machine learning in climate science

# High-recall causal discovery in autocorrelated time series

Bedartha Goswami

Journal Club

6 July 2021

# High-recall causal discovery for autocorrelated time series with latent confounders

**Andreas Gerhardus**
German Aerospace Center
Institute of Data Science
07745 Jena, Germany
andreas.gerhardus@dlr.de

**Jakob Runge**
German Aerospace Center
Institute of Data Science
07745 Jena, Germany
jakob.runge@dlr.de

## Paper

SCIENCE ADVANCES | RESEARCH ARTICLE

RESEARCH METHODS

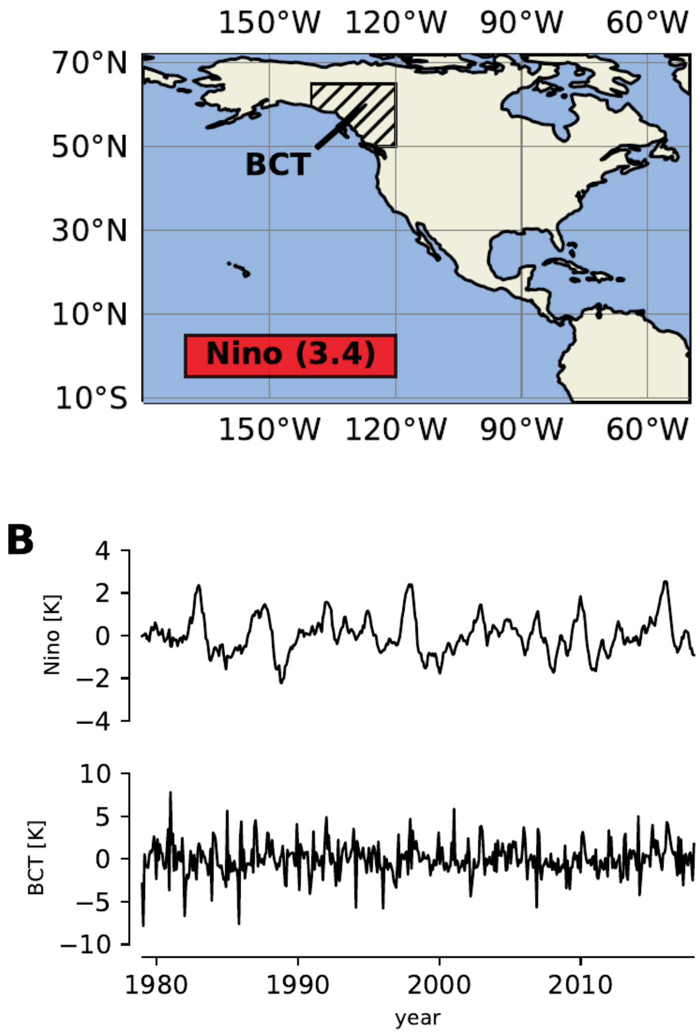
# Detecting and quantifying causal associations in large nonlinear time series datasets

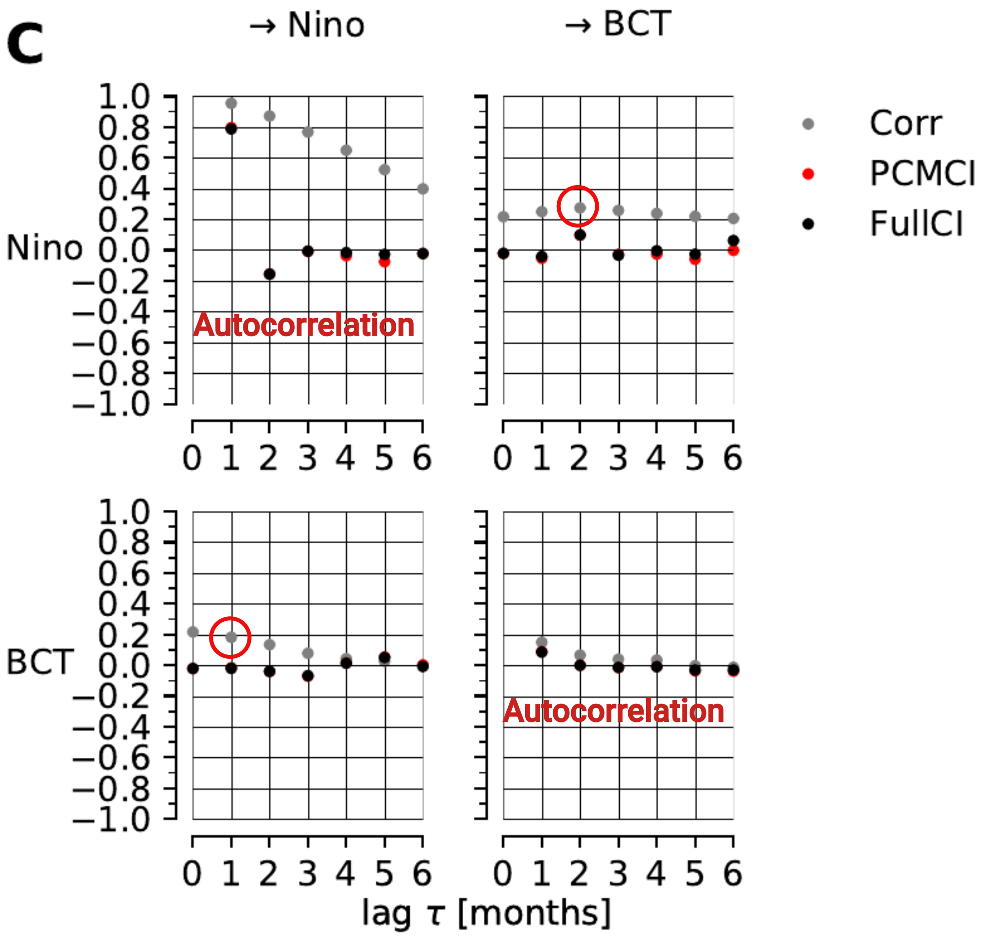
Jakob Runge[1,2]*, Peer Nowack[2,3,4], Marlene Kretschmer[5]†, Seth Flaxman[4,6], Dino Sejdinovic[7,8]

Paper

## Toy example

- Causal link between Niño 3.4 and British Columbia
- Test the impact of dimensionality and effect size
  - Dimensionality: Introduce artifical time series into the data set
  - Effect size: Impact of the causal link measured by partial correlation
- Power of the test: Percent times where the test was able to detect true links

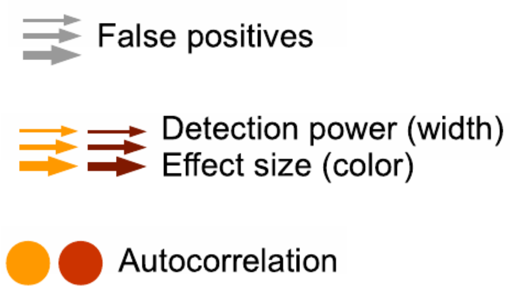







## Challenge with higher dimensionality and effect size

Effect size ~ 0.3 at lag = 2 (using correlation)

## Toy example

- Causal link between Niño 3.4 and British Columbia
- Test the impact of dimensionality and effect size
  - **Dimensionality: Introduce artifical time series into the data set**
  - Effect size: Impact of the causal link measured by partial correlation
- **Power of the test: Percent times where the test was able to detect true links**

**Test 1**

$$Z_t = 2 \cdot \mathrm{Nino}_{t-1} + \eta_t^Z$$

**Test 2**

$$W^i \ (i = 1, \ldots, 6)$$

$$W_t^i = a^i W_{t-1}^i + c\, W_{t-2}^{i-1} + \eta_t^i \text{ for } i = 2, 4, 6$$

$$W_t^i = a^i W_{t-1}^i + \eta_t^i \text{ for } i = 1, 3, 5$$

## Full Conditional independence test (FullCI)

- Test for conditional independence between X and Y
- For the link X → Y
  - Fit a linear autoregressive model for Y(t) dependent on all past variables of Y, i.e., Y(t-1), Y(t-2), … and X, i.e., X(t-1), X(t-2), …
  - Estimate which autoregressive coeficien %ts are significantly different from zero

**Test 1**

$$Z_t = 2 \cdot \mathrm{Nino}_{t-1} + \eta_t^Z$$

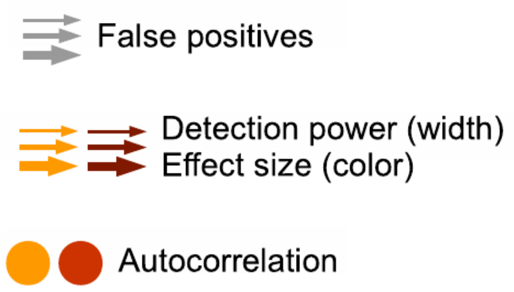
**Test 2**

$$W^i \; (i = 1, \ldots, 6)$$

$$W_t^i = a^i W_{t-1}^i + c\, W_{t-2}^{i-1} + \eta_t^i \text{ for } i = 2, 4, 6$$

$$W_t^i = a^i W_{t-1}^i + \eta_t^i \text{ for } i = 1, 3, 5$$

Effect size ~ 0.3 at lag = 2 (using correlation)
Effect size ~ 0.1 at lag = 2 (using FullCI)

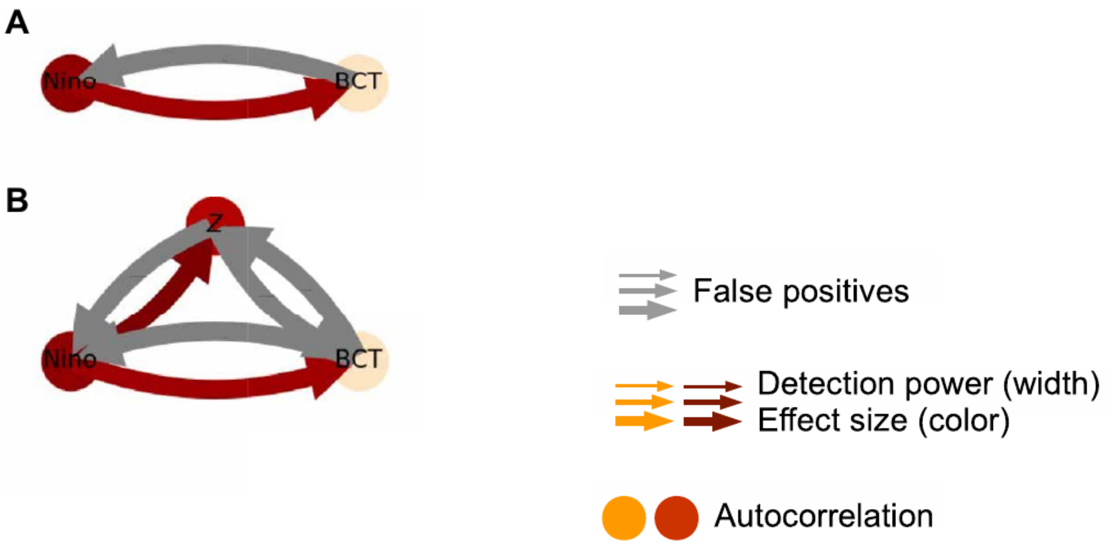






## Challenge with higher dimensionality and effect size

Effect size ~ 0.3 at lag = 2 (using correlation)
Effect size ~ 0.1 at lag = 2 (using FullCI)

Effect size ~ 0.09 at lag = 2 (using Full CI)
Power = 53% (85% if Z is independent of Niño)





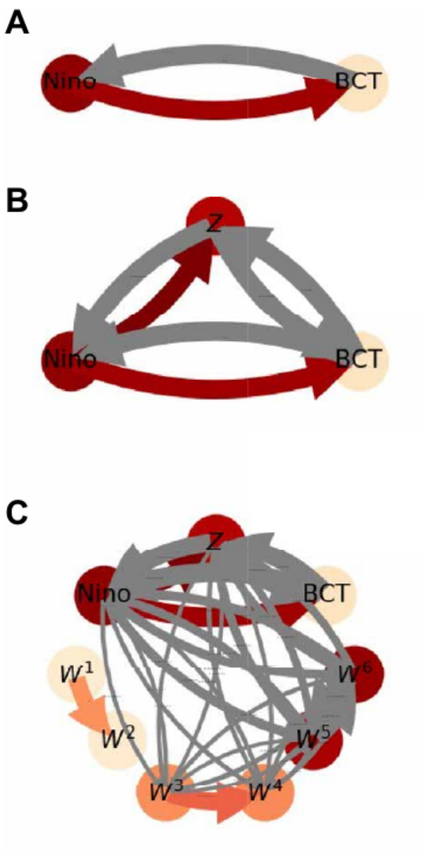

Bedartha Goswami

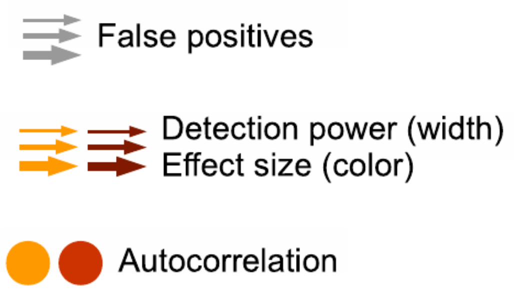
## Challenge with higher dimensionality and effect size

Effect size ~ 0.3 at lag = 2 (using correlation)
Effect size ~ 0.1 at lag = 2 (using FullCI)

Effect size ~ 0.09 at lag = 2 (using Full CI)
Power = 53% (85% if Z is independent of Niño)

Effect size ~ 0.09 at lag = 2 (using Full CI)
Power = 40%






## Challenge with higher dimensionality and effect size

Consider the *N*-dimensional system (i.e., in our case, a system that has been obsrved at *N* spatial locations)

$$\mathbf{X}_t = (X_t^1, \ldots, X_t^N)$$

where each $X^j_t$ evolves in time according to some function of the past states of all locations (incl. itself)

$$X_t^j = f_j(\mathcal{P}(X_t^j), \eta_t^j)$$

potentially nonlinear functional dependency

mutually indepdendent dynamical noise

causal "parents" of $\mathbf{X}^j_t$

A causal link $X_{t-\tau}{}^i \rightarrow X_t{}^j$ exists iff $X_{t-\tau}{}^i \in \mathcal{P}(X_t{}^j)$

Equivalently, the causal link $X_{t-\tau}{}^i \rightarrow X_t{}^j$ is defined as

$$X_{t-\tau}^i \not\perp\!\!\!\perp X_t^j \mid \mathbf{X}_t^- \setminus \{X_{t-\tau}^i\}$$

## PCMCI consists of two steps

- PC step:
  - Identify relevant conditions (i.e. parents) of every variable $X^j_t$, i.e. estimate $\widehat{\mathcal{P}}(X^j_t)$
- MCI step:
  - Momentary Conditional Independence
  - Test whether

$$X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \widehat{\mathcal{P}}\left(X^j_t\right) \setminus \left\{X^i_{t-\tau}\right\}, \widehat{\mathcal{P}}\left(X^i_{t-\tau}\right)$$

## PC step

- Initialize preliminary parents: $\widehat{\mathcal{P}}(X_t^j) = (\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \ldots, \mathbf{X}_{t-\tau_{\max}})$

- First iteration, $p = 0$
  - Conduct unconditional independence tests
  - Remove $X^i_{t-\tau}$ if from the parents of $X^j_t$ if the null hypothesis that $X^i_{t-\tau}$ and $X^j_t$ are unconditionally independent cannot be rejected at significance level $\alpha_{\mathrm{PC}}$

- Next iterations, p → p + 1
  - Sort parents of $X^j_t$ according to magnitude of test statistic (e.g., absolute partial correlation)
  - Conduct conditional independence tests $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \mathcal{S}$; where $S$ is is the set of strongest parents
  - Remove those parents that whose conditional independence cannot be rejected

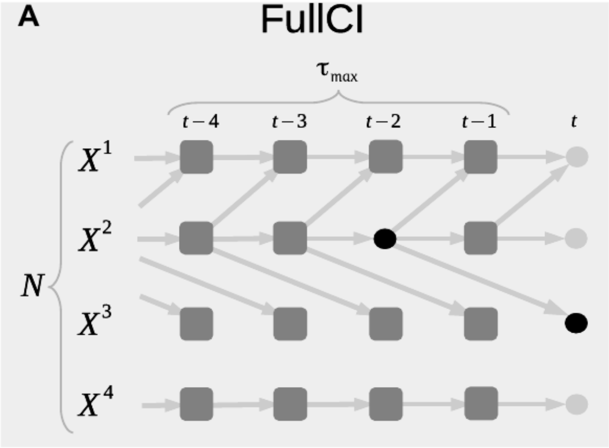
## MCI step

- Use the set of parents identified from the PC step
- For the link $X_{t-\tau}{}^{i} \rightarrow X_t{}^{j}$
- Instead of the initial definition of a causal link

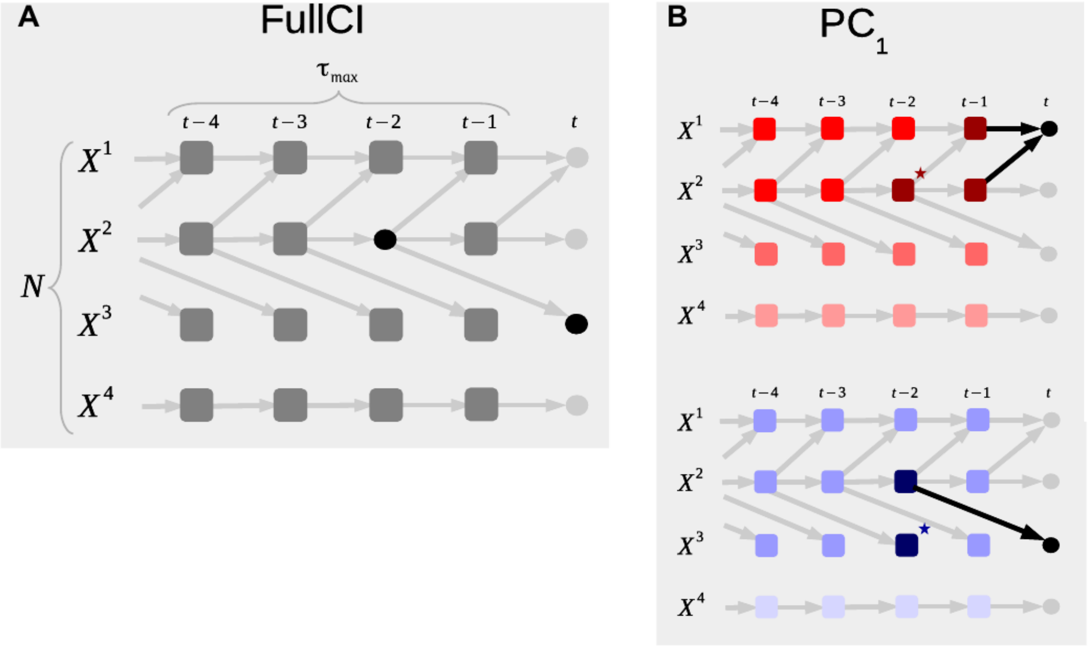
$$X_{t-\tau}^{i} \not\!\perp\!\!\!\perp X_t^{j} \mid \mathbf{X}_t^{-} \setminus \{X_{t-\tau}^{i}\}$$

- Use the more efficient causality condition

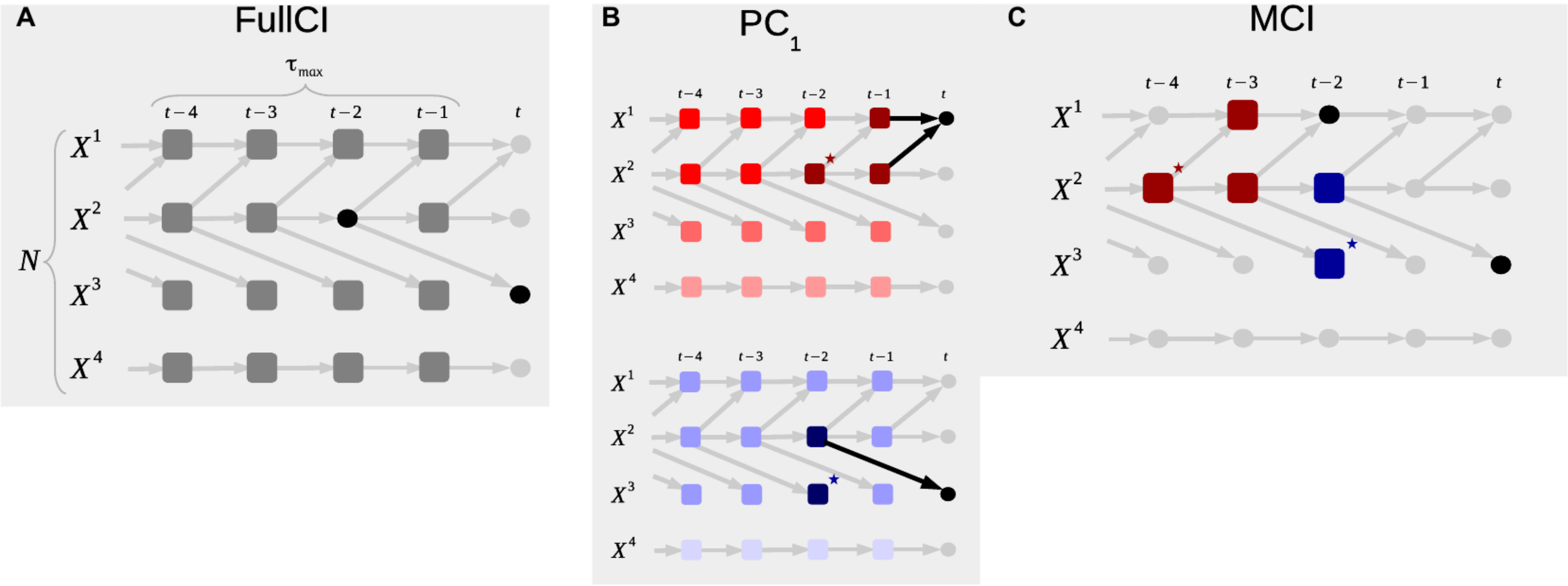
$$X_{t-\tau}^{i} \not\!\perp\!\!\!\perp X_t^{j} \mid \widehat{\mathcal{P}}\left(X_t^{j}\right) \setminus \left\{X_{t-\tau}^{i}\right\}, \widehat{\mathcal{P}}\left(X_{t-\tau}^{i}\right)$$



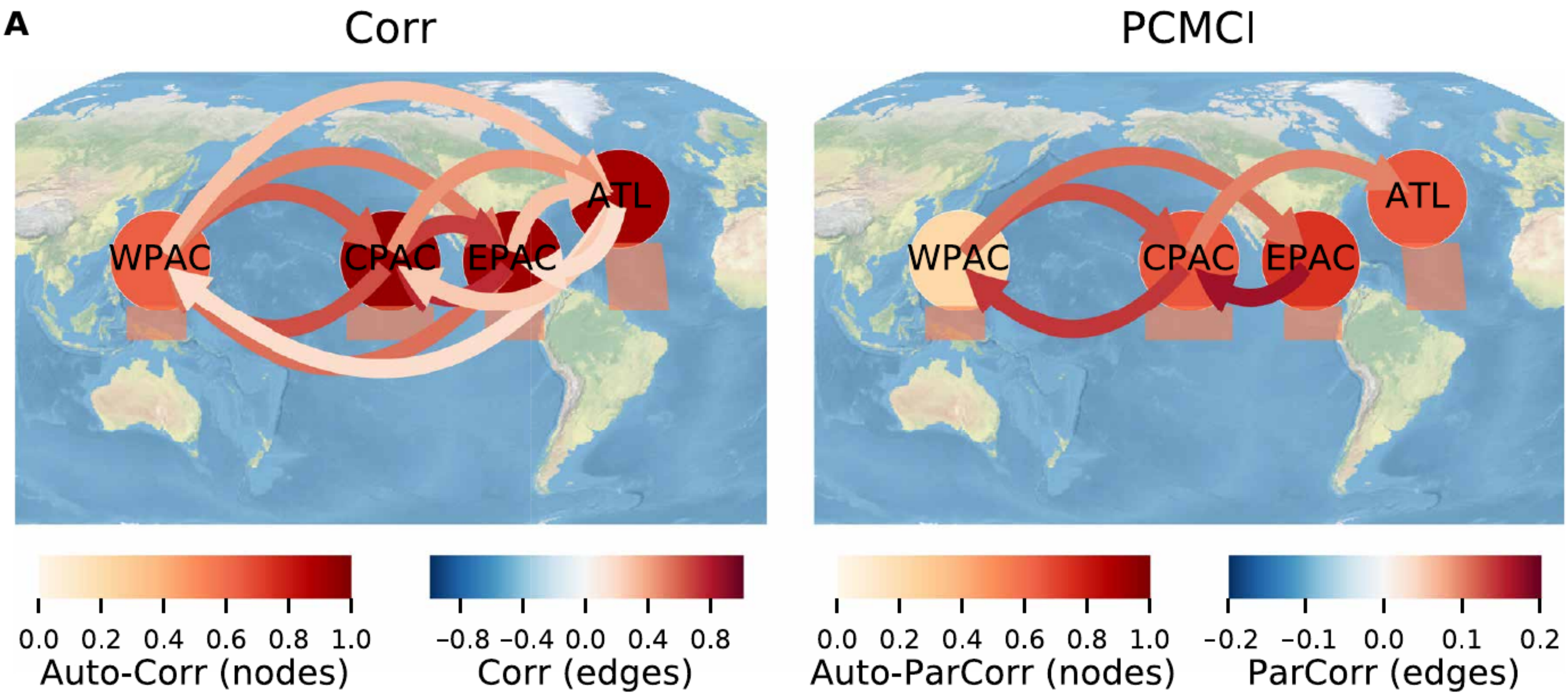

## PCMCI: Toy example




## PCMCI: Toy example → PC Step




## PCMCI: Toy example → MCI Step




PCMCI applied to monthly surface pressure anomalies (1948–2012) from western Pacific (WPAC), central Pacific (CPAC), estearn Pacific (EPAC), and tropical Atlantic (ATL)

## Summary

- Dimensionality reduces the power of causal discovery tests
- Authors propose a two-step method PCMCI to reliably detect causal links
- The PC step iteratively removes independent parents from each node in a time series graphical model
- The MCI step considers the final (converged) set of parents from the PC step and estimates causal links based on a momentary conditional independence test
- Results show reliable results in synthetic and climate examples