


Explicit Granger causality in kernel Hilbert spaces

Diego Bueso ^{*}, Maria Piles , and Gustau Camps-Valls 
Image Processing Laboratory (IPL), Universitat de València, 4610 València, Spain

 (Received 14 May 2020; accepted 28 October 2020; published 2 December 2020)

Granger causality (GC) is undoubtedly the most widely used method to infer cause-effect relations from observational time series. Several nonlinear alternatives to GC have been proposed based on kernel methods. We generalize kernel Granger causality by considering the variables' cross-relations explicitly in Hilbert spaces. The framework is shown to generalize the linear and kernel GC methods and comes with tighter bounds of performance based on Rademacher complexity. We successfully evaluate its performance in standard dynamical systems, as well as to identify the arrow of time in coupled Rössler systems, and it is exploited to disclose the El Niño–Southern Oscillation phenomenon footprints on soil moisture globally.

DOI: [10.1103/PhysRevE.102.062201](https://doi.org/10.1103/PhysRevE.102.062201)

I. INTRODUCTION

Establishing causal relations between random variables from observational data is perhaps the most important challenge in today's science, from Earth sciences [1] and neurosciences [2] to social sciences [3]. Granger causality (GC) [4] was introduced as a first attempt to formalize quantitatively the causal relation between time series and is the most widely used method. The intuition behind GC is to test whether the past of X helps in predicting the future of Y from its past alone. GC implicitly tells us about the concept of *information* using *forecasting*. Other methods rely on similar concepts of *information flow* and *predictability*: connections can be established between GC and transfer entropy [5], directed information [6], convergent cross-mapping [7], and Liang's measure of information flow [8] and with the graphical causal model perspective [9].

Noting the strong linearity assumption in GC [10], nonlinear extensions of GC have been proposed, and many discussions of the validity of nonparametric test statistics and nonlinear GC models exist in the literature [11,12]. Several studies propose replacing the linear autoregressive (AR) models with neural networks or random forests as forecasting methods: while improved efficiency and model versatility are achieved, there is no principled statistical test to assess GC causality. A solid and mathematically sound approach comes from the field of kernel methods [13], which allows the development of nonlinear models from linear ones, while still resorting to linear algebra operations. Kernel methods have been widely used for regression, classification, and dimensionality reduction. GC with kernels was originally introduced in [14]. The method assumed a particular class of functions and an additive interaction between them. An alternative kernel-based test in combination with a filtering approach was later introduced in [15]. In all these studies, the AR models use kernel-based regression for stacking the involved variables in

input spaces. This approach, however, is limited, as it does not consider nonlinear cross-relations between X and Y explicitly in Hilbert spaces.

We here introduce explicit feature maps and corresponding kernel functions that account for nonlinear cross-relations in kernel space [16]. We demonstrate that the cross-kernel methodology generalizes linear and kernel GC methods, come with statistical guarantees, and yield enhanced detection power.

II. NONLINEAR GRANGER CAUSALITY WITH KERNELS

GC first builds univariate and bivariate autoregressive models—(i) $y_{t+1} = \sum_{p=0}^P a_p y_{t-p} + \varepsilon_t^y$ and (ii) $y_{t+1} = \sum_{p=0}^P a_p y_{t-p} + \sum_{q=0}^Q b_q x_{t-q} + \varepsilon_t^{y|x}$ —and then computes a GC test as the ratio of model fitting errors, $\delta_{x \rightarrow y} = \log(\mathbb{V}[\varepsilon_t^y]/\mathbb{V}[\varepsilon_t^{y|x}])$, where the residual errors are defined for the unrestricted ε_t^y and restricted $\varepsilon_t^{y|x}$ cases separately, and \mathbb{V} represents the variance operator. Time embeddings P and Q are selected by cross-validation or sensible statistical criteria. Regressors are defined as $\mathbf{y}_t = [y_t, y_{t-1}, \dots, y_{t-P}]^T$ and $\mathbf{x}_t = [x_t, x_{t-1}, \dots, x_{t-Q}]^T$, and vector coefficients $\mathbf{a} = [a_1, \dots, a_P]^T$ and $\mathbf{b} = [b_1, \dots, b_Q]^T$ are typically estimated by least squares.

A. Feature maps and kernel functions

The linear GC formulation can be generalized to the nonlinear case using elements of the theory of reproducing kernel Hilbert spaces (RKHS) [13]. Let us assume the existence of a Hilbert space \mathcal{H} equipped with an inner product where samples in \mathcal{X} are mapped in by means of a feature map $\phi: \mathcal{X} \rightarrow \mathcal{H}$, $\mathbf{x}_i \mapsto \phi(\mathbf{x}_i)$, $1 \leq i \leq n$. The similarity between the elements in \mathcal{H} can be estimated using its associated dot product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ via RKHS, $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, such that pairs of points $(\mathbf{x}, \mathbf{x}') \mapsto k(\mathbf{x}, \mathbf{x}')$. Therefore, one can estimate similarities in \mathcal{H} without the explicit definition of the feature map ϕ and, hence, without even having access to the points in \mathcal{H} .

^{*}diego.bueso@uv.es

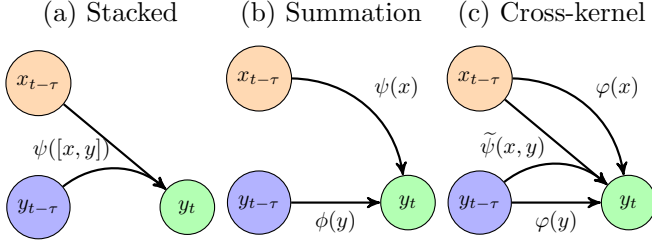


FIG. 1. Representation of the GC model for the different kernel functions. Each model encodes explicit relations of past states with future ones. Note the difference between the regular map $\psi([x, y])$ working with the concatenation of time series in the input domain and the joint map $\tilde{\psi}(x, y)$ working with the concatenation of maps of x and y in Hilbert spaces.

An important concept in kernel methods is the representer or Riesz' representation theorem [17,18]. The *representer theorem* gives us the general form of the solution to the common loss formed by a cost (loss, energy) term and a regularization term.

In an RKHS \mathcal{H} , there exists a (kernel) function $k(\cdot, \cdot)$ such that $f(\mathbf{x}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i)$, $\alpha_i \in \mathbb{R}$, $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_n]^T \in \mathbb{R}^n$, which is a linear combination of kernel functions. This property has been widely used to develop kernel methods for classification, clustering, and regression [13]. Defining the regularized least squares functional, $L(y, \hat{y}) = \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2 + \lambda \|f\|_{\mathcal{H}}^2$ leads to the kernel ridge regression (KRR) method [13,19], which has a convenient analytic solution, $\boldsymbol{\alpha} = (\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{y}$, where λ is the regularization term, \mathbf{I} is the identity matrix, and \mathbf{K} is the kernel matrix with entries $k(\mathbf{x}_i, \mathbf{x}_j) \in \mathbb{R}$. The KRR method is the preferred kernel method for nonlinear GC because of its simplicity (only one hyperparameter is involved) and good results in practice [14,15]. In kernel GC, however, an important aspect has been largely disregarded: the proper definition of the mapping function that gives rise to the kernel function itself. Next, we formalize the field of kernel GC by proposing an explicit definition of the cross-terms between variable X and variable Y in Hilbert spaces.

1. Stacked kernel

The standard kernel GC (KGC) approach considers a straightforward approach to AR modeling with kernels [14,15] [see Fig. 1(a)]. The method essentially defines two feature maps, ϕ and ψ , to an RKHS \mathcal{H} endorsed with reproducing kernels k and ℓ , which \mathbf{y}_t and the concatenation $\mathbf{z}_t = [\mathbf{y}_t, \mathbf{x}_t] \in \mathbb{R}^{P+Q}$ are mapped to, respectively. This leads to the kernel regression models (i) $y_{t+1} = \mathbf{a}_H^T \phi(\mathbf{y}_t) + \varepsilon_t^y$ and (ii) $y_{t+1} = \mathbf{b}_H^T \psi(\mathbf{z}_t) + \varepsilon_t^{y|x}$, where now $\mathbf{a}_H, \mathbf{b}_H \in \mathbb{R}^{H \times 1}$. Now, by using the representer's theorems [17,18] on the model weights defined in RKHS, $\mathbf{a}_H = \Phi^T \boldsymbol{\alpha}$ and $\mathbf{b}_H = \Psi^T \boldsymbol{\beta}$, where $\Phi, \Psi \in \mathbb{R}^{n \times H}$, the AR models can be defined in terms of kernel functions only: $y_{t+1} = \boldsymbol{\alpha}^T \mathbf{k}_t + \varepsilon_t^y$, and $y_{t+1} = \boldsymbol{\beta}^T \boldsymbol{\ell}_t + \varepsilon_t^{y|x}$, respectively, where $\mathbf{k}_t = [k(\mathbf{y}_1, \mathbf{y}_t), \dots, k(\mathbf{y}_n, \mathbf{y}_t)]^T$ and $\boldsymbol{\ell}_t = [\ell(\mathbf{z}_1, \mathbf{z}_t), \dots, \ell(\mathbf{z}_n, \mathbf{z}_t)]^T$ contain all evaluations of the kernel functions, k and ℓ , at time t , which act as similarity measures between the input feature vectors. Importantly, note that since

data are mapped to the same Hilbert space \mathcal{H} , the same kernel function and parameters are used for both k and ℓ .

2. Summation kernel

An alternative to the stacked approach builds *implicit* AR models in RKHS [14] such that $y_{t+1} = \mathbf{a}_H^T \phi(\mathbf{y}_t) + \varepsilon_t^y$ and $y_{t+1} = \mathbf{b}_H^T \psi(\mathbf{z}_t) + \varepsilon_t^{y|x}$, which leads to the kernel AR models $y_{t+1} = \boldsymbol{\alpha}^T \mathbf{k}_t + \varepsilon_t^y$ and $y_{t+1} = \boldsymbol{\beta}^T \boldsymbol{\ell}_t + \varepsilon_t^{y|x}$, where now $\boldsymbol{\ell}_t := [\ell(\mathbf{x}_1, \mathbf{x}_t), \dots, \ell(\mathbf{x}_n, \mathbf{x}_t)]^T$. The summation kernel is more appropriate when large-time embeddings P and Q are needed to capture long-term memory processes, since it avoids constructing large-dimensional feature vectors \mathbf{z} by concatenation [cf. Fig. 1(b)]. However, the cross-information between X and Y is missing [16].

3. Explicit cross-kernel

In order to account for cross-correlations in Hilbert space, we explicitly define two feature maps: the standard individual map ϕ and the joint feature mapping ψ for the second AR model, $y_{t+1} = \mathbf{a}_H^T \phi(\mathbf{y}_t) + \varepsilon_t^y$ and $y_{t+1} = \mathbf{b}_H^T \psi(\mathbf{x}_t, \mathbf{y}_t) + \varepsilon_t^{y|x}$, where the joint map is defined by construction as $\tilde{\psi}(\mathbf{x}_t, \mathbf{y}_t) := [\mathbf{A}_1 \phi(\mathbf{y}_t), \mathbf{A}_2 \phi(\mathbf{x}_t), \mathbf{A}_3 (\phi(\mathbf{y}_t) + \phi(\mathbf{x}_t))]^T$, where ϕ is a nonlinear feature map into an RKHS \mathcal{H} , and $\mathbf{A}_i, i = 1, 2, 3$, are three linear transformations from \mathcal{H} to \mathcal{H}_i . The induced joint kernel function readily becomes

$$\begin{aligned} n((\mathbf{x}_t, \mathbf{y}_t), (\mathbf{x}'_t, \mathbf{y}'_t)) &= \tilde{\psi}(\mathbf{x}_t, \mathbf{y}_t)^T \tilde{\psi}(\mathbf{x}'_t, \mathbf{y}'_t) \\ &= \boldsymbol{\varphi}(\mathbf{y}_t)^T \mathbf{R}_1 \boldsymbol{\varphi}(\mathbf{y}'_t) + \boldsymbol{\varphi}(\mathbf{x}_t)^T \mathbf{R}_2 \boldsymbol{\varphi}(\mathbf{x}'_t) \\ &\quad + \boldsymbol{\varphi}(\mathbf{y}_t)^T \mathbf{R}_3 \boldsymbol{\varphi}(\mathbf{x}'_t) + \boldsymbol{\varphi}(\mathbf{x}_t)^T \mathbf{R}_3 \boldsymbol{\varphi}(\mathbf{y}'_t) \\ &= n_1(\mathbf{y}_t, \mathbf{y}'_t) + n_2(\mathbf{x}_t, \mathbf{x}'_t) + n_3(\mathbf{y}_t, \mathbf{x}'_t) \\ &\quad + n_4(\mathbf{x}_t, \mathbf{y}'_t), \end{aligned}$$

where $\mathbf{R}_1 = \mathbf{A}_1^T \mathbf{A}_1 + \mathbf{A}_3^T \mathbf{A}_3$, $\mathbf{R}_2 = \mathbf{A}_2^T \mathbf{A}_2 + \mathbf{A}_3^T \mathbf{A}_3$, and $\mathbf{R}_3 = \mathbf{A}_3^T \mathbf{A}_3$. Note that the new kernel function considers cross-term relations between the time series through kernels n_3 and n_4 and still works with the original time embeddings. Besides, there is no need to explicitly use the same kernel function or parameters. We now show that the cross-kernel GC (XKGC) method generalizes previous KGC methods and comes with statistical guarantees [see Fig. 1(c)].

B. Statistical characterization with Rademacher complexity

Let us now characterize the generalization capabilities of the proposed cross-kernel using the notion of Rademacher complexity, which is perhaps the most useful measure used in the theoretical analysis and design of kernel algorithms [20,21]. Rademacher complexity roughly states that one can infer (measure) an upper bound on the generalization performance of a given class by its ability to fit random data. The theory makes use of the Rademacher variables and produces a measure of capacity called the Rademacher complexity.

In what follows we give a bound of performance for the general case of compositions of kernels as in our proposed framework. Generalization bounds based on Rademacher complexity [22,23] provide a strong theoretical foundation for a family of learning kernel algorithms based on convex combinations of base kernels, as in our case. Let us define

a sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$ generated by a distribution D on a set \mathcal{X} , a family of functions $\mathbf{H} = \{h : \mathcal{X} \rightarrow \mathbb{R}\}$, and a loss function $L : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$. The goal is to find the h hypothesis in \mathbf{H} with a small generalization error with respect to the target $f(\mathbf{x})$, $R_D(h) = \mathbb{E}_{\mathbf{x} \sim D}[L(h(\mathbf{x}), f(\mathbf{x}))]$, empirically estimated as $R_D(h) = \frac{1}{n} \sum_{i=1}^n L(h(\mathbf{x}_i), y_i)$, e.g., the least squares $L(y, \hat{y}) = (y - \hat{y})^2$.

Theorem 1: Generalization bound with finite \mathbf{H} . Assuming a finite hypothesis set, \mathbf{H} , and that L is bounded by ε , then for any $\delta > 0$, with probability at least $1 - \delta$,

$$R(h) \leq \widehat{R}(h) + \varepsilon \sqrt{\frac{\log |\mathbf{H}| + \log(2/\delta)}{2n}},$$

which can be particularized for the squared loss.

Theorem 2: Least-squares kernel regression bound. Let $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a positive definite kernel and $\phi : \mathcal{X} \rightarrow \mathcal{H}$ be a feature map associated with k . Let the class of functions $\mathbf{H} = \{\mathbf{x} \mapsto \mathbf{w}^\top \phi(\mathbf{x}) : \|\mathbf{w}\| \leq \Lambda\}$. Assume that both the kernel and the function are bounded; $k(\mathbf{x}, \mathbf{x}) \leq M^2$ and $|f(\mathbf{x})| \leq \Lambda M$ for all $\mathbf{x} \in \mathcal{X}$. The generalization risk R is bounded by the empirical risk \widehat{R} as follows. For $\delta > 0$, with probability at least $1 - \delta$ over random draws of samples of size n , in the sample S for every $h \in \mathbf{H}$ satisfies

$$R(h) \leq \widehat{R}(h) + \frac{8M^2\Lambda^2}{\sqrt{n}} \left(1 + \frac{1}{2} \sqrt{\frac{\log(1/\delta)}{2}}\right),$$

and for every $h \in \mathbf{H}$

$$R(h) \leq \widehat{R}(h) + \frac{8M^2\Lambda^2}{\sqrt{n}} \left(\sqrt{\frac{\text{Tr}[\mathbf{K}]}{nM^2}} + \frac{3}{4} \sqrt{\frac{\log(2/\delta)}{2}}\right).$$

This follows from the generalization bound with a finite \mathbf{H} and the bound on the Rademacher complexity of kernel hypotheses [20,23]. Let us particularize this result for the cross-kernel and assess its generality.

Theorem 3: Cross-kernel Rademacher complexity bounds. The KRR function class h uses the squared loss for $\widehat{R}(h)$. Let us assume a radial basis function kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / (2\sigma^2))$ so $k(\mathbf{x}, \mathbf{x}) = 1$, and let $\gamma \in [0, 1]$ and $\beta \in [\gamma, 1]$. The Rademacher complexity regression minimum bound for the cross-kernel is

$$R_{\text{cross}}(h) \leq \widehat{R}(h) + \frac{8\|h\|^2}{\sqrt{n}} \left(\sqrt{\frac{1+\gamma}{1+\beta}} + \frac{3}{4} \sqrt{\frac{\log(2/\delta)}{2}}\right).$$

Proof. The Rademacher complexity for a sum of N kernels K_i can be easily bounded as $\widehat{R}(h) = \sqrt{N} \widehat{R}(h_i)$, $i = 1, \dots, N$. It is easy to see that $M^2 = 2(1 + \beta)$ and the $\text{Tr}[\mathbf{K}] = 2n(1 + \gamma)$ for the cross-kernel. The result follows from substituting them in Theorem 1. Since for the stacked kernel, $M^2 = 1$ and $\text{Tr}[\mathbf{K}] = n$, and for the summation $M^2 = 2$ and $\text{Tr}[\mathbf{K}] = 2n$, it follows that $R_{\text{cross}}(h) \leq R_{\text{sum}}(h) = R_{\text{stacked}}$. Note that for $\gamma = \beta$, i.e., when X and Y convey correlated information, the cross-kernel bound converges to the stacked and the summation bounds. Interestingly, since $\gamma \leq \beta$, the cross-kernel bound will always be tighter than the stacked and summation bounds, which are confirmed experimentally (see Table I).

TABLE I. Complexity terms for logistic maps.

	Stacked	Summation	Cross-kernel
R	0.4390	0.4361	0.4352
\widehat{R}	0.6283	0.5001	0.4992

III. EXPERIMENTS

In all our experiments, we used the radial basis function kernel function and the KRR method. Hyperparameters (regularization term λ and kernel length scale σ) were selected by cross-validation. The statistical test of robustness was computed as in KGC [15], and the threshold was set to the highest causal strength estimated from 100 surrogate time series [24]. For the sake of reproducibility, code snippets and demos are provided in XKGC [25]. We compare GC, KGC (stacked and summation kernels are theoretically identical and deemed similar in practice in low-dimensional settings), and XKGC in all experiments.

A. Nonlinear coupled system

Let us first consider a bivariate system with strongly coupled, nonlinear, and autoregressive relations defined as $x_{t+1} = 3.4x_t(1 - x_t^2) \exp(-x_t^2) + \varepsilon_t^x$ and $y_{t+1} = 3.4y_t(1 - y_t^2) \exp(-y_t^2) + \frac{x_t y_t}{2} + \varepsilon_t^y$, where ε is white Gaussian noise with mean 0 and variance 0.4. The causal direction is $x \rightarrow y$, the opposite direction being anticausal. Standard GC, KGC, and XKGC were run on a set of $n = 4000$ samples and repeated 10 000 times. Figure 2 shows the histogram of the estimated causality index. Results reveals the insensitivity of linear GC to the causal direction and the high false-positive rate of KGC, while XKGC shows a higher detection power and lower rates of false positives and true negatives.

B. Logistic maps

The second example considers the standard system of two logistic maps, defined as $x_{t+1} = 1 - 1.8x_t^2$ and $y_{t+1} = (1 - \alpha)(1 - 1.8y_t^2) + \alpha(1 - 1.8x_t^2)$, where $\alpha \in [0, 1]$ controls the coupling strength. The causal relationship implemented is $X \rightarrow Y$, and the challenge is to assess the detection power of methods without introducing any external variable, just using X and Y . We analyze segments of length $n = 2000$ and fixed

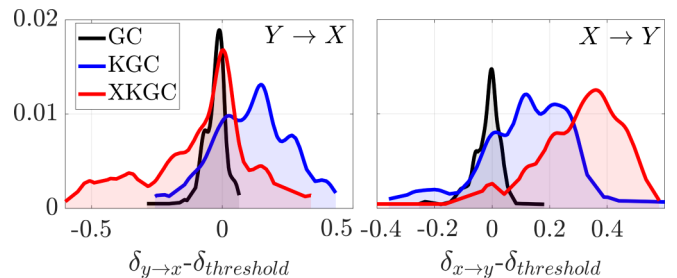


FIG. 2. Significance of the positive and negative cases detected for the coupled AR system. Histogram of the difference between the estimated causality index δ and the associated threshold $\delta_{\text{threshold}}$ are shown for each method and direction.

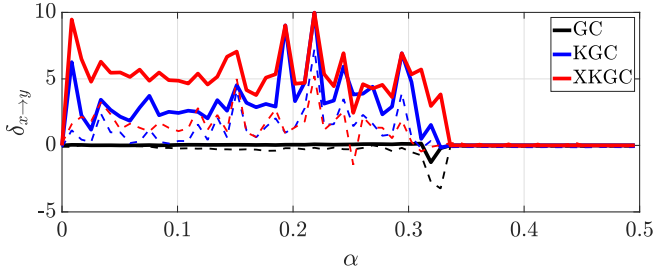


FIG. 3. Causality index δ estimated for each method (solid lines) and their associated thresholds (dashed lines) for the coupled logistic maps as a function of the coupling parameter α .

$p = 2$. Figure 3 shows the prediction skills for varying α . Note that the system is completely synchronized at $\alpha = 0.37$. The XKGC method shows improved detection power in the whole solution range.

We confirmed empirically the theoretical results in Sec. II B for the system of two logistic maps with $\alpha = 0.1$. Empirical results of the Rademacher complexity bounds are provided in Table I, where the cross-kernel achieves tighter bounds.

C. The arrow of time

Let us now exemplify the performance of the proposed methods in the challenging problem of detecting the arrow of time from bivariate time series. This is a mostly academic question that has captured attention in the physics literature [26], where both theoretical [27] and experimental [28] results have recently confirmed identifiability.

We study the coupled Rössler system, which encompasses the prediction of the causal direction between two variables as well as the identification of the direction of time. The Rössler systems were originally introduced in the 1970s as prototype equations for the study of continuous-time chaos. The bivariate system studied in our paper was extracted from [29]. The unidirectional bivariate coupled Rössler system was tested for a coupling parameter of $\epsilon = 0.07$, delay parameter $\delta = 0$, and 6000 samples. Characteristic parameters for each system and initial conditions remain as in the original work. We estimated the detection power in both forward and backward propagation by just flipping the time series.

Figure 4 shows that the physical nature of the coupling system emerges as a forward propagation and with $X \rightarrow Y$.

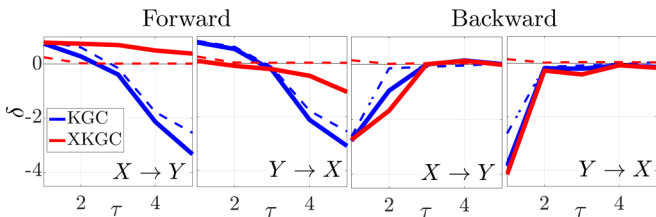


FIG. 4. Arrow of time and the causal direction for a coupled Rössler system. Causal strength estimated for each method (solid lines) and associated thresholds (dashed lines) for forward and backward propagation directions versus the time delay τ .

XKGC is the only method that can reconstruct the causal direction over the time delay order τ properly. This suggests that the proposal captures an extra variability of the coupled system, which in turn helps causal inference.

IV. CAUSAL FOOTPRINTS OF EL NIÑO ON SOIL MOISTURE

Causal discovery in Earth and climate sciences is a fundamental topic of research, as it allows systematic hypothesis testing, model-data intercomparison, and discovery of patterns and causal links in observational data [1]. The challenges are multifaceted: Earth data show spatiotemporal dimensions, complex nonlinear dynamics, and teleconnections. We here tackle the problem of inferring causal links between El Niño–Southern Oscillation (ENSO) and soil moisture (SM) globally. ENSO is a coupled ocean-atmosphere phenomenon, which manifests as a quasiperiodic fluctuation in sea surface temperature and air pressure in the equatorial Pacific Ocean. Although the exact causes initiating warm or cool ENSO events are not fully understood, the two components of ENSO—sea surface temperature and atmospheric pressure—are strongly related [30]. During the ENSO event, the atmospheric current of equatorial Walker circulation stops and changes its westward propagation for a more eastward direction. This transition occurs between 3 and 4 months before the rise in the sea surface temperature and is related to the normal ocean-atmosphere coupling affects the propagation of the low-pressure centers over the tropical regions, influencing temperature and precipitation across the globe. ENSO is hence strongly connected with global dry-wet anomalies, mostly over the tropics [32], but also over supratropical regions [33]. ENSO and SM are connected by the atmospheric current and their causal relation is interrupted by its variations.

Our causal analysis aims to uncover (spatially explicit) dry and wet patterns and identify footprints of ENSO on SM using satellite-based measurements. We use global SM maps from the ESA [34] Soil Moisture and Ocean Salinity (SMOS) mission and time series of the ENSO4 [35] climate index for the period 2010–2017. The information for the data set we use here is introduced in Appendix A. Our analysis focuses on the 2015–2016 ENSO event, which had a strong impact on the atmosphere circulation [36], being one of the three strongest El Niño events on record and the one of the longest duration [37]. ENSO events can emerge with different spatial patterns (called flavors as well) [38], which dominate the evolution of the ocean-atmosphere feedbacks and therefore the teleconnection patterns [39].

Previous studies have shown that interannual variability in SM reflected known ENSO teleconnection patterns [40,41]. However, these works focused only on association (correlation), not on predictability (Granger causality). In this work, we are concerned with two important questions of the ENSO-SM coupled system: (1) Can we identify the different phases of the ENSO event and its transitions from the neutral state from purely observational data? and (2) What part of the SM interannual signal is Granger caused by ENSO? Answering these two questions allows us to revisit the ENSO-SM tele-

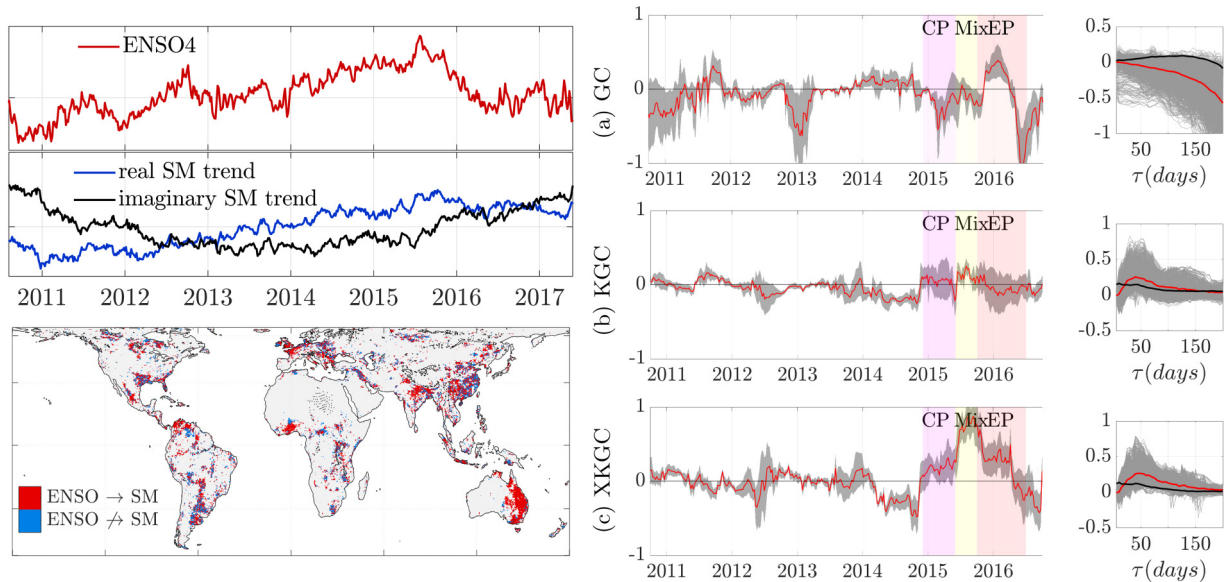


FIG. 5. Top left: ENSO4 (red) and estimated SM interannual complex components (blue and black) studied. Right panel: GC analysis of ENSO and SM. Left plots show δ (red line) with (a) GC, (b) KGC, and (c) XKGC estimated with a 2-year moving window, its variability (shaded), and highlight the three phases in the 2016 ENSO event. CP, Central Pacific; EP, East Pacific; Mix, mixture of both cases). Right plots show the distribution (gray) and averages (red) of δ and thresholds (black) for lag τ and 100 runs. Bottom left: Spatial distribution of causal footprints of ENSO on SM obtained by XKGC.

connection map, to find unreported footprints of ENSO on SM globally.

The information in the global data cubes of SM needs to be summarized in the so-called ‘modes of variability,’ i.e., spatially and temporally intrinsic components describing regional subprocess interactions. This is typically done with dimensionality reduction methods, such as the principal component analysis (PCA)–empirical orthogonal function. PCA can only achieve a *linear* and *orthogonal* feature representation, which is not appropriate to deal with the highly nonlinear and interdependent nature of Earth observational data. As an alternative, a complex-valued nonlinear PCA was applied here to extract the dominant modes of global SM variability for the study period [42]. The method allows us to extract nonlinear features that have independent spatial and temporal components in the complex domain. The so-called ROCK (rotated complex kernel)-PCA method performs the eigendecomposition of a kernel matrix using data in the complex domain after applying the Hilbert transform [43] and further rotating with a Promax transform [44]. Complex-valued processes return us more useful components, for example, the interpretation of phase-modulation decomposition versus only the real part returned by regular PCA. In addition, the nonlinear nature of ROCK-PCA allows us to better capture feature relations. We provide source code of the method in ROCK-PCA [42,45].

In our experiments, we focus on the extracted interannual component of the global soil moisture satellite data, which represents 10.2% of the total variance. We use the temporal feature to estimate the link with ENSO and its spatial representation to identify the regions where the mode is relevant, i.e., its spatial amplitude is higher than the median of the spatial amplitudes with one positive standard deviation. The interannual SM component is lag-correlated (80 days) with ENSO4, $\rho \sim 0.8$, and cointegrated [see Fig. 5(a)]. A map of

the ENSO-SM causality index and details on its spatialization are provided in Appendix B.

We analyzed the causal relation over a 2-year moving window to deal with nonstationarity [46] and studied the model’s δ sensitivity following [47]. We studied the model’s sensitivity by jittering parameters for each trained model resulting from a different combination of window and time embedding [47]. The model was also trained for several time embeddings to find the optimum time delay of the variability shared between the signals. For each trained model (window and time embedding) we have also estimated the sensibility of the causal index introducing a slight perturbation of the model parameters [47]. Results in Fig. 5 (right panel) reveal clear differences between the linear and the nonlinear $\delta_{\text{ENSO} \rightarrow \text{SM}}$ both across time and per time embedding. KGC and XKGC yield similar results yet differ in the magnitude of the captured variability. Note that XKGC more clearly differentiates the three phases and the atmospheric disruption before ENSO rises. The impact of ENSO on the spatial distribution of the SM interannual trend is analyzed in Fig. 5 (bottom left). On average, XKGC results indicate that about 50% of the SM interannual variability is caused by ENSO. Regions where the SM interannual variability can be predicted by ENSO reproduce the well-known ENSO-induced precipitation patterns and teleconnections [48–51]. While some regions are clearly dominated by ENSO (e.g., Australia), others can only partially be explained by it (e.g., Gulf of Mexico, Southeast Asia), probably due to the influence of other atmospheric hydroclimatic patterns such as the Madden-Julian Oscillation [52]. Notably, XKGC uncovers the impact of ENSO in yet unreported areas (e.g., northwestern Europe). Investigating additional causes dominating SM interannual variability and the emergence of potential new teleconnection patterns is recommended for future research.

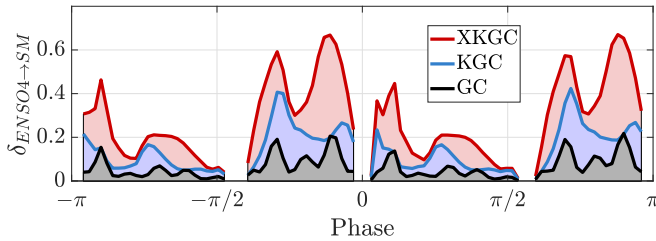


FIG. 6. Dependence of $\delta_{\text{ENSO4} \rightarrow \text{SM}}$ on the phase. Note that all methods preserve the noncausal phase dependence. The causality index-phase relation is cyclic with the π period.

V. CONCLUSIONS

We have considered the problem of Granger causality and proposed a kernel-based framework that generalizes linear GC and KGC approaches. The theory of reproducing kernel functions allows us to derive different nonlinear algorithms while still resorting to linear algebra operations. The methodology copes with nonlinear relationships more efficiently and comes with statistical guarantees.

The methodology outperformed linear and nonlinear counterparts in standard dynamical systems, the arrow of time problem, and a real Earth system science problem. We expect that the generalized kernel Granger causality framework introduced here paves the way to enhanced models through the appropriate definition of kernel functions that account for signal characteristics explicitly, from correlated noise to complex-valued signals and spatiotemporal structures, just to name a few of the pressing challenges in many fields of science.

ACKNOWLEDGMENTS

This work was partly supported by the European Research Council (ERC) under the ERC-CoG-2014 SEDAL project (647423) and project RTI2018-096765-A-100 (MCIU/AEI/FEDER, UE).

APPENDIX A: SMOS AND ENSO DATA SETS

We use global soil moisture maps from the ESA SMOS mission [available at Barcelona Expert Center (BEC)]; [34]. Since its launch in 2010, SMOS provides global maps of the Earth's surface soil moisture (top 5 cm) every 3 days with a spatial resolution of ~ 50 km and a target accuracy of $0.04 \text{ m}^3 \cdot \text{m}^{-3}$. We selected the first 7 years of SMOS observa-

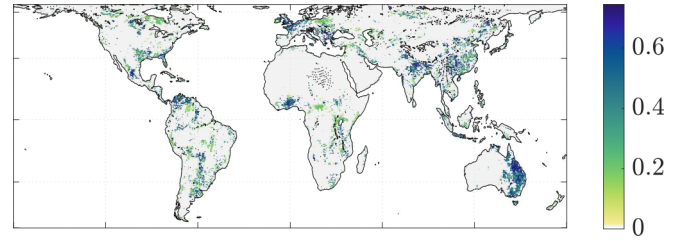


FIG. 7. Global distribution of the ENSO-SM causality index $\delta_{\text{ENSO4} \rightarrow \text{SM}}$. A greater δ indicates that a higher percentage of SM interannual variability can be explained by ENSO.

tions, after its commissioning phase (from May 2010 to May 2017) and focus on the transition of the 2016 ENSO event. As suggested in [41], ascending and descending daily orbits were temporally averaged and 5-day bins were constructed to ensure enough coverage and smooth spatiotemporal transitions; pixels with less than 30% temporal coverage and latitudes higher than 60° were not considered. Alongside the SM data, we use time series of the ENSO4 climate index from The Royal Netherlands Meteorological Institute (KNMI; [35]), which is calculated daily based on sea surface temperature (SST) anomalies averaged across the central equatorial Pacific Ocean (5N-5S, 160E-150W). ENSO4 time series were temporally averaged into 5-day bins for this study.

APPENDIX B: SPATIALIZATION OF THE CAUSALITY INDEX

The ROCK-PCA method extracts time series and spatial components in the complex domain, which can be connected in the phase space, i.e., the phase of the spatial component is the phase of the time series for each pixel. Hence, searching for the dependence of δ on the phase (mixture of real and imaginary time-series components as shown in Fig. 6), we can spatialize our results. This phase dependence is the temporal mean for each moving window δ estimation.

The relation between δ and each pixel using the transformation of the curve in Fig. 6 allows us to obtain the spatial map of δ , which is shown in Fig. 7 for the XKGC method. Regions with causal representation but without SM variability (e.g., deserts) are masked in Fig. 7. Interestingly, regions with clearly differentiated δ levels emerge. This map can be interpreted as a forecast skill, where differences in skill are caused by the different underlying mechanisms involved in the ENSO and SM relation in each region. Most of the causal regions are represented over the tropics as we expect, but other supratropical regions, such as Northwest Europe, emerge.

-
- [1] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Clymour, M. Kretschmer, M. Mahecha, J. Muñoz-Mari *et al.*, *Nat. Commun.* **10**, 2553 (2019).
- [2] A. T. Reid, D. B. Headley, R. D. Mill, R. Sanchez-Romero, L. Q. Uddin, D. Marinazzo, D. J. Lurie, P. A. Valdés-Sosa, S. J. Hanson, B. B. Biswal *et al.*, *Nat. Neurosci.* **22**, 1751 (2019).
- [3] M. M. Marini and B. Singer, *Sociol. Methodol.* **18**, 347 (1988).
- [4] C. W. J. Granger, *Econometrica* **37**, 424 (1969).
- [5] T. Schreiber, *Phys. Rev. Lett.* **85**, 461 (2000).
- [6] J. Massey, in *The International Symposium on Information Theory and Its Applications (ISITA)* (Citeseer, Waikiki, Hawaii, 1990), pp. 303–305.
- [7] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, *Science* **338**, 496 (2012).

- [8] X. S. Liang, *Phys. Rev. E* **94**, 052201 (2016).
- [9] H. White, K. Chalak, X. Lu *et al.*, in *Neural Information Processing Systems Mini-Symposium on Causality in Time Series* (Association for Computing Machinery, New York, 2011), pp. 1–29.
- [10] M. Eichler, in *Proceedings from the 13th International Congress of Logic, Methodology and Philosophy of Science* (North-Holland, Amsterdam, 2007), pp. 1–12.
- [11] N. Cartwright, *Hunting Causes and Using Them: Approaches in Philosophy and Economics* (Cambridge University Press, Cambridge, UK, 2007).
- [12] C. Diks and M. Wolski, *J. Appl. Econometr.* **31**, 1333 (2016).
- [13] J. Rojo-Álvarez, M. Martínez-Ramón, J. Muñoz-Marí, and G. Camps-Valls, *Digital Signal Processing with Kernel Methods* (Wiley & Sons, Chichester, UK, 2018).
- [14] N. Ancona, D. Marinazzo, and S. Stramaglia, *Phys. Rev. E* **70**, 056221 (2004).
- [15] D. Marinazzo, M. Pellicoro, and S. Stramaglia, *Phys. Rev. Lett.* **100**, 144103 (2008).
- [16] M. Martínez-Ramón, J. L. Rojo-Álvarez, G. Camps-Valls, A. Navia-Vázquez, E. Soria-Olivas, and A. R. Figueiras-Vidal, *IEEE Trans. Neur. Networks* **17**, 1617 (2007).
- [17] F. Riesz and B. S. Nagy, *Functional Analysis* (Frederick Ungar, New York, 1955).
- [18] G. Kiefer and G. Wahba, *J. Math. Anal. Appl.* **33**, 82 (1971).
- [19] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis* (Cambridge University Press, Cambridge, UK, 2004).
- [20] P. L. Bartlett and S. Mendelson, *J. Mach. Learn. Res.* **3**, 463 (2002).
- [21] V. Koltchinskii, D. Panchenko *et al.*, *Ann. Stat.* **30**, 1 (2002).
- [22] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, *J. Machine Learn. Res.* **5**, 27 (2004).
- [23] Y. Ying and C. Campbell, in *International Conference on Machine Learning, Quebec, Canada* (Association for Computing Machinery, New York, 2009).
- [24] R. Q. Quiroga, A. Kraskov, T. Kreuz, and P. Grassberger, *Phys. Rev. E* **65**, 041903 (2002).
- [25] <http://isp.uv.es/code/xkgc.zip>.
- [26] H. Reichenbach, *The Direction of Time, Vol. 65* (University of California Press, Berkeley, 1991).
- [27] S. Bauer, B. Schölkopf, and J. Peters, in *International Conference on Machine Learning, New York* (Association for Computing Machinery, New York, 2016), pp. 2043–2051.
- [28] M. Paluš, A. Krakovská, J. Jakubík, and M. Chvosteková, *Chaos* **28**, 075307 (2018).
- [29] M. Paluš and M. Vejmelka, *Phys. Rev. E* **75**, 056211 (2007).
- [30] J. D. Neelin, D. S. Battisti, A. C. Hirst, F.-F. Jin, Y. Wakata, T. Yamagata, and S. E. Zebiak, *J. Geophys. Res. [Oceans]* **103**, 14261 (1998).
- [31] G.-Y. Yang and B. Hoskins, *J. Atmos. Sci.* **70**, 3513 (2013).
- [32] K. E. Trenberth, G. W. Branstator, D. Karoly, A. Kumar, N.-C. Lau, and C. Ropelewski, *J. Geophys. Res. [Oceans]* **103**, 14291 (1998).
- [33] M. A. Alexander, I. Bladé, M. Newman, J. R. Lanzante, N.-C. Lau, and J. D. Scott, *J. Clim.* **15**, 2205 (2002).
- [34] <http://bec.icm.csic.es/>.
- [35] <https://www.knmi.nl/home>.
- [36] P. A. Newman, L. Coy, S. Pawson, and L. R. Lait, *Geophys. Res. Lett.* **43**, 8791 (2016).
- [37] P. Zhai, R. Yu, Y. Guo, Q. Li, X. Ren, Y. Wang, W. Xu, Y. Liu, and Y. Ding, *J. Meteorol. Res.* **30**, 283 (2016).
- [38] N. C. Johnson, *J. Clim.* **26**, 4816 (2013).
- [39] M. Wiedermann, J. F. Siegmund, J. F. Donges, J. Kurths, and R. V. Donner, Differential imprints of distinct enso flavors in global extreme precipitation patterns, *Geophys. Res. Abstr.* **19**, EGU2017-12639 (2017).
- [40] D. G. Miralles, M. J. V. D. Berg, J. H. Gash, R. M. Parinussa, R. A. M. D. Jeu, H. E. Beck, T. R. H. Holmes, C. Jiménez, N. E. C. Verhoest, W. A. Dorigo *et al.*, *Nature Climate Change* **4**, 122 (2013).
- [41] M. Piles, J. Ballabrera-Poy, and J. Muñoz-Sabater, *Remote Sens.* **11**, 95 (2019).
- [42] D. Bueso, M. Piles, and G. Camps-Valls, *IEEE Trans. Geosci. Remote Sens.* **58**, 5752 (2020).
- [43] C. Rusu, P. Kuosmanen, and J. Astola, in *Proceedings of The 2005 International TICSP Workshop on Spectral Methods and Multirate Signal Processing, Riga, Latvia* (Tampere International Center for Signal Processing, Tampere, Finland, 2005), pp. 79–84.
- [44] A. E. Hendrickson and P. O. White, *Br. J. Stat. Psych.* **17**, 65 (1964).
- [45] <https://github.com/DiegoBueso/ROCK-PCA>.
- [46] M. F. Stuecker, F.-F. Jin, and A. Timmermann, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 13490 (2015).
- [47] N. Nicolaou and T. G. Constantinou, *Front. Neuroinformat.* **10**, 19 (2016).
- [48] A. Dai and T. M. L. Wigley, *Geophys. Res. Lett.* **27**, 1283 (2000).
- [49] B. Lyon and A. G. Barnston, *J. Clim.* **18**, 5095 (2005).
- [50] S.-W. Yeh, W. Cai, S.-K. Min, M. J. McPhaden, D. Dommenget, B. Dewitte, M. Collins, K. Ashok, S.-I. An, B.-Y. Yim *et al.*, *Rev. Geophys.* **56**, 185 (2018).
- [51] H.-M. Kim and P. Webster, in *Seminar on Seasonal Prediction: Science and Applications, 3–7 September 2012* (ECMWF, Shinfield Park, Reading, UK, 2013), pp. 25–34.
- [52] Y. Tang and B. Yu, *J. Geophys. Res. Atmos.* **113**, D14 (2008).