Table S.I)
	CCM[5]	Nearest neighbour metric		ℓ_p, may depend on state space / distribution	ℓ_2
		$T_{\max}$	Max. segment length	Convergence: compute ρ for $T \uparrow T_{\max}$[5]	$T_{\max} = T$
		n_T	No. segments of size T	ρ values averaged across n_T segments, size T	$n_T = 40$
		ρ_∞	Converged CCM value	$\rho_{T_{\max}}$ in Ref 36 or fit exponential regression[37]	$\rho_{T_{\max}}$ (if $\downarrow$ holds)
		δ_ρ	Convergence tolerance	'Converged' if $\rho_\infty - \rho_{m+2} > \delta_\rho$	$\delta_\rho = 0.05$

TABLE II. Brief summary of the characteristics of each numerical simulation model system and parameters (equations 1-4). The difference between identical and non-identical Hénon bidirectional maps is the value of b_y ($b_y = b_x$ for identical maps and $b_y < b_x$ for non-identical maps). In each simulation, the first 10^5 iterations were discarded as transients (10^4 for linear process). Each simulation is initialised randomly but seeded for reproducibility. The coupling parameters were incremented by 0.01 in all cases, for each of 10 independent runs and all indices. We use the following abbreviations in this table: I - identical maps; NI - non-identical maps; L & S - linear and stochastic; NL & D & C - non-linear, deterministic and chaotic.

Simulation	Coupling	Dynamics	$T = 10^p$	Simulation model parameters	Coupling strength
Linear process	$X \leftarrow Y$	L & S	$p = 4$	$b_x = 0.8$, $b_y = 0.4$, $\sigma_x^2 = \sigma_y^2 = 0.2$	$\lambda \in [0,1]$
Ulam lattice	$X \rightarrow Y$	NL & D & C	$p = 3,5$	$N_L = 100$ (size of lattice)	$\lambda \in [0,1]$
Hénon uni-d	$X \rightarrow Y$	NL & D & C	$p = 3,4,5$	$a = 1.4$, $b_x = 0.3$, $b_y = 0.3$	$\lambda \in [0,1]$
Hénon bi-d (I, NI)	$X \leftrightarrow Y$	NL & D & C	$p = 4$	$a = 1.4$, $b_x = 0.3$, $b_y = 0.3$ or 0.1	$\lambda_{xy}, \lambda_{yx} \in [0,0.4]$

partly due to their use of fuzzy c-means. Little detail is provided about their implementation of this but it may perhaps be that an early stopping criteria sometimes forces a 'poor quality' clustering. Further, we found that the performance of NLGC in Hénon bidirectional map simulations improved significantly with a different set of NLGC parameters (e.g. $P = 50$ instead of $P = 10$), though we do not present these alternate results.

For the **linear process** (LP), the simplest simulation model, all indices show very strong positive correlation in the $Y \rightarrow X$ direction (Figure 3). In the reverse $X \rightarrow Y$ direction, TE and CTIR both decrease with increasing λ, the cross mapped indices all show a marked increase and the remaining indices are approximately zero for all λ. This gives rise to patterns of positive and negative correlation between pairs of methods. As each x_t or y_t is a sum of Gaussian variables, we can derive theoretical values for Shannon entropy and, consequently, for the information theory methods (see **Supplementary Materials**). Figure 3 shows that TE (KSG) reliably estimates the 'true' transfer entropy but TE (H) significantly
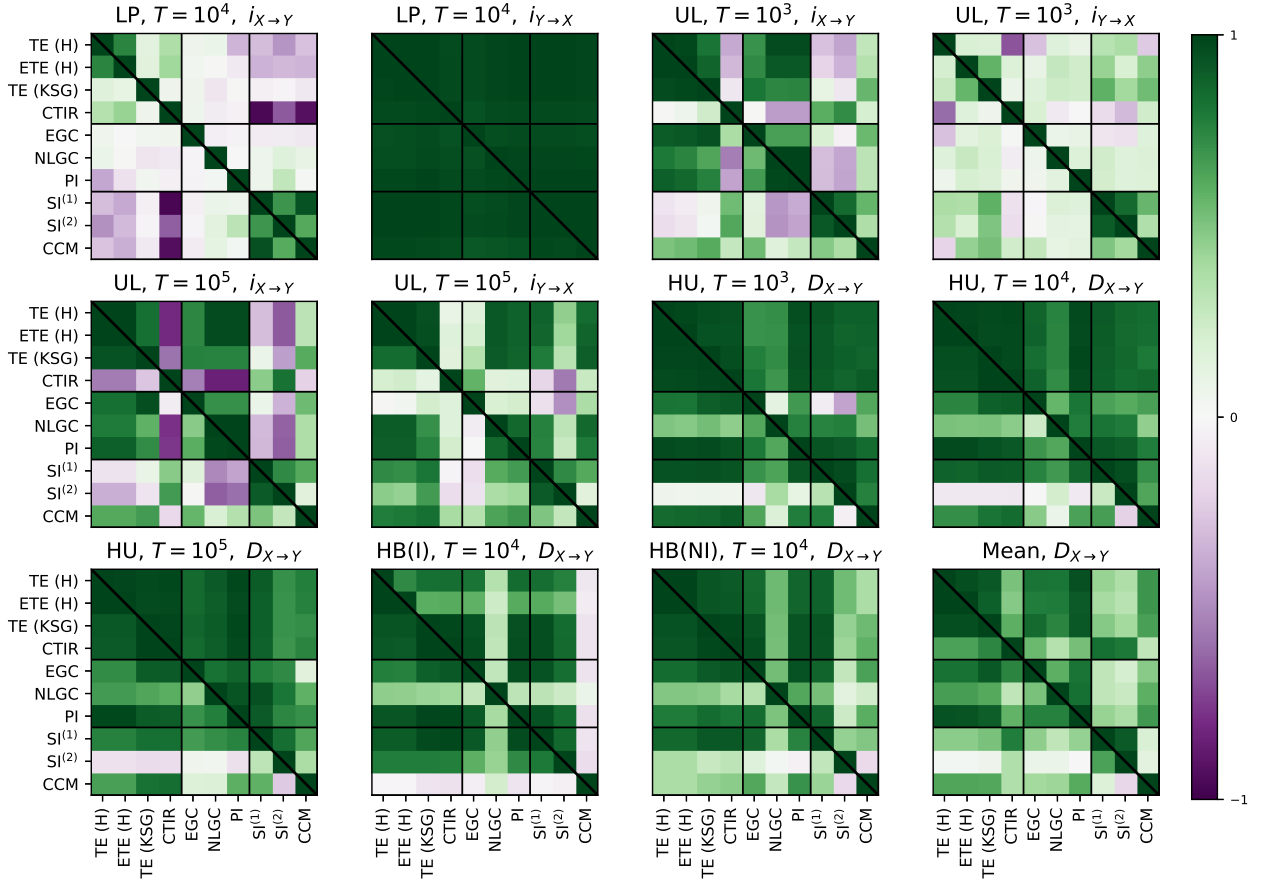
FIG. 2. Correlations between each of the causality indices, for all simulations: linear process (LP), Ulam lattice (UL) and Hénon unidirectional maps (HU), identical/non-identical Hénon bidirectional maps maps (HB (I)/HB (NI)). For several of the simulated systems, simulations were repeated for increasing data size T. In each subplot, the lower left half below the diagonal shows the Pearson correlation between each pair of indices (across all runs and values of λ) and the upper right half shows the rank-based Spearman correlation. Both $i_{X \to Y}$ and $i_{Y \to X}$ are shown for LP and UL simulations but only $D_{X \to Y} = i_{X \to Y} - i_{Y \to X}$ was computed for HU and HB. In the final bottom right subplot, we average correlations in $D_{X \to Y}$ for each simulation, weighting each simulation equally.

underestimates the theoretical values. This is a fundamental flaw that undermines any other advantageous properties of TE (H). Though computed CTIR values match the theoretical values, it is clearly negative in $X \to Y$ and as such does not reflect the causal structure of the system. Increasing the size of the data T alters the value of TE (H) here, but TE (KSG) remains accurate as T increases. However, in all other simulations, TE (H) is more robust to increasing data size.

The **Ulam lattice** (UL) chains together unidirectional coupled chaotic Ulam maps. For large N_L, the causal influence from Y to X is negligible. UL exhibits synchronisation for $\lambda \approx 0.18$, 0.82, where cause and effect variables are indistinguishable from each other, e.g. the system converges to a two state attractor. As a result, most indices either have values approximately equal to zero or suffer from high variance numerical instabilities. Outside of these regions of synchronisation, the information theoretic methods and regression based indices show reasonable consistency (Figures 2 and 4). The exception is CTIR, which slowly decreases as λ increases for $T = 10^5$, albeit still correctly identifying the direction of in-

formation flow. ETE (H) successfully corrects for the small sample effects that give rise to these spurious positive TE (H) results in the $Y \to X$ direction when $T = 10^3$. Both similarity indices fail to identify any causal structure in the UL. For CCM, whilst the net directed index $D_{X \to Y}$ increases with λ, it is negative for $\lambda < 0.5$, and so misidentifies the direction of causality. The positive correlations between methods in $i_{Y \to X}$ for $T = 10^5$ occur due to a very slight peak in value at $\lambda \approx 0.5$ for nearly all methods (except CTIR and EGC).

Synchronisation occurs in the range $\lambda \in [0.7, 1]$ for **Hénon unidirectional** maps (Figure 4). All indices are consistent and perform reasonably well and we do not observe the noisy fluctuations seen in Ref 19, apart from for EGC when $\lambda > 0.7$. It is notable that about half the indices (TE (H), ETE (H), EGC, NLGC, PI) are fairly consistent even as the order of magnitude of the data size T is increased, whilst the values of the other indices (TE (KSG), CTIR, SI, CCM) increase, in some cases tripling in value. There is a strong degree of similarity between all methods for the **Hénon bidirectional** maps (Figure 5). In HB (I) simulations, the exceptions to this are
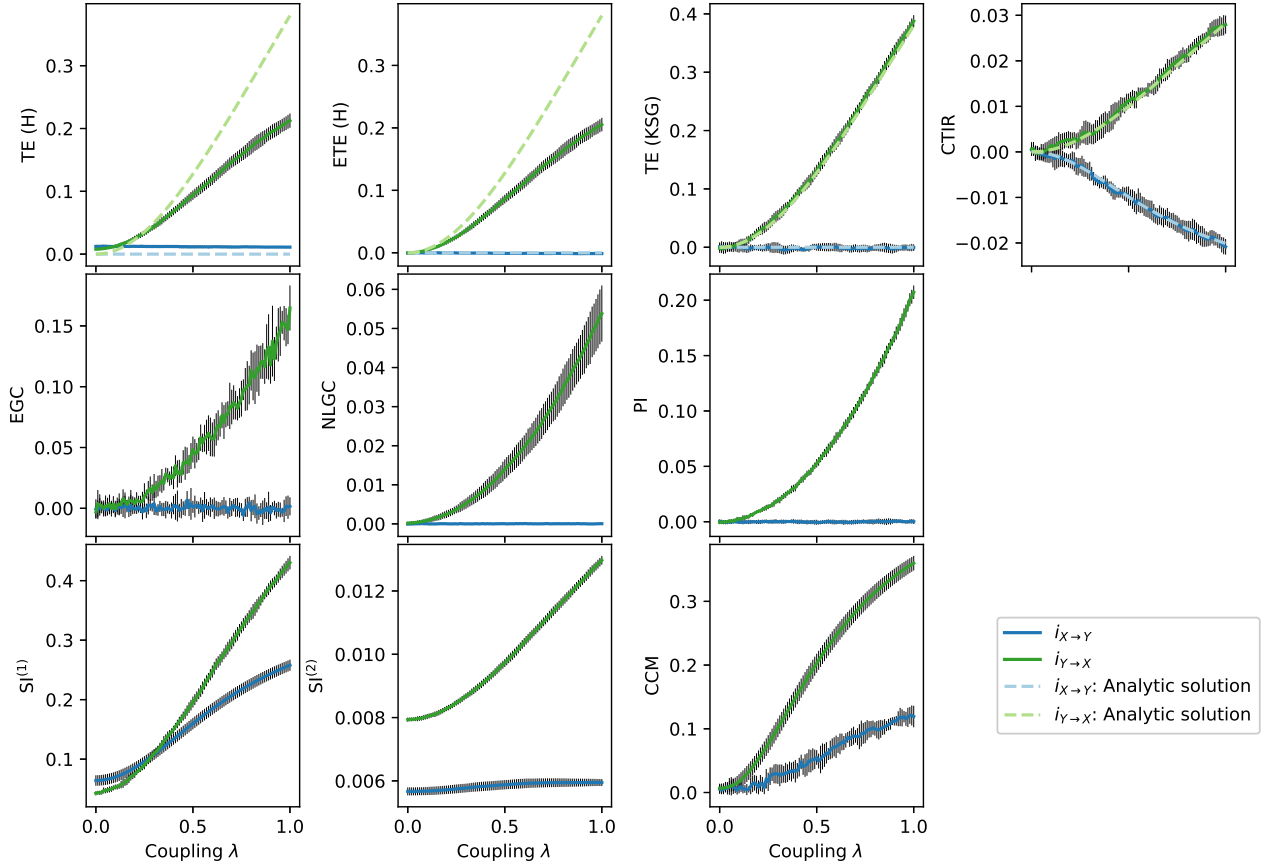
FIG. 3. Linear Gaussian processes with $T = 10^4$ data points and unidirectional $(Y \rightarrow X)$ coupling. Error bars report ± 1 empirical standard deviation from mean values, after 10 independent simulations from the LP system. Simulation parameters are given in Table II and parameters for each causality index are given in Table S.I. We have derived analytic solutions for all the information theoretic indices: TE (H), ETE (H), TE (KSG) and CTIR, which are shown as dashed lines. Note that for TE (KSG) and CTIR, the computational results match the analytic solutions almost exactly, and the dashed lines overlay the solid.

NLGC and CCM, though we found that setting a larger number of RBFs resulted in better performance for NLGC. For CCM, the direction of causality is sometimes incorrect and the reason for this is unclear, though this may also be a result of poor parameter choices. We observe expected symmetry in the values of λ, and synchronisation in the region approximately equal to $\{(\lambda_{xy}, \lambda_{yx}): \lambda_{xy} + \lambda_{yx} > 0.28\}$. There are a small number of points in which numerical instabilities are present in all indices, but the consistency across all indices suggests that these are isolated points in which the system converges to some limit cycle or attractor. There are more differences between methods in HB (NI) results. Significant numerical instabilities occur in EGC, particularly when the system is in a state of synchrony: the region approximately equal to $\{(\lambda_{xy}, \lambda_{yx}): 0.05 < \lambda_{xy} < 0.15, 0.1 < \lambda_{yx} < 0.28\}$ Outside of this region, EGC is broadly similar to the information theoretic indices, which are highly correlated, and to a slightly lesser degree with SI$^{(1)}$ and the PI. In contrast, NLGC, SI$^{(2)}$ and CCM have unusual results. The first of these is negative almost everywhere (even in a repeat analysis with more RBF kernels) and the latter two are mostly non-negative, and

moreover the regions with the most extreme values occur in quite different places in all three.

Computational burden

An important consideration in selecting a suitable method is any trade-off between performance and computational efficiency. The most significant factor in this is often how the algorithmic cost of each method scales with increasing data size T. Table S.II shows the mean and standard deviations of the time taken to compute each index and simulation. TE (H)/ETE (H) is the fastest in almost all cases, even though this calculation also includes 10 reshuffles and recomputations for ETE (H). CCM, EGC and TE (KSG) are similarly among methods with smaller computational cost. Several methods have extreme values in UL simulations with $T = 10^5$, particularly CTIR and PI, but this is distorted by difficulties in computation when the system is in a synchronised state. We observed a marked difference in computational cost for NLGC when using k-means for clustering instead of fuzzy c-means
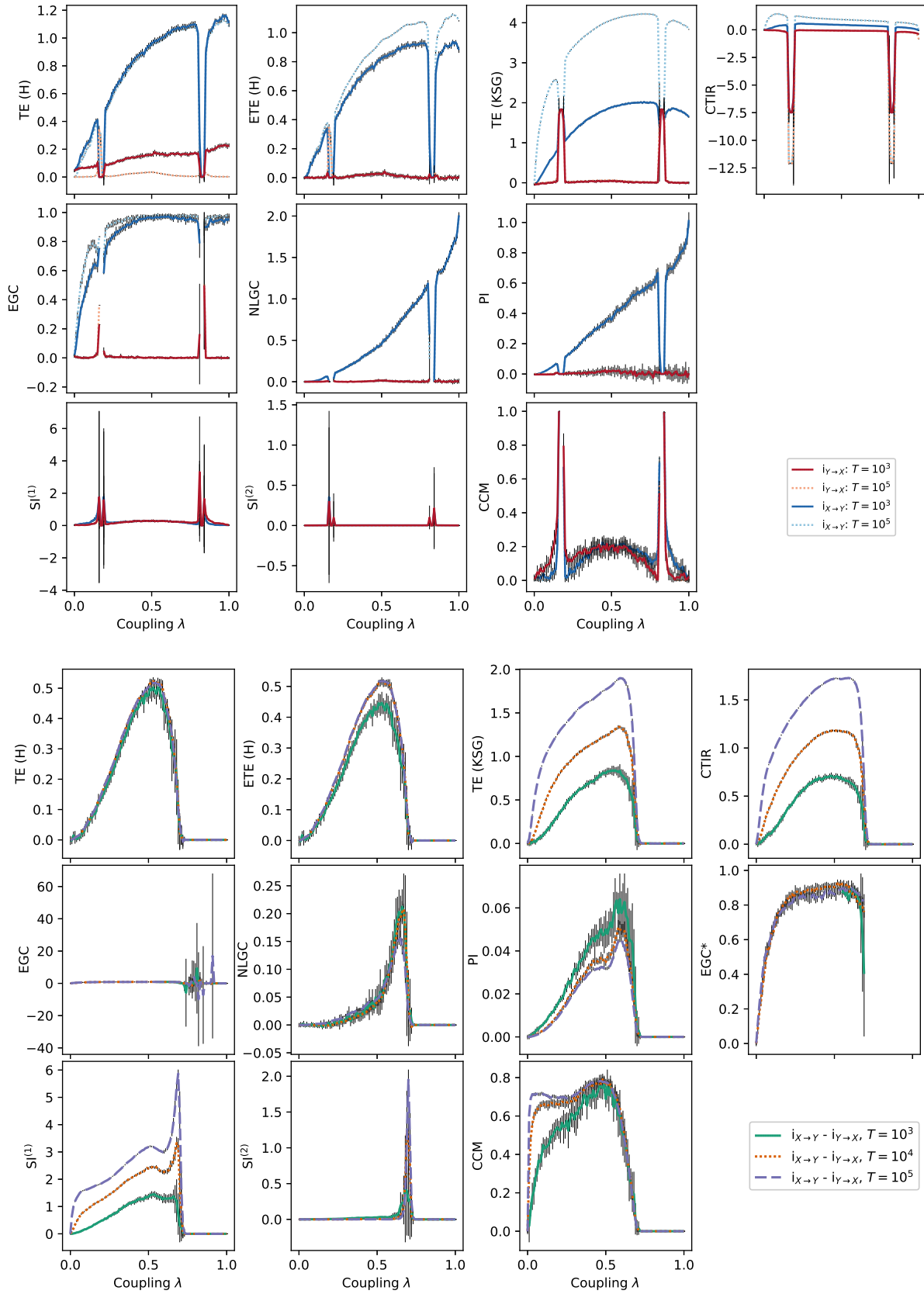
FIG. 4. Ulam lattice (top half) and Hénon unidirectional map (bottom half) simulation results, both with $(X \rightarrow Y)$ coupling, with varying T. For the latter, we show the net directed index $D_{X \rightarrow Y} = i_{X \rightarrow Y} - i_{Y \rightarrow X}$. Error bars report ± 1 standard deviation from mean values, after 10 independent simulations of each map. Simulation parameters are given in Table II and parameters for each causality index are given in Table S.I. Due to extreme results in the EGC index when the system exhibits synchrony ($\lambda > 0.7$ for HU), we set these values set to NA and repeat the plot (EGC*).
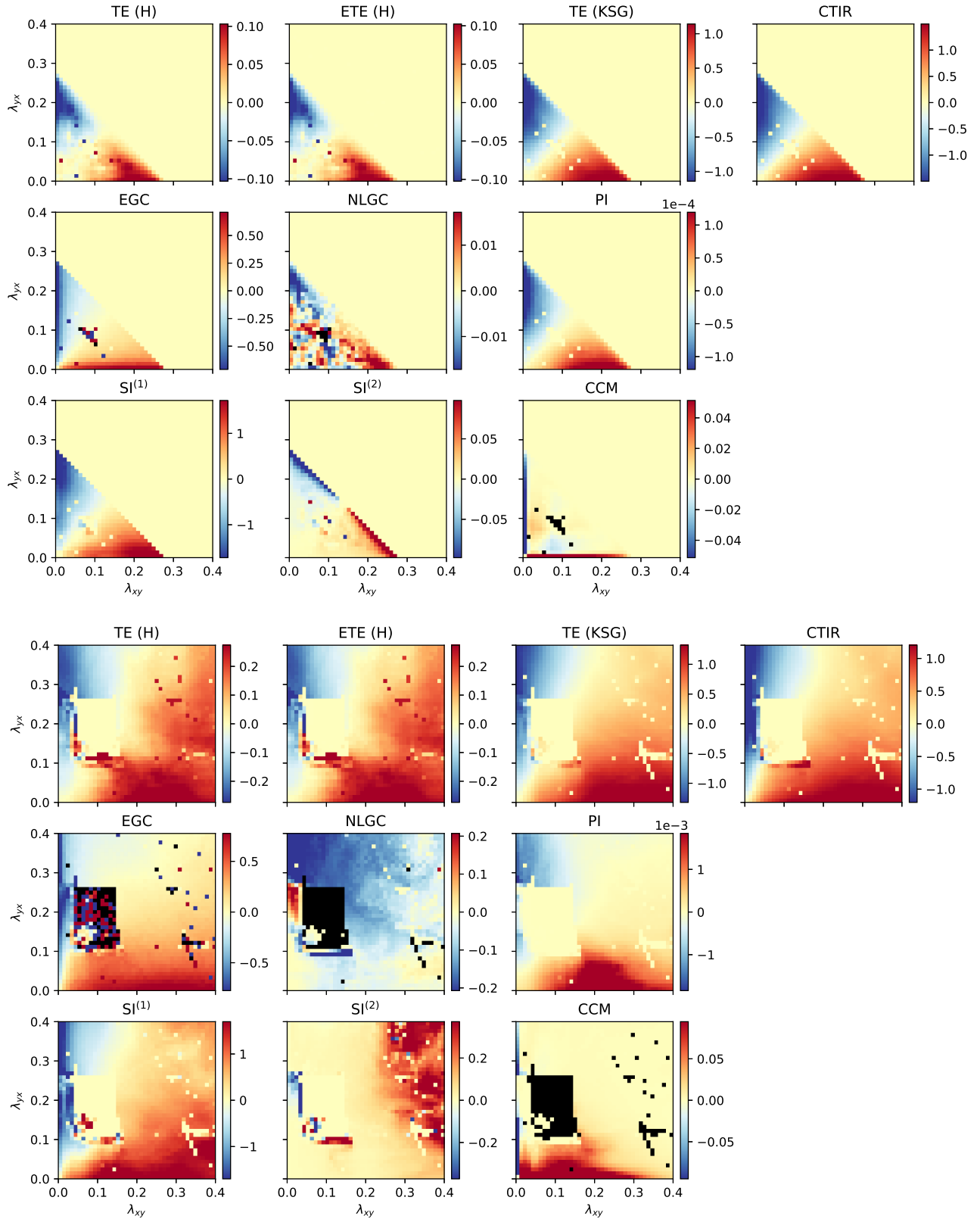
FIG. 5. Hénon map simulation results, with $T = 10^4$ data points and bidirectional coupling, for both identical maps (top half) and non-identical maps (bottom half). We show the net directed index $D_{X \to Y} = i_{X \to Y} - i_{Y \to X}$, averaged across 10 independent simulations. Colorbar limits are capped by percentiles, so as to minimise the effect of extreme values in data visualisation. We set these limits as $\pm \max(|p_1|, |p_{99}|)$ for identical maps and $\pm \max(|p_5|, |p_{95}|)$ for non-identical maps, where p_i is the ith percentile of results for that index. Simulation parameters are given in Table II and parameters for each causality index are given in Table S.I.

and we suggest that *k*-means is more suitable here. Our computation was done in a high performance CPU computing cluster using SkyLake 6140 with 18 core 2.3GHz processors and 384GB of RAM. Although the computational times we report may be slightly faster than on a laptop computer with less processing power, we did not observe any substantial difference when internally comparing run times.

Real-world relevant data issues

Next, we investigated the sensitivity of all causality indices to a number of modifications mirroring issues that often arise in real-world data. We choose UL with $T = 10^3$ to illustrate the effects of these transformations, as LP is too simplistic a model to give sufficient insight. We keep all simulation parameters and causality index parameters the same. In Table III, we summarise the means $\hat{\mu}$ and standard deviations $\hat{\sigma}$ of the directed indices $D_{X \to Y}$ for each λ, which are normalised by their deviation from the 'base' UL results and averaged over all λ. This normalisation allows us to compare across methods that take values in different ranges.

1.   Data availability

Many of the indices, with the exception of TE (KSG) and CTIR, remain consistent with increasing data size T, whilst at the same time exhibit decreasing variance. These results reinforce the similar observations in HU maps. The large increases in the value of the two methods mentioned is concerning and represent a drawback of both methods that should be acknowledged in applications of these approaches. It is unclear whether there is convergence to some 'correct' value as the amount of data increases or whether both are unbounded as $T \to \infty$, but initial computations do not support the former (not shown). Though the $D_{X \to Y}$ values from both transfer entropy methods are highly correlated, they are both estimates of the same quantity and it is difficult to reconcile their different magnitudes, particularly as we have already seen significant underestimation in TE (H) for LP simulations.

2.   Standardisation and scaling

In the second set of experiments, we perform three tests: standardising both series by their sample mean and standard deviation in the first, and separately scaling each unstandardised time series by a factor of 10 (Figure S2). For the Ulam lattice system, sample means for both X and Y are typically between 0.4 and 0.7 and standard deviations are both approximately equal to 1.2 (except when the system is in synchrony). Several methods are invariant under linear scaling or shifting of the original series X and Y, including cross mapping approaches. Information theoretic measures are also invariant in theory, but the KSG algorithm, based on *k*-nearest neighbours, does not retain this property. Similarly, EGC relies on a neighbourhood size parameter, and scaling the data without

changing this parameter accordingly can result in insufficient points available for the locally linear regressions, as is observed when either X or Y is scaled by 10. The directed index for both NLGC and PI has vastly inflated magnitude when Y is scaled by 10. With this in mind, we recommend standardisation or normalisation of the data before employing these methods.

3.   Rounding error and missing data

We perform three experiments to investigate rounding error, first rounding each time series separately to 1 decimal place and then rounding both to 2 decimal places (Figure S3). TE and both GC extensions have similar performance to the baseline in all cases, whilst CCM suffers the most. In two experiments with missing data of 10% and 20%, all methods appear robust to this.

4.   Noisy data

In the case of the earlier LP simulations, Gaussian noise forms an integral component of the system itself and the theoretical expression for TE shows that this depends only on the ratio of the variances σ_x / σ_y (see **Supplementary Material**). However, this noise is inherent in the simulation process (i.e. it does not arise in observation of the system). In our UL experiments, we added additional Gaussian noise *after* simulation. The inclusion of this 'observation' noise does not alter the state of the system or the information flow between variables but it does obscure the causal structure. In the first of these experiments (Figure S3), in which we added small variance Gaussian noise ($\sigma_G = 0.1$), the amplitude of this noise is an order of magnitude less than the original UL values and the inclusion of this noise has a small effect for all indices. In the latter experiments, we added Gaussian noise (with $\sigma_G = 1$) to each variable individually and the effect is more pronounced. NLGC performs best in general and appears very resilient to noise added to Y (effect variable), though it drops slightly in value when Gaussian noise is added to X (cause variable). It is interesting that the two SI have quite different results, with $SI^{(1)}$ more robust to noise, though both methods are not in general able to successfully identify the direction of causality.

DISCUSSION

In-depth comparative studies of this kind are relatively rare in the mathematical literature (examples include Refs. 17 and 39), particularly in evaluating performance of methods for estimating a concept, such as causality, that does not have a consistent, fundamental mathematical definition. Even without this, causal inference has a huge importance in how we can model, predict and exploit real-world applications from many scientific disciplines. Asymmetric bivariate causal inference is the first key step to providing this insight into interactions between components in complex networks. In the con-

TABLE III. Summary of all results from experiments into the effects of data size, scaling, rounding, missing data and Gaussian noise. Taking the Ulam lattice ($T = 10^3$) as a baseline, we recompute the causality indices across 10 independent experimental runs for each $\lambda \in [0,1]$ as before. For each λ, we compute the mean, $\hat{\mu}$, and standard deviation, $\hat{\sigma}$, of directed indices $D_{X \to Y}$. We subsequently compute the deviation from the baseline μ and σ, reporting the average of these over the λ values (excluding the few λ values where the system exhibits general synchronisation). We normalise deviations between μ and $\hat{\mu}$ by the absolute value of μ, with $f(\mu,\hat{\mu}) = \langle \mu - \hat{\mu} \rangle / \langle |\mu| \rangle$ and $g(\sigma,\hat{\sigma}) = \langle \hat{\sigma} \rangle / \langle \sigma \rangle$, where $\langle \cdot \rangle$ is the mean over λ. If the modified simulation returns the same values as the baseline, then $f = 0$ and $g = 1$. All entries except in the column for the baseline $T = 10^3$ report these f and g values.

Method	Baseline $T = 10^3$		Data size $T = 10^5$	Scaling Stand. X,Y	$D_{10X \to Y}$	$D_{X \to 10Y}$	Rounding 1 d.p. X	1 d.p. Y	2 d.p. X,Y	Missing data 10% NA X,Y	20% NA X,Y	Gaussian noise $\sigma_G^2 = 0.1$ X,Y	$\sigma_G^2 = 1$ X	$\sigma_G^2 = 1$ Y
EGC	$\langle \mu \rangle = 0.840$	$f(\mu,\hat{\mu})$	-0.064	0.036	0.207	-0.071	0.031	0.112	0.004	-0.027	-0.040	0.533	0.981	0.950
	$\langle \sigma \rangle = 0.021$	$g(\sigma,\hat{\sigma})$	0.660	1.004	1.425	0.691	0.959	0.946	0.970	**1.023**	**1.033**	1.025	0.473	0.598
NLGC	$\langle \mu \rangle = 0.610$	$f(\mu,\hat{\mu})$	0.013	0.299	2.905	-101.849	**0.000**	**-0.003**	0.001	-0.008	-0.020	0.031	0.740	**-0.007**
	$\langle \sigma \rangle = 0.023$	$g(\sigma,\hat{\sigma})$	0.089	0.608	64.469	126.734	**1.000**	0.954	0.994	1.353	1.906	1.023	1.345	2.325
PI	$\langle \mu \rangle = 0.380$	$f(\mu,\hat{\mu})$	**0.002**	0.293	1.576	-98.727	0.950	-0.951	-0.016	**0.001**	**0.005**	0.011	0.617	0.019
	$\langle \sigma \rangle = 0.032$	$g(\sigma,\hat{\sigma})$	0.094	0.681	69.340	66.364	0.918	1.051	**1.001**	1.214	1.511	**1.007**	2.041	2.120
TE (H)	$\langle \mu \rangle = 0.675$	$f(\mu,\hat{\mu})$	-0.158	**0.000**	**0.000**	**0.000**	0.011	0.030	**0.000**	0.071	0.159	0.026	0.786	0.731
	$\langle \sigma \rangle = 0.019$	$g(\sigma,\hat{\sigma})$	0.085	**1.000**	**1.000**	**1.000**	1.004	**0.994**	1.013	1.313	1.633	1.112	**1.015**	0.920
ETE (H)	$\langle \mu \rangle = 0.674$	$f(\mu,\hat{\mu})$	-0.158	**0.000**	**0.000**	**0.000**	0.014	0.026	0.000	0.075	0.155	0.026	0.748	0.774
	$\langle \sigma \rangle = 0.019$	$g(\sigma,\hat{\sigma})$	0.083	**1.000**	**1.000**	**1.000**	0.992	0.994	1.009	1.296	1.634	1.109	0.947	0.854
TE (KSG)	$\langle \mu \rangle = 1.509$	$f(\mu,\hat{\mu})$	-1.348	**0.000**	0.269	0.609	0.095	-0.134	-0.029	0.111	0.216	0.306	0.841	0.863
	$\langle \sigma \rangle = 0.025$	$g(\sigma,\hat{\sigma})$	0.120	1.071	0.869	1.026	1.232	1.320	1.042	1.197	1.430	1.028	1.017	**0.962**
CTIR	$\langle \mu \rangle = 0.462$	$f(\mu,\hat{\mu})$	-1.226	**0.000**	0.128	0.713	0.299	-0.355	-0.026	0.083	0.161	0.273	0.826	0.848
	$\langle \sigma \rangle = 0.014$	$g(\sigma,\hat{\sigma})$	0.110	1.016	0.898	0.920	1.159	1.209	1.042	1.076	1.258	0.958	0.942	0.872
SI$^{(1)}$	$\langle \mu \rangle = 0.001$	$f(\mu,\hat{\mu})$	0.015	**0.000**	**0.000**	**0.000**	0.549	-0.556	0.005	-0.013	0.018	-0.007	**-0.061**	0.066
	$\langle \sigma \rangle = 0.029$	$g(\sigma,\hat{\sigma})$	0.097	**1.000**	**1.000**	**1.000**	1.363	1.355	1.053	1.084	1.219	0.895	0.705	0.693
SI$^{(2)}$	$\langle \mu \rangle = 0.000$	$f(\mu,\hat{\mu})$	0.029	**0.000**	**0.000**	**0.000**	-3.105	3.121	-0.001	-0.037	0.017	**0.000**	-8.256	7.736
	$\langle \sigma \rangle = 0.000$	$g(\sigma,\hat{\sigma})$	**0.000**	**1.000**	**1.000**	**1.000**	1.394	1.399	1.074	1.666	2.630	0.968	2.334	2.018
CCM	$\langle \mu \rangle = 0.001$	$f(\mu,\hat{\mu})$	0.031	**0.000**	**0.000**	**0.000**	-1.249	1.289	-0.005	-0.009	-0.025	0.013	0.151	-0.075
	$\langle \sigma \rangle = 0.047$	$g(\sigma,\hat{\sigma})$	0.103	**1.000**	**1.000**	**1.000**	1.115	1.105	0.740	1.090	1.250	1.010	0.944	0.959

text of causality indices, review papers[15–17] have previously had a narrower focus in some manner, for example on only one group of methods or on a few bivariate methods and their multivariate extensions. We follow the template of Ref 19 in reviewing methods drawn from diverse mathematical foundations, but we extend this review with additional methods and crucially we investigate the impact of common issues that are relevant to real-world data. In reproducing and updating their work, we are also able to resolve some computational stability issues and comment on the computational costs of each method, whilst we also make our code publicly available for other researchers to develop further.

Further work

A primary concern in causality inference is the difficulties with model misspecification, specifically causal identification in multivariate systems. Omission of confounding variables can create spurious false-positive causal relationships. There may also be redundancy across multiple variables that provide similar information to the effect variable or sets of variables that interact synergistically such that their combined causal influence is greater than the 'sum of their parts'. These are key concerns outlined in Ref 16 and, consequently, results from bivariate indices cannot be definitively interpreted as the ex-

istence of a fundamental direct causal relationship between two variables[29]. A key avenue for further work is to advance this analysis beyond a bivariate setting by including possible confounding variables, in line with conditional extensions to Granger causality[9,20,40,41] and transfer entropy[42]. Recent work with graphical models of multivariate systems[2] is an important step towards high-dimensional causal identification.

Separate univariate embedding is not without some limitations and is not necessarily the optimal multivariate embedding. Aside from in Ref 43, mixed embeddings are as yet uncommon in causality estimation. There is not yet a theoretical framework for longitudinal data that is recorded non-simultaneous and irregularly. Often, a typical workflow for such data involves pre-processing to transform the data into a multivariate series with constant time intervals. However, many imputation methods result in significant and poorly quantified biases in information content and flow, which inevitably propagate through to estimates of causality, and more work is needed to explicitly factor this into a causal inference framework.

Summary and recommendations

Each causality index has strengths and weaknesses, and there is no single method whose all-round performance ex-

ceeds all others. Transfer entropy and Granger causality have long been regarded as the leading methods for systems that contain a small number of variables, and these have had wide applications[8,9,14]. Transfer entropy has the distinct advantage that it is built upon the principles of Shannon entropy in a well-established and universal information theoretic framework. It performs solidly throughout, though there is some tension between algorithms for TE, with the estimates rarely in complete agreement. We have shown that a histogram fixed partition approach is biased even in the simplest model, despite TE (H) having general consistency and computationally efficiency. Therefore, we recommend the KSG algorithm for transfer entropy computation, unless perhaps data is extremely scarce ($T < 10^3$). However, there are some unanswered concerns about TE (KSG), particularly that it appears to increase in magnitude as more data is available. TE (KSG) also suffers in performance when data is unequally scaled, due to the resultant difficulties with identifying unique nearest neighbours. CTIR, whilst sometimes not wholly dissimilar in value from TE, did not seem to offer any obvious advantage to compensate for its much higher computational cost or occasional unusual behaviour. Vanilla GC is widely favoured but has restrictive assumptions and is ill-suited to complex nonlinear problems. Of the two nonlinear extensions to Granger causality, Lungarella et al.[19] appear to prefer EGC. Some of the computational challenges and numerical instability that they experienced with NLGC may have been a result of their choice of a fuzzy c-means for determining RBF kernels, and alternate parameter choices appear to resolve some of their concerns. We find that NLGC is one of the most robust methods to rounding error, missing data and Gaussian noise. They rightly note that "If the rank of the data is small, kernel based methods tend to overfit"[19], but we did not observe any issues with this in our simulation experiments. Predictability improvement (PI) likewise performed solidly, and has a slight advantage amongst the regression based indices in that it that it is perhaps less reliant on parameter choices. Finally, dynamical systems theory offers a different insight into causal inference that should not be readily dismissed despite our mixed results here, even though the deterministic simulation models appeared to be well-suited to the underlying theory. Convergent cross mapping is a more recent and popular method, and this offered a broad improvement on the similarity indices (SI), which did not consistently identify the strength or direction of causality. However, CCM too did not always manage to determine the correct causal flow in our simulations.

We have highlighted the value of a standardisation preprocessing step in in avoiding algorithmic issues, which is also important in comparing results from different data for each method. Rounding error gives rise to practical issues within the implementation of several of the algorithms. For instance, in k-nearest neighbour approaches it is typically assumed that the distances between pairs of points are unique and not discrete. Subsequent edge cases can be treated by adding random noise with low amplitude to the data before estimating the causal relationships[27], though propagation of this noise to final estimates is something that should be analysed. Likewise, many existing implementations of the methods are not equipped to handle missing data (e.g. Refs. 42 and 44). We believe this is broadly straightforward to implement across all indices, as it can be handled exclusively within the time-delay embedding vector step, by performing an embedding and then removing any embedding vectors missing at least one component. As Mönster et al.[37] put it, "Noise in real-world data is ubiquitous, the inclusion of noise in model investigations has been largely ignored". Added Gaussian noise leads to the biggest changes in value for most methods, particularly noise at observation in the causal variable. However, provided the magnitude of noise is small compared to the values themselves, all methods perform adequately.

On the basis of this work, we conclude that the strongest choice for identifying and quantifying bivariate causal relationships is, in our view, either **transfer entropy (KSG)** or **nonlinear Granger causality**. Predictability improvement is a reasonable alternative and perhaps the next best candidate. A more cautious approach may involve using more than one method, from different theoretical backgrounds. Where possible, it is advantageous to identify a base case for the system, which subsequent results can be reliably compared against. For new methodologies, we recommend investigation into the real-world issues we have discussed.

CODE AND DATA AVAILABILITY

Our code is openly available at the GitHub repository `https://github.com/tedinburgh/causality-review` and Ref 45. The data that support the findings of this study are openly available at the same repository. A CODECHECK certificate is available confirming that the computations underlying this article could be independently executed: doi.org/10.5281/zenodo.4720843.

Existing open-access code for some indices include repositories for information theory and transfer entropy: IDTxl[42] v1.1, PyIF[46]; and for convergent cross mapping: pyEDM[44] v1.7.4. We also adapted fuzzy c-means code based on Ref 47. We checked our results for transfer entropy and convergent cross mapping against those from IDTxl and pyEDM respectively. All code in our repository and in these others is Python.

SUPPLEMENTARY MATERIALS

Our supplementary materials contains additional tables and figures. Tables S.I and S.II show full parameter choices and computational time requirements of each method respectively. Figures S1-3 show the results of real-world relevant transformation experiments. The supplementary materials also contain theoretical results for information theoretic measures in the linear process simulation.

ACKNOWLEDGMENTS

[1] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," Econometrica **37**, 424–438 (1969).

[2] J. Runge, "Causal network reconstruction from time series: From theoretical assumptions to practical estimation," Chaos **28**, 075310 (2018).

[3] M. Eichler, "Graphical modelling of multivariate time series," Probab. Theory Related Fields **153**, 233–268 (2012).

[4] J. Aldrich, "Correlations genuine and spurious in Pearson and Yule," Stat. Sci. **10**, 364–376 (1995).

[5] G. Sugihara, R. May, H. Ye, C.-H. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems," Science **338**, 496–500 (2012).

[6] C. A. Sims, "Money, income, and causality," Am. Econ. Rev. **62**, 540–552 (1972).

[7] T. Schreiber, "Measuring information transfer," Phys. Rev. Lett. **85**, 461–464 (2000).

[8] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," Econometrica **37**, 424–438 (1969).

[9] J. Geweke, "Inference and causality in economic time series models," in *Handbook of Econometrics*, Vol. 2 (Elsevier, 1984) pp. 1101–1144.

[10] D. D. Zhang, H. F. Lee, C. Wang, B. Li, Q. Pei, J. Zhang, and Y. An, "The causality analysis of climate change and large-scale human crisis," Proc. Natl. Acad. Sci. U. S. A. **108**, 17296–17301 (2011).

[11] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," Sci Adv **5**, 4996 (2019).

[12] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, E. H. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler, "Inferring causation from time series in earth system sciences," Nat. Commun. **10**, 2553 (2019).

[13] C. M. Gray, P. König, A. K. Engel, and W. Singer, "Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties," Nature **338**, 334–337 (1989).

[14] A. K. Seth, A. B. Barrett, and L. Barnett, "Granger causality analysis in neuroscience and neuroimaging," J. Neurosci. **35**, 3293–3297 (2015).

[15] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya, "Causality detection based on information-theoretic approaches in time series analysis," Phys. Rep. **441**, 1–46 (2007).

[16] M. Eichler, "Causal inference with multiple time series: principles and problems," Philos. Trans. A Math. Phys. Eng. Sci. **371**, 20110613 (2013).

[17] A. Papana, C. Kyrtsou, D. Kugiumtzis, and C. Diks, "Simulation study of direct causality measures in multivariate time series," Entropy **15**, 2635–2661 (2013).

[18] S. Palachy, "Inferring causality in time series data - towards data science," https://towardsdatascience.com/inferring-causality-in-time-series-data-b8b75fe52c46 (2019), accessed: Aug 28, 2020.

[19] M. Lungarella, K. Ishiguro, Y. Kuniyoshi, and N. Otsu, "Methods for quantifying the causal structure of bivariate time series," Int. J. Bifurcat. Chaos **17**, 903–921 (2007).

[20] Y. Chen, G. Rangarajan, J. Feng, and M. Ding, "Analyzing multiple nonlinear time series with extended Granger causality," Phys. Lett. A **324**, 26–35 (2004).

[21] N. Ancona, D. Marinazzo, and S. Stramaglia, "Radial basis function approach to nonlinear Granger causality of time series," Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **70**, 056221 (2004).

[22] U. Feldmann and J. Bhattacharya, "Predictability improvement as an asymmetrical measure of interdependence in bivariate time series," Int. J. Bifurcat. Chaos **14**, 505–514 (2004).

[23] R. Marschinski and H. Kantz, "Analysing the information flow between financial time series," The European Physical Journal B - Condensed Matter and Complex Systems **30**, 275–281 (2002).

[24] 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).

[25] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel method for nonlinear granger causality," Phys. Rev. Lett. **100**, 144103 (2008).

[26] M. Paluš, V. Komárek, Z. Hrncír, and K. Sterbová, "Synchronization as adjustment of information rates: detection from bivariate time series," Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **63**, 046211 (2001).

[27] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," Physical Review E **69**, 006138–006154 (2004).

[28] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick 1980* (Springer Berlin Heidelberg, 1981) pp. 366–381.

[29] J. Arnhold, P. Grassberger, K. Lehnertz, and C. E. Elger, "A robust method for detecting interdependences: application to intracranially recorded EEG," Physica D **134**, 419–430 (1999).

[30] J. Bhattacharya, E. Pereda, and H. Petsche, "Effective detection of coupling in short and noisy bivariate data," IEEE Trans. Syst. Man Cybern. B Cybern. **33**, 85–95 (2003).

[31] M. B. Kennel, R. Brown, and H. D. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," Phys. Rev. A **45**, 3403–3411 (1992).

[32] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," Phys. Rev. A Gen. Phys. **33**, 1134–1140 (1986).

[33] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," J. R. Stat. Soc. Series B Stat. Methodol. **63**, 411–423 (2001).

[34] J. Rissanen, "Modeling by shortest data description," Automatica **14**, 465–471 (1978).

[35] P. Hall and E. J. Hannan, "On stochastic complexity and nonparametric density estimation," Biometrika **75**, 705–714 (1988).

[36] A. T. Clark, H. Ye, F. Isbell, E. R. Deyle, J. Cowles, G. D. Tilman, and G. Sugihara, "Spatial convergent cross mapping to detect causal relationships from short time series," Ecology **96**, 1174–1181 (2015).

[37] D. Mønster, R. Fusaroli, K. Tylén, A. Roepstorff, and J. F. Sherson, "Inferring causality from noisy time series data," arXiv (2016), https://arxiv.org/pdf/1603.01155, arXiv:1603.01155 [nlin.CD].

[38] M. Hénon, "A two-dimensional mapping with a strange attractor," Commun. Math. Phys. **50**, 69–77 (1976).

[39] C. S. Cutts and S. J. Eglen, "Detecting pairwise correlations in spike trains: an objective comparison of methods and application to the study o retinal waves," J. Neurosci. **34**, 14288–14303 (2014).

[40] E. Siggiridou and D. Kugiumtzis, "Granger causality in multivariate time series using a Time-Ordered restricted vector autoregressive model," IEEE Trans. Signal Process. **64**, 1759–1773 (2016).

[41] S. Guo, A. K. Seth, K. M. Kendrick, C. Zhou, and J. Feng, "Partial Granger causality—eliminating exogenous inputs and latent variables," Journal of Neuroscience Methods **172**, 79–93 (2008).

[42] J. T. Lizier, "JIDT: An information-theoretic toolkit for studying the dynamics of complex systems," Frontiers in Robotics and AI **1**, 11 (2014).

[43] I. Vlachos and D. Kugiumtzis, "Nonuniform state-space reconstruction and coupling detection," Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **82**, 016207 (2010).

[44] J. Park, C. Smith, G. Sugihara, and E. Deyle, "EDM: Empirical dynamic modelling ('pyEDM'). Python package version 1.7.0." (2020), https://github.com/SugiharaLab.

[45]T. Edinburgh, "Bivariate causality indices review: code, data and figures," https://doi.org/10.5281/zenodo.4746192 (2021).

[46]R. Brunner, K. Ikegwu, J. Trauger, and T. Trauger, "PyIF," https://github.com/lcdm-uiuc/PyIF (2019), accessed: Sep 05, 2020.

[47]A. Nour Jamal El-Din and O. Aljabasini, "Kernel Granger Causality," https://github.com/ITE-5th/fuzzy-clustering (2018), accessed: Jan 15, 2021.

[48]F. Perez and B. E. Granger, "IPython: A system for interactive scientific computing," Computing in Science Engineering **9**, 21–29 (2007).

[49]J. D. Hunter, "Matplotlib: A 2D graphics environment," Computing in Science Engineering **9**, 90–95 (2007).

[50]C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. Del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," Nature **585**, 357–362 (2020).

[51]M. Davis, "Palettable," https://github.com/jiffyclub/palettable (2012), accessed: Mar 09, 2020.

[52]W. McKinney, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference* (SciPy, 2010).

[53]G. Van Rossum and F. L. Drake Jr, *Python Tutorial* (Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995).

[54]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).

[55]P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: fundamental algorithms for scientific computing in python," Nat. Methods **17**, 261–272 (2020).

[56]S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, Vol. 57 (2010) p. 61.