Reconstructing regime-dependent causal relationships from observational time series

Reconstructing regime-dependent causal relationships from observational time series

Cite as: Chaos **30**, 113115 (2020); https://doi.org/10.1063/5.0020538
Submitted: 01 July 2020 . Accepted: 19 October 2020 . Published Online: 06 November 2020

Elena Saggioro, Jana de Wiljes, Marlene Kretschmer, and Jakob Runge

Reconstructing regime-dependent causal relationships from observational time series

Cite as: Chaos **30**, 113115 (2020); doi: 10.1063/5.0020538
Submitted: 1 July 2020 · Accepted: 19 October 2020 ·
Published Online: 6 November 2020

Elena Saggioro,[1,a] Jana de Wiljes,[2,b] Marlene Kretschmer,[3,c] and Jakob Runge[4,d]

AFFILIATIONS

[1]Department of Mathematics and Statistics, University of Reading, Reading RG6 6AX, United Kingdom
[2]Institute for Mathematics, University of Potsdam, D-14476 Potsdam, Germany
[3]Department of Meteorology, University of Reading, Reading RG6 6AX, United Kingdom
[4]German Aerospace Center, Institute of Data Science, 07745 Jena, Germany

[a]**Author to whom correspondence should be addressed:** e.saggioro@pgr.reading.ac.uk
[b]wiljes@uni-potsdam.de
[c]m.j.a.kretschmer@reading.ac.uk
[d]Jakob.Runge@dlr.de

ABSTRACT

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 few 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. Often dynamical systems feature transitions 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 detect regime-dependent causal relations, we combine the conditional independence-based PCMCI method [based on a condition-selection step (PC) followed by the momentary conditional independence (MCI) test] with a regime learning optimization approach. PCMCI allows for causal discovery from high-dimensional and highly correlated time series. Our method, Regime-PCMCI, is evaluated on a number of numerical experiments demonstrating that it can distinguish regimes with different causal directions, time lags, and sign of causal links, as well as changes in the variables' autocorrelation. Furthermore, Regime-PCMCI is employed to observations of El Niño Southern Oscillation and Indian rainfall, demonstrating skill also in real-world datasets.

Regime-dependent non-stationarity is a ubiquitous feature of physical systems, especially prominent in atmospheric sciences. This dependence can be looked at as an intermittent change in relationships defining the dynamics of a multivariate system, each of which can be described as a time series causal network. In this work, we develop a novel algorithm to detect regime-dependent causal relations that combines the constrained-based causal discovery algorithm PCMCI with a regime assigning linear optimization algorithm. Our method, Regime-PCMCI, is evaluated on a number of numerical experiments and demonstrates high performance in detecting a variety of regime-dependent features. Finally, Regime-PCMCI is applied to observations of El Niño Southern Oscillation and Indian rainfall, demonstrating skill in detecting well-known seasonal regimes in a real-world dataset.

I. INTRODUCTION

Understanding causal relationships[1,2] among different processes is a ubiquitous task in many scientific disciplines as well as engineering (e.g., in the context of climate research,[3–8] econometrics,[9,10] molecular,[11] and animal group[12] dynamics). 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.[8] Based on that graph, the processes that generated the data can be modeled in the framework of structural causal models (SCMs)[1] to further understand causal relations, predict the effect of interventions, and for forecasting.

Today's ever-growing abundance of time series datasets promises many application scenarios for the now numerous data-driven causal discovery methods. But many challenges emerging from the dynamic nature of such datasets have not yet been met. Furthermore, causal knowledge cannot be gained from data alone and each method comes with its particular set of assumptions[2] about properties of the underlying processes and the observed data. Runge *et al.*[8] recently provided an overview of current methodological frameworks, their application scenarios, and open challenges.

One such issue is posed by time-varying causal relationships, a frequent feature of both natural and artificial systems. In Earth sciences, for example, the dominant causal relationship between soil moisture and air temperature periodically reverses due to land-atmosphere feedbacks;[13] in animal group dynamics, the leader–follower role of an individual often mutates in time;[14] and in econometrics, the direction of influence between stock markets and macroeconomic variables is often dynamical.[15]

A feature commonly observed in non-stationary dynamical systems is regime dependence. Regime dependence means that the causal relationships between the considered processes vary depending on some prevailing background regime that may be modeled as switching between different states. Furthermore, often such regimes have strong persistence, that is, they operate and affect causal relations on much longer timescales 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.[16,17] It is, however, generally assumed that ENSO does only marginally affect Indian rainfall in winter.[18] 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 timescale (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 Runge *et al.*[8] In general, causal (graph) discovery methods can be classified into classical Granger causality approaches,[9,19] constraint-based causal network learning algorithms,[2] score-based Bayesian network learning methods,[20,21] structural causal models,[22,23] and state-space reconstruction methods.[24,25]

Here, we focus on the constraint-based framework, which has the advantage that it can flexibly account for linear and nonlinear causal relations and different data types (continuous and categorical, univariate and multivariate). In particular, we adopt the PCMCI algorithm[26] [based on a condition-selection step (PC) followed by the momentary conditional independence (MCI) test] to reconstruct time series causal graphs. PCMCI is an adaptation of the constraint-based PC algorithm (named after its inventors Peter Spirtes and Clark Glymour[2]) that addresses autocorrelation of time series via the use of a momentary conditional independence (MCI) test. PCMCI yields high detection power also in high-dimensional and strongly autocorrelated time series settings (see Sec. III A and Runge *et al.*[26] for more details). However, one of the general assumptions of PCMCI (as well as of other causal discovery algorithms) is stationarity, i.e., that the existence or absence of a causal link does not change over the considered time series segment.[27] 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 that constitute a particular case of latent confounding.

Some recent work addresses causal discovery in the presence of non-stationarity. Malinsky and Spirtes[28] model non-stationarity in the form of (continuous) stochastic trends in a linear autoregressive framework. Zhang *et al.*[29] 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 modeled as latent drivers. Peters, Bühlmann, and Meinshausen[30] and Christiansen and Peters[31] assumed that known non-stationary regimes are exploited to estimate causal relations also in the presence of general latent confounders. Furthermore, in the context of continuously varying causality, methods based on information transfer metrics have been proposed. In the field of animal group dynamics, for instance, detection of time-varying leader–follower relationships is achieved with the use of a time dependent transfer entropy.[23,32,33] Applied to non-stationary climate systems, Hagan *et al.*[34] proposed a Kalman filter estimate of the time-varying parameters for the Liang–Kleeman information flow. Benefits of these methods are the treatment of non-linearity[32] and the identification of both timing and frequency of interactions.[34] However, in these approaches, only bivariate influences are modeled, i.e., the effect of a third variable Z on the estimated effect of X onto Y cannot be accounted for. The practical extension to high-dimensional systems and short time series also remains hard to address.

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 (RAMs) were introduced already in 1990.[35] These can yield physically well interpretable results that, however, require well-chosen ancillary variables and a seasonal index that are not learned from data. Thus, RAM requires *a priori* knowledge of the regimes, which one often aims to learn rather than enforce. In the context of discrete state spaces, regime-dependent causal discovery has been considered in Gerber and Horenko[11] for Boolean variables. Non-stationary Boolean network models have also been considered in Porfiri and Marin[23]—specifically, the approach is to fit an appropriate parameterization of associate transition probabilities. Another approach

that has been proposed to model time dependent Granger (non-) causality is based on a Markov Switching VAR ansatz with an economics application in mind.[10] Specifically, the regime assignments are computed by sampling from a Markov chain. Further methods have been proposed to obtain time step specific bivariate Granger Causality from partitioning the time-series into regular time windows.[36,37]

A more general framework to handle discrete regimes is the Markov-switching ansatz of de Wiljes *et al.*,[38] which flexibly models regime dependence utilizing the assumption of a finite number of regimes and a level of persistence in the transitions between different regimes. This ansatz has been successfully realized in combination with many different model assumptions (e.g., see Refs. 39 and 40). Here, we want to explore it for causal networks by combining it with PCMCI,[26] a constraint-based time series causal discovery method.[2] We call our method Regime-PCMCI.

The remainder of the paper is structured as follows: In Sec. II, the underlying mathematical problem, concepts, and key assumptions are formalized, and a motivating example is discussed to provide some intuition. Our novel method Regime-PCMCI is then presented in Sec. III. These theoretical and algorithmic parts are complemented by a thorough numerical investigation of the proposed method in various artificial settings in Sec. IV. Finally, in Sec. V, Regime-PCMCI is applied to a real-world dataset 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 realization 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 realizations. We assume that the underlying process is modeled 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 $\mathscr{D}^j$, and the sets $\mathscr{P}_t^j \subset (X_{t-1}, X_{t-2}, \ldots)$ define the causal parents of X_t^j. Note that we assume lagged causal relationships, but this is not a necessity since there exist causal discovery algorithms that can deal with contemporaneous causal links[41] and also hidden confounders. 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 1 (Sec. II B).

The problem setting considered in this manuscript is of the nature of the following inverse problem:

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

with $\widehat{\mathbf{G}}_t = [\widehat{g}_t^1, \ldots, \widehat{g}_t^{N_X}]$, where $\widehat{g}_t^j$ belongs to an appropriate function space for each t and j. $\tau_{\max}$ is the maximum considered time lag. 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 Sec. II A, we will discuss the particular structure of the parameters Θ_t we are interested in. Please refer to Table I for a summary of the notation used throughout the text.

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.[1,2] For time series, we use the concept of time series graphs. The nodes in the time

TABLE I. Notation used throughout the paper.

List of notations	Model	Parameters	
$\{X_t\}_{t\in[0,T]}$	Stochastic process	N_X	Dimension of stochastic process
η_t^j	Noise variable of component X_t^j	T	Time length of stochastic process
$\mathscr{D}^j$	Stationary noise distribution	N_K	Number of regimes
g_t^j	Structural causal model function	N_C	Max switches for each regime
$\mathscr{P}_t^j$	Causal parents of X^j, time dependent	N_M	Regime average persistence
$\mathbf{x}_t$	Realization of X_t	$\tau_{\max}$	Maximum causal time lag
$\widehat{\mathbf{G}}_t$	Operator in inverse problem	CI test	Conditional independence test
$\hat{g}_t^j$	Components of operator $\widehat{\mathbf{G}}_t$	α_{PC}	Significance level for PC_1 step
Θ_t	Unknown parameter in inverse problem	α	Link significance level
$\mathbf{L}(\Gamma, \mathscr{P}, \Phi)$	Cost functional	N_Q	Number of optimization iterations
$\Gamma(t)$	Regime-assigning process	N_A	Number of annealings
Φ_t	Linear link coefficients, time dependent	N_R	Number of realizations for a toy example
$\gamma_k(t)$	Regime-assigning process for regime k	$\mathcal{N}(0, \{\sigma^2\}^{\mathrm{ref}})$	Ground-truth Gaussian noise distribution
$\Phi_k^j(i, \tau)$	Linear link coefficient in regime k	$\{\Phi_k^j(i, \tau)\}^{\mathrm{ref}}$	Ground-truth linear link coefficient
$\mathscr{P}_k^j$	Causal parents of X^j in regime k	N_{para}	Number of model parameters
Υ_k	Collection of time steps associated with regime k	$\hat{\mathbf{x}}_{k,t}$	Reconstructed time-series for regime k

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 $X_{t-\tau}^i$ and X_t^j for a time lag $\tau > 0$ and a given t are connected by a lag-specific directed link, denoted $X_{t-\tau}^i \to X_t^j$, when $X_{t-\tau}^i \in \mathscr{P}_t^j$ for a particular t. If a SCM is not given, another way to define links is that $X_{t-\tau}^i$ is not conditionally independent of X_t^j given the past of all variables, defined by $X_{t-\tau}^i \not\perp\!\!\!\perp X_t^j \,|\, \mathbf{X}_t^- \backslash \{X_{t-\tau}^i\}$, with $\not\perp\!\!\!\perp$ denoting the absence of a (conditional) independence.[27] Widely used in constraint-based methods, note that this definition cannot be used to define directed contemporaneous links. We denote the maximum ground-truth time lag of any parent as $\tau_{\max}^{\mathscr{P}}$. For a more detailed introduction, the reader is referred to Runge *et al.*[26] In the following, we will use graphs and networks interchangeably.

The collection of parent sets for all components at time t is denoted $\mathscr{P}_t = \{\mathscr{P}_t^1, \ldots, \mathscr{P}_t^{N_X}\}$. 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 $g_t^j(\mathscr{P}_t^j, \eta_t^j)$ in SCM (1) corresponding to these links. We will assume a known function class $\widehat{\mathbf{G}}_t(\ldots; \Phi_t)$ of unknown coefficients $\Phi_t = \{\Phi_t^1, \ldots, \Phi_t^{N_X}\}$ that are going to be inferred via

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

In other words, for a given time series, $\mathbf{x}_t \in \mathbb{R}^{N_X}$ with $t \in [0, T]$ and known function class $\widehat{\mathbf{G}}_t$ the aim is to find the unknown parameters $\Theta_t = [\mathscr{P}_t, \Phi_t]$.

Note that we will specifically focus on linear function classes, as discussed in Sec. III.

B. Persistence

As mentioned above, in many application areas, non-stationarity may be modeled not in the form of abrupt or continuous changes, but via piece-wise stationary regimes.[11,42,43] These regimes will further exhibit a certain persistent behavior. In order to capture non-stationary systems with these properties, we will restrict our inference to regime-dependent persistent dynamics.

Assumption 1. *Denote the causal parents and functional dependency of a given variable j for a regime k as $\mathscr{P}_t^j = \mathscr{P}_k^j$ and $g_t^j(\mathscr{P}_t^j, \eta_t^j) = g_k^j(\mathscr{P}_k^j, \eta_t^j)$. We call a regime (N_M, N_K)-persistent if the parents and functional dependencies are stationary for an average of N_M consecutive time steps t and that there is a finite number of regimes on the whole time domain, i.e., $k \in \{1, \ldots, N_K\}$.*

The persistency enters here via the regime average persistence N_M, which also naturally implies a finite number of regimes $N_K \leq T/N_M$.

Under Assumption 1, the considered inverse problem (3) reduces to finding the unknown parameters $\Theta_t = [\Gamma(t), \mathscr{P}, \Phi]$ comprising (1) a set of regimes' network parameters

$$\mathscr{P}, \Phi = \{\mathscr{P}_1, \ldots, \mathscr{P}_{N_K}, \Phi_1, \ldots, \Phi_{N_K}\}$$

and, to encode their time dependence, (2) the change points between the regimes given by the regime-assigning process

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

with $\Gamma(t) \in [0, 1]^{N_K \times T}$. For example, component k of the regime-assigning process can be of the form $\gamma_k = (0, 1, 1, \ldots 0, 1) \in [0, 1]^T$, indicating that regime k is active for all time steps for which $\gamma_k(t) = 1$.

C. Optimization problem

Finally, in order to solve the inverse problem (3) under the persistency Assumption 1, we can define a cost functional

$$\mathbf{L}(\Gamma, \mathscr{P}, \Phi) = \sum_{t=0}^{T} \sum_{k=1}^{N_K} \gamma_k(t) d(\mathbf{x}_t - \widehat{\mathbf{G}}_t(\mathscr{P}_k; \Phi_k)) \tag{4}$$

subject to constraints

$$\sum_{k=1}^{N_K} \gamma_k(t) = 1 \quad \forall\, t, \text{ with } \gamma_k(t) \in [0, 1] \tag{5}$$

and

$$\sum_{t=1}^{T-1} |\gamma_k(t+1) - \gamma_k(t)| \leq N_C \quad \forall k, \tag{6}$$

where d is a distance measure such as the squared Euclidean distance $\|\cdot\|_2^2$ and $\gamma_k(t)$ can be regarded as the weight of the k regime-specific network at each time t.

This learning approach is based on the ideas first proposed in Horenko[39] and later extended to many different models.[38] The format of $\mathbf{L}(\Gamma, \mathscr{P}, \Phi)$ relies on the assumption that the system associated with the considered data exhibits metastability in time (see Assumption 1 that translates in the summation over k, controlled by the regime number N_K). The desired level of persistence enters the functional in the form of a regularization [see Constraint (6), controlled by parameter N_C]. An alternative option is to add a regularization term that enforces some form of smoothness of Γ (e.g., Tikhonov regularization[44]).

The tuning parameter N_C is related to the *average* regime duration of N_M time steps of Assumption 1 as follows: an average regime duration of N_M in all N_K regimes is implemented by choosing $N_C \approx T/(N_M N_K)$. Importantly, note that the regularization (6) ensures an average persistence on each regime without imposing that *individual* regime durations are a constant; in fact, they can be fully irregular within the bound of performing at maximum N_C switches. This regime learning method thus provides a simple, flexible, and computationally tractable strategy to go beyond the assumption of fixed length for each regime duration often employed in previous methods (e.g., Refs. 37 and 33).

Note that, in practice, it is reasonable to assume that an estimate of the average regime switching time N_M is available, consistent with a typical timescale of the application domain. The choice of parameters N_K and $N_C(N_M)$ will be discussed in Sec. III D.

D. Motivating example

Before we introduce our novel regime detecting the 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

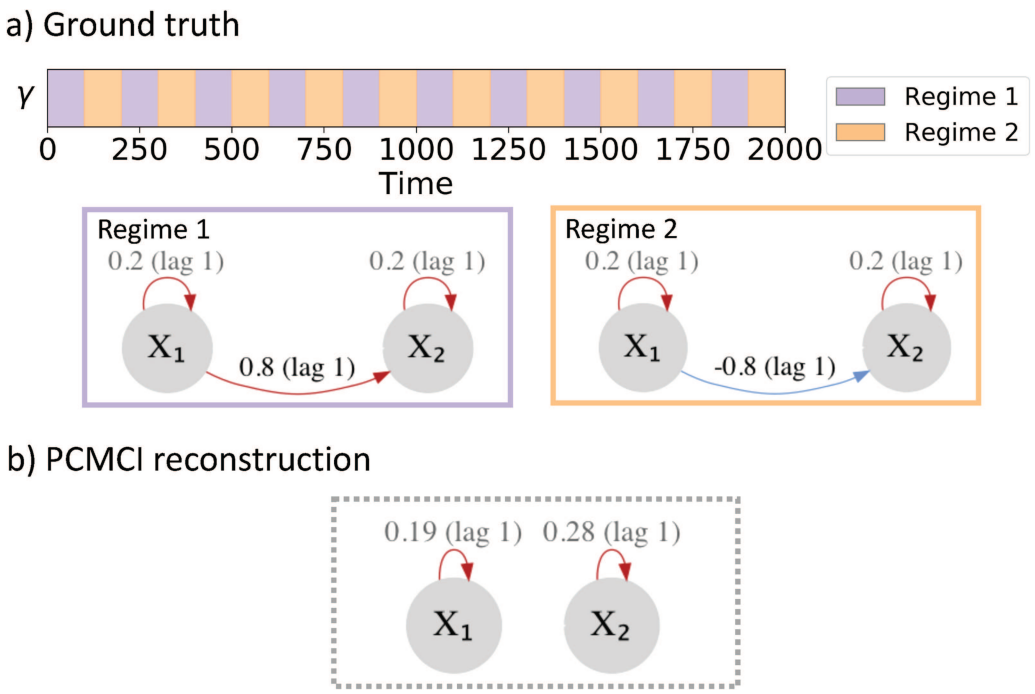

FIG. 1. Motivating example. (a) Regime-dependent ground truth: regime-assigning process and regime-dependent networks. The links are labeled with the associated linear coefficient $\Phi_k^j(i,\tau)$ and lag τ. The sign of the coefficient is highlighted by the color (red for positive, blue for negative). (b) Network reconstruction with PCMCI estimated from the whole time series, i.e., if links are wrongly assumed to be stationary.

series X^1 and X^2 and the associated causal graphs as shown in Fig. 1(a). 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 in time. 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 method 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 a linear version of PCMCI on the whole time sample would give a network of disconnected variables [Fig. 1(b)].

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 (not shown). To deal with such problems automatically, our algorithm needs to learn both the regimes and the regime-dependent causal relations.

III. METHOD

The proposed approach, Regime-PCMCI, is designed to solve the optimization problem (4) by alternating between learning the regimes and learning 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[26] as a well-tested method that adapts the constraint-based causal discovery framework to the time series case.

In the following, we focus on a pure linear setting, which is a reasonable assumption in many application areas.[45,46] This implies that the function class $\widehat{\mathbf{G}}_t(\mathscr{P}_t;\Phi_t)$ in the inverse problem (3) is assumed to be linear in the parents' variables with linear coefficients Φ_t.

A. Causal discovery

The constraint-based framework has the advantage that it can flexibly account for various functional causal relations and different data types (continuous and categorical, univariate and multivariate) since it is based on discovering causal links by means of conditional independence (see link definition in Sec. II A). Two variables X and Y are conditionally independent given a (potentially multivariate) variable Z, denoted $X \perp\!\!\!\perp Y|Z$, if

$$p(x,y|z) = p(x|z)p(y|z), \tag{7}$$

where p denotes the associated probability density functions.

To practically test this relationship, there exist a large variety of conditional independence tests; see Runge *et al.*[26,27] for a discussion. If relationships are assumed linear, as is the case of the present work, partial correlation can be used such that it can be shown that $X \not\perp\!\!\!\perp Y|Z$ if the partial correlation between X and Y conditioned on Z is significantly different from 0.

As mentioned in Sec. I A, PCMCI is based on the constraint-based PC algorithm[2] combined with the momentary conditional independence (MCI) test. It consists of two stages. (1) At first, the

PC_1 condition pre-selection method is run to identify relevant conditions $\widehat{\mathscr{B}}_t^j$ for all time series variables X_t^j. More specifically, PC_1 is a Markov set discovery algorithm based on the PC-stable algorithm[47] that removes irrelevant conditions for each of the N_X variables by iterative independence testing. (2) Then, the MCI test is performed, to confirm whether $X_{t-\tau}^i \to X_t^j$ by means of testing

$$\text{MCI:} \quad X_{t-\tau}^i \not\perp\!\!\!\perp X_t^j \,|\, \widehat{\mathscr{B}}_t^j \setminus \{X_{t-\tau}^i\}, \widehat{\mathscr{B}}_{t-\tau}^j. \tag{8}$$

Thus, MCI conditions on both the (potential) parents of X_t^j and the time-shifted parents of $X_{t-\tau}^i$. These two stages serve the following purposes. PC_1 acts as a filter to remove irrelevant lagged conditions (up to some $\tau_{\max}$) for each variable. 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. More precisely, while the conditioning on the parents of X_t^j (the potential effect) is sufficient to establish conditional independence in the infinite sample limit (Markov property), the additional condition on the lagged parents (parents of $X_{t-\tau}^i$, the potential cause) leads to a test that is better suited for autocorrelated data.

A causal interpretation of the relationships estimated with PCMCI comes from the standard assumptions in the constraint-based framework,[2,26,27] namely, causal sufficiency, the causal Markov condition, faithfulness, non-contemporaneous effects, and stationarity within the regimes as further discussed below. As demonstrated in Runge *et al.*,[26] PCMCI has high detection power and controlled false positives also in high-dimensional and strongly autocorrelated time series settings.

The main free parameters of PCMCI are the chosen conditional independence test, the maximum time lag $\tau_{\max}$, and the significance levels α in MCI and α_{PC} in PC_1. We discuss the selection of these parameters in Sec. III D. In terms of conditional independence test, note that PCMCI can be used in combination with linear or nonlinear tests and can therefore extract also non-linear causal relationships. In this work, we focus on linear systems and thus use PCMCI in conjunction with partial correlation.

In each iterative step of our approach, PCMCI is applied to the sample subset of the time series pertaining to the estimated k regime. Given a significance level α, the output of PCMCI is the set of parents $\mathscr{P}_k = \{\mathscr{P}_k^1, \ldots, \mathscr{P}_k^{N_X}\}$ for all time series variables for that regime,

$$\mathscr{P}_k^j = \{X_{t-\tau}^i : \text{p value}_{k,\text{MCI}}(X_{t-\tau}^i, X_t^j) \le \alpha\} \quad \forall k, j. \tag{9}$$

Based on these parents and associated causal links, causal effects Φ_k that quantify the strength of a link can be estimated. Details of the regime-specific PCMCI fit are found in Sec. III B 1.

B. Regime-dependent causal discovery

The Regime-PCMCI algorithm iterates over two major estimation steps: (step 1) causal discovery to obtain $\mathscr{P}_k$ and fit the coefficients Φ_k and (step 2) regime learning to update the regime variable Γ.

To find good estimates of the parameters and the regime variable optimizing the cost functional (4), this two-step procedure is necessary. In fact, there are generally no analytic solutions to the problem available due to the complexity of the cost functional. Fixing one variable to estimate the other allows in both cases to solve the individual optimization step via linear programming. In Theorem 2.1 of Ref. 39 and Sec. II B in Ref. 48 it is shown that these types of algorithms monotonously decrease the value of $\mathbf{L}$. It is important to note, however, that due to the non-convexity of the underlying problem the algorithm can be caught in regions of local minima. This issue is addressed via additional simulated annealing steps as discussed in more detail in Sec. III B 2.

In the following, q indicates the current iteration. The superscript (q) is added combined with brackets to the variables updated in each loop. The details of the consecutive subroutines are laid out below.

1. Step 1: Causal discovery and model estimation

The first step is to estimate a set of parents $\{\mathscr{P}_k\}^{(q)}$ and coefficients $\{\Phi_k\}^{(q)}$ with $k \in \{1, \ldots, N_K\}$ on the basis of a fixed $\{\Gamma(t)\}^{(q)}$ obtained in step 2 of the previous iteration (first and second bullets in q-loop of of Algorithm 1). In the first iteration, the regimes are assigned randomly. $\{\mathscr{P}_k\}^{(q)}$ and $\{\Phi_k\}^{(q)}$ are estimated on the basis of a subset of the time series $\mathbf{x}_t$ with

$$t \in \{\Upsilon_k\}^{(q)} := \left\{t : \{\gamma_k(t)\}^{(q)} \ge 0.5\right\} \tag{10}$$

for each regime k. The regime-dependent parents $\{\mathscr{P}_k\}^{(q)}$ are estimated via PCMCI.

As stated at the beginning of Sec. III, to solve Eq. (3), we assume a linear functional relationship that relates each variable to its parents $\mathscr{P}_k$. It implies that coefficients Φ_k can be estimated from the following regression model for each fixed k:

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

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

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

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., $\Phi_k^j(i,\tau) := 0$ for $X_{t-\tau}^i \notin \mathscr{P}_k^{j,(q)}$.

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 third bullet in q-loop of

Algorithm 1). In agreement with the cost functional (4), the following optimization problem needs to be solved: find

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

subject to the constraints (5) and (6), and where for each $k \in \{1, \ldots, N_K\}$

$$\hat{x}^j_{k,t} = \sum_{X^i_{t-\tau} \in \mathscr{P}^j_k} \Phi^j_k(i,\tau) x^i_{k,t-\tau} \quad \text{for } t \in \{1, \ldots, T\}. \qquad (14)$$

Since the first $\tau_{\max}$ time steps cannot be predicted, we choose to set those to $\hat{x}^j_{k,t} = x^j_{k,t}$ and to not consider this portion of the time series in the algorithm evaluation. This step can be solved with standard optimization linear programming routines.

In order to search for the global minimum of this non-convex problem, 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 an optimal fit. Note that the individual annealing steps are *embarrassingly parallelizable*.

C. Reconstruction of time series

A single prediction from Eq. (14) can be derived as the weighted sum over k

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

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

D. Parameter selection

Regime-PCMCI involves a number of parameters that need to be chosen. They can be separated into parameters of the causal discovery method PCMCI and those of the regime learning part.

The main free parameters of PCMCI are the chosen conditional independence test, the maximum time lag $\tau_{\max}$, and the significance levels α in MCI and α_{PC} in PC_1. α_{PC} should be regarded as a hyperparameter and can be chosen based on model-selection criteria such as the Akaike information criterion (AIC)[49] or cross-validation. $\tau_{\max}$ could be incorporated into this model selection. But since PCMCI is not very sensitive to this parameter[26] (as opposed to, e.g., Granger causality), its choice can be based on lagged correlation functions, see Runge *et al.*[26] for a discussion. The choice of conditional independence test is a modeling assumption guided by the assumed nonlinearity of the underlying process and also finite sample considerations. Finally, α is chosen based on the desired level of false positives.

The two free parameters in the regime learning step are the bound on the number of switches N_C and the number of regimes N_K. Usually, N_C can be reasonably inferred from the application and given the number of regimes, as explained in Sec. II C. Here, we assume that N_C is known. Note that since N_C bounds the maximum number of switches between regimes, i.e., the optimal reconstructed number can be lower, and does not constrain the individual regime durations, its choice can account for a degree of error. Yet 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 dataset and thus the underlying dynamics well. One way to assess this balance heuristically is to employ an information criterion (IC),[50] which has been derived in the context of regression models and since been adapted to various other model scenarios including graphs.[51]

An IC is designed to capture the goodness of fit penalized by the number of parameters in order to prefer models with as few parameters as possibles to avoid overfitting (parsimony). Here, the number of parameters is defined as

$$N_{\text{para}} = (N_K - 1)N_C + \sum_{k=1}^{N_K} \sum_{j=1}^{N_X} |\mathscr{P}^j_k|. \qquad (16)$$

Algorithm 1. Regime-PCMCI.

Input:
- Time series $\mathbf{x}_t \in \mathbb{R}^{N_X}$ with $t \in \{1, \ldots, T\}$
- Parameters:
  - Number of assumed regimes N_K
  - Maximum number of transitions within a single regime N_C
  - Maximum time lag $\tau_{\max}$
  - Functional model $\hat{\mathbf{G}}$, here linear
  - Conditional independence test according to $\hat{\mathbf{G}}$, here partial correlation
  - Significance level α (and α_{PC} for PC_1 step)
  - Annealing steps N_A
  - Number of optimization 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**
    Causal discovery and model estimation:
- Infer parents $\{\mathscr{P}_k\}^{(q)}$ by means of PCMCI run on subset $\left\{\mathbf{x}_t : t \in \{\Upsilon_k\}^{(q)}\right\}$ for each k
- Fit model coefficients $\{\Phi_k\}^{(q)}$ via Eq. (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 Eq. (14).
    Fit regime-assigning process:
- Update $\{\Gamma\}^{(q+1)}$ solving Eq. (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 effects Φ_k for every $k \in \{1, \ldots, N_K\}$

The first term in Eq. (16) relates to the number of parameters required to describe Γ, which can be fully determined via the change points. The second term in Eq. (16) counts the number of relevant parents, or equivalently the non-zero coefficients $\Phi_k^j(i,\tau)$. Here, we use the corrected Akaike information criterion (AICc) first proposed in Hurvich and Tsai[52] to estimate N_K, assuming known N_C. Note that we use the corrected version of the original AIC[49] to correct for small samples sizes relative to the number of parameters

$$AICc = -2\log(\mathscr{L}) + 2N_{\text{para}} + \frac{2N_{\text{para}}(N_{\text{para}}+1)}{T - N_{\text{para}} - 1}, \tag{17}$$

where $\mathscr{L}$ is the maximum value of the likelihood function for the model one assumes for the residuals (see Metzner *et al.*[48] for a more detailed discussion). Note that the AICc also depends on N_C (as it enters the number of parameters N_{para}) and it is in general possible to simultaneously estimate N_K and N_C.[38,40] The choice of N_K is numerically investigated in Sec. IV C.

The number of iteration steps N_Q should be chosen to ensure that the optimization process converges. In our experiments, we found with exploratory testings that N_Q shows convergence after about 10–20 iterations for all examples investigated. The number of annealing steps N_A should be chosen to ensure spanning a large number of local solutions to this non-convex optimization problem [Eq. (4)]. Computational time will set a limit to a too high parameter. Note, however, that the annealing part is embarrassingly parallelizable.

IV. NUMERICAL INVESTIGATION

In the following, we investigate the performance of Regime-PCMCI by means of several toy examples. The artificial data are 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 linear SCM time series model:

$$x_t^j = \sum_{k=1}^{N_K} \{\gamma_k(t)\}^{\text{ref}} \sum_{X_{t-\tau}^i \in \mathscr{P}_k^j} \{\Phi_k^j(i,\tau)\}^{\text{ref}} x_{t-\tau}^i + \eta_t^j,$$
$$\eta_t^j \sim \mathscr{N}(0, \{\sigma^2\}^{\text{ref}}) \tag{18}$$

with predefined $\{\Gamma(t)\}^{\text{ref}}$, $\{\Phi_k\}^{\text{ref}}$, and $\{\sigma^2\}^{\text{ref}}$. Note that here we numerically investigate equally distributed noises for all variables ($\eta_t^j \sim \mathscr{D} \forall j$), but we refer the interested reader to Appendix A for a treatment of heterogeneous noise distributions. Note that the reference set of parents is specified by the non-zero coefficients $\{\Phi_k^j(i,\tau)\}^{\text{ref}}$.

A. Low dimensional data with two underlying regimes

First, we focus on a simple setting of two regimes, i.e., $\{N_K\}^{\text{ref}} = 2$, 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 $X_{t-\tau}^i \to X_t^j$ will be called auto-links or auto-dependencies for $i = j$ and cross-links for $i \neq j$. We consider the following scenarios as summarized in Table II: 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 (linear coefficient 0.2), which is a realistic yet challenging assumption for many algorithms.

1. Experiment settings

We design five toy models, in network terms, correspond-

TABLE II. Artificial model configurations for different low dimensional experiments with $N_K = 2$ underlying regimes.

Example	$k=1$	$k=2$	$\{\Phi_1^j(i,\tau)\}^{\text{ref}}$	$\{\Phi_2^j(i,\tau)\}^{\text{ref}}$
Arrow direction	$X^1 \to X^2$	$X^1 \leftarrow X^2$	$\{\Phi_1^2(1,1)\}^{\text{ref}} = 0.8$	$\{\Phi_2^1(2,1)\}^{\text{ref}} = 0.8$
			$\{\Phi_1^1(1,1)\}^{\text{ref}} = 0.2$	$\{\Phi_2^1(1,1)\}^{\text{ref}} = 0.2$
			$\{\Phi_1^2(2,1)\}^{\text{ref}} = 0.2$	$\{\Phi_2^2(2,1)\}^{\text{ref}} = 0.2$
Causal effect	$X^1 \xrightarrow{\lvert a\rvert} X^1$	$X^1 \xrightarrow{\lvert b\rvert} X^1$	$\{\Phi_1^1(1,1)\}^{\text{ref}} = 0.8$	$\{\Phi_2^1(1,1)\}^{\text{ref}} = 0.1$
			$\{\Phi_1^2(2,1)\}^{\text{ref}} = 0.4$	$\{\Phi_2^2(2,1)\}^{\text{ref}} = 0.4$
Lag	$X^1 \xrightarrow{\tau=1} X^2$	$X^1 \xrightarrow{\tau=2} X^2$	$\{\Phi_1^2(1,1)\}^{\text{ref}} = 0.8$	$\{\Phi_2^2(1,2)\}^{\text{ref}} = 0.8$
			$\{\Phi_1^1(1,1)\}^{\text{ref}} = 0.2$	$\{\Phi_2^1(1,1)\}^{\text{ref}} = 0.2$
			$\{\Phi_1^2(2,1)\}^{\text{ref}} = 0.2$	$\{\Phi_2^2(2,1)\}^{\text{ref}} = 0.2$
Sign X^1	$X^1 \xrightarrow{\lvert a\rvert} X^1$	$X^1 \xrightarrow{-\lvert a\rvert} X^1$	$\{\Phi_1^1(1,1)\}^{\text{ref}} = 0.8$	$\{\Phi_2^1(1,1)\}^{\text{ref}} = -0.8$
			$\{\Phi_1^2(2,1)\}^{\text{ref}} = 0.2$	$\{\Phi_2^2(2,1)\}^{\text{ref}} = 0.2$
Sign X^1X^2	$X^1 \xrightarrow{\lvert a\rvert} X^2$	$X^1 \xrightarrow{-\lvert a\rvert} X^2$	$\{\Phi_1^2(1,1)\}^{\text{ref}} = 0.8$	$\{\Phi_2^2(1,1)\}^{\text{ref}} = -0.8$
			$\{\Phi_1^1(1,1)\}^{\text{ref}} = 0.2$	$\{\Phi_2^1(1,1)\}^{\text{ref}} = 0.2$
			$\{\Phi_1^2(2,1)\}^{\text{ref}} = 0.2$	$\{\Phi_2^2(2,1)\}^{\text{ref}} = 0.2$

TABLE III. Method parameters for low dimensional examples with $N_K = 2$ underlying regimes.

CI test	$\tau_{\max}$	α	α_{PC}	N_K	N_C	N_Q	N_A
ParCorr	3	0.01	0.2	2	40	20	50

ing to different sets of parents defined via the reference parameters $\{\Phi_k^j(i,\tau)\}^{\mathrm{ref}}$ given in columns 4–5 of Table II. Furthermore, synthetic regime-assigning processes $\{\Gamma(t)\}^{\mathrm{ref}}$ are generated for all examples. More specifically, $\{\gamma_1(t)\}^{\mathrm{ref}}$ is designed to consist of 41 alternating windows, i.e., $\{N_C\}^{\mathrm{ref}} = 40$ regime transitions. The length of these windows is randomly selected to be between 70 and 100 and the constraint (5) imposes $\{\gamma_2(t)\}^{\mathrm{ref}} = 1 - \{\gamma_1(t)\}^{\mathrm{ref}}$. The final length of the time series is capped at $T = 3000$ to ensure equally long regime assignment time series.

Then, an artificial time series $\mathbf{x}_t$ via (18) with $\{\sigma^2\}^{\mathrm{ref}} = 1$ is generated. Note that the stochastic process (18) can be exactly

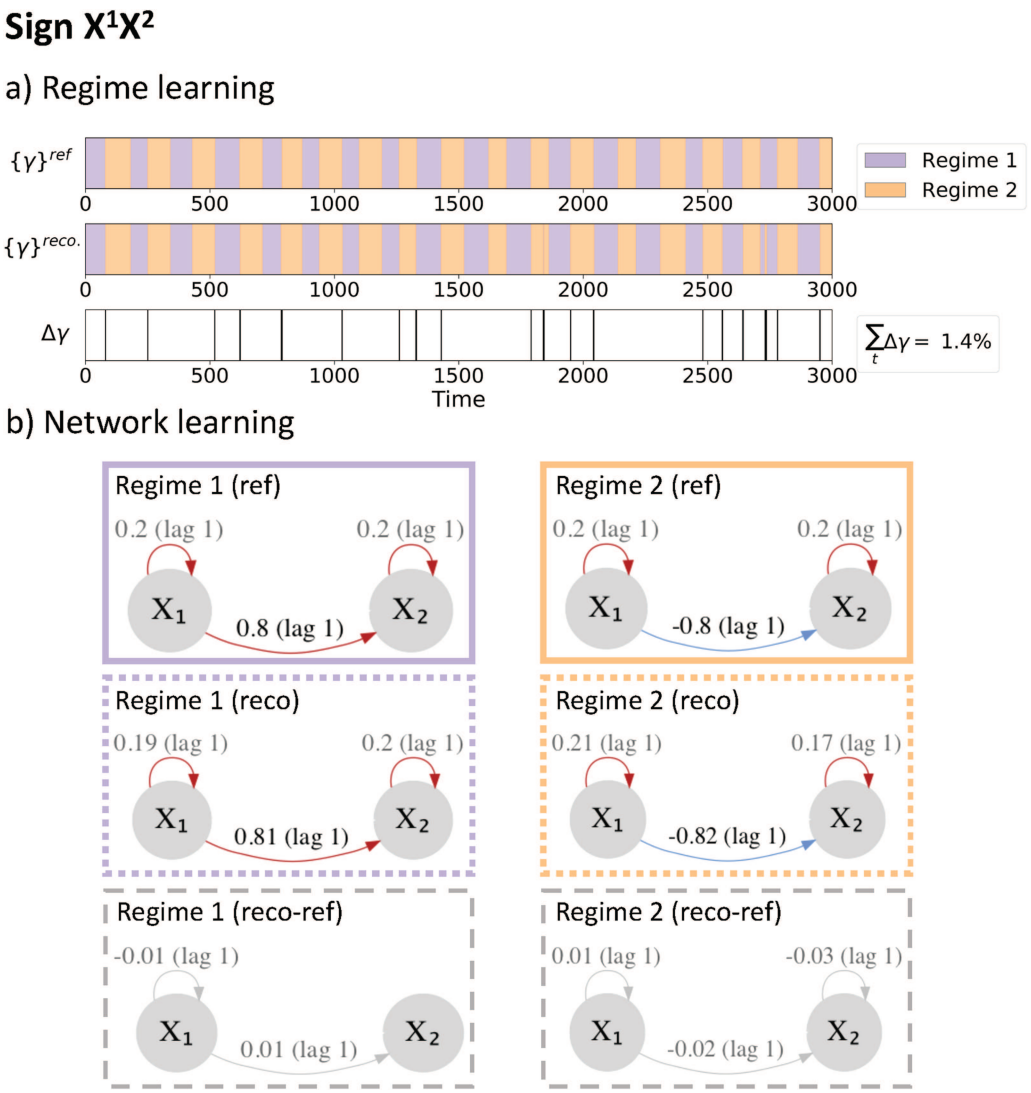

FIG. 2. Example case *Sign* X^1X^2. (a) The ground-truth regime-assigning process, $\{\gamma\}^{ref}$ (top), the Regime-PCMCI reconstructed process, $\{\gamma\}^{reco.}$ (middle), and the difference between the two, $\Delta\gamma$ (bottom). (b) The ground-truth networks for each regime (top), the Regime-PCMCI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

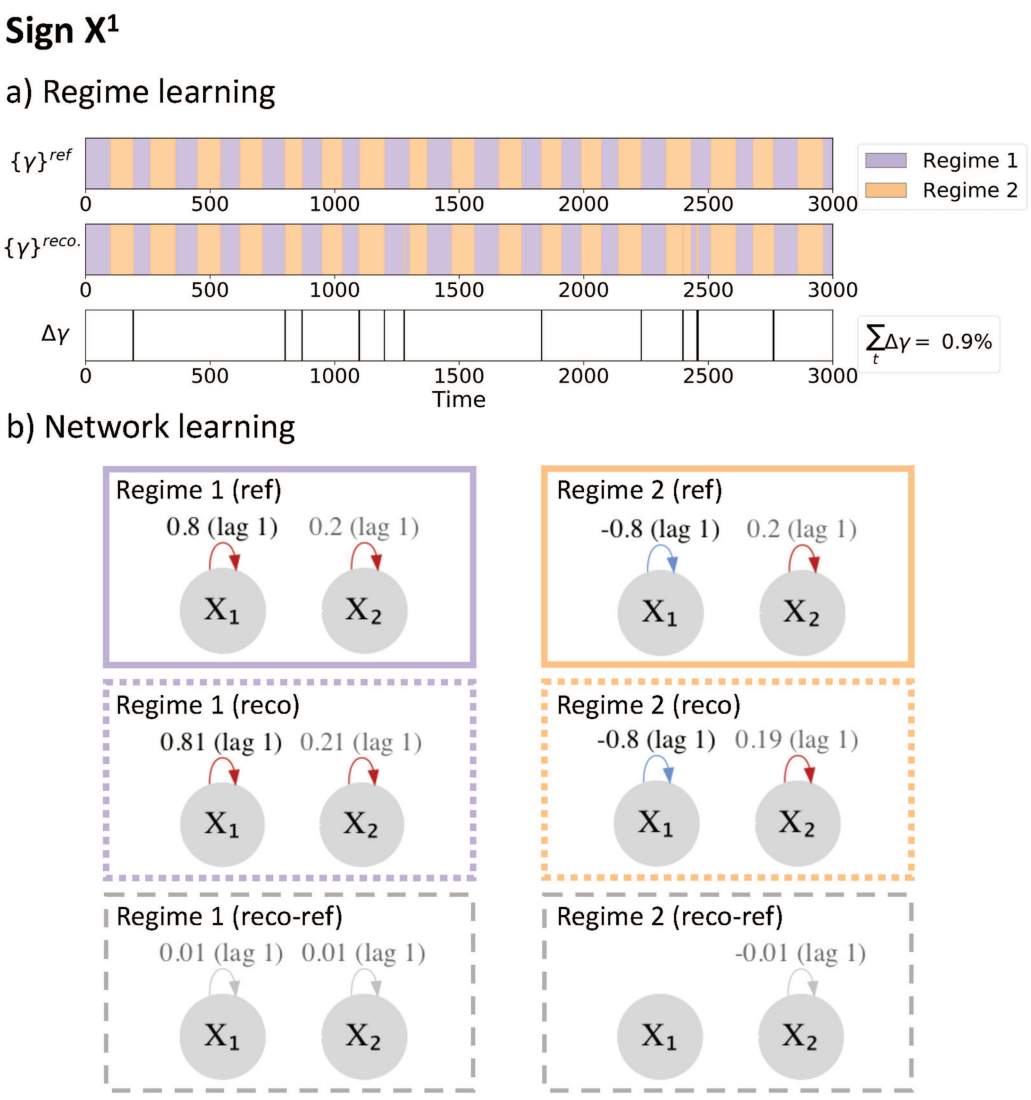

FIG. 3. Example case *Sign* X^1. See description in Fig. 2.

reconstructed via the coefficients $\{\Phi_k^j(i,\tau)\}^{\text{ref}}$, their activation $\{\Gamma(t)\}^{\text{ref}}$, and a specific realization of the innovation term η_t^j.

The PCMCI parameters are chosen as follows: partial correlation as a conditional independence test, $\alpha = 0.01$, $\alpha_{\text{PC}} = 0.2$ as recommended in Runge *et al.*[53] $\tau_{\max} = 3$, and masking type "y" (see the documentation of **tigramite** for the definition of masking types). The number of regimes was set to $N_K = 2$, and the maximum number of regime transitions is $N_C = 40$, i.e., correct guess on number of regimes and switches (model selection for N_K is investigated in Sec. IV C). The number of iterations is $N_Q = 20$, and the number of annealings is $N_A = 50$. A summary of the parameters is shown in Table III. We generate $N_R = 100$ time series realizations for each example.

2. Results

The ability of the proposed method to recover the networks and the regimes on the basis of the artificially designed time series are presented in the following. Figures 2–6 present results for each case in Table II, focusing on one of the N_R synthetic datasets. Table IV and Fig. 7 show summary statistics over all 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 panels a and b in Fig. 2; in the middle part of both panels, their Regime-PCMCI reconstruction is shown; and in the bottom part, the difference between reconstructed and true regimes is presented to visually inspect the accuracy. The reconstructed regime-assigning process for each regime matches the truth in 98.6% of time steps

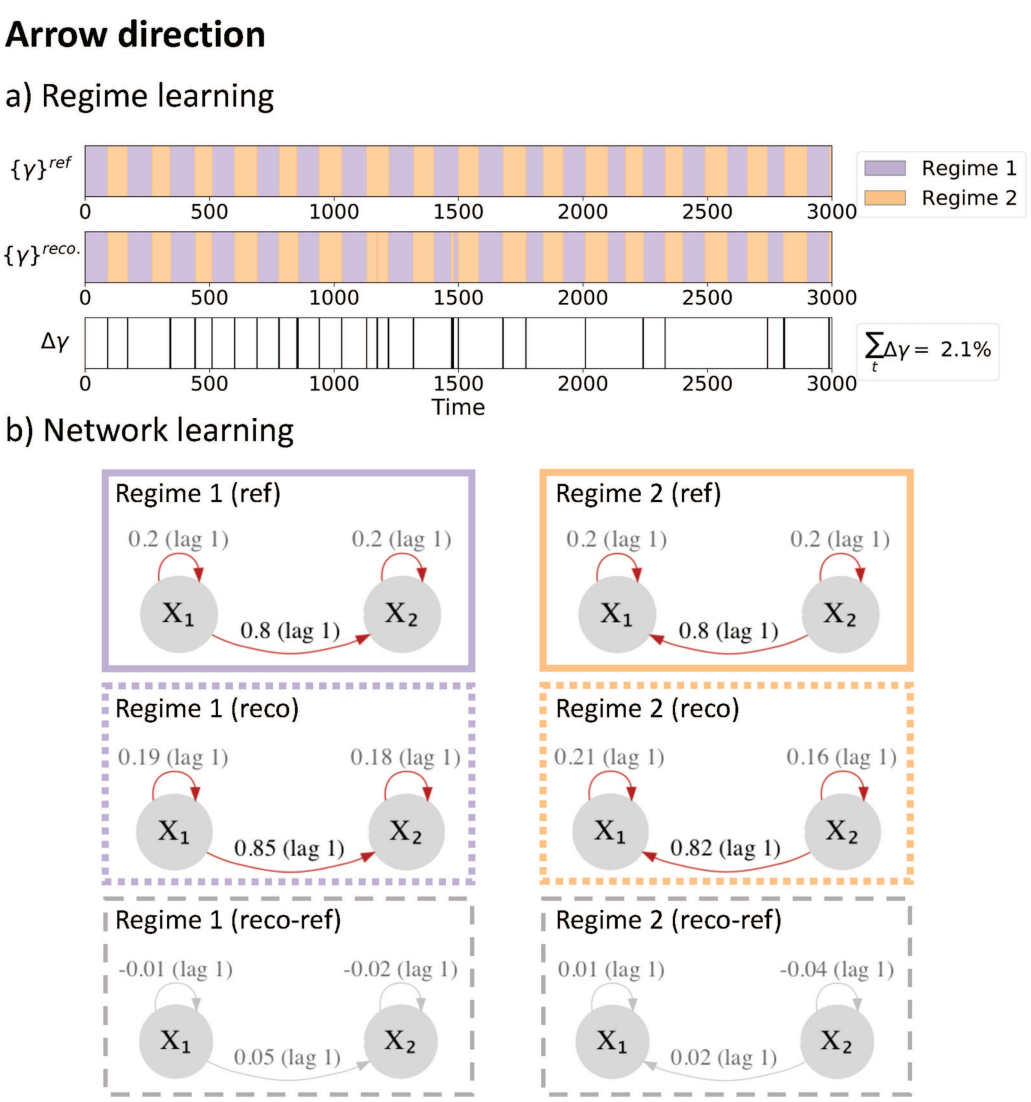

FIG. 4. Example case *Arrow direction*. See description in Fig. 2.

(97% average value over N_R, see Table IV). The corresponding networks have all and only the correct links (TPR = 0.99 and FPR = 0.01 average value over N_R); their linear causal effect is also well estimated with each link correct up to ± 0.02 [N_R-averaged error per link is 0.028 (9%)].

The other four cases are presented in Figs. 3–6. The case *causal effect*, and to a lesser extent *lag change*, are hardest to detect. This is because the difference between the individual regimes and a mixed state of the two is not very large and thus the detection is more challenging. This adds to the general challenge of non-convexity of the functional we are optimizing, which we mitigate by the annealing steps as mentioned in Sec. III D. A similar challenge is found for some high-dimensional runs for which we refer to Sec. IV D.

The average accuracy of Regime-PCMCI is estimated from N_R=100 synthetic datasets per each example and is presented in Fig. 7 and Table IV.

For a compact overview of the results and to facilitate the comparison between examples, Fig. 7 focuses on two key statistics: the precision of the reconstructed regime-assigning process (lightblue box plot, $\Delta\gamma$ %) and the precision of the reconstructed links' causal effects (pink box blot, $\Delta\Phi$%). $\Delta\gamma$% 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). $\Delta\Phi$% is the average difference between the reconstructed linear coefficient and the reference values of the ground-truth links expressed as percentage, i.e., each difference is weighted by the absolute value of the ground-truth coefficient. The precise definition of

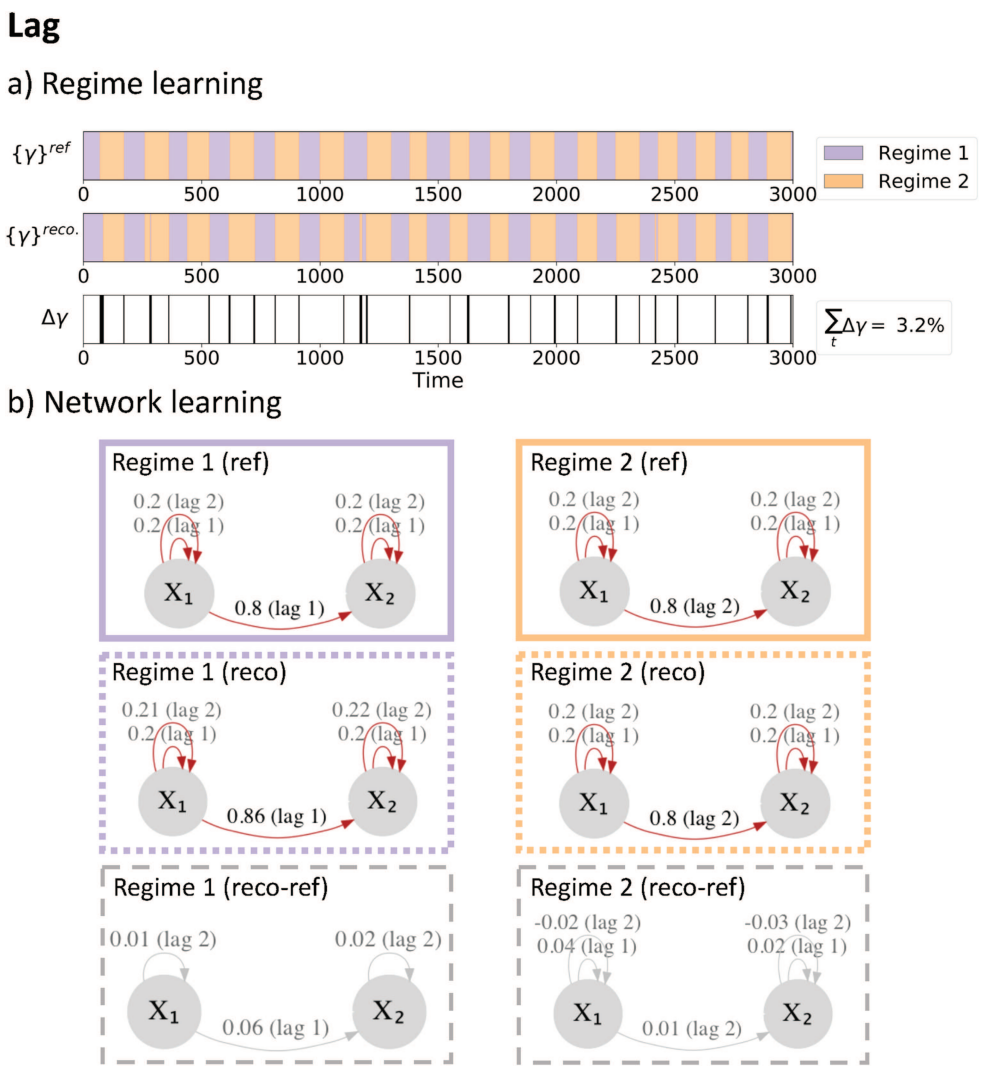

FIG. 5. Example case *Lag*. See description in Fig. 2.

the statistics can be found in Appendix B. These quantities provide a summary of the reconstructed regimes' accuracy, previously shown in the bottom parts of Figs. 2–6, and are presented for all N_R runs. The examples are ranked in the order of decreasing performance.

As already seen from inspection of single runs, change of *arrow direction* and of *causal effect* are the hardest to detect (the highest error in Fig. 7). It is also clear that a worsening performance in regime detection results in a higher network error. Except for *causal effect*, regime error $\Delta\gamma\%$ is between 1% and 7% and network error $\Delta\Phi\%$ is between 1% and 10%. Note that the average $\Delta\Phi\%$ (dark pink cross) is very close to $\Delta\Phi^{\text{ref}}\%$ (black dot), error for PCMCI runs with the ground-truth regime variable known (but causal structures unknown). This reference value sets the optimal baseline to which compare Regime-PCMCI performance and is met by our algorithm in all but one example.

Table IV shows a more detailed summary of the results over the N_R realizations. The estimation errors are presented in terms of the regime-assigning process (the second column), the network structure (third to sixth columns), the causal effects of links (seventh to tenth columns) and the overall reconstructed time series (the last column). The second column is $\Delta\gamma\%$, the average percentage of wrongly estimated time steps per regime, introduced above. In terms of networks, the link detection performance is evaluated via the true positive (TPR) and false positive rates (FPR). Furthermore, we compare these with the reference FPR and TPR (superscript *ref*) if PCMCI is run with the ground-truth regime variable known (but causal structure unknown). The accuracy in links' causal effects is

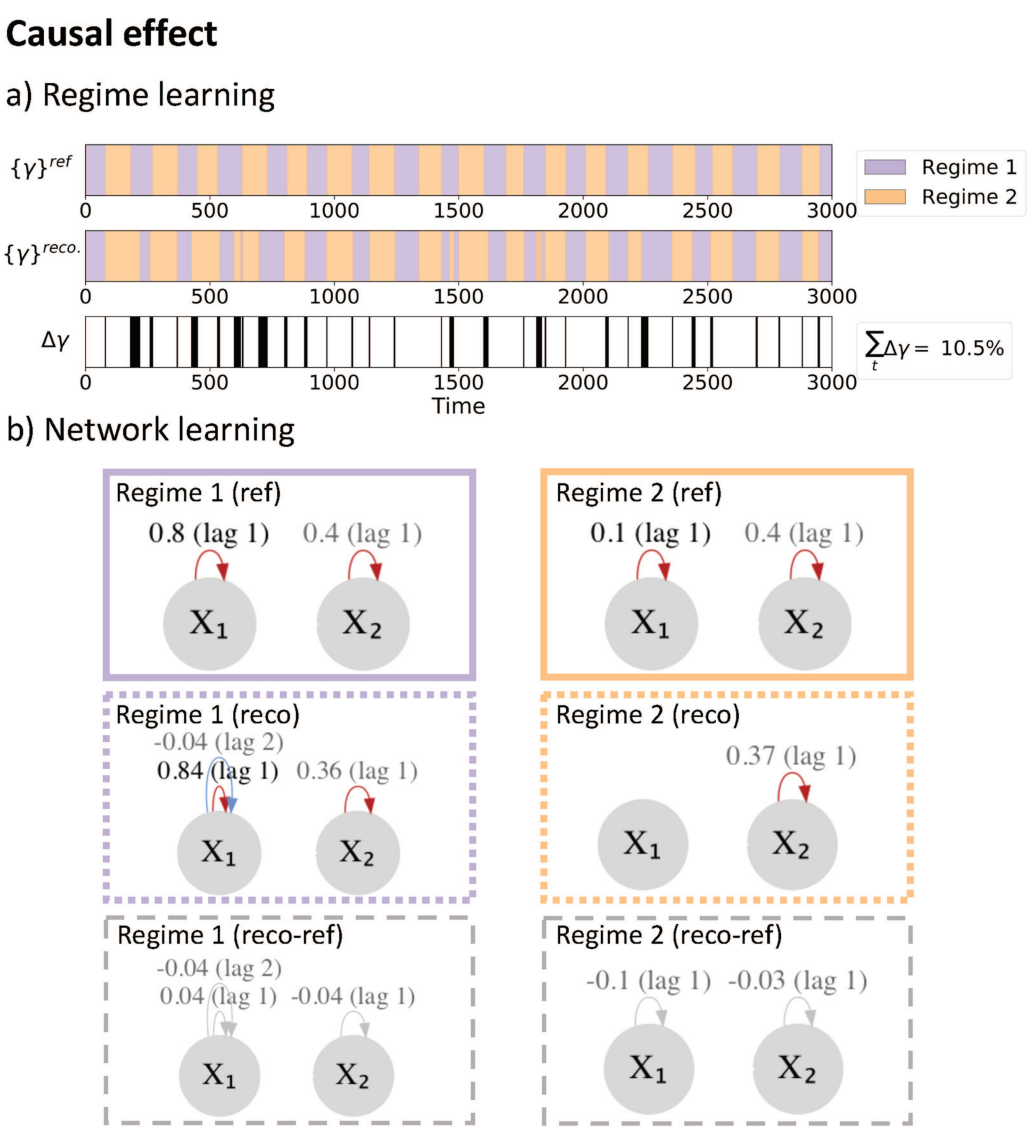

FIG. 6. Example case *Causal effect*. See description in Fig. 2.

assessed via $\Delta\Phi$, the average difference between the reconstructed linear coefficient and the reference values of the ground truth links, and in the percentage version ($\Delta\Phi$%, see above). 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. (4) and 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 Appendix B.

In summary, Table IV shows that

- $\Delta\gamma$%: on average, the regime-assigning process is reconstructed correctly in ∼94% of the time steps for all cases except *causal effect*. The *causal effect* and *lag* examples are the hardest to infer, with the 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 regimes. Therefore, the algorithm struggles to decide which time steps belong to which regime since they could fit both to some degree. Yet, also for *causal effect* there are seven instances where $\Delta\gamma < 15\%$ (one presented in Fig. 6) and those, as expected from PCMCI, give a 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 shows that runs that end up in mixed states can still fit the data quite well. Also, we notice that the causal effect setup reaches local minima in 16% of the 100 runs, thus in 84% of the runs the algorithm cannot easily

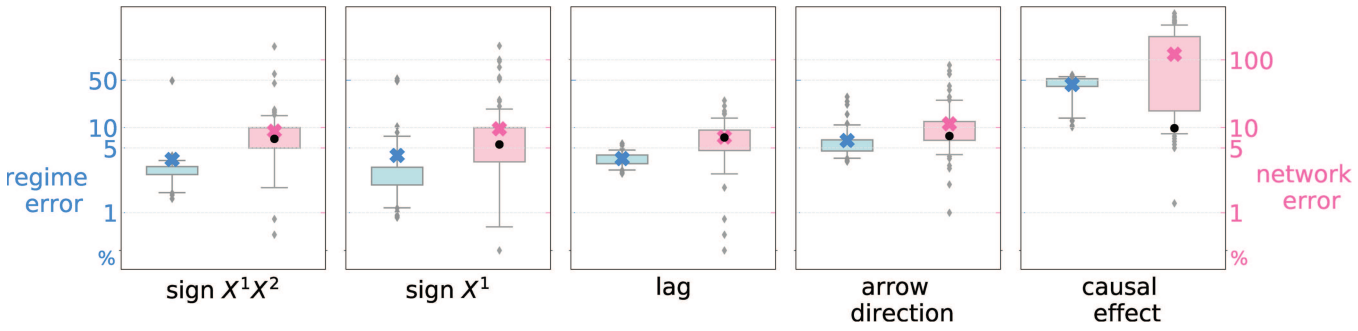

FIG. 7. A summary of the general performance of the Regime-PCMCI for the examples described in Sec. IV A. The performance skill is separated in *regime error* (defined as the average percentage of wrongly estimated time steps $\Delta\gamma$%, lightblue box plot) and *network error* (defined as the relative percentage difference between reconstructed and reference links' coefficients $\Delta\Phi$%, pink box plot). Box plots summarize the distribution of $N_R = 100$ runs (boxes between 25 and 75 percentiles and whiskers between 5 and 95 percentiles), and the mean is marked with a cross. The optimal baseline for the network fit, $\Delta\Phi^{\text{ref}}$%, is marked with a black dot (see definition in the text). Note the symlog scale on the y-axes. Synthetic datasets are generated according to Table II for $N_R = 100$ random ground-truth regime-assigning processes.

find a stable solution which 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, dynamicwise, are strong enough to be detectable.
- FPR: Ideally, the false positive rate should be upper-bounded by $\alpha = 0.01$. This is also the case if we assume the correct regimes (see column 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 the causal effects case, we expect to obtain also good results for the linear coefficients. This is indeed the case, as the difference is of order 10^{-2}, implying a relative error of about 10%. Also, this matches very closely the optimal baseline for the network fit $\Delta\Phi^{\text{ref}}$%.

B. Low dimensional data with three underlying regimes

To illustrate how Regime-PCMCI deals with more than two regimes, we also considered a toy time series based on three different causal regimes. It is, of course, possible to consider the case $N_K > 3$, yet in applications it is often desirable to infer a few prominent and 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 (parsimony).

The artificial time series is generated via a regime-dependent causal graph that is designed by combining two of the regimes settings presented in Sec. IV A, namely, *sign* X^1X^2 change and *arrow inversion* (for details, see Table V). The regime-assigning reference process $\{\Gamma\}^{\text{ref}}$ is generated by randomly choosing between different persistence lengths of 60, 70, and 80 time steps and iterating for 20 times. The algorithm is run with free parameters in Table VI.

Figure 8 shows the results. There are only minimal deviations from the true reference values, which confirms that the proposed method is capable to deal with $N_K > 2$. This also holds for the summary results over $N_R = 100$ runs presented in Table VII. 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 $N_K = 3$.

C. Regime parameter selection

We investigate how parameter selection of the number of regimes affects the results by means of the AICc scores defined in (17). We investigate two test scenarios of $\{N_K\}^{\text{ref}} = 2, 3$ for a selection of the examples defined in Secs. IV A and IV B. The PCMCI parameters are as in Secs. IV A and IV B, while $N_R = 29$, $N_Q = 20$

TABLE IV. Results for $N_K = 2$ experiments averaged over $N_R = 100$ realizations generated for each example described in Table II.

Example	$\Delta\gamma$%	TPR_{all}	$\text{TPR}^{\text{ref}}_{\text{all}}$	FPR_{all}	$\text{FPR}^{\text{ref}}_{\text{all}}$	$\Delta\Phi$	$\Delta\Phi^{\text{ref}}$	$\Delta\Phi$ %	$\Delta\Phi^{\text{ref}}$ %	$\hat{\epsilon}$
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 V. Artificial model configuration for a low dimensional example with $N_K = 3$ underlying regimes.

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 $X^1X^2$and *arrowdirection*	$X^1 \xrightarrow{\lvert a\rvert} X^2$	$X^1 \xrightarrow{-\lvert a\rvert} X^2$	$X^2 \xrightarrow{\lvert a\rvert} X^1$	$\{\Phi_1^2(1,1)\}^{\text{ref}} = 0.8$ $\{\Phi_1^1(1,1)\}^{\text{ref}} = 0.2$ $\{\Phi_1^2(2,1)\}^{\text{ref}} = 0.2$	$\{\Phi_2^2(1,1)\}^{\text{ref}} = -0.8$ $\{\Phi_2^1(1,1)\}^{\text{ref}} = 0.2$ $\{\Phi_2^2(2,1)\}^{\text{ref}} = 0.2$	$\{\Phi_3^1(2,1)\}^{\text{ref}} = 0.8$ $\{\Phi_3^1(1,1)\}^{\text{ref}} = 0.2$ $\{\Phi_3^2(2,1)\}^{\text{ref}} = 0.2$

and $N_A = 20$. The resulting AICc values are displayed in Fig. 9. The N_C value is changed adaptively for each N_K to ensure a similar N_M value for the different number of regimes, i.e.,

$$N_C(N_K) = \{N_C^{\text{ref}}\}\{N_K\}^{\text{ref}}/N_K \tag{19}$$

for $N_K > \{N_K\}^{\text{ref}}$. The choice of N_C is based on N_M which in real life application can be chosen according to the considered processes and data as a good estimate of the timescale of regime changes is often available. Nevertheless, it is also possible to chose N_C via an information criterion simultaneously with N_K (e.g., in the context of regime-dependent clustering Falkena *et al.*[40] or regime-dependent Markov regression de Wiljes *et al.*[38]) The reference value for the number of switches is on average (due to randomization of $\{\Gamma\}^{\text{ref}}$) $\{N_C^{\text{ref}}\} = 40$ for both $\{N_K\}^{\text{ref}} = 2, 3$.

We note that the lowest N_K at which the AICc plateaus is the ground-truth one. The plateau itself occurs due to the fact that only the links with non-zero causal effect values are counted toward the number of parameters. Thus, a higher number of regimes N_K does not necessarily result in an increase of the total number of parameters. In other words, the penalization is not becoming stronger with higher values of N_K. Concluding, it is clearly visible that no significant improvement is gained by increasing the number of N_K beyond the reference number of regimes. Since the entry point to the plateau reveals the reference number of regimes, it seems possible to face scenarios where the true number of regimes is unknown.

D. High-dimensional linear network

In this section, the algorithm is evaluated on high-dimensional datasets, 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 network structures are randomly generated from a family of linear networks defined via the parameters shown in Table VIII, where L is the number of randomly drawn cross-variable links with random coefficients from the third column. Note that each variable is also auto-linked at lag 1 with coefficient randomly drawn from the fourth column. The time series $\mathbf{x}_t \in \mathrm{R}^{10}$ are generated with model (18) and for $N_R = 70$ realizations. Regime-PCMCI is then run with the settings shown in Table IX.

TABLE VI. Method parameters for a low dimensional example with $N_K = 3$ underlying regimes.

CI test	$\tau_{\max}$	α	α_{PC}	N_K	N_C	N_Q	N_A
ParCorr	3	0.01	0.2	3	40	20	50

The results are shown in Table X, which is structured like Table IV except for TPR and FPR being estimated for the cross-variables links thus focusing on the connections between variables. All links are considered in $\Delta\Phi$. 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 $\Delta\gamma = 11.7\%$ (the second row in Table X). The other seven runs are responsible for most of the deviation of the average statistics from the reference values (the 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 the regime-assigning process $\Delta\gamma$, meaning that we cannot use a filtering on $\hat{\varepsilon}$ to find the best performing runs. This behavior can be explained from the tendency of the algorithm to still over-fit when too many degrees of freedom are available, as well as from the complexity of distinguishing different causal effects (a challenge already manifested in the *causal effect* case).

E. Computational complexity

Table XI shows some indicators of the performance of the method: the fraction of N_R runs that correspond to a (local) minima, the average number of q-iterations needed to reach a local minima and the runtime for the whole N_R set of runs (the code run parallel over the N_A annealings and using 4–6 CPUs per job).

Most of the examples reach local minima in more than 50% of the N_R runs, while the percentage is very low for the *causal effect* (the second column). We note that examples with a high percentage of local minima correspond also to quick convergence in terms of iteration steps (the third column). They are also associated with better regimes reconstruction (see Tables IV, VII, and X), confirming that a clear cost functional minimum (as shown from the second and third columns) is linked to better detection. Finally, the runtime is quite fast: the low dimensional examples take between 10 and 20 min for $N_K = 2$ and 45 min for $N_K = 3$ to complete 100 runs. The high-dimensional example takes just below 3 h for 70 runs.

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 Sec. I between the winter and summer months, i.e., the background regimes, and to detect a reported link from ENSO to AIR during summer.

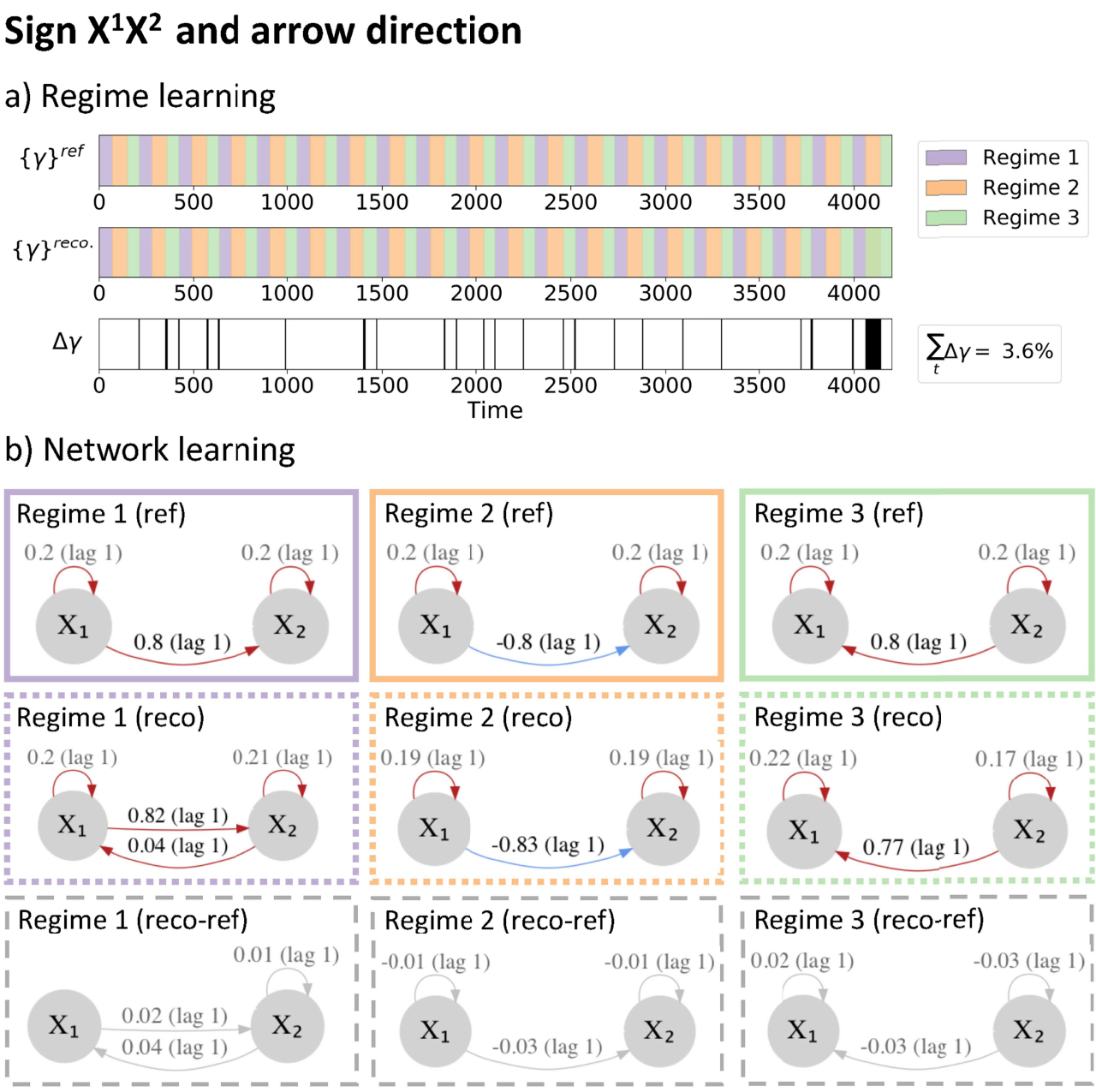

FIG. 8. Example with $N_k = 3$ regimes for the case *Sign* X^1X^2 and *arrow direction*. See description in Fig. 2 but with three regimes.

This example can be considered a difficult case since the expected signal from ENSO to AIR is likely small compared to natural variability.[16] Furthermore, climate data are typically very noisy with causal relationships being diluted by other, often unknown processes given a complex coupled climate system.[43]

Our input data consist of monthly observations of ENSO and AIR, for the years 1871–2016, 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).[54,55] Data for AIR anomalies (with the climatology subtracted) are provided by the Indian Institute of Tropical Meteorology (IITM).[56,57]

We choose the following parameters of Regime-PCMCI: for the regime part, we set $N_K = 2$ and $N_C = 292$, which is equivalent to

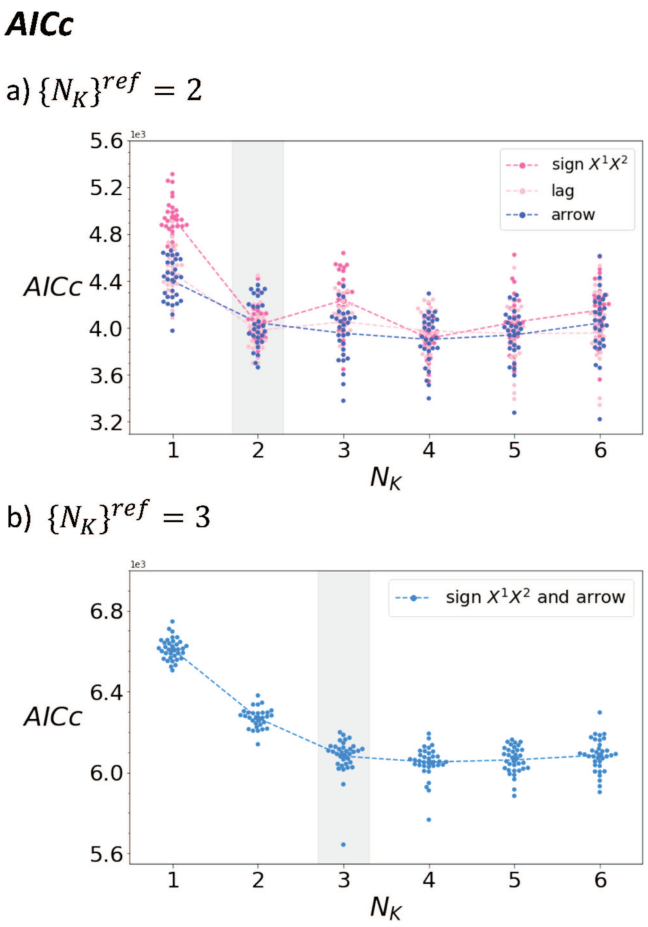

FIG. 9. Numerical investigation of AICc values for runs with different N_K and (a) $\{N_K\}^{\text{ref}} = 2$ for three network examples (*sign* X^1X^2, *arrow*, and *lag* change) and (b) $\{N_K\}^{\text{ref}} = 3$ for the *sign* X^1X^2 and *arrow* change example. In each example, individual dots represent the value attained by the $N_R = 29$ runs, and the dashed line goes through the mean values of each set. The vertical gray bar highlights the ground-truth number of regimes $\{N_K\}^{\text{ref}}$.

TABLE VII. Results for $N_K = 3$ experiments averaged over $N_R = 100$ realizations generated for each example described in Table V.

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

TABLE VIII. High-dimensional network parameters.

N_X	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

assuming two seasons per year. For the PCMCI settings, we use a significance level $\alpha = 0.01$ ($\alpha_{\text{PC}} = 0.2$). Furthermore, we use a maximum time lag of 2 months, i.e., $\tau_{\max} = 2$. The optimization is run $N_A = 100$ annealing times, to span many local minima, with each annealing allowed for up to $N_Q = 100$ iteration steps to converge.

Among the annealing steps, which correspond to different random initial guesses on the regime-assigning process Γ, some clearly performed better in terms of fitting the data. We estimate the average prediction error associated with each annealing, $\hat{\varepsilon}$ (B 4), and Fig. 10(a) shows it for all annealings (ranked according to $\hat{\varepsilon}$). A red box highlights the top performing cluster (13 runs).

All of the top 13 annealings 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 annealings and plot links that surpass a strength of 0.1.

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 [Fig. 10(c)]. This negative dependence is well documented in the literature.[16] During regime 2, in contrast, ENSO and AIR are, on average, almost independent, with only a very weak link (−0.05, not shown) detected from AIR to ENSO. More importantly, our results indicate a clear seasonal dependence. Figure 10(d) shows the number of months assigned to each regime (normalized by the number one would expect on the hypothesis of no seasonality, see the figure caption). A clear peak in summer months is found for regime 1. More precisely, most of the months between June and 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.

A method to detect long-term changes of this summer teleconnection has recently been proposed by Bódai *et al.*[58] using ensemble climate models. The additional dimension provided by the ensemble members' allows to compute year-dependent correlations to infer inter-annual changes. In contrast, our method uses a single realization of the dynamics (e.g., observations) to still obtain time-dependent statistics (networks), although in a finite number (N_K). Note that long-term changes may still be detectable with Regime-PCMCI either as long-term changes to the persistence/start of a regime each year or with the emergence of a regime in a specific time period.

Overall, these results are promising and show the potential of Regime-PCMCI to detect regime-dependent causal structures in a

TABLE IX. Method parameters for high-dimensional experiments with two underlying regimes.

CI test	$\tau_{\max}$	α	α_{PC}	N_K	N_C	N_Q	N_A
ParCorr	4	0.05	0.2	2	49	30	50

TABLE X. Results for high-dimensional experiments over $N_R = 70$ realizations generated for each example described in Table VIII.

Selection	$\Delta\gamma\%$	TPR_{cros}	$\text{TPR}^{\text{ref}}_{\text{cros}}$	FPR_{cros}	$\text{FPR}^{\text{ref}}_{\text{cros}}$	$\Delta\Phi$	$\Delta\Phi^{\text{ref}}$	$\Delta\Phi$ %	$\Delta\Phi^{\text{ref}}$ %	$\hat{\epsilon}$	no. of runs
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

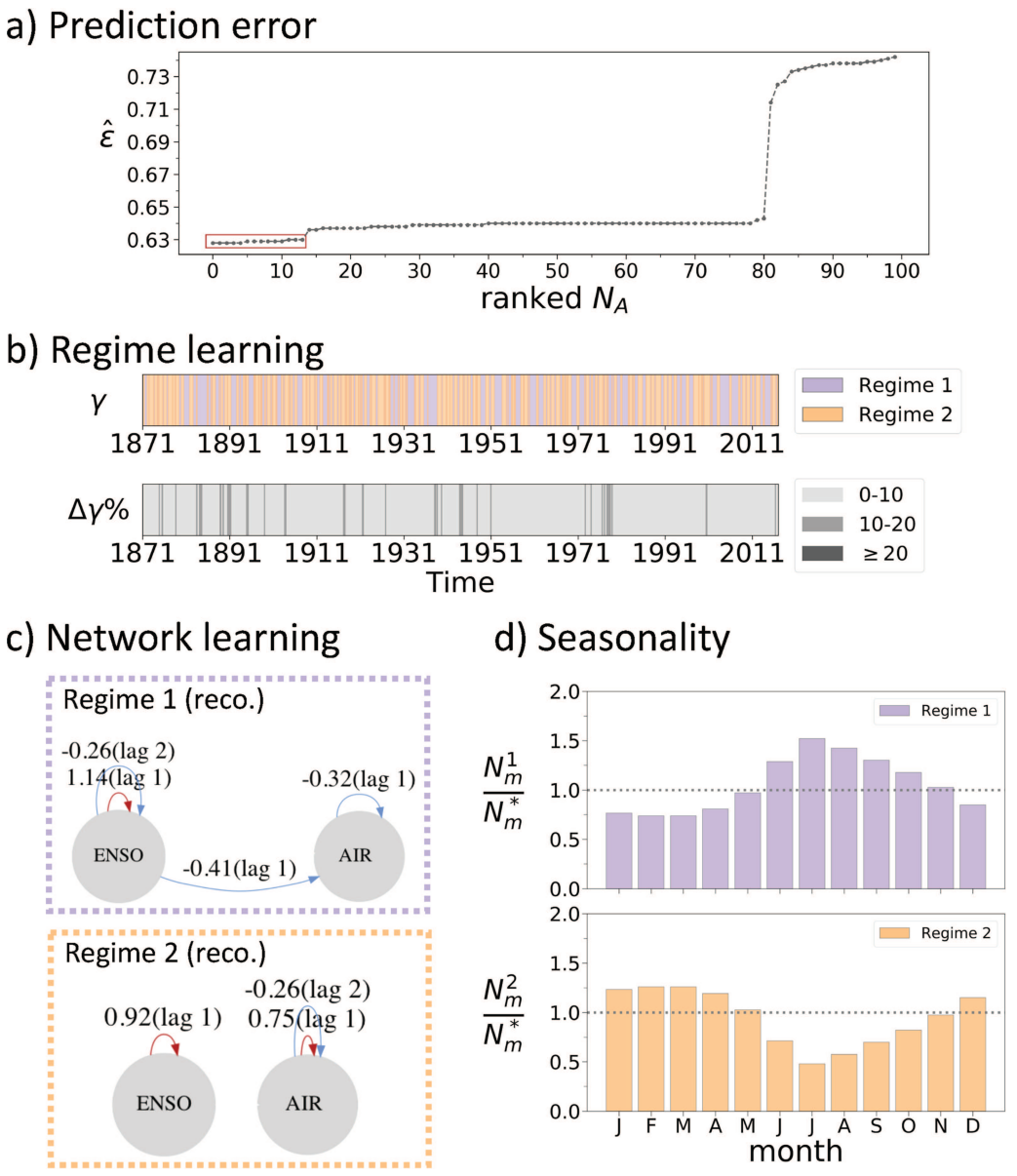

FIG. 10. Climate example. (a) Prediction error for each annealing step in the ascending order, lowest 13 annealings highlighted in red box. All the other panels refer to this selection. (b) Regime learning: regime-assigning process corresponding to the best annealing (rank 0) (top) and departure from this estimate of the remaining best 12 annealings (in percentage difference). (c) Network learning: mean networks per regime, each causal effect is the mean of the corresponding coefficient in the individual 13 annealings. (d) Seasonality of the regimes: Number of years per month m assigned to each regime (N^k_m), normalized by N^*_m, which refers to the expected number of months assigned to a given regime if one assumes equal probability $1/N_K$ of assigning a month to one of the two regimes. Thus, here, $N^*_m = 13 \cdot T/(12 * N_K)$.

TABLE XI. Summary performance statistics of all examples. The third column is the average value over the respective N_R.

Example	No. of local minima/N_R (%)	Iterations to minima	Runtime (s)
Arrow direction	92	7	600
Causal effect	16	13	970
Lag	60	11	1130
Sign X^1	52	12	970
Sign X^1X^2	70	9	700
Sign X^1X^2 and *arrow*	56	10	2670
High dimensional	92	6	10 780

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 (N_C and N_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 CONCLUSIONS

Causal discovery is emerging as an important framework across many disciplines in science and engineering, but each discipline has particular challenges that novel methods need to address.[4] We introduced a novel method, Regime-PCMCI, to learn regime-dependent causal relations, overcoming one of the key drawbacks of current causal recovery methods. The performance of Regime-PCMCI was analyzed for many different artificially generated causal scenarios and for varying regimes showing that the method covers a wide range of settings (see Figs. 2–5, 7 and Table IV). The performance of the algorithm is maintained also for high-dimensional settings with 10 variables (see Table X) as well as for more than two regimes (see Fig. 8 and Table VII). We found limitations of the method for the case where only the causal effect strength of a link changes between regimes (see Fig. 6), which seems to be hard to detect with our optimization scheme and requires further investigation. Furthermore, the capability of Regime-PCMCI was verified by means of a well-documented climate example using real data of ENSO and Indian rainfall (see Fig. 10). Overall, the proposed method presents itself as a promising approach in the context of non-stationary causal links manifested in regime changes in time.

Note that a causal interpretation of estimated links in our observational causal discovery framework still assumes causal sufficiency, that is, no unobserved common causes. However, estimated non-causality (zero coefficients) does not require this assumption[27] and can be interpreted as an absence of a causal relation already under the weaker faithfulness assumption.[26] While for PCMCI asymptotic consistency was shown,[26] this is a more difficult task for Regime-PCMCI and deferred to further research.

There are several interesting aspects that could be explored in the future, building on the present work. These extensions can build on other causal discovery algorithms or extensions of PCMCI in the causal discovery step of our method. For example, the PCMCI algorithm allows for nonlinear causal links[26] and thus a nonlinear extension of the Regime-PCMCI is a logical next step, e.g., Gaussian processes are used to estimate g_k^j, then the Gaussian Process Distance Correlation (GPDC) test (see Runge *et al.*[26]) could potentially be used. In that, yet the Regime-PCMCI version would require a different cost functional and optimization approach. Recent extensions of PCMCI to the case of not only lagged, but also contemporaneous causal relations can also be integrated.[41] Moreover, potentially it is also possible to better capture the *causal effect* case and it might be possible to learn a regime dependence of the noise term.

With respect to applications in climate science, it would be interesting to utilize the proposed method to study other links in the climate system that are likely regime dependent, but less understood than the presented El Niño-Indian rainfall example. Since Regime-PCMCI is formulated in general terms that are not only specific to climate datasets, problems of causal non-stationarity in other application areas could be explored.

AUTHORS' CONTRIBUTIONS

E.S., J.d.W., M.K., and J.R. designed the research; E.S. mainly performed the research; and E.S., J.d.W., M.K., and J.R. analyzed the results and wrote the manuscript.

ACKNOWLEDGMENTS

E.S. was supported by the Centre for Doctoral Training in Mathematics of Planet Earth, UK EPSRC funded (Grant No. EP/L016613/1). M.K. and J.d.W. have been partially funded by the ERC Advanced Grant ACRCC (Grant No. 339390). The research of J.d.W. has been partially funded by Deutsche Forschungsgemeinschaft (DFG) (Grant No. SFB1294/1-318763901) and by the Simons CRM Scholar-in-Residence Program. M.K. has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement (No. 841902). The optimization software by Gurobi™ was used for this work. The authors would like to thank Giorgia di Capua for discussions on the climate example.

NOMENCLATURE

AIC	Akaike information criterion)
AICc	Corrected Akaike information criterion
ENSO	El Niño Southern Oscillation
FPR	False positive rate
MCI	Momentary conditional independence
PCMCI	Causal discovery method[26]
RAM	Regime-dependent autoregressive model
SCM	Structural causal model
TPR	True positive rate

APPENDIX A: HETEROGENEOUS NOISE

In the general framework laid out in Eq. (1), the noise variables η_t^j are only assumed to be jointly independent and stationary, each distributed according to a distribution $\mathscr{D}^j$. Given that the primary focus of this work is to detect regime-dependent causal structures rather than noise structures, the effective choice for noise distributions used to generate the data is a Gaussian with unit variance $\eta_t^j \sim \mathcal{N}(0, 1)$ for all variables j (Sec. IV).

TABLE XII. Results, for example, *sign* X_1X_2 averaged over $N_R = 100$ realizations, for each noise variances combination described in Appendix A 1.

Case	$\Delta\gamma$%	$\mathrm{TPR}_{\mathrm{all}}$	$\mathrm{TPR}^{\mathrm{ref}}_{\mathrm{all}}$	$\mathrm{FPR}_{\mathrm{all}}$	$\mathrm{FPR}^{\mathrm{ref}}_{\mathrm{all}}$	$\Delta\Phi$	$\Delta\Phi^{\mathrm{ref}}$	$\Delta\Phi$ %	$\Delta\Phi^{\mathrm{ref}}$ %
$\sigma_2 = 0.25$	0.3	1.0	1.0	0.01	0.01	0.010	0.010	5.2	5.2
$\sigma_2 = 0.5$	0.8	1.0	1.0	0.02	0.01	0.013	0.013	5.3	5.2
$\sigma_2 = 2.0$	24.0	0.85	1.0	0.02	0.01	0.16	0.03	36.0	9.0

Yet, this simplification does not necessarily represent the variability of processes in real-world scenarios. Here, the performance of the proposed Regime-PCMCI is exemplified for Gaussian noises with variable-specific variances $\eta_t^j \sim \mathcal{N}(0, \sigma_j^2)$ (see Appendix A 1) and noises from two different distributions, Gaussian and uniform (see Appendix A 2).

1. Gaussian noises with variable-specific variances

The data are generated from model (18) with example *sign change* X^1X^2 coefficients. The noise terms η_j are Gaussian distributed with a fixed variance for variable X_1, $\mathscr{D}^1 = \mathcal{N}(0, 1)$, and three different cases for variable X_2, $\mathscr{D}^2 = \mathcal{N}(0, \sigma_2^2)$ with $\sigma_2 = 0.25$, $\sigma_2 = 0.5$, and $\sigma_2 = 2.0$. The Regime-PCMCI results, averaged over 100 different realizations of the regime-assigning processes, are presented in Table XII. The algorithm performs very well in the first two cases (average regime detection error $\Delta\gamma \leq 1\%$). This is to be expected since a smaller noise in X^2 allows for a better fit of the data. The latter case is harder to infer since the noise on X^2 is very large compared to the deterministic signal (here $\Delta\gamma \simeq 25\%$).

2. Different noise distributions

The data are generated from model (18) with example *sign change* X^1X^2 coefficients. The noise terms η_j are set to follow completely different distributions: variable X^1 is associated with a unit variance Gaussian noise, $\mathscr{D}^1 = \mathcal{N}(0, 1)$, and variable X^2 with uniformly distributed noise between ± 1.5, $\mathscr{D}^2 = \mathcal{U}(-1.5, 1.5)$. The Regime-PCMCI results, averaged over 100 different realizations of the regime-assigning processes, are presented in Table XIII. This scenario gives results comparable to the ones presented in the paper for the same example, i.e., $\Delta\gamma \simeq 3\%$.

To summarize, the results show that Regime-PCMCI can deal with specific heterogeneous noise distributions, even belonging to different families of distributions. Since the optimization method acts on regression residuals, we can speculate that we expect good performance as long as the noise terms are not too large in their magnitude and are not too skewed. An elaborate study of these conclusions and an investigation of the potential for generalization of the method to more extreme noise distributions is an interesting research aspect for the future.

APPENDIX B: DEFINITION OF RESULT STATISTICS

The definitions for the statistics presented in Tables IV, VII, and X are outlined as follows:

1. Regime-assigning process

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

2. Link detection

TPR

$$\mathrm{TPR} = \frac{\mathrm{TP_X}}{\mathrm{P_X}}.$$

Over the cross-variables links (in Table X):

$$\mathrm{TP_{cros}} = |\{(i,j,\tau) : \{\Phi_k^j(i,\tau)\}^{reco.} \neq 0 \,\&\, \{\Phi_k^j(i,\tau)\}^{ref} \neq 0 \,\&\, i \neq j\}|,$$

$$\mathrm{P_{cros}} = |\{(i,j,\tau) : \{\Phi_k^j(i,\tau)\}^{ref} \neq 0 \,\&\, i \neq j\}|.$$

And over all links (in Tables IV and VII):

$$\mathrm{TP_{all}} = |\{(i,j,\tau) : \{\Phi_k^j(i,\tau)\}^{reco.} \neq 0 \,\&\, \{\Phi_k^j(i,\tau)\}^{ref} \neq 0\}|,$$

$$\mathrm{P_{all}} = |\{(i,j,\tau) : \{\Phi_k^j(i,\tau)\}^{ref} \neq 0\}|.$$

FPR

$$\mathrm{FPR} = \frac{\mathrm{FP_X}}{\mathrm{N_X}}.$$

Over the cross-variables links (in Table X):

$$\mathrm{FP_{cros}} = |\{(i,j,\tau) : \{\Phi_k^j(i,\tau)\}^{reco.} \neq 0 \,\&\, \{\Phi_k^j(i,\tau)\}^{ref} = 0 \,\&\, i \neq j\}|,$$

$$\mathrm{N_{cros}} = |\{(i,j,\tau) : \{\Phi_k^j(i,\tau)\}^{ref} = 0 \,\&\, i \neq j\}|.$$

TABLE XIII. Results, for example, *sign* X_1X_2 averaged over $N_R = 100$ realizations, for different noise distributions described in Appendix A 2.

Case	$\Delta\gamma$%	$\mathrm{TPR}_{\mathrm{all}}$	$\mathrm{TPR}^{\mathrm{ref}}_{\mathrm{all}}$	$\mathrm{FPR}_{\mathrm{all}}$	$\mathrm{FPR}^{\mathrm{ref}}_{\mathrm{all}}$	$\Delta\Phi$	$\Delta\Phi^{\mathrm{ref}}$	$\Delta\Phi$ %	$\Delta\Phi^{\mathrm{ref}}$ %
Gauss, Unif	3	1.0	1.0	0.01	0.01	0.025	0.019	8.0	7.0

And over all links (in Tables IV and VII):

$$\mathrm{FP_{all}} = |\{(i,j,\tau) : \{\Phi^j_k(i,\tau)\}^{reco.} \neq 0 \,\&\, \{\Phi^j_k(i,\tau)\}^{ref} = 0\}|,$$

$$\mathrm{N_{all}} = |\{(i,j,\tau) : \{\Phi^j_k(i,\tau)\}^{ref} = 0\}|.$$

3. Link coefficients

$$\Delta\Phi = \frac{1}{N_K}\sum_{k=1}^{N_K} \frac{\sum_j \sum_{X^i_{t-\tau}\in\mathscr{P}^j_k} |\{\Phi^j_k(i,\tau)\}^{reco.} - \{\Phi^j_k(i,\tau)\}^{ref}|}{\sum_j |\mathscr{P}^j_k|},$$

can also be computed as average *percentage* error per regime,

$$\Delta\Phi(\%) = \frac{1}{N_K}\sum_{k=1}^{N_K} \frac{\sum_j \sum_{X^i_{t-\tau}\in\mathscr{P}^j_k} \frac{|\{\Phi^j_k(i,\tau)\}^{reco.} - \{\Phi^j_k(i,\tau)\}^{ref}|}{\{\Phi^j_k(i,\tau)\}^{ref}}}{\sum_j |\mathscr{P}^j_k|} \times 100.$$

4. Prediction error

$$\hat{\varepsilon} \equiv \frac{1}{N_X T}\sum_t \sum_j |\{x^j(t)\}^{ref} - \{x^j(t)\}^{reco.}| \approx \sqrt{\frac{\mathbf{L}}{N_X \cdot T}},$$

with **L** defined in Eq. (4).

DATA AVAILABILITY

The data that support the analysis of the first four sections of this study have been synthetically generated by the authors and can be fully reproduced using the equations and parameters described in the article. The data that support the findings of the last section of this study are openly available in the KNMI Climate Explorer at https://climexp.knmi.nl/, Refs. 55 and 57. PCMCI is part of the open-source Python package tigramite available at https://github.com/jakobrunge/tigramite, Ref. 59.

REFERENCES

[1] J. Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge University Press, New York, 2000).
[2] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search* (MIT Press, Boston, 2000).
[3] I. Ebert-Uphoff and Y. Deng, "Causal discovery for climate research using graphical models," J. Clim. **25**, 5648–5665 (2012).
[4] J. Runge, V. Petoukhov, and J. Kurths, "Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models," J. Clim. **27**, 720–739 (2014).
[5] X. S. Liang, "Unraveling the cause-effect relation between time series," Phys. Rev. E **90**, 52150 (2014).
[6] M. Kretschmer, D. Coumou, J. F. Donges, and J. Runge, "Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation," J. Clim. **29**, 4069–4081 (2016).
[7] 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," in *Nonlinear and Stochastic Climate Dynamics* (Cambridge University Press, 2017), pp. 184–208.
[8] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, E. H. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler, "Inferring causation from time series in earth system sciences," Nat. Commun. **10**, 2553 (2019).
[9] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," Econometrica **37**, 424–438 (1969).
[10] M. Droumaguet, A. Warne, and T. Woźniak, "Granger causality and regime inference in Markov switching VAR models with Bayesian methods," J. Appl. Econom. **32**, 802–818 (2016).
[11] S. Gerber and I. Horenko, "On inference of causality for discrete state models in a multiscale context," Proc. Natl. Acad. Sci. U.S.A. **111**, 14651–14656 (2014).
[12] A. Strandburg-Peshkin, D. Papageorgiou, M. C. Crofoot, and D. R. Farine, "Inferring influence and leadership in moving animal groups," Philos. Trans. R. Soc. B Biol. Sci. **373**, 20170006 (2018).
[13] S. I. Seneviratne, T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling, "Investigating soil moisture-climate interactions in a changing climate: A review," Earth Sci. Rev. **99**, 125–161 (2010).
[14] S. Nakayama, J. L. Harcourt, R. A. Johnstone, and A. Manica, "Initiative, personality and leadership in pairs of foraging fish," PLoS ONE **7**, 1–7 (2012).
[15] G. Muradoglu, F. Taskin, and I. Biga, "Causality between stock returns and macroeconomic variables in emerging markets," Russ. East Eur. Finance Trade **36**, 33–53 (2000).
[16] P. J. Webster and T. N. Palmer, "The past and the future of El Niño ," Nature **390**, 562–564 (1997).
[17] J. Shaman and E. Tziperman, "Summertime enso–north african–asian jet teleconnection and implications for the indian monsoons," Geophys. Res. Lett. **34**, L11702, https://doi.org/10.1029/2006GL029143 (2007).
[18] 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," Clim. Dyn. **44**, 1003–1015 (2015).
[19] L. Barnett and A. K. Seth, "Granger causality for state space models," Phys. Rev. E **91**, 040101 (2015).
[20] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, Cambridge, 2010).
[21] D. M. Chickering, "Learning equivalence classes of Bayesian-network structures," J. Mach. Learn. Res. **2**, 445–498 (2002).
[22] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorixthms* (MIT Press, Cambridge, MA, 2017), pp. 1214–1216.
[23] M. Porfiri and M. R. Marin, "Inference of time-varying networks through transfer entropy, the case of a Boolean network model," Chaos **28**, 103123 (2018).
[24] G. Sugihara, R. May, H. Ye, C.-H. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems," Science **338**, 496–500 (2012).
[25] J. Arnhold, P. Grassberger, K. Lehnertz, and C. Elger, "A robust method for detecting interdependences: Application to intracranially recorded EEG," Physica D **134**, 419–430 (1999).
[26] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," Sci. Adv. **5**, eaau4996 (2019).
[27] J. Runge, "Causal network reconstruction from time series: From theoretical assumptions to practical estimation," Chaos **28**, 075310 (2018).
[28] D. Malinsky and P. Spirtes, "Learning the structure of a nonstationary vector autoregression," Proc. Mach. Learn. Res. **89**, 2986–2994 (2019).
[29] K. Zhang, B. Huangy, J. Zhang, C. Glymour, and B. Schölkopf, "Causal discovery from Nonstationary/heterogeneous data: Skeleton estimation and orientation determination," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (ACM, 2017), pp. 1347–1353.

[30]J. Peters, P. Bühlmann, and N. Meinshausen, "Causal inference by using invariant prediction: Identification and confidence intervals," J. R. Stat. Soc. Ser. B (Stat. Methodol.) **78**, 947–1012 (2016).
[31]R. Christiansen and J. Peters, "Switching regression models and causal inference in the presence of discrete latent variables," J. Mach. Learn. Res. **21**, 41:1–41:46 (2020).
[32]V. Mwaffo, J. Keshavan, T. Hedrick, and S. Humbert, "Detecting intermittent switching leadership in coupled dynamical systems," Sci. Rep. **8**, 10338 (2018).
[33]S. Butail and M. Porfiri, "Detecting switching leadership in collective motion," Chaos **29**, 011102 (2019).
[34]T. Hagan, D. Fiifi, W. Guojie, S. X. Liang, and D. A. J. Han, "A time-varying causality formalism based on the Liang–Kleeman information flow for analyzing directed interactions in nonstationary climate systems," J. Clim. **32**, 7521–7537 (2019).
[35]F. Zwiers and H. V. Storch, "Regime-dependent autoregressive time series modeling of the southern oscillation," J. Clim. **3**, 1347–1363 (1990).
[36]M. Zanin and D. Papo, "Detecting switching and intermittent causalities in time series," Chaos **27**, 047403 (2017).
[37]M. Jiang, X. Gao, H. An, H. Li, and B. Sun, "Reconstructing complex network for characterizing the time-varying causality evolution behavior of multivariate time series," Sci. Rep. **7**, 10486 (2017).
[38]J. de Wiljes, L. Putzig, and I. Horenko, "Discrete nonhomogeneous and non-stationary logistic and Markov regression models for spatiotemporal data with unresolved external influences," Comm. App. Math. Comp. Sci. **9**(1), 1–46 (2014).
[39]I. Horenko, "Finite element approach to clustering of multidimensional time series," SIAM J. Sci. Comput. **32**, 62–83 (2010).
[40]S. Falkena, J. de Wiljes, A. Weisheimer, and T. Shepherd, "Revisiting the identification of wintertime atmospheric circulation regimes in the Euro-Atlantic sector," Q. J. R. Meteorol. Soc. **146**, 2801–2814 (2020).
[41]J. Runge, "Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets," in *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, edited by D. Sontag and J. Peters (AUAI Press, 2020).
[42]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).
[43]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," Geophys. Res. Lett. **44**, 201–211, https://doi.org/10.1002/2017GL075483 (2017).
[44]A. N. Tikhonov, A. Goncharsky, V. V. Stepanov, and A. G. Yagola, *Numerical Methods for the Solution of Ill-Posed Problems* (Springer, 1995).
[45]G. A. F. Seber and A. J. Lee, *Linear Regression Analysis*, 2nd ed. (Wiley, 2014).
[46]D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. (Wiley, 2012).
[47]D. Colombo and M. H. Maathuis, "Order-independent constraint-based causal structure learning," J. Mach. Learn. Res. **15**, 3921–3962 (2014).
[48]P. Metzner, L. Putzig, and I. Horenko, "Analysis of persistent non-stationary time series and applications," Commun. Appl. Math. Comp. Sci. **7**, 175–229 (2012).
[49]H. Akaike, "Information theory as an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory, Tsahkadsor, Armenia, 2–8 September 1971*, edited by B. N. Petrov and F. Csáki (Akadémiai Kiadó, Budapest, 1973), pp. 267–281.
[50]K. Burnham and D. Anderson, *Model Selection and Multimodel Inference* (Springer, 2002).
[51]B. Shipley and J. C. Douma, "Generalized AIC and chi-squared statistics for path models consistent with directed acyclic graphs," Ecology **101**, e02960 (2020).
[52]M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," Biometrika **76**, 297–307 (1989).
[53]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," Phys. Rev. E **86**, 061121 (2012).
[54]B. Huang, P. W. Thorne, V. F. Banzon, T. Boyer, G. Chepurin, J. H. Lawrimore, M. J. Menne, T. M. Smith, R. S. Vose, and H.-M. Zhang, "Extended reconstructed sea surface temperatures version 5 (ERSSTv5): Upgrades, validations, and intercomparisons," J. Clim. **20**, 8179–8205 (2017).
[55]See climexp.knmi.nl/getindices.cgi?WMO=NCDCData/ersst_nino3.4a_rel&STATION=NINO3.4_rel for data used for the Nino 3.4 index (ENSO in text).
[56]D. R. Kothawale and M. Rajeevan, "Monthly, seasonal and annual rainfall time series for all-India, homogeneous regions and meteorological subdivisions: 1871–2016," Technical report, IITM, Pune, India, August 2017.
[57]See climexp.knmi.nl/getindices.cgi?WMO=IITMData/ALLIN&STATION=All-India_Rainfall&TYPE=p for data used for All India rainfall (AIR in text).
[58]T. Bódai, G. Drótos, M. Herein, F. Lunkeit, and V. Lucarini, "The forced response of the El Niño–Southern oscillation–Indian monsoon teleconnection in ensembles of earth system models," J. Clim. **33**, 2163–2182 (2020).
[59]PCMCI code is part of the open-source Python package tigramite, https://github.com/jakobrunge/tigramite.