Reconstructing regime-dependent causal relationships from observational time series

Elena Saggioro,[1] Jana de Wiljes,[2] Marlene Kretschmer,[3] and Jakob Runge[4]
[1])*University of Reading, Department of Mathematics and Statistics, Whiteknights, PO Box 220, Reading RG6 6AX, UK (*e.saggioro@pgr.reading.ac.uk*)*
[2])*Universität Potsdam, Institut für Mathematik, Karl-Liebknecht-Str. 24/25, D-14476 Potsdam, Germany (*wiljes@uni-potsdam.de*)*
[3])*University of Reading, Department of Meteorology, Whiteknights, PO Box 220, Reading RG6 6AX, UK (*m.j.a.kretschmer@reading.ac.uk*)*
[4])*Deutsches Zentrum fr̈ Luft und Raumfahrt, Institut für Datenwissenschaften Datenmanagement und -analyse, Mälzerstr. 3, D-07745 Jena, Germany (*Jakob.Runge@dlr.de*)*

(Dated: 25 June 2020)

Inferring causal relations from observational time series data is a key problem across science and engineering whenever experimental interventions are infeasible or unethical. Increasing data availability over the past decades has spurred the development of a plethora of causal discovery methods, each addressing particular challenges of this difficult task. In this paper we focus on an important challenge that is at the core of time series causal discovery: regime-dependent causal relations. Many dynamical systems feature transitions in time, depending on some, often persistent, unobserved background regime, and different regimes may exhibit different causal relations. Here, we assume a persistent and discrete regime variable leading to a finite number of regimes within which we may assume stationary causal relations. To allow for flexible linear and nonlinear, high-dimensional analysis settings, we utilize the constraint-based PCMCI causal discovery method, and combine it with a regime assigning linear optimisation, inspired by the regime learning in non-stationary Markov regression or clustering, to detect regime-dependent causal relations. Our method, Regime-PCMCI, is evaluated on a number of numerical experiments demonstrating that it can distinguish regimes with different causal directions, time lags, effects and sign of causal links, as well as changes in the variables' autocorrelation. Further, Regime-PCMCI is employed to observations of El Niño Southern Oscillation and Indian rainfall, demonstrating skill also in real-world data sets.

Keywords. causal discovery, time series, non-stationarity, regime-dependence, high dimensionality, climate research

I. INTRODUCTION

Understanding causal relationships among different processes is an ubiquitous task in many scientific disciplines as well as engineering (e.g., in the context of climate process[1], econometrics[2] or molecular dynamics[3]). Yet, the common approach to gaining causal knowledge by conducting experiments is often infeasible or unethical, for example in Earth sciences. All that is often given is a set of time series describing these processes with no specific knowledge about the direction and form of their causal relationships available. The challenge, termed causal discovery, is then to reconstruct the underlying graph of causal relationships from time series data[4]. Based on that graph the processes that generated the data can then be modelled in the framework of structural causal models (SCMs,[5]) to further understand causal relations, predict the effect of interventions, and for forecasting.

Today's ever-growing abundance of time series data sets promises many application scenarios for data-driven causal discovery methods, but many challenges emerging from the dynamic nature of such data sets have not yet been met. Further, causal knowledge cannot be gained from data alone and each method comes with its particular set of assumptions[6] about properties of the underlying processes. See[4] for an overview over methodological frameworks, challenges, and application scenarios.

A particular and wide-spread challenge is regime-dependence, a common property of nonlinear dynamical systems that can also be described as one form of non-stationary behaviour. Regime-dependence means that the causal relationships between the considered processes vary depending on some prevailing background regime that may be modelled as switching between different states. Further, often such regimes have strong persistence, that is, they operate and affect causal relations on much longer time scales than the causal relations among the individual processes. In the climate system, for instance, several cases of such regime-dependencies exist. For example, rainfall in India in summer is known to be influenced by the so-called El Niño Southern Oscillation (ENSO), an important mode of variability in the tropical Pacific affecting the large-scale atmospheric circulation and thereby weather patterns around the globe[7,8]. It is, however, generally assumed that ENSO does only marginally affect Indian rainfall in winter[9]. Thus, the causal relationships between ENSO and rainfall over India change dependent on the season that here defines the background-regime and operates on a longer time scale (several months) than the causal relations among ENSO and Indian rainfall (several weeks).

A. Existing work

Causal discovery has seen a steep rise with a plethora of novel approaches and methods in recent years. Each approach has different underlying assumptions and targets different real world challenges as discussed in[4]. In general, causal (network) discovery methods can be classified into classical Granger causality approaches[10,11], constraint-based causal network learning algorithms[6], score-based Bayesian network learning methods[12,13], structural causal models[3,14], and state-space reconstruction methods[15,16].

Here we focus on the constraint-based framework which has the advantage that it can flexibly account for nonlinear causal relations and different data-types (continuous and categorical, univariate and multivariate). PCMCI adapts this framework for the time series case yielding high detection power and controlled false positives also in high-dimensional and strongly autocorrelated time series settings. However, one of the general assumptions of PCMCI (as well as of other causal discovery algorithms) is stationarity, i.e., that at least the existence or absence of a causal link does not change over the considered time series[17]. While known changes in the background signal can be accounted for by restricting the time series to the stationary regimes, PCMCI cannot handle unknown background regimes.

Some recent work addresses causal discovery in the presence of non-stationarity. The authors in[18] model non-stationarity in the form of (continuous) stochastic trends in a linear autoregressive framework.[19] account for non-stationarity in the more general constraint-based framework. However, both address the case of a (smoothly) varying continuous background variable that continuously changes causal relations among the observed variables. This means that these methods will not output regime-dependent causal graphs, but a "summary" graph that accounts for regimes modelled as latent drivers. In[20,21] assumed known non-stationary regimes are exploited to estimate causal relations also in the presence of general latent confounders.

Currently few methods exist that address the case of a discrete regime variable leading to distinct causal regimes that may be physically interpreted. For example, in the climate science context, regime-dependent autoregressive models (RAM) were introduced already in 1990[22]. These can yield physically well interpretable results that, however, require well-chosen ancillary variables and a seasonal index which are not learned from data. Thus, RAM not only they requires a priori knowledge of the regimes, which one often aims to learn rather than enforce. Furthermore, the autoregressive framework only permits linear relationships. In the context of discrete state spaces regime dependent causal discovery has been considered in[3]. An another approach that has been proposed to model time dependent Granger (non) causality is based on Markov Switching VAR ansatz with a economics application in mind[2]. Specifically the regime assignments are computed by sampling from a Markov chain.

A more general framework to handle discrete regimes is the Markov-switching ansatz of[23], which flexibly models regime-dependence utilizing the assumption of a finite number of regimes and a level of persistency in the transitions between different regimes. The key underlying assumption is that that the system exhibits some form of persistence. This ansatz has been successfully realised in combination with many different model assumptions (e.g., see[24]) here we want to explore it for causal networks and combine it with PCMCI[25], a constraint-based time series causal discovery method[6]. We call our method Regime-PCMCI.

The remainder of the paper is structured as follows: First, in section I A we discuss existing methods for regime-dependent causal discovery. Second, in section II the underlying mathematical problem, concepts, and key assumptions are formalised, and a motivating example is discussed to provide some intuition. Our novel method Regime-PCMCI is then presented in section III. These theoretical and algorithmic parts are complemented by a thorough numerical investigation of the proposed method in various artificial settings in section IV. Finally, in section V, Regime-PCMCI is applied to a real-world data set from climate science, addressing the changing relationships of ENSO and rainfall over India.

II. PROBLEM SETTING

Let $\{X_t\}_{t\in\mathbb{Z}}$ be a sequence of real-valued N_X dimensional random variables $X_t \in \mathbb{R}^{N_X}$ where t is associated with time. A realisation over the time interval $[0,T]$ of this stochastic process is denoted $\{\mathbf{x}_t\}_{t\in[0,T]}$ and we assume that it is possible to obtain observations of these realisations. We assume that the underlying process is modelled by a regime-stationary discrete-time structural causal model (SCM)

$$X_t^j = g_t^j(\mathscr{P}_t^j, \eta_t^j) \quad \text{with } j = 1,\ldots,N_X\ . \tag{1}$$

Here the measurable functions g_t^j depend non-trivially on all their arguments, the noise variables η_t^j are jointly independent and are assumed to be stationary, i.e., $\eta_t^j \sim \mathscr{D}^j$ for all t for some distribution $\mathscr{D}$, and the sets $\mathscr{P}_t^j \subset (X_{t-1}, X_{t-2}, \ldots)$ define the causal parents of X_t^j. Here we assume lagged relationships, but this is not a necessity. In contrast to approaches assuming stationarity, both g_t^j and $\mathscr{P}_t^j$ are allowed to depend on regimes in time as further formalized in Assumption II B. Then the problem setting considered in this manuscript is of the nature of the following inverse problem

$$\mathbf{x}_t = \widehat{\mathbf{G}}_t\left(\mathbf{x}_{t-1},\ldots,\mathbf{x}_{t-\tau_{\max}};\Theta_t\right) \tag{2}$$

with $\widehat{\mathbf{G}}_t = [\widehat{g}_t^1,\ldots,\widehat{g}_t^{N_X}]$ where $\widehat{g}_t^j$ belong to an appropriate functions space for each t and i. In other words, the aim is to fit a set of unknown parameters Θ_t on the basis of an observed time series $\{\mathbf{x}_t\}_{t\in[0,T]}$. In the next section we will discuss the particular structure of the parameters Θ_t we are interested in.

A. Causal Graphs

Representing causal relations between different processes as graphs (also referred to as networks) is common practice in the context of causal inference and causal discovery[5,6]. For time series, we use the concept of time series graphs. The nodes in the time series graph associated with the SCM (1) are the individual time-dependent variables X_t^j with $j = 1,\ldots,N_X$

at each time $t \in \mathbb{Z}$. Variables $X^i_{t-\tau}$ and X^j_t for a time lag $\tau > 0$ and a given t are connected by a lag-specific directed link "$X^i_{t-\tau} \to X^j_t$" if $X^i_{t-\tau} \in \mathscr{P}^j_t$ for a particular t. We denote the maximum time lag of any parent as $\tau_{\max}$.

For a more detailed introduction the reader is referred to[25]. In the following we will use graphs and networks interchangeably.

The collection of parent sets for all components at time t is denoted $\mathscr{P}_t = \{\mathscr{P}^1_t, \dots, \mathscr{P}^{N_X}_t\}$. This set of parents is part of the unknown parameters we want to infer. Note that their dimensionality is assumed finite, but not known a priori. The other quantity of interest is the functional form of the causal relations $g^j_t(\mathscr{P}^j_t, \eta^j_t)$ in SCM (1) corresponding to these links which we here restrict to an appropriate function class as modelled in Eq. (2). If we assume linear functions with coefficients Φ^i_t, then the inverse problem Eq. (2) simplifies to

$$\mathbf{x}_t = \widehat{\mathbf{G}}_t(\mathscr{P}_t;\Phi_t) \tag{3}$$

Thus for a given time series $\mathbf{x}_t \in \mathbb{R}^N$ and with $t \in [0,T]$ and functional $\mathbf{G}_t$ the aim is to find the unknown parameters $\Theta_t = [\mathscr{P}_t, \Phi_t]$.

B. Persistence

As mentioned above, in many application areas non-stationarity may be modelled not in form of abrupt or continuous changes, but via piece-wise constant regimes[3,26,27]. These regimes will further exhibit a certain persistent behaviour. In order to capture non-stationary systems with these properties we will restrict our inference to regime-dependent persistent dynamics.

Assumption: *Denote the parents and functional dependency of a given variable j for a regime k as $\mathscr{P}^j_t = \mathscr{P}^j_k$ and $g^j_t(\mathscr{P}^j_t, \eta^j_t) = g^j_k(\mathscr{P}^j_k, \eta^j_t)$. We call a regime persistent if the parents and functional dependencies are stationary for an average of N_M consecutive time steps t. Further, we assume that there is a finite number of regimes on the whole time domain, i.e., $k \in \{1, \dots, N_K\}$.*

Note that persistence enters here via a regime average persistence N_M, which naturally implies a finite number of regimes $N_K \leq T/N_M$.

Under Assumption II B the considered linear inverse problem (3) reduces to finding a set of parameters

$$\{\mathscr{P}_1, \dots, \mathscr{P}_{N_K}, \Phi_1, \dots, \Phi_{N_K}\}.$$

and the change points between the regimes given by the regime assigning process

$$\Gamma(t) = [\gamma_1(t), \dots, \gamma_{N_K}(t)].$$

C. Motivating example

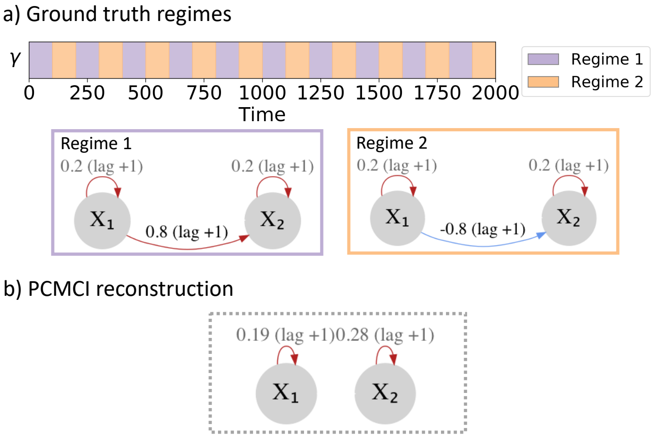

FIG. 1. (a) Regime dependent ground truth: regime-assigning process and regime-dependent networks. The links are labelled with the associated linear coefficient and the lag, $\Phi^j_k(i,\tau)$ (l+τ), and the sign of the coefficient is highlighted with the color (red for positive, blue for negative). (b) Network reconstruction with PCMCI fitted on the whole time series, i.e. links are assumed to be stationary.

Before we introduce our novel regime detecting causal discovery algorithm, we illustrate the underlying challenges of causal discovery in the face of regime-dependence by giving a simple example. Consider the case of two background regimes and two time-series X^1 and X^2 and the associated causal graphs as shown in Fig. 1. Variable X^1 linearly influences X^2 but the sign changes in time, alternating between a positive (during regime 1) and a negative (during regime 2) influence. Here the two regimes alternate equidistantly. The cross-correlation of X^1 and X^2 over the whole time-period is zero because the opposite sign effects cancel each other out in the linear regression. Thus, any linear causal discovery algorithms would fail in detecting the influence of X^1 on X^2 when no a priori knowledge on the two background regimes exists. For example, applying PCMCI on the whole time sample would give a network of disconnected variables (Figure 1, top-right).

In contrast, if the regimes are known and PCMCI is applied to samples from both regimes separately, the positive and negative links are correctly detected. To deal with such problems automatically, our algorithm needs to learn both the regimes as well as the regime-dependent causal relations.

III. METHOD

Our approach is designed to alternate between learning the regimes and the causal graphs for each regime in an iterative fashion. In principle, any causal discovery method that yields a causal graph can be used. Here we chose PCMCI[25] as a method that adapts the constraint-based causal discovery framework to the time series case.

A. Causal discovery

The constraint-based framework has the advantage that it can flexibly account for nonlinear causal relations and different data-types (continuous and categorical, univariate and multivariate) since it is based on conditional independence defined as follows.

Definition: *Two variables X and Y are conditionally independent given a (potentially multivariate) variable Z if*

$$X \perp\!\!\!\perp Y | Z \Leftrightarrow p(x,y|z) = p(x|z)p(y|z) \qquad (4)$$

where p denotes associated probability density functions.

There exist a large variety of conditional independence tests, see[25] for a discussion. If relationships are assumed linear, as is the case of the present work, a simple partial correlation can be used.

As is explained in detail in[25], PCMCI is based on a variant of the PC algorithm (names after its inventors Peter Spirtes and Clark Glymour[6]) combined with the momentary conditional independence (MCI) test. It consists of two stages: (i) PC_1 condition selection to identify relevant conditions $\widehat{\mathscr{P}}_t^j$ for all time series variables X_t^j and (ii) the MCI test to test whether $X^i_{t-\tau} \to X^j_t$ with

$$\text{MCI:} \quad X^i_{t-\tau} \not\perp\!\!\!\perp X^j_t \mid \widehat{\mathscr{P}}^j_t \setminus \{X^i_{t-\tau}\}, \widehat{\mathscr{P}}^i_{t-\tau}. \qquad (5)$$

Thus, MCI conditions on both the parents of X_t^j and the time-shifted parents of $X^i_{t-\tau}$. These two stages serve the following purposes: PC_1 is a Markov set discovery algorithm based on the PC-stable algorithm[28] that removes irrelevant conditions for each variable by iterative conditional independence testing. A liberal significance level α_{PC} in the tests lets PC_1 adaptively converge to typically only few relevant conditions that include the causal parents with high probability, but might also include some false positives. The MCI test then addresses false positive control for the highly-interdependent time series case, which is why we chose it here. A causal interpretation of the relationships estimated with PCMCI comes from the standard assumptions in the constraint-based framework[6,17], namely causal sufficiency, the Causal Markov condition, Faithfulness, non-contemporaneous effects, and stationarity within the regimes as further discussed below. As demonstrated in[25], PCMCI has high detection power and controlled false positives also in high-dimensional and strongly autocorrelated time series settings.

The main free parameter of PCMCI are the chosen conditional independence test, the significance levels α in MCI and α_{PC} in PC_1, the latter should be regarded as a hyper-parameter and can be chosen based on model-selection criteria such as the Akaike Information Criterion (AIC)[29] or cross-validation.

PCMCI is applied to sample subsets of the time series pertaining to an estimated regime k in an iterative step of our method. Given a significance level α, the output of PCMCI is a the set of parents $\mathscr{P}_k$ for all time series variables for that regime.

B. Regime learning

Given an estimated set of parents $\mathscr{P}_k$, the regimes are learned assuming a particular non-stationary setting of finite metastable regimes as defined in Assumption II B. This learning approach is based on ideas first proposed in[24] and later extended to many different models[23].

In the following we focus on the linear setting. To learn the regime parameters for the inverse problem (3) introduced in Section II,

$$\Phi_k \text{ for every } k \in \{1, \ldots, N_K\} \qquad (6)$$

given $\mathscr{P}_k$ (the output from PCMCI), we define a cost functional

$$\mathbf{L}(\Gamma, \mathscr{P}) = \sum_{t=0}^{T} \sum_{k=1}^{N_K} \gamma_k(t) d(\mathbf{x}_t - \mathbf{G}_t(\mathscr{P}_k; \Phi_k)) \qquad (7)$$

subject to constraints

$$\sum_{k=1}^{N_K} \gamma_k(t) = 1 \quad \forall\, t, \text{ with } \gamma_k(t) \in [0,1] \qquad (8)$$

and

$$\sum_{t=1}^{T-1} |\gamma_k(t+1) - \gamma_k(t)| \leq N_C \quad \forall k \qquad (9)$$

where d is a distance measure such as the squared euclidean distance $\|\cdot\|_2$ and γ is a regime assigning process describing the weight of the individual networks at each time t.

The format of $\mathbf{L}(\Gamma, \mathscr{P})$ relies on the assumption that the system associated with the considered data exhibits metastability in time (see Assumption II B, that translates in the summation over k). Note that the persistence enters the functional in form of a regularization (see Constraint 9). An alternative option is to add a regularisation term that enforces so form of smoothness of Γ (e.g., Tikhonov regularisation[30]).

Here we choose the free tuning parameter N_C so that the average time to be in a regime is approximately N_M. For $K = 2$ for example, it has to be approximately $N_C \approx T/N_M$. Note that in practice, the average regime switching time N_M might not be exactly known. However, we expect in many application areas that prior domain knowledge on reasonable time scales of regime switching is available. The tuning of parameters, including choices of value N_K, will be discussed in Section IV B 3.

C. Pseudocode

The Regime-PCMCI algorithm iterates over two key estimation steps where q indicates the current iteration. Note that (q) is added as a superscript combined with brackets to the variables changing with each loop. The details of the consecutive subroutines are laid out below.

1. Step 1: Causal discovery for learning the parents

The first step is to find a set of parents $\{\mathscr{P}_k\}^{(q+1)}$ and coefficients $\{\Phi_k\}^{(q+1)}$ with $k \in \{1,\ldots,K\}$ on the basis of a fixed $\{\Gamma(t)\}^{(q)}$ obtained in step 2 of the previous iteration (see lines of Algorithm 1 and Section III C 2). The coefficients $\{\mathscr{P}_k\}^{(q+1)}$ and $\{\phi_k\}^{(q+1)}$ are estimated on the basis of a subset of the time series $\mathbf{x}_t$ with

$$t \in \{\Upsilon_k\}^{(q)} := \Big\{t : \{\gamma_k(t)\}^{(q)} \geq 0.5\Big\} \qquad (10)$$

for each regime k. The regime dependent parents set $\mathscr{P}_k$ is computed via PCMCI. Here we choose partial correlation as a conditional independence test and the PCMCI hyperparameter $\alpha_{\mathrm{PC}} = 0.2$ as recommended in[31]. Further, we consider parents that are significant at $\alpha = 0.01$ (for $N_X = 2$) and $\alpha = 0.05$ (for $N_X = 15$).

a. Fit of coefficients To obtain the reconstructed time-series $\hat{\mathbf{x}}_t$ an estimate of the coefficients Φ_k characterising the assumed functional relationship $\mathbf{G}$ between each variables and its detected predictors $\mathscr{P}_k$ (for each regime) has to be computed. Here we assume that the functionals g_k^j are linear which yields that the coefficients can be estimated via an appropriate regression that assumes the following holds for each fixed k:

$$x_t^j = \sum_{(i,\tau)\in\mathscr{P}_k^j} \{\Phi_k^j(i,\tau)\}^{(q)} x_{t-\tau}^i + \varepsilon_t^j \qquad (11)$$

for $t \in \{\Upsilon_k\}^{(q)}$. In other words for every $k \in \{1,\ldots,N_K\}$ the following optimisation has to be solved

$$\{\Phi_k^j(i,\tau)\}^{(q)} = \arg\min \left\| x_t^j - \sum_{(i,\tau)\in\mathscr{P}_k^j} \{\Phi_k^j(i,\tau)\}^{(q)} x_{t-\tau}^i \right\|_2^2 \qquad (12)$$

for $t \in \{\Upsilon_k\}^{(q)}$. Note that the coefficients not indicated as relevant via the parent set are defined to be zero, i.e., $\Phi_k^j(\tau,i) := 0$ for $(\tau,i) \notin \mathscr{P}_k^j$.

2. Step 2: Regime learning

Step 2 is to determine an optimal regime assigning process $\{\Gamma_t\}^{(q+1)} \in [0,1]^{N_K\times T}$ given the current estimates $\{\mathscr{P}_k\}^{(q)}$ for the parents and $\{\phi_k\}^{(q)}$ coefficients (see lines of Algorithm 1). For this the following optimisation problem needs to be solved

$$\{\Gamma_t\}^{(q+1)} = \arg\min \sum_{k=1}^{N_K}\sum_{t=1}^{T} \gamma_k(t) \left\| \mathbf{x}_t - \{\hat{\mathbf{x}}_{k,t}\}^{(q)} \right\|_2^2 \qquad (13)$$

subject to the constraints (8) and (9), and where

$$\hat{x}_{k,t}^j = \sum_{(i,\tau)\in\mathscr{P}_k^j} \Phi_k^j(i,\tau) x_{k,t-\tau}^i \quad \text{for } t \in \{1,\ldots,T\}. \qquad (14)$$

Since the first $\tau_{\max}$ time steps cannot be predicted, we choose to set those to $\hat{x}_{k,t}^j = x_{k,t}^j$ and to not consider this portion of the time series in the algorithm evaluation.

In order to search for the global minimum, the algorithm is run for a number N_A of different initializations of $\{\Gamma\}^{(0)}$ (annealing). The annealing run with the lowest cost functional objective is chosen as optimal fit. Note that the individual annealing steps are *embarrassingly parallelizable*.

Algorithm 1 Method

```
Input:
  • time series $\mathbf{x}_t \in \mathbb{R}^{N_X}$ with $t \in \{1,\ldots,T\}$
  • Set parameters:
      – number of assumed regimes $N_K$
      – maximum number of transitions within a single regime $N_C$

      – maximal lag $\tau_{\max}$ for each regime
      – functional model $\mathbf{G}$
      – conditional independence measure according to $\mathbf{G}$ (e.g. partial correlation $\rho$ for linear $\mathbf{G}$)
      – significance level $\alpha$
      – type of masking ‘$y$’

      – annealing steps $N_A$
      – number of optimisation iterations $N_Q$

for $a = 0 : N_A$ do
    Initialize random $\{\Gamma\}^{(0)} \in [0,1]^{N_K\times T}$
    for $q = 0 : N_Q$ do
        Fit network:
        • Infer parents $\{\mathscr{P}_k\}^{(q)}$ by means of PCMCI run on subset $\Big\{\mathbf{x}_t : t \in \{\Upsilon_k\}^{(q)}\Big\}$ for each $k$
        • Fit model coefficients $\{\Phi_k\}^{(q)}$ via (12) for each $k$, and use them to generate $k$ reconstructed time-series $\{\hat{\mathbf{x}}_{k,t}\}^{(q)}$ defined for every $t \in \{1,\ldots,T\}$ according to (14).
        Fit regime assigning process:
        • Update $\{\Gamma\}^{(q+1)}$ solving 13.
          Break if $\{\Gamma\}^{(q+1)} = \{\Gamma\}^{(q)}$ (a local or global minimum is reached)
    end for
end for

Output:
  • $\Gamma = [\gamma_1(t),\ldots,\gamma_{N_K}(t)]^\dagger \in [0,1]^{N_K\times T}$
  • causal parents $\mathscr{P}_k$ and causal effect $\Phi_k$ for every $k \in \{1,\ldots,N_K\}$
    return
```

D. Reconstruction of time series

A prediction from Eq 14 can be build as the weighted sum over k

$$\hat{x}_t^{*j} = \sum_{k=1}^{N_K} \lceil \gamma_k(t) \rfloor \hat{x}_{k,t}^j \quad \text{for } t \in \{1,\ldots,T\}. \qquad (15)$$

But note this is never used in the code (only 14 via its presence in 13 is used).

E. Consideration on nonlinearity

It is important to mention that the choice of functions g_k^j in the learning problem (2) should be determined according to the considered applications and on assumptions on the data. Further, the conditional independence test used in PCMCI should cover at least an equally expressive functional dependency class. For example, if Gaussian processes are used to estimate g_k^j, then the Gaussian Process Distance Correlation (GPDC) test (see[25]) can be used.

Consequently, a nonlinear version of the presented Regime-PCMCI would require a different cost functional. The complexity of the assumed model would increase significantly due to the two-fold presence of non-linearity (one through the regime-dependence and the other one via nonlinear causal relations). Therefore, we here restricted ourselves to linear functions g_k^j. Addressing nonlinearity in combination with the considered non-stationarity will be explored in subsequent research.

IV. NUMERICAL INVESTIGATION

In the following we investigate the performance of Regime-PCMCI by means of several toy examples. The artificial data is designed to test the methods robustness and accuracy with respect to various potential scenarios that could occur in real applications. At first low dimensional ($N_X = 2$) causal relations are studied as the results can be interpreted more easily. Next, we also consider higher dimensional settings ($N_X = 10$). The reference time series are generated with the following variant of SCMs:

$$\begin{aligned} x_t^j &= \sum_{k=1}^{K} \{\gamma_k(t)\}^{\text{ref}} \sum_{(\tau,i)\in\mathscr{P}_k^j} \{\Phi_k^j(i,\tau)\}^{\text{ref}} x_{t-\tau}^i + \{\varepsilon_t^j\}, \\ \varepsilon_t^j &\sim \mathscr{N}(0,\{\Sigma\}^{\text{ref}}) \end{aligned} \tag{16}$$

with $x_t^j = \varepsilon_t^j$ and predefined $\{\Gamma(t)\}^{\text{ref}}$, $\{\Phi_k\}^{\text{ref}}$, $\{\Xi\}^{\text{ref}}$ and $\{\Sigma\}^{\text{ref}}$. Note that the reference set of parents is specified by the non-zero coefficients $\{\Phi_k^j(i,\tau)\}^{\text{ref}}$ defining the causal child-parent links $(i,\tau) \in \{\mathscr{P}_k^j\}^{\text{ref}}$.

A. Low dimensional data with two underlying regimes

First we focus on a simple setting of two regimes, i.e. $N_K = 2$, and a two dimensional underlying process $X_t \in \mathbb{R}^2$ (i.e., $N_X = 2$). Our aim is to test the performance of Regime-PCMCI for different elemental features that can change between regimes. For brevity, links $X_{t-\tau}^i \to X_t^i$ will be called auto links and $X_{t-\tau}^i \to X_t^j$ cross links. We consider the following scenarios as summarised in Table I: sign change of coefficient (in auto link and cross variables link), lag change (in cross link), coefficient change (in auto link) and child-parent inversion defined via an assortment of linear functions and associated coefficients. In all examples, each variable is also auto-linked at lag 1, which is a realistic yet challenging assumption for many algorithms. At first we will describe the specific design of the toy data sets and the settings of the algorithms runs. Then the results obtained via the proposed methods are going to be compared to the reference values.

1. *Experiment settings*

We design five toy models, in network terms, corresponding to different sets of parents defined via the references parameters $\{\Phi_k^j(i,\tau)\}^{\text{ref}}$ given in columns 4 to 5 of Table I. Further, synthetic regime assigning processes $\{\Gamma(t)\}^{\text{ref}}$ are generated for all examples. More specifically, $\{\gamma_1(t)\}^{\text{ref}}$ is designed to consist of 41 alternating windows, i.e., $\{N_C\}^{\text{ref}} = 40$ regime transitions. The regime assignment is indicated by setting it to 1 (active regime) and 0 (inactive regime). The length of these windows is randomly selected to be between 70 and 100 and the constraint (8) imposes $\{\gamma_2(t)\}^{\text{ref}} = 1 - \{\gamma_1(t)\}^{\text{ref}}$. The final length of the time series is capped at $T = 3,000$ to ensure equally-long regime assignment time-series.

Then an artificial time series $\{\mathbf{x}_t\}^{\text{ref}}$ via (16) with $\{\Sigma_{jj}\}^{\text{ref}} = 1$ can be generated. Note that the stochastic process (16) can be exactly reconstructed via the coefficients $\{\Phi_k^j(i,\tau)\}^{\text{ref}}$, their activation $\{\Gamma(t)\}^{\text{ref}}$ and a specific realisation of the innovation term $\{\varepsilon_t^j\}^{\text{ref}}$.

The PCMCI parameters are chosen as follows: $\alpha = 0.05$, PC-$\alpha = 0.2$, $\tau_{\max} = 3$ and masking type 'y' (see the documentation of `tigramite` for the definition of masking types). The regime parameter is $N_K = 2$ and the max regime transitions $N_C = 40$, i.e., correct guess on number of regimes and switches (model selection in terms N_K is discussed in Section IV B 3). The number of iterations is $N_Q = 20$ and the annealing are $N_A = 50$. A summary of the parameters is shown in Table II. The reconstructed time-series is generated via (15) where coefficients $\{\Phi_k\}$ are estimated for each variable via multiple linear regression on the detected parents.

The random procedure of generating $\gamma_k(t)$, and the associated its $\mathbf{x}_t$, is repeated $N_R = 100$ times for each example, thus testing the algorithm on a variety of data for each family of time series (robustness to data).

2. *Results*

The ability of the proposed method to recover the networks and the path assigning process on the basis of the artificially designed time series are presented in the following. Figures 2-6 present results for each case in Table I, focusing on one of the N_R synthetic data sets. Table V shows statistics from the N_R runs.

The case *sign* X^1X^2 is discussed in detail. The ground-truth regime evolution and networks are shown in the top part of

example	$k=1$	$k=2$	$\{\Phi_1^j(i,\tau)\}^{\text{ref}}$	$\{\Phi_2^j(i,\tau)\}^{\text{ref}}$
arrow direction	$X^1 \rightarrow X^2$	$X^1 \leftarrow X^2$	$\{\Phi_1^2(1,1)\}^{\text{ref}} = 0.8$ $\{\Phi_1^1(1,1)\}^{\text{ref}} = 0.2$ $\{\Phi_1^2(2,1)\}^{\text{ref}} = 0.2$	$\{\Phi_2^1(2,1)\}^{\text{ref}} = 0.8$ $\{\Phi_2^1(1,1)\}^{\text{ref}} = 0.2$ $\{\Phi_2^2(2,1)\}^{\text{ref}} = 0.2$
causal effect	$X^1 \xrightarrow{\lvert a\rvert} X^1$	$X^1 \xrightarrow{\lvert b\rvert} X^1$	$\{\Phi_1^1(1,1)\}^{\text{ref}} = 0.8$ $\{\Phi_1^2(2,1)\}^{\text{ref}} = 0.4$	$\{\Phi_2^1(1,1)\}^{\text{ref}} = 0.1$ $\{\Phi_2^2(2,1)\}^{\text{ref}} = 0.4$
lag	$X^1 \xrightarrow{\tau=1} X^2$	$X^1 \xrightarrow{\tau=2} X^2$	$\{\Phi_1^2(1,1)\}^{\text{ref}} = 0.8$ $\{\Phi_1^1(1,1)\}^{\text{ref}} = 0.2$ $\{\Phi_1^2(2,1)\}^{\text{ref}} = 0.2$	$\{\Phi_2^2(1,2)\}^{\text{ref}} = 0.8$ $\{\Phi_2^1(1,1)\}^{\text{ref}} = 0.2$ $\{\Phi_2^2(2,1)\}^{\text{ref}} = 0.2$
sign X^1	$X^1 \xrightarrow{\lvert a\rvert} X^1$	$X^1 \xrightarrow{-\lvert a\rvert} X^1$	$\{\Phi_1^1(1,1)\}^{\text{ref}} = 0.8$ $\{\Phi_1^2(2,1)\}^{\text{ref}} = 0.2$	$\{\Phi_2^1(1,1)\}^{\text{ref}} = -0.8$ $\{\Phi_2^2(2,1)\}^{\text{ref}} = 0.2$
sign X^1X^2	$X^1 \xrightarrow{\lvert a\rvert} X^2$	$X^1 \xrightarrow{-\lvert a\rvert} X^2$	$\{\Phi_1^2(1,1)\}^{\text{ref}} = 0.8$ $\{\Phi_1^1(1,1)\}^{\text{ref}} = 0.2$ $\{\Phi_1^2(2,1)\}^{\text{ref}} = 0.2$	$\{\Phi_2^2(1,1)\}^{\text{ref}} = -0.8$ $\{\Phi_2^1(1,1)\}^{\text{ref}} = 0.2$ $\{\Phi_2^2(2,1)\}^{\text{ref}} = 0.2$

TABLE I. Artificial model configurations for different experiments.

N_K	N_C	α	$\tau_{\max}$	mask	N_Q	N_A	N_R
2	40	0.01	3	‘y’	20	50	100

TABLE II. Algorithm setting for PCMCI for runs on low dimensional data with two underlying regimes.

panels *a* and *b* in Figure 2; in the middle part of both panels their Regime-PCMCI reconstruction is shown; in the bottom part the difference between true and reconstructed regimes are presented to visually inspect the accuracy. The reconstructed regime assigning process for each regime matches the truth in 99.6% of time steps (97% average value over N_R). The corresponding networks have all and only the correct links (average network TPR = 0.99 and FPR = 0.01); their linear causal effect is also well estimated with each link correct up to ±0.02 (average error per link is 0.028 (9%)).

The other four cases are presented in Figures 3-6. The *causal effect* example results being the hardest to detect, as it is further exposed by the average over the N_R data sets. The reason if probably traceable to weak identifiability of the model based on the data. In *causal effect* - and in *lag change* to a lesser extent - the difference between the individual regimes and a mixed state of the two is not so dramatic and thus the identification results harder. This adds to the general challenge of non-convexity of the functional we are optimising for, that we mitigate for via the annealing steps. A similar challenge of identifiability is found for some high dimensional runs for which we refer to Section IV C.

Table V records the regime-averaged result for each case after repeating the Regime-PCMCI discovery for N_R different ground-truth regime evolution. The estimation errors are presented in terms of regime assigning process (second column), network structure (third to sixth) and causal effect of links (last three columns). The second column, $\Delta\gamma\%$, is the average percentage of wrongly estimated time steps per regime (the lower the better, note that this value is the same for $k = 1,2$ by construction). In terms of networks, the performance in link detection is evaluated via the true positive rates (TPR) and false positive rates (FPR) and compared with the reference value, i.e. what is obtained if PCMCI is run on the ground-truth regime data (superscript *ref*). The accuracy in links' causal effects is assessed via $\Delta\Phi$, the average difference between the reconstructed linear coefficient and the reference values of the ground truth links. $\Delta\Phi$ is also expressed as percentage, i.e. each difference is weighted by the absolute value of the ground truth coefficient. The last column, $\hat{\varepsilon}$, is the expected prediction error per variable and per time step and is computed as $\hat{\varepsilon} = \sqrt{\mathbf{L}/N_X T}$ with $\mathbf{L}$ defined in Eq 7 and with N_X and T referring to the number of variables, here two, and the length of the time series respectively. The precise definition of all the above statistics can be found in the Appendix.

In summary, Table V shows that :

- $\Delta\gamma\%$: on average, the regime assigning process is reconstructed correctly in $\sim 94\%$ of the time steps for all cases except *causal effect*. The *causal effect* and *lag* examples are the hardest to infer, with causal effect being particularly deficient. In these examples a mixed-regime state (e.g. arising from assigning a considerable fraction of wrong time steps to a regime) is still quite close to any of the true the case. Therefore the algorithm might struggle to decide which time steps belong to which regime, since they could fit both to some degree. Yet, there are 7 instances where $\Delta\gamma < 15\%$ (one presented in Figure 6) and those, as expected from PCMCI, give very good network fit. We notice that these runs do not correspond to the lowest objective values of the N_R set (i.e. better fit) which proves that runs that end up in mixed states can still fit the data quite

example	$\Delta\gamma\%$	TPR_{all}	$\text{TPR}^{\text{ref}}_{\text{all}}$	FPR_{all}	$\text{FPR}^{\text{ref}}_{\text{all}}$	$\Delta\Phi$	$\Delta\Phi^{\text{ref}}$	$\Delta\Phi$ %	$\Delta\Phi^{\text{ref}}$ %	$\hat{\varepsilon}$
arrow direction	3.0	1.0	1.0	0.02	0.01	0.021	0.020	7.0	7.0	0.76
causal effect	43.0	0.81	0.98	0.11	0.01	0.286	0.020	120.0	10.0	0.68
lag	6.0	0.98	1.0	0.04	0.01	0.027	0.018	11.0	8.0	0.68
sign X^1	4.0	0.98	1.0	0.03	0.01	0.033	0.016	10.0	6.0	0.65
sign X^1X^2	3.0	0.99	1.0	0.01	0.01	0.028	0.019	9.0	7.0	0.75

TABLE III. Results from $N_R = 100$ data generated per each examples described in Table I. Mean is taken over the $N_K = 2$ regimes and over N_R.

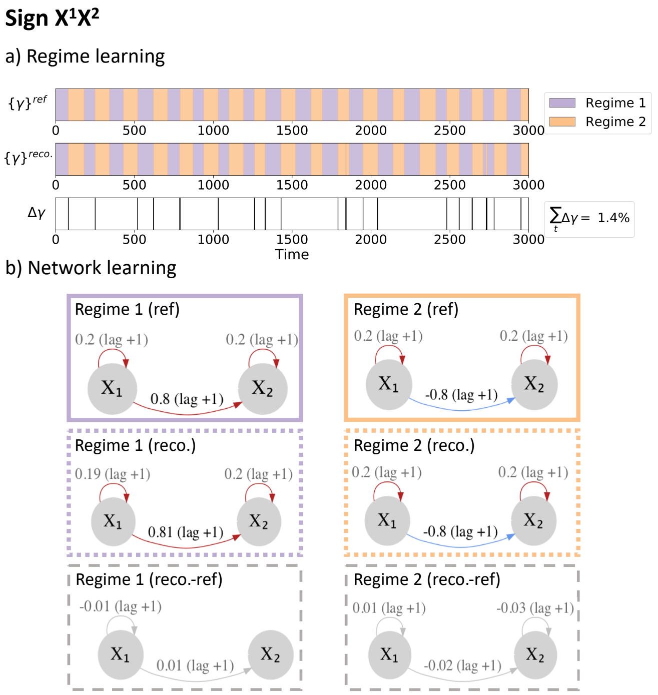

FIG. 2. *Sign* X^1X^2. (a) The ground-truth regime-assigning process, $\{\gamma\}^{ref}$ (top), the Regime-PCMCI reconstructed process, $\{\gamma\}^{reco.}$ (middle) and the difference between the two, $\Delta\gamma$ (bottom). (b) The ground-truth networks for each regime (top), the Regime-PCMCI reconstructed networks (middle) and the difference between the two (bottom). The links are labelled with the associated linear coefficient and the lag, $\Phi^j_k(i,\tau)$ (l+τ), and the sign of the coefficient is highlighted with the color (red for positive, blue for negative).

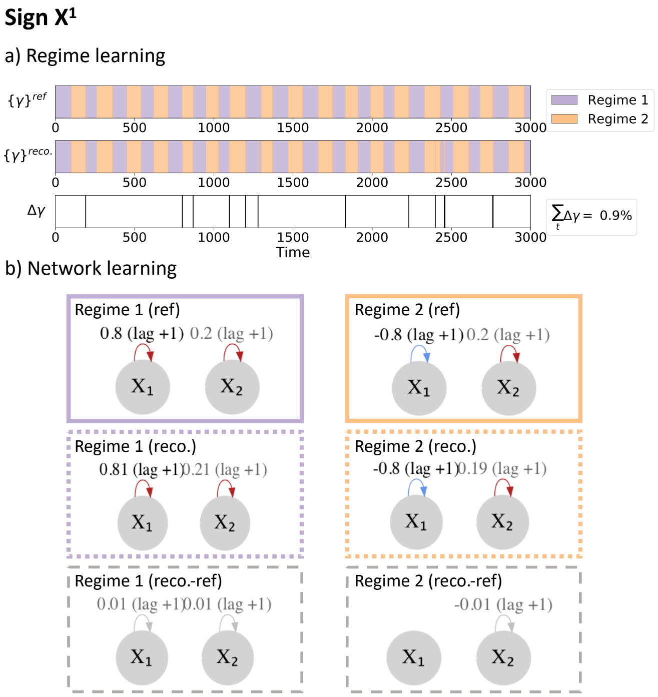

FIG. 3. *Sign* X^1. Constructed as Figure 2.

well. Also, we notice that causal effect setup reaches local minima in 16% of the 100 runs, thus in most of the runs the algorithm cannot easily find a stable solution and points at a weaker confidence in the output.

- TPR: despite some errors in reconstructing the regime assigning process, the TPR is always very close to 1. This can indicate that the true signals, dynamic wise, are strong enough to be detectable.

- FPR: Ideally the false positive rate should be upper-bounded by α. This is also the case if we assume the correct regimes (see columns FPR^{ref}). However, if the regimes are learned, in most of the examples the FPR value is higher due to errors in learning the regimes. If a wrong regime is learned, then both false positives and false negatives can occur. False negatives, i.e., missing links in the PC_1 step of PCMCI can lead to false positives in the MCI step.

- $\Delta\Phi$%: Errors in parents' detection (either due to false positives (FPR) or to false negatives (missed links, FNR = 1-TPR)) surely impact the estimation of link effects. Since the TPR and FPR are good, except for causal effects, we expect to obtain also good results for the linear coefficients. This is indeed the case, as the difference in each entry is order 10^{-2}. Put in the context of the true coefficients, the relative error is of about 10%.

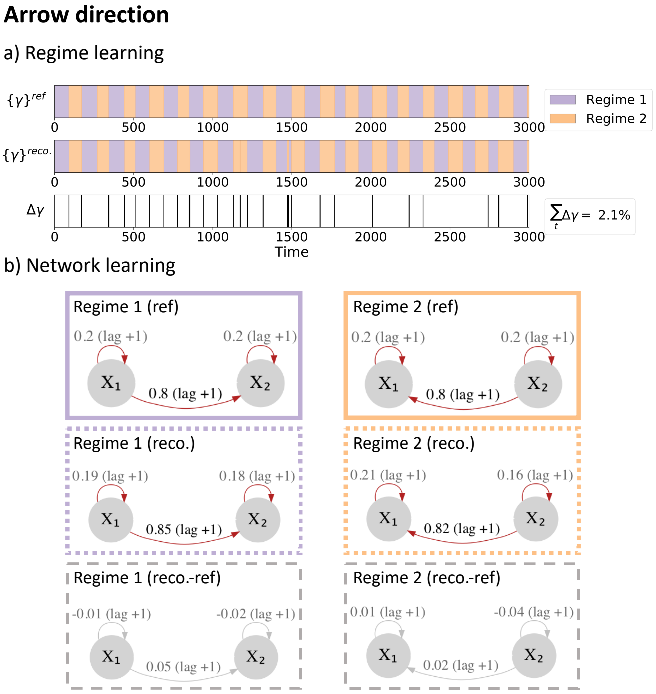

FIG. 4. *Arrow direction*. Constructed as Figure 2.

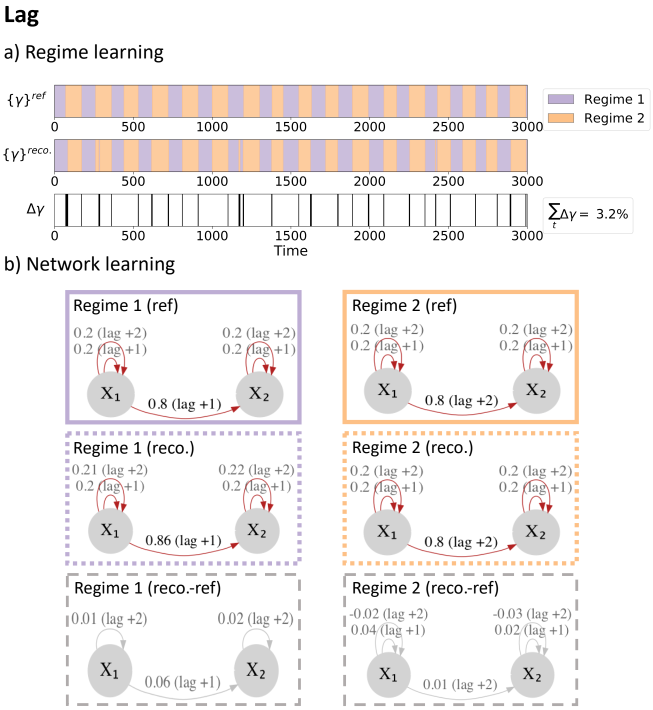

FIG. 5. *Lag*. Constructed as Figure 2.

B. Low dimensional data with three underlying regimes

In order to illustrate how the algorithm deals with more than two regimes we also considered a toy time series with based on 3 different causal regimes. It is of course possible to consider the case $N_K > 3$, yet in most application it is often desirable to infer a few prominent and very relevant regimes rather than having too many that are not interpretable anymore. In other words the aim is to avoid overfitting and to increase the information gain by reducing the complexity of the assumed model.

1. Experiment settings

The artificial time series is generated via a regime dependent causal graph that is designed by combining two of the regimes settings we had in IV A, namely *sign* X^1X^2 change and *arrow inversion* (for details see Table IV). The regime assigning references process $\{\Gamma\}^{\text{ref}}$ is generate by randomly choosing between different persistence lengths 60, 70 and 80 and iteration over it for 20 times.

2. Results

In Figure 7 one can see the obtained regime assignment and coefficient ϕ_k compared to the values used to generate the data. There are only minimal deviations from the references values which confirms that the proposed method is capable to deal with $N_K > 2$. This is also projected in the averaged results obtained for $N_R = 100$ runs presented in Table V. Yet it is important to note that we chose a combination of causal graphs that performed well for $N_K = 2$, i.e, causal effect changes would also be difficult to detect for $N_K = 3$.

3. Model selection

Determining a suitable choice of the unknown number of regimes N_K is a difficult task. In particular it is hard to find the right balance between avoiding to overfit and to choose appropriately complex models to describe a specific data set and thus the underlying dynamics well. One way to assess this balance heuristically is to employ an information criterion (IC)[32] which has been derived in the context of regression models and since been adapted to various other model scenarios including graphs[33]. An IC is designed to capture the goodness of fit penalised by the number of parameters in order to prefer models with as few parameters as possibles, to avoid over fitting. Here the number of parameters is defined via

$$N_{\text{para}} = (N_K - 1)N_C + \sum_{k=1}^{N_K}\sum_{j=1}^{N_X} |\mathscr{P}_k^j|. \qquad (17)$$

The first term in (17) relates to the number of parameters required to describe Γ which can be fully determined via the change points. The second term in (17) counts the number of relevant parents; in other words the non-zero coefficients $\Phi_k^j(i,\tau)$. Due to the fact that only the links are counted towards the number of parameters a higher number of regimes N_K does not necessarily result in an increase of the total num-

example	$k=1$	$k=2$	$k=3$	$\Phi_1^j(i,\tau)^{\text{ref}}$	$\{\Phi_2^j(i,\tau)\}^{\text{ref}}$	$\{\Phi_3^j(i,\tau)\}^{\text{ref}}$
sign X^1X^2	$X^1 \xrightarrow{\lvert a\rvert} X^2$	$X^1 \xrightarrow{-\lvert a\rvert} X^2$	$X^2 \xrightarrow{\lvert a\rvert} X^1$	$\{\Phi_1^2(1,1)\}^{\text{ref}}=0.8$	$\{\Phi_2^2(1,1)\}^{\text{ref}}=-0.8$	$\{\Phi_3^1(2,1)\}^{\text{ref}}=0.8$
and *arrow*				$\{\Phi_1^1(1,1)\}^{\text{ref}}=0.2$	$\{\Phi_2^1(1,1)\}^{\text{ref}}=0.2$	$\{\Phi_3^1(1,1)\}^{\text{ref}}=0.2$
direction				$\{\Phi_1^2(2,1)\}^{\text{ref}}=0.2$	$\{\Phi_2^2(2,1)\}^{\text{ref}}=0.2$	$\{\Phi_3^2(2,1)\}^{\text{ref}}=0.2$

TABLE IV. Artificial model configuration for an example of $N_K=3$

$\Delta\gamma\%$	TPR_{all}	$\text{TPR}_{\text{all}}^{\text{ref}}$	FPR_{all}	$\text{FPR}_{\text{all}}^{\text{ref}}$	$\Delta\Phi$	$\Delta\Phi^{\text{ref}}$	$\Delta\Phi$ %	$\Delta\Phi^{\text{ref}}$ %	$\hat{\varepsilon}$
4.0	0.98	1.0	0.05	0.01	0.033	0.020	10.0	7.0	0.5

TABLE V. Results from $N_R=100$ data generated per each examples described in Table IV. Mean is taken over the $N_K=3$ regimes and over N_R.

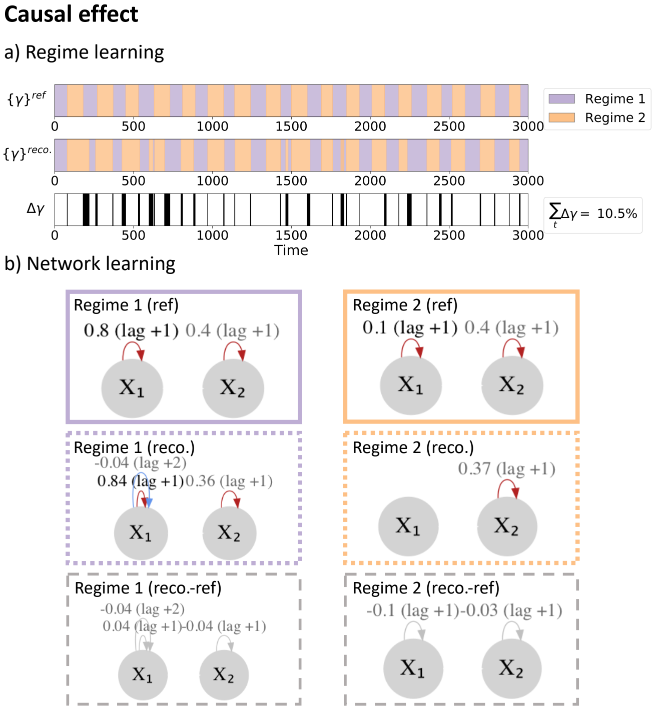

FIG. 6. *Causal effect*. Constructed as Figure 2.

ber of parameters. Further it is important to mention that the objective value (7) decreases for increasing N_C. This effect is natural to optimisation procedures which unless forced via a constraint such as the persistency (see (8) and (9)) or restricted number of parameters prefer the best fit in terms of the underlying cost function. Due to this fact we weight $N_C=\{N_C^{\text{ref}}\}$ with $\{N_K\}^{\text{ref}}/N_K$ for $N_K>\{N_K\}^{\text{ref}}$ while we consider model selection with respect to N_K. Here we will use the corrected Akaike Information criterion (AICc) first proposed in[34] to estimate N_K. Note that we use the corrected version of the original AIC[29] to correct for small samples sizes relative to the number of parameters

$$AICc = -2\log(\mathscr{L}) + 2N_{\text{para}} + \frac{2N_{\text{para}}(N_{\text{para}}+1)}{N_T - N_{\text{para}} - 1}$$

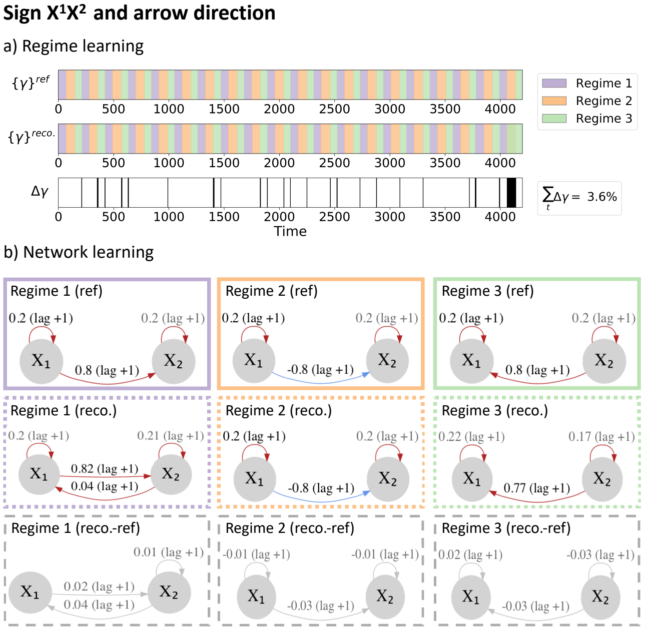

FIG. 7. $N_k=3$ case, *Sign* X^1X^2 *and arrow direction*. Constructed as Figure 2 but with 3 regimes.

and $\mathscr{L}$ is the maximum value of the likelihood function for the model one assumes for the residuals (see[35] for a more detail discussion). The resulting AICc values for two test scenarios, $\{N_K=\}^{\text{ref}}=2,3$, with $N_R=29$, $N_Q=20$ and $N_A=20$ are displayed in Figure 8. We note that the lowest N_K at which the *AICc* plateaus is the ground-truth one.

C. High dimensional linear network

In this section the algorithm is briefly tested on high-dimensional data sets, with each dataset consisting of $N_X=10$ interacting variables. The background regimes are generated with two regular alternating regimes of 300 time steps each, for a total length $T=15,000$. The networks' structured are randomly generated from a family of linear networks defined via the parameters shown in Table VI, where L is the number

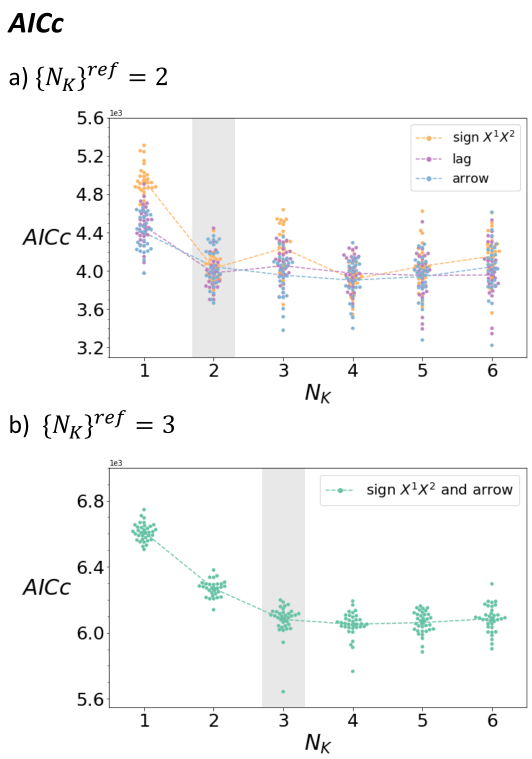

FIG. 8. *AICc* values for runs with different N_K values and (a) $\{N_K\}^{\text{ref}} = 2$ for three networks examples (*sign* X^1X^2*, arrow* and *lag* change) and (b) $\{N_K\}^{\text{ref}} = 3$ for the *sign* X^1X^2 *and arrow* change example. In each example, individual dots represent the value attained by the $N_R = 29$ runs, and the dashed line goes through the mean values of each set. The vertical grey bar highlights the ground-truth number of regimes $\{N_K\}^{\text{ref}}$.

of randomly drawn cross variable links with random coefficients from the fouth column. Note that each variable is also auto-linked at lag 1 with coefficient randomly drawn from the second column. The time series $\mathbf{x}_t \in \text{R}^{10}$ is generated following the Eq (16). The procedure is repeated for $N_R = 70$ times to ensure robustness of the results to data.

N	L	$\Phi^j_k(i,\tau)$	$\Phi^i_k(i,\tau)$	max lag
10	30	[-0.4, 0.4]	[0.2,0.5, 0.9]	3

TABLE VI. High dimensional network parameters

Finally, Regime-PCMCI is run with the setting shown in Table VII

N_K	N_C	α	$\tau_{\max}$	mask	N_Q	N_A	N_R
2	49	0.05	4	‘y’	30	50	70

TABLE VII. Algorithm setting for runs on high dimensional data with two underlying regimes.

The results are shown in Table VIII, which is structured as Table V. Regime-PCMCI performs very well even in this challenging setting. Notably, individual runs can perform extremely well, with $\Delta\gamma$ reaching as low as 0.02%, and a total of 53 runs below total average of $\Delta\gamma = 11.7\%$ (second row in table). The other 7 runs are responsible most of the deviation of the average statistics from the reference values (first row).

As in the *causal effect* case, there is a mismatch between runs with the lowest prediction errors $\hat{\varepsilon}$ and the lowest error on regime-assigning process $\Delta\gamma$, i.e. we cannot use a filtering on $\hat{\varepsilon}$ to find the best performing runs. This behaviour can be explained as the tendency of the algorithm to still over-fit when too many degrees of freedom are available, as well as the complexity of distinguishing different links causal effects (a challenge already manifested in the *causal effect* case).

D. Computational complexity

Table IX shows some indicators of performance of the algorithm: the fraction of N_R runs that correspond to a (local) minima, the percentage of annealing per each run that reach a minima and the corresponding average number of q-iterations to get there. The mean value of the prediction error across all N_R.

V. A REAL-WORLD EXAMPLE: THE EFFECT OF EL NIÑO SOUTHERN OSCILLATION ON INDIAN RAINFALL

We finally test the performance of Regime-PCMCI on real-world data, and apply it to address the non-stationary relationship of El Niño Southern Oscillation (ENSO) and all-India rainfall (AIR) mentioned in the introduction. We are interested if, for given time-series of ENSO and AIR, our method is able to distinguish between the winter and summer months, i.e. the background-regimes, and to detect a reported link from ENSO to AIR during summer.

This example can be considered a difficult case as the expected signal form ENSO to AIR is likely small compared to natural variability[7]. Further, climate data is typically very noisy with causal relationships being diluted by other, often unknown processes given a complex and fully coupled climate system[27].

Our input data consist of monthly observations of ENSO and AIR, for the years 1871 to 2016, resulting in two time-series consisting of 1740 monthly values each. More precisely, ENSO is represented by the so-called relative Nino3.4 index provided by the National Oceanic and Atmospheric Administration (NOAA)[36]. Data for AIR anomalies (relative to climatology) are provided by the Indian Institute of Tropical Meteorology (IITM)[37]

As free parameters of Regime-PCMCI we chose $K = 2$ regimes to be detected and $C = 292$, which is equivalent to assuming two seasons per year. For the PCMCI settings, we chose the significance level $\alpha = 0.01$. Further, we chose a maximum time-lag of two months, i.e. $\tau_{\max} = 2$, and set mask type ’y’. The optimisation is attempted $N_A = 100$ annealing

selection	$\Delta\gamma\%$	$\mathrm{TPR}_{\mathrm{cros}}$	$\mathrm{TPR}^{\mathrm{ref}}_{\mathrm{cros}}$	$\mathrm{FPR}_{\mathrm{cros}}$	$\mathrm{FPR}^{\mathrm{ref}}_{\mathrm{cros}}$	$\Delta\Phi$	$\Delta\Phi^{\mathrm{ref}}$	$\Delta\Phi$ %	$\Delta\Phi^{\mathrm{ref}}$ %	$\hat{\varepsilon}$	number
all	11.7	0.94	1.0	0.18	0.08	0.059	0.005	16.0	1.5	0.85	70
$\Delta\gamma < 11.7$ %	0.19	1.0	1.0	0.08	0.07	0.006	0.005	1.8	1.5	0.70	53

TABLE VIII. Results from $N_R = 100$ data generated per each examples described in Table IV. Mean is taken over the $N_K = 3$ regimes and over N_R.

example	n. local minima (%)	iterations to minima	runtime (s)
arrow direction	92 (98 %)	7	600
causal effect	16 (32 %)	13	970
lag	60 (848 %)	11	1,130
sign X^1	52 (74 %)	12	970
sign X^1X^2	70 (93 %)	9	700
sign X^1X^2 and *arrow*	56 (80 %)	10	2,670
high dimensional	65 (97 %)	6	10,780

TABLE IX. Run performance of all examples examples. Average over respective N_R.

times, to span many local minima, with each annealing allowed up to $N_Q = 100$ iteration to converge.

Among the annealing steps, which correspond to different random initial guess on the regime-assigning process Γ, some clearly performed better in terms of fitting the data. We define the average prediction error associated with one annealing, $\hat{\varepsilon}$, defined in Section IV A. Figure 9(top) shows the average prediction error for all the annealings (ranked according to $\hat{\varepsilon}$), with a red box highlighting the top performing cluster (13).

All of the top 13 annealing find a link from ENSO to AIR during one of their two regimes only (for simplicity hereafter called regime 1). In the following we present results averaged over these annealing.

The causal link from ENSO to AIR in regime 1 has an average standardized linear effect of -0.4, meaning that a one standard deviation increase in ENSO results in a reduction of 0.4 standard deviations in AIR. This negative dependence is well documented in the literature[7]. During regime 2, in contrast, ENSO and AIR are, on average, almost independent, with only a very weak link (-0.05) detected from AIR to ENSO.

More importantly, our results indicate a clear seasonal dependence. Figure 9 shows the number of months assigned to each regime (normalised by the number one would expect on the hypothesis of no seasonality, see figure caption). A clear peak in summer months is found for regime 1. More precisely, most of the months between June to September are assigned to regime 1 (70%). These are the months in which the Indian summer Monsoon is active and for which a robust influence from ENSO has been shown. In contrast, months assigned to regime 2 are predominantly winter months (60% of all December to March months). Thus, despite the relatively weak mean causal effect of ENSO on AIR during summer, and the large inter-annual variability, our algorithm successfully reconstructed this well-documented relationship given all-year time-series of ENSO and AIR.

Overall, these results are promising and show the potential of Regime-PCMCI to understand and detect regime-dependent causal structures in a system as complex as the climate system. On the other hand, it also shows that domain knowledge is required to assure a suitable choice of parameters (C and K) and an interpretation of the results. This is yet a common caveat to many data-driven approaches, which we nevertheless want to stress strongly.

VI. DISCUSSION AND CONCLUSION

A novel Regime-PCMCI algorithm that overcomes one of the key drawbacks of many causal recovery methods by allowing to learn non-stationary causal relations has been introduced. The performance of the technique is analysed for many different artificially generated causal scenarios and for varying regimes. Except in the context of identifiability issues that might require more and distinct samples (see Figure 6) the results are impressively accurate for all settings (see Figures 2-5 and Table V). The good performance of the algorithm is maintained even for high dimensional state spaces (see Table VIII) as well as for more than two regimes (see Figure 7 and Table III). This thorough investigation of different scenarios by means of toy models allowed us to reveal the strength and also the limits of the proposed algorithm which is a valuable asset when the method is applied to real data sets from various application areas. Further the capability of the Regime-PCMCI is verified by means of a well understood real data set of ENSO and Indian rainfall (see Figures 9). Concluding the proposed approach presents a promising approach in the context of nonlinear causal links manifested in regime changes in time.

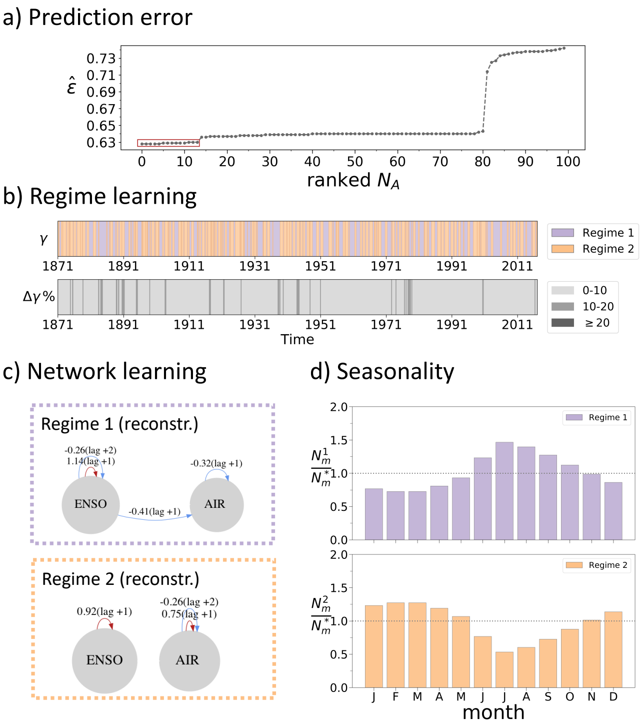

FIG. 9. Prediction error for each annealing step in ascending order, lowest 13 annealings highlighted in red box. All the other panels refer to this selection (a). Regime learning (b): regime-assigning process corresponding to the best annealing (rank 0) (top) and departure from this estimate of the remaining best 12 annealings (in percentage difference). Network learning (c): mean networks per regime, each causal effect is the mean of of the corresponding coefficient in the individual 13 annealings. Seasonality of the regimes (d): Number of months assigned to each regime (N_m^k). The values is normalised by a factor N_m^* corresponding to the expected number of months assigned to a given regime under the null hypothesis of no seasonality (months assigned to each regime with probability $1/N_K$) ($N_m^* = 13 \cdot T/(12 * N_K)$).

A. Outlook

There are several interesting aspects that could be explored in the future by building on the fundament laid out here. As already discussed in Section III E the stationary PCMCI algorithm allows for nonlinear causal links and a nonlinear extension of the Regime-PCMCI and through investigation of its properties is a potential next step. Further it would be possible to learn the structure of the noise term and allow for non-stationary noise. In terms of application it would be highly interesting to utilise the proposed method for observations not as well understood as the presented El Niño-Indian rainfall scenario.

Further note that a causal interpretation of estimated links in our framework still assumes causal sufficiency, that is, no unobserved common causes. However, estimated non-links do not require this assumption. Our approach could be extended by combining latent causal discovery methods with our regime assignment procedure instead of PCMCI.

VII. AUTHOR'S CONTRIBUTION

E.S., J.R., M.K. and J.dW. designed the research, E.S. performed the research, E.S., J.R., M.K. and J.dW. analyzed the results and wrote the manuscript.

ACKNOWLEDGEMENT

ES was supported by the Centre for Doctoral Training in Mathematics of Planet Earth, UK EPSRC funded (grant EP/L016613/1). MK and JdW have been partially funded by the ERC Advanced Grant ACRCC (grant 339390). The research of JdW has been partially funded by Deutsche Forschungsgemeinschaft (DFG) - SFB1294/1 - 318763901 and by the Simons CRM Scholar-in-Residence Program. MK has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement [No 841902]. The optimisation software by Gurobi$^{\text{TM}}$was used for this work.

[1] I. Horenko, S. Gerber, T. J. O'Kane, J. S. Risbey, and D. P. Monselesan., "On inference and validation of causality relations in climate teleconnections," Nonlinear and Stochastic Climate Dynamics (2016).

[2] j. . J. y. . . v. . . n. . . p. . . Matthieu Droumaguet Anders Warne Tomasz Woźniak, title = Granger Causality and Regime Inference in Markov Switching VAR Models with Bayesian Methods, .

[3] S. Gerber and I. Horenko, "On Inference of causality for discrete state models in a multiscale context," PNAS **111**, 14651–14656 (2014).

[4] 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," Nature Communications **10**, 2553 (2019).

[5] J. Pearl, Causality: Models, Reasoning, and Inference (Cambridge University Press, New York, NY, 2000).

[6] P. Spirtes, C. Glymour, and R. Scheines, Causation, Prediction, and Search (MIT Press, Boston, 2000).

[7] P. J. Webster and T. N. Palmer, "The past and the future of el niño," Nature **390**, 562—-564 (1997).

[8] J. Shaman and E. Tziperman, "Summertime enso–north african–asian jet teleconnection and implications for the indian monsoons," Geophysical Research Letters **34** (2007), 10.1029/2006GL029143, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2006GL029143.

[9] I. Pal, A. W. Robertson, U. Lall, and M. A. Cane, "Modeling winter rainfall in northwest india using a hidden markov model: understanding occurrence of different states and their dynamical connections," Climate Dynamics **44**, 1003—-1015 (2015).

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

[11] L. Barnett and A. K. Seth, "Granger causality for state space models," Phys. Rev. E **91**, 040101 (2015), arXiv:1501.06502.

[12] D. Koller and N. Friedman, MIT Press (MIT Press, Cambridge, 2010) arXiv:arXiv:1011.1669v3.

[13] D. M. Chickering, "Learning Equivalence Classes of Bayesian-Network Structures," J. Mach. Learn. Res. **2**, 445–498 (2002).

[14] J. Peters, D. Janzing, and B. Schölkopf, Elements of causal inference: foundations and learning algorithms (MIT Press, Cambridge, MA, 2017) pp. 1214–1216.

[15] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems." Science (80-.). **338**, 496–500 (2012).

[16] J. Arnhold, P. Grassberger, K. Lehnertz, and C. Elger, "A robust method for detecting interdependences: application to intracranially recorded EEG,"

Phys. D Nonlinear Phenom. **134**, 419–430 (1999), arXiv:9907013 [chao-dyn].

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

[18] D. Malinsky and P. Spirtes, “Learning the structure of a nonstationary vector autoregression,” in The 22nd International Conference on Artificial Intelligence and Statistics (2019) pp. 2986–2994.

[19] K. Zhang, B. Huangy, J. Zhang, C. Glymour, and B. Schölkopf, “Causal discovery from Nonstationary/heterogeneous data: Skeleton estimation and orientation determination,” in Proc. Int. Jt. Conf. Artif. Intell. (California, 2017) pp. 1347–1353, arXiv:15334406.

[20] J. Peters, P. Bühlmann, and N. Meinshausen, “Causal inference by using invariant prediction: identification and confidence intervals,” J. R. Stat. Soc. Ser. B **78**, 947–1012 (2016), arXiv:1501.01332.

[21] R. Christiansen and J. Peters, “Switching regression models and causal inference in the presence of discrete latent variables,” J. Mach. Learn. Res (2020).

[22] F. Zwiers and H. V. Storch, “Regime-Dependent Autoregressive Time Series Modeling of the Southern Oscillation,” Journal of Climate **3**, 1347–1363 (1990).

[23] J. de Wiljes, L. Putzig, and I. Horenko, “Discrete nonhomogeneous and nonstationary logistic and markov regression models for spatiotemporal data with unresolved external influences,” Communications in Applied Mathematics and Computational Science **284**, 184–193 (2014).

[24] I. Horenko, “Finite Element Approach to Clustering of Multidimensional Time Series,” SIAM J. Sci. Comp. **32**, 62–83 (2010).

[25] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, “Detecting and quantifying causal associations in large nonlinear time series datasets,” Science Advances **5** (2019), 10.1126/sciadv.aau4996.

[26] D. M. C. F. J. Risbey, T. O’Kane and I. Horenko, “Metastability of northern hemisphere teleconnection modes,” J. Atmos. Sci. **72**, 35–54 (2015).

[27] P. D. Williams, M. J. Alexander, E. A. Barnes, A. H. Butler, H. C. Davies, C. I. Garfinkel, Y. Kushnir, T. P. Lane, J. K. Lundquist, O. Martius, R. N. Maue, W. R. Peltier, K. Sato, A. A. Scaife, and C. Zhang, “A census of atmospheric variability from seconds to decades,” Geophysical Research Letters **44**, 201–211 (2017).

[28] D. Colombo and M. H. Maathuis, “Order-Independent Constraint-Based Causal Structure Learning,” J. Mach. Learn. Res. **15**, 3921–3962 (2014).

[29] H. Akaike, “Information theory and an extension of the maximum likelihood principle ,” 2nd International Symposium on Information Theory (1973).

[30] A. N. Tikhonov, A. Goncharsky, V. V. Stepanov, and A. G. Yagola, Numerical Methods for the Solution of Ill-Posed Problems (Springer, 1995).

[31] J. Runge, J. Heitzig, N. Marwan, and J. Kurths, “Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy,” Physical Review E **86**, 061121 (2012), arXiv:1210.2748.

[32] K. Burnham and D. Anderson, Model Selection and Multimodel Inference (Springer, 2002).

[33] B. Shipley and J. C. Douma, “Generalized aic and chi-squared statistics for path models consistent with directed acyclic graphs,” Ecology **101**, e02960 (2020).

[34] M. Hurvich and C.-L. Tsai, “Regression and time series model selection in small samples,” Biometrika **76**, 297–307 (1989).

[35] .

[36] `http://climexp.climexp-knmi.surf-hosted.nl/getindices.cgi?WMO=NCDCData/ersst_nino3.4a_rel&STATION=NINO3.4_rel&TYPE=i&id=someone@somewhere`.

[37] `http://climexp.climexp-knmi.surf-hosted.nl/getindices.cgi?WMO=IITMData/ALLIN&STATION=All-India_Rainfall&TYPE=p&id=someone@somewhere`.

Appendix A: Definition of result statistics

The definition for the statistics presented in Tables V, III and VIII is outlined in the following.

1. Regime assigning process

$$\Delta\gamma(\%) = \frac{\sum_{t=\tau_{\max}}^{T} |\{\gamma_k(t)\}^{reco.} - \{\gamma_k(t)\}^{ref}|}{T-\tau_{max}} \times 100\% \tag{A1}$$

2. Links’ detection

TPR

$$\mathrm{TPR} = \frac{\mathrm{TP_X}}{\mathrm{P_X}} \tag{A2}$$

Over the cross-variables links (in Tables VIII):

$$\begin{aligned} \mathrm{TP_{cros}} &= |\{(i,j,\tau) : \{\Phi_k^j(i,\tau)\}^{reco.} \neq 0 \ \ \& \\ &\qquad \{\Phi_k^j(i,\tau)\}^{ref} \neq 0 \ \ \& \ \ i \neq j\}| \qquad \text{(A3)} \\ \mathrm{P_{cros}} &= |\{(i,j,\tau) : \{\Phi_k^j(i,\tau)\}^{ref} \neq 0 \ \ \& \ \ i \neq j\}| \end{aligned}$$

And over all links (in Tables V and III):

$$\begin{aligned} \mathrm{TP_{all}} &= |\{(i,j,\tau) : \{\Phi_k^j(i,\tau)\}^{reco.} \neq 0 \ \ \& \\ &\qquad \{\Phi_k^j(i,\tau)\}^{ref} \neq 0\}| \qquad \text{(A4)} \\ \mathrm{P_{all}} &= |\{(i,j,\tau) : \{\Phi_k^j(i,\tau)\}^{ref} \neq 0\}| \end{aligned}$$

FPR

$$\mathrm{FPR} = \frac{\mathrm{FP_X}}{\mathrm{N_X}} \tag{A5}$$

Over the cross-variables links (in Tables VIII):

$$\begin{aligned} \mathrm{FP_{cros}} &= |\{(i,j,\tau) : \{\Phi_k^j(i,\tau)\}^{reco.} \neq 0 \ \ \& \\ &\qquad \{\Phi_k^j(i,\tau)\}^{ref} = 0 \ \ \& \ \ i \neq j\}| \qquad \text{(A6)} \\ \mathrm{N_{cros}} &= |\{(i,j,\tau) : \{\Phi_k^j(i,\tau)\}^{ref} = 0 \ \ \& \ \ i \neq j\}| \end{aligned}$$

And over all links (in Tables V,III):

$$\begin{aligned} \mathrm{FP_{all}} &= |\{(i,j,\tau) : \{\Phi_k^j(i,\tau)\}^{reco.} \neq 0 \ \ \& \\ &\qquad \{\Phi_k^j(i,\tau)\}^{ref} = 0\}| \qquad \text{(A7)} \\ \mathrm{N_{all}} &= |\{(i,j,\tau) : \{\Phi_k^j(i,\tau)\}^{ref} = 0\}| \end{aligned}$$

3. Links’ coefficients

$$\Delta\Phi = \frac{1}{N_K} \sum_{k=1}^{N_K} \frac{1}{\sum_j |\mathscr{P}_k^j|} \sum_j \sum_{(i,\tau)\in\mathscr{P}_k^j} |\{\Phi_k^j(i,\tau)\}^{reco.} - \{\Phi_k^j(i,\tau)\}^{ref}| \tag{A8}$$

Can be also computed as average *percentage* error per regime:

$$\Delta\Phi(\%) = \frac{1}{N_K} \sum_{k=1}^{N_K} \frac{1}{\sum_j |\mathscr{P}_k^j|} \sum_j \sum_{(i,\tau)\in\mathscr{P}_k^j} \frac{|\{\Phi_k^j(i,\tau)\}^{reco.} - \{\Phi_k^j(i,\tau)\}^{ref}|}{\{\Phi_k^j(i,\tau)\}^{ref}} \times 100\% \tag{A9}$$

4. Prediction error

$$\hat{\varepsilon} \equiv \frac{1}{N_X T}\sum_t \sum_j |\{x^j(t)\}^{ref} - \{x^j(t)\}^{reco.}| = \sqrt{\frac{L}{N_X \cdot T}} \quad \text{(A10)}$$

Abbreviations	
AIC	Akaike Information criterion
AICc	corrected Akaike Information criterion
ENSO	El Nino Southern Oscillation
FPR	false positive rate
MCI	momentary conditional independence
MLR	multi linear regression
PCMCI	causal discovery algorithm
RAM	Regime-dependent Autoregressive Model
SCM	structural causal model
TPR	true positive rate

TABLE X. Abbreviations used throughout the manuscript.

List of notation	
$\{X_t\}_{t\in\mathbb{Z}}$	Stochastic Process
N_X	Spatial dimension of $\{X_t\}$
N_K	Number of regimes
N_C	Bound for switches of $\gamma_k(t)$ for each k
N_Q	Number of iteration steps
N_A	Number of annealing steps
N_R	Number of runs for random initial values
N_{para}	Number of parameters
α	link confidence level
$\mathscr{P}_t^j$	parents of component X_t^j
$\Gamma(t)$	regime assigning process
Φ_t	causal effect parameters, time dependent
Υ_k	collection of specific time steps, dependent on regime
$\mathbf{x}_t$	time series

TABLE XI. Notation used throughout the manuscript.