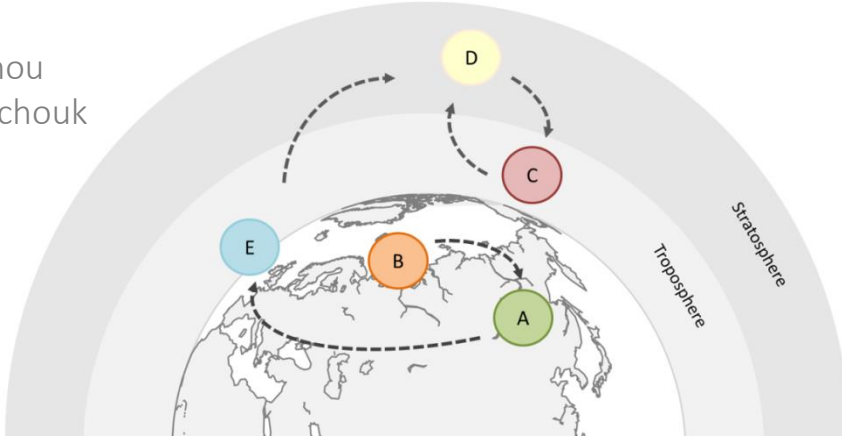
ECMWF ANNUAL SEMINAR 2019

ANALYZING CAUSAL PATHWAYS OF THE STRATOSPHERIC POLAR VORTEX

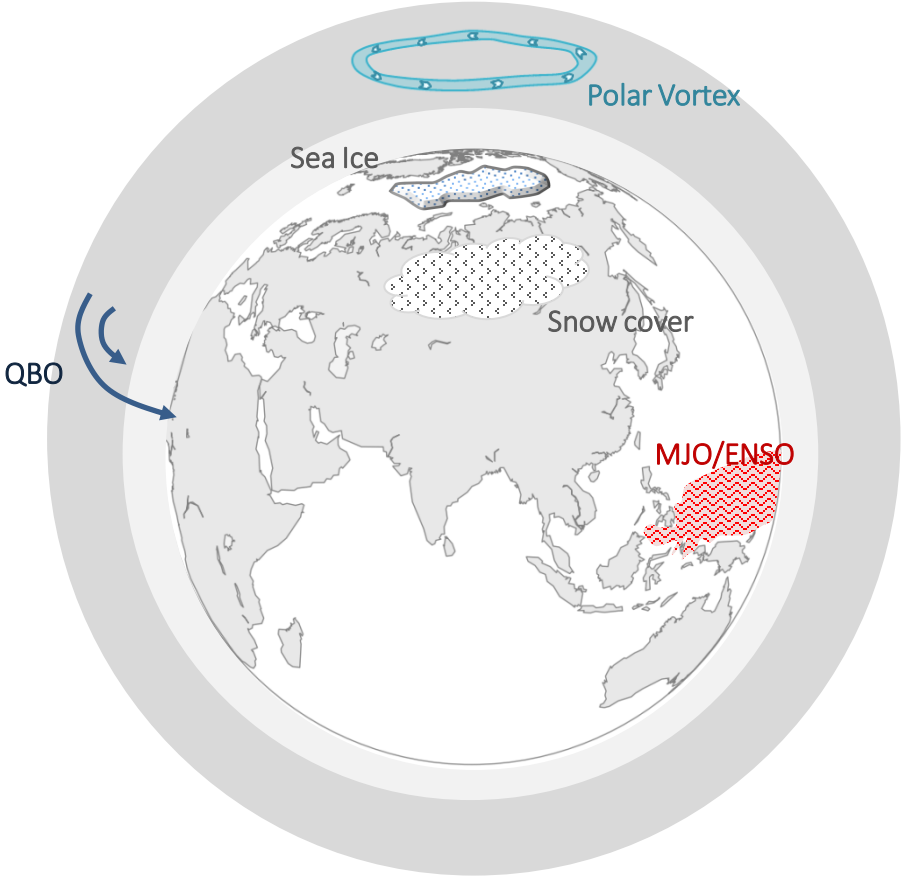
USING MACHINE LEARNING TOOLS

Marlene Kretschmer
Jakob Runge, Dim Coumou
Ted Shepherd, Inna Polichtchouk
Antje Weisheimer

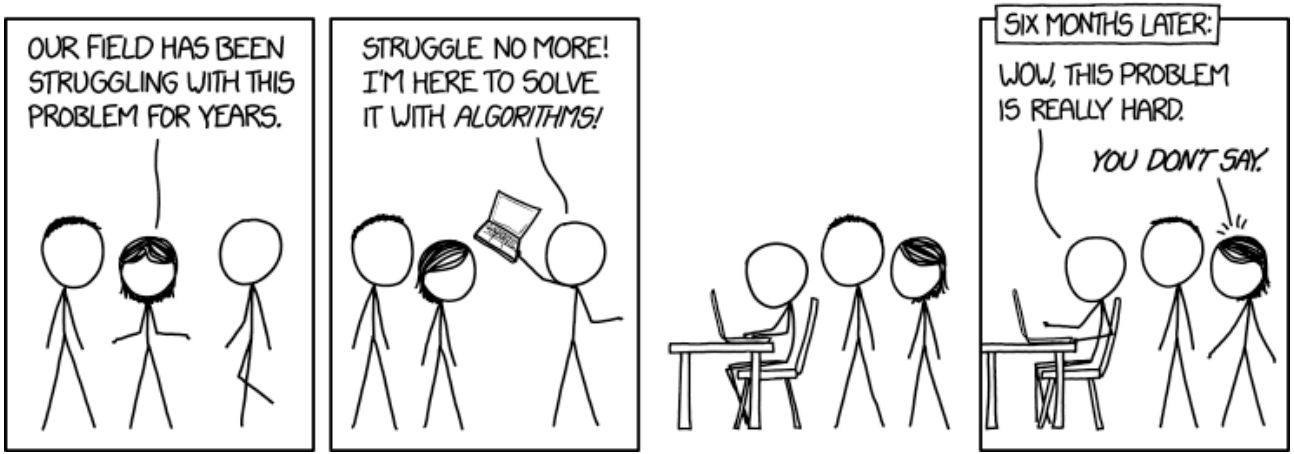


MOTIVATION

- Slow moving components of the climate system are potential sources of predictability on S2S time-scales
- However, the exact interactions are generally not well understood
- Separating the signal (i.e. the teleconnection pattern) from the noise (i.e. chaotic atmospheric variability) remains challenging



Machine Learning vs Climate Physics?

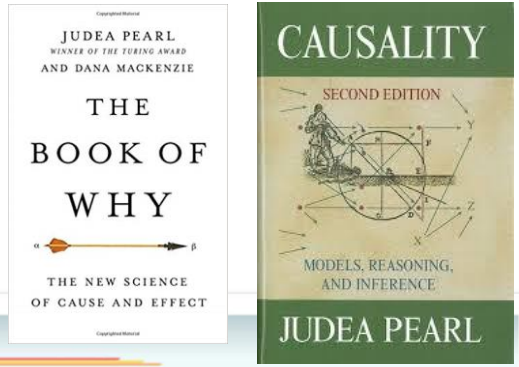


Machine Learning is not the answer to everything but might help to get an <u>additional</u> perspective on climate data

Xkcd #1381

"THE CAUSAL REVOLUTION"

- Assessing causality has long been neglected in statistics
- "Causal Revolution":
  rigorous mathematical framework to identify and quantify cause-effect relationships
- These concepts are also more and more applied to climate science




nature COMMUNICATIONS

PERSPECTIVE

https://doi.org/10.1038/s41467-019-10105-3 OPEN

Inferring causation from time series in Earth system sciences

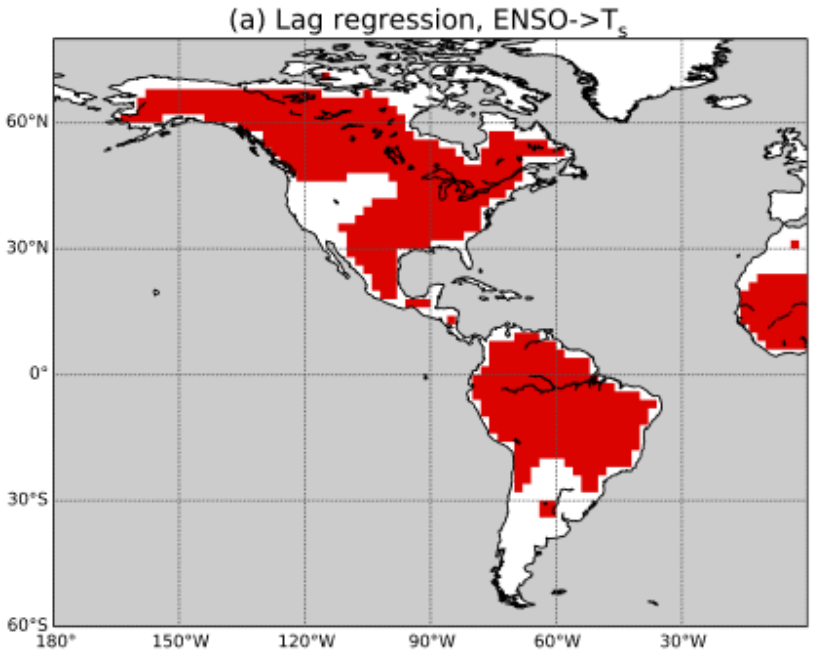
Jakob Runge[1,2], Sebastian Bathiany[3,4], Erik Bollt[5], Gustau Camps-Valls[6], Dim Coumou[7,8], Ethan Deyle[9], Clark Glymour[10], Marlene Kretschmer[8], Miguel D. Mahecha[11], Jordi Muñoz-Marí[6], Egbert H. van Nes[4], Jonas Peters[12], Rick Quax[13,14], Markus Reichstein[11], Marten Scheffer[4], Bernhard Schölkopf[15], Peter Spirtes[10], George Sugihara[9], Jie Sun[5,16], Kun Zhang[10] & Jakob Zscheischler[17,18,19]

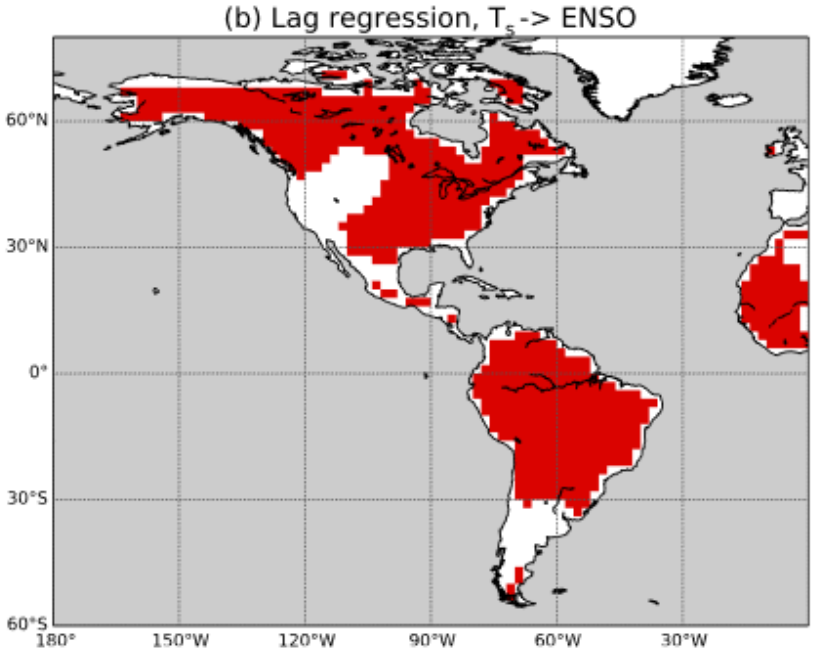
Outline

Why? What? How?

WHY SHOULD WE USE CAUSAL DISCOVERY ALGORITHMS IN CLIMATE SCIENCE?

MOTIVATING EXAMPLE





Auto-correlation Affects the Correlation strength

$$X_t = a\, X_{t-1} + \varepsilon^X_t$$

$$Y_t = b\, Y_{t-1} + c\, X_{t-1} + \varepsilon^Y_t$$

X and Y are both auto-correlated, Y additionally depends on X

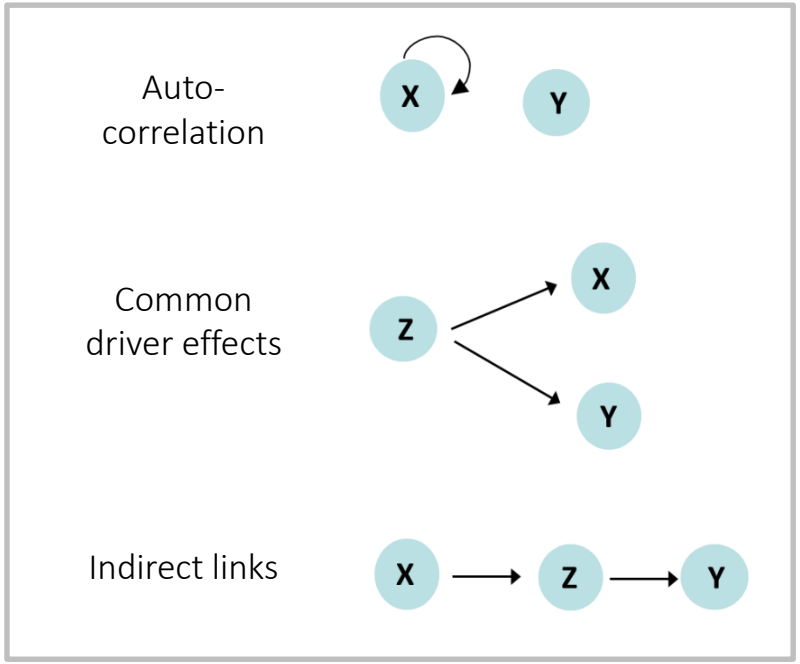
$$r(X_{t-\tau}, Y_t) = \begin{cases} \dfrac{a^{|\tau|} c \sigma_X^2}{(1-a^2)(1-ab)} \Big/ \sqrt{\Gamma_X \Gamma_Y} & \text{for} \quad \tau \leq 0 \\ \dfrac{c \sigma_X^2 [a^{|\tau|}(1-ab) - b^{|\tau|}(1-a^2)]}{(1-a^2)(a-b)(1-ab)} \Big/ \sqrt{\Gamma_X \Gamma_Y} & \text{for} \quad \tau > 0 \end{cases}$$

$$\text{with} \quad \Gamma_X = \frac{\sigma_X^2}{1-a^2}, \quad \Gamma_Y = \frac{c^2 \sigma_X^2 (1+ab) + \sigma_Y^2 (1-a^2)(1-ab)}{(1-a^2)(1-b^2)(1-ab)}$$

Correlation depends not only on c but also on a and b

Climate data is often strongly auto-correlated!

Typical Challenges with Climate Data



- A simple bi-variate correlation analysis is often too limited
- Multi-variate statistical frameworks are needed

Causal discovery algorithms can help

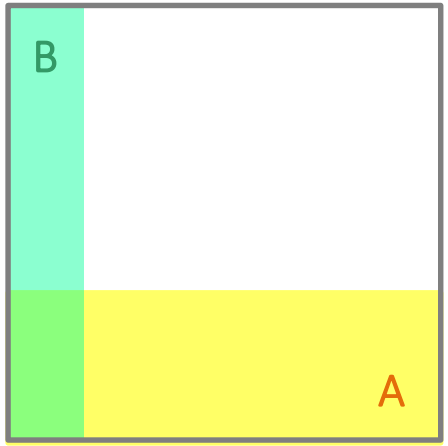
WHAT ARE CAUSAL DISCOVERY ALGORITHMS?

Conditional Independence

A and B are (*unconditionally*) **independent**

$\Leftrightarrow$

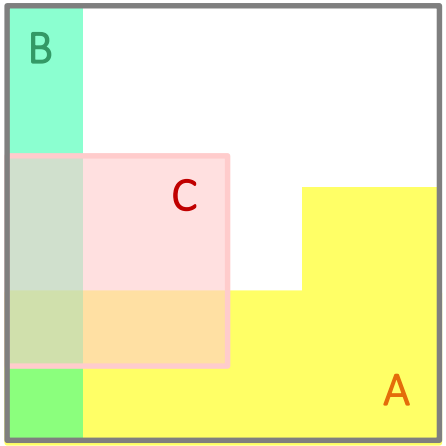
$P(A, B) = P(A)\,P(B)$



A and B are **conditionally independent given C**

$\Leftrightarrow$

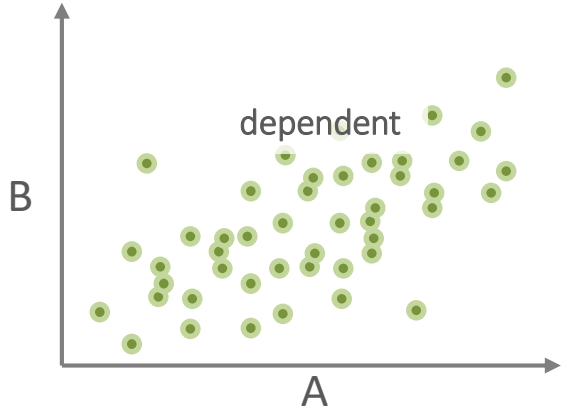
$P(A, B \mid C) = P(A \mid C)\;\; P(B \mid C)$



Measuring Conditional (In)Dependence

Correlation

r(A,B)

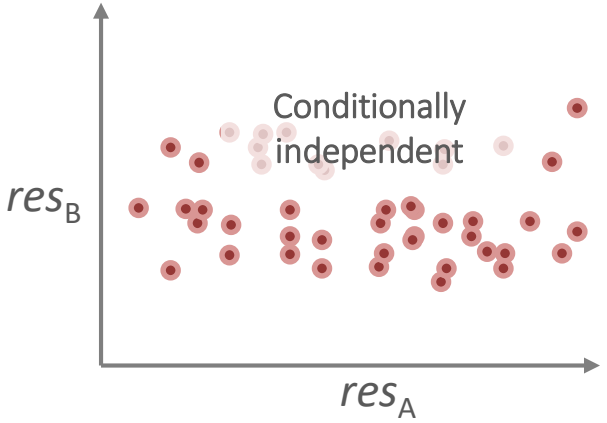


Partial Correlation

r(A, B | C)

$A = a\,C + res_A$

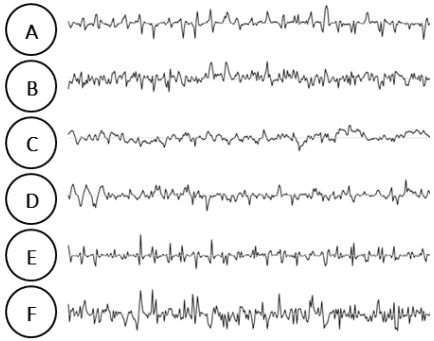
$B = b\,C + res_B$



Reconstructing Causal Relationships from Data

Input

Time-series



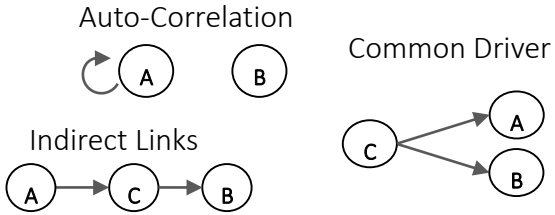
PCMCI

PCMCI: Adapted version of PC Algorithm

Spirtes & Glymour (1991), Runge et al. (2014, 2018)

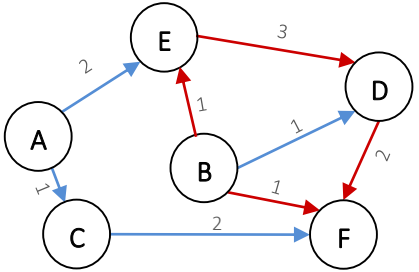
$r(A_{t-\tau}, B_t \mid$ *Iterate through combinations of conditions*$)$

Identify Spurious Correlations

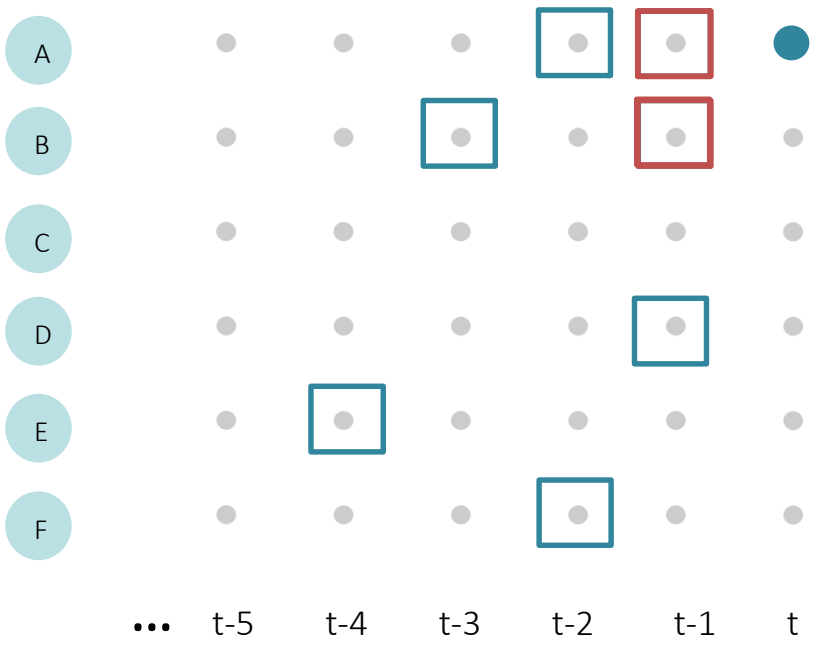


Output

Causal Effect Network



PCMCI



Step 1: Estimate Parents

- Number of conditions = 0
  - $r(B_{t-1}, A_t)$, … , $r(F_{t-5}, A_t)$

  $P_0 = \{A_{t-1}, B_{t-1}, B_{t-3}, A_{t-2}, D_{t-1}, E_{t-4}, F_{t-2}\}$

- Number of conditions = 1
  - $r(B_{t-1}, A_t \mid A_{t-1})$, …, $r(F_{t-2}, A_t \mid A_{t-1})$
  - Then condition on B_{t-1}
  - …..

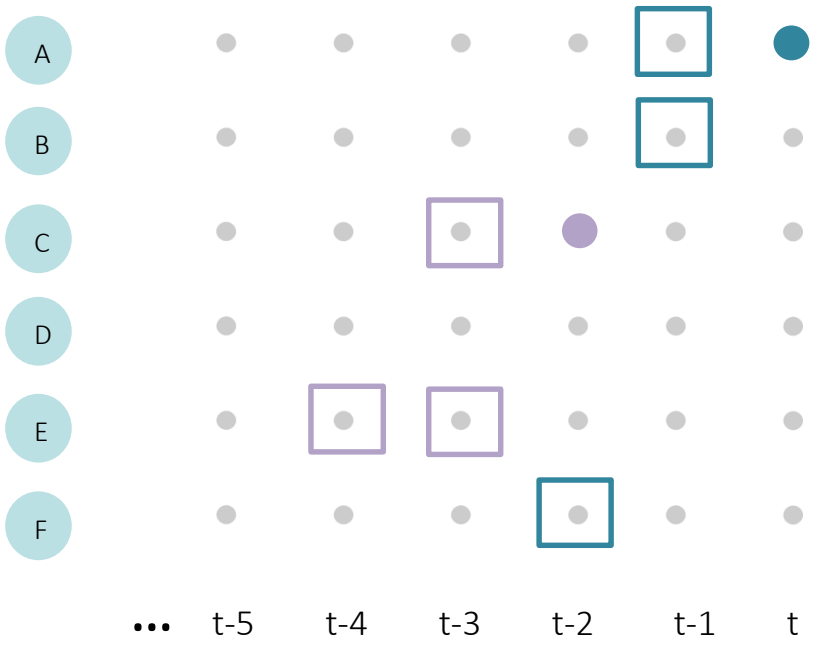
  $P_1 = \{A_{t-1}, B_{t-1}, B_{t-3}, F_{t-2}\}$

- Number of conditions = 2
  - $r(B_{t-3}, A_t \mid A_{t-1}, B_{t-1})$
  - ….

  $P_2 = \{A_{t-1}, B_{t-1}, F_{t-2}\}$

- Algorithm converged : $\mathcal{P}_A = P_2$

PCMCI

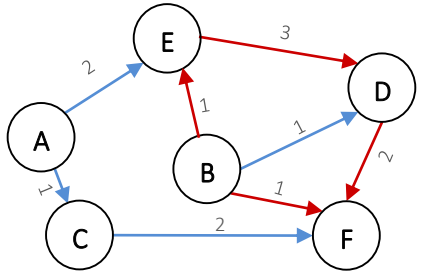


Step 2: Estimate Link Strength

What is the influence of C on A at lag 2?

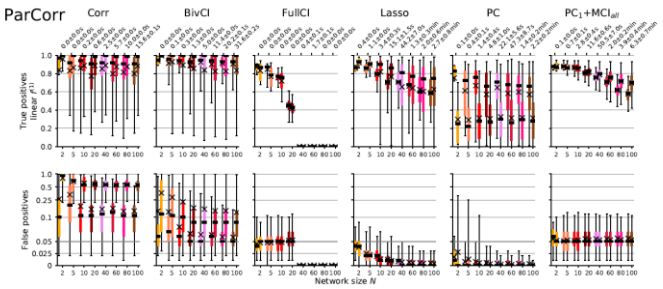
- $p(C_{t-2}, A_t \mid \boldsymbol{\mathcal{P}}_A, \boldsymbol{\mathcal{P}}_C)$

Step 3: Visualize links



THE PCMCI ALGORITHM

- Several publications on topic discussing the underlying assumptions, numerical tests on generic data, real-world examples, comparison with other methods…
- Implemented in python
  (check out the tutorial on github)
- Also non-linear metrics implemented



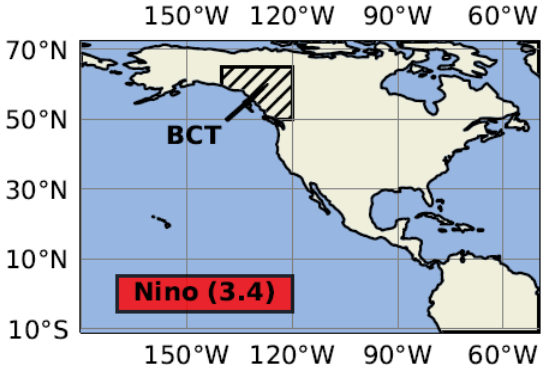
Causal discovery with TIGRAMITE

TIGRAMITE is a time series analysis python module. It allows to reconstruct graphical models (conditional independence graphs) from discrete or continuously-valued time series based on a causal discovery algorithm and create high-quality plots of the results. This tutorial explains the main features and gives walk-through examples. It covers:

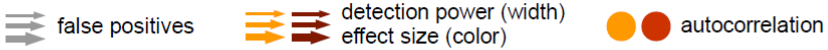
1. Basic usage
2. Plotting
3. Nonlinear conditional independence tests
4. Symbolic time series
5. Causal assumptions
6. Missing values and masking
7. Parallization
8. PCMCI vs Granger causality
9. Causal effects and mediation
10. Optimal predictor selection

Runge et al., *Science Advances* (in revision), Runge et al. *Chaos* (2018)

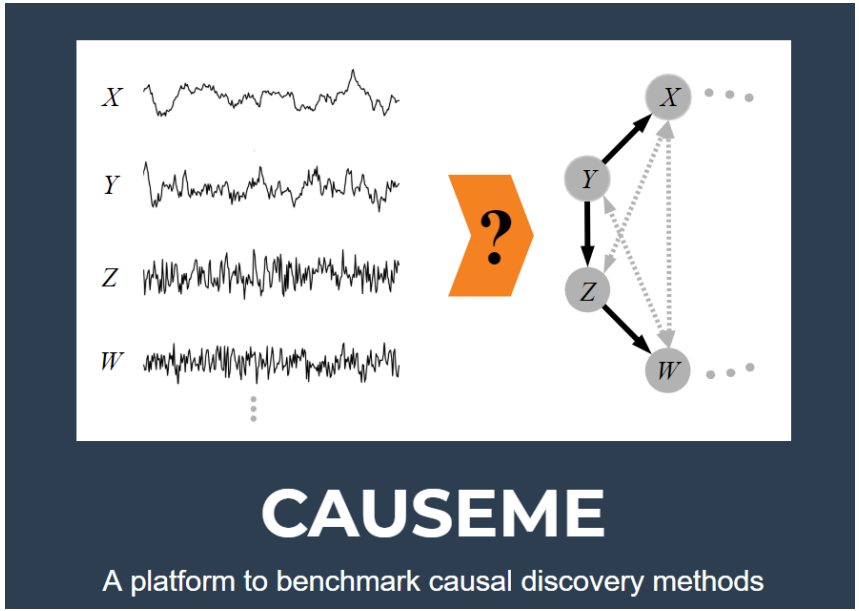
EXAMPLE (WITH KNOWN GROUND TRUTH)



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



CAUSEME.NET



Don't just label me. Cause Me!

How to get involved?

1. Develop a new method and assess its performance on available datasets.
2. Provide synthetic or real world multivariate time series data with known causal ground truth.

HOW CAN WE USE CAUSAL DISCOVERY ALGORITHMS IN CLIMATE SCIENCE?

EXAMPLE 1: HYPOTHESIS TESTING

Autumn

Early Winter

Late Winter

Kim et al. (2014)	Cohen et al. (2014)
Low Barents Kara sea ice (BK-SIC)	High Eurasian snow cover (EA-snow)
↓	↓
Ural Blocking (Ural-SLP)	Siberian Blocking (Sib- SLP)

↓

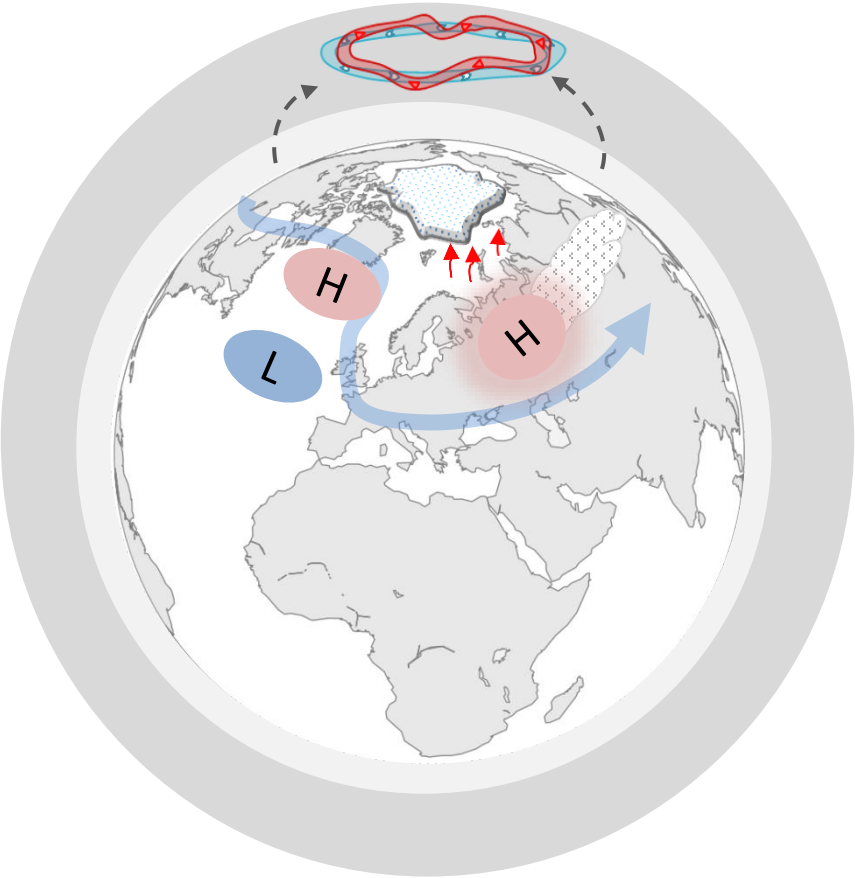
Enhanced vertical wave activity flux (v-flux)

↓

Weak Polar Vortex (PoV)

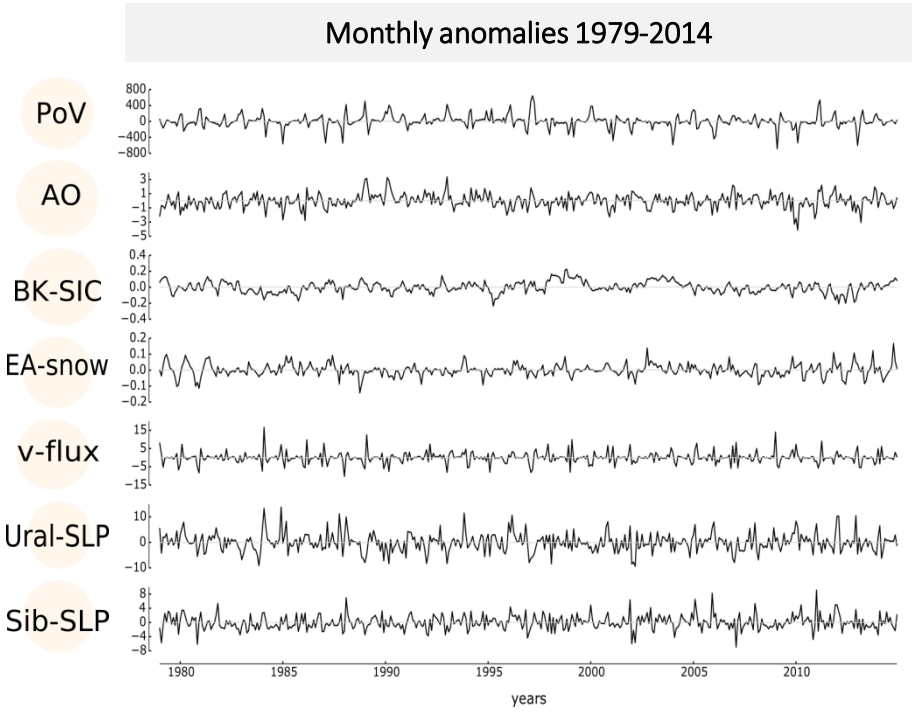
↓

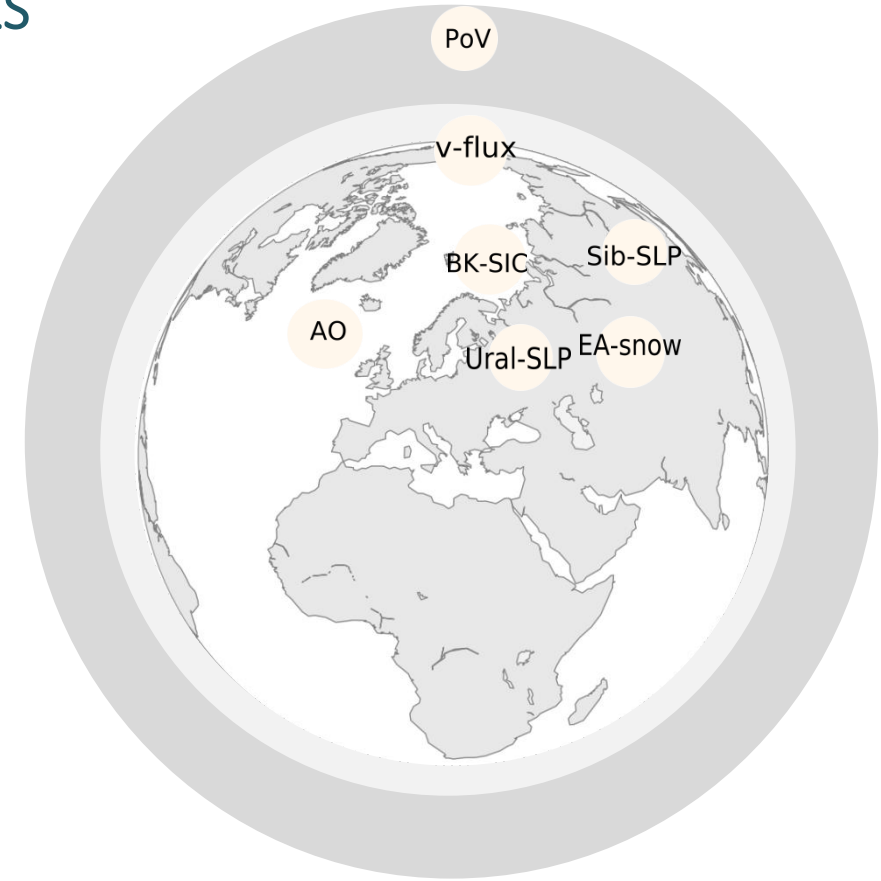
Negative Arctic Oscillation (AO)



Kretschmer et al. *J CLim* (2016)

TIME-SERIES OF INVOLVED PROCESSES





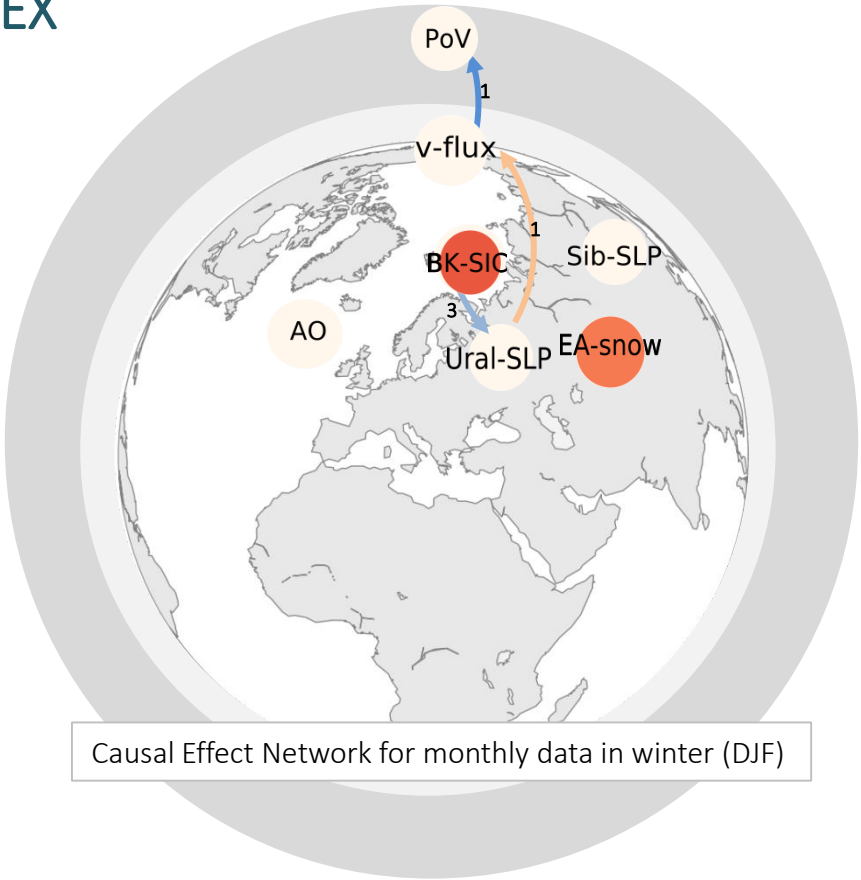
Kretschmer et al. *J CLim* (2016)

Arctic Drivers of the Polar Vortex

Key Findings

- Low Barents and Kara sea ice in autumn leads to a weakening of the polar vortex in winter
- Role of Eurasian snow cover less robust

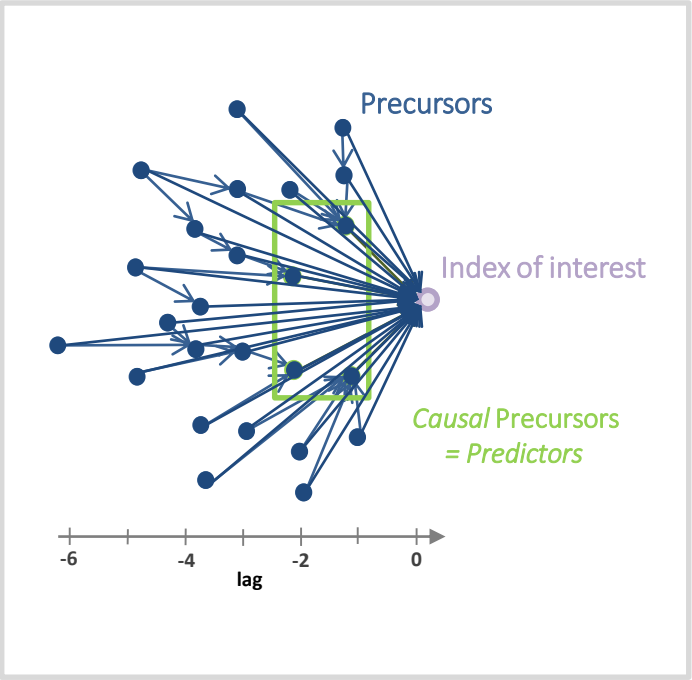
Zhang et al. (2017), MC Kenna (2018), Screen et al (2017), Kim et al. (2014), Hoshi et al. (2017)…

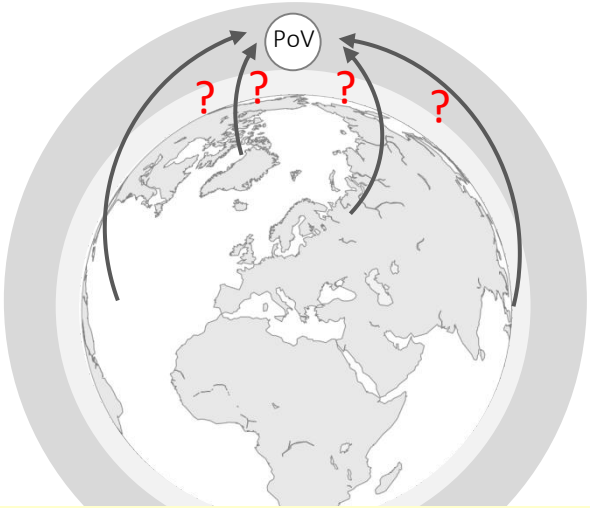


Causal Effect Network for monthly data in winter (DJF)

Kretschmer et al. *J CLim* (2016)

EXAMPLE 2: STATISTICAL FORECASTING





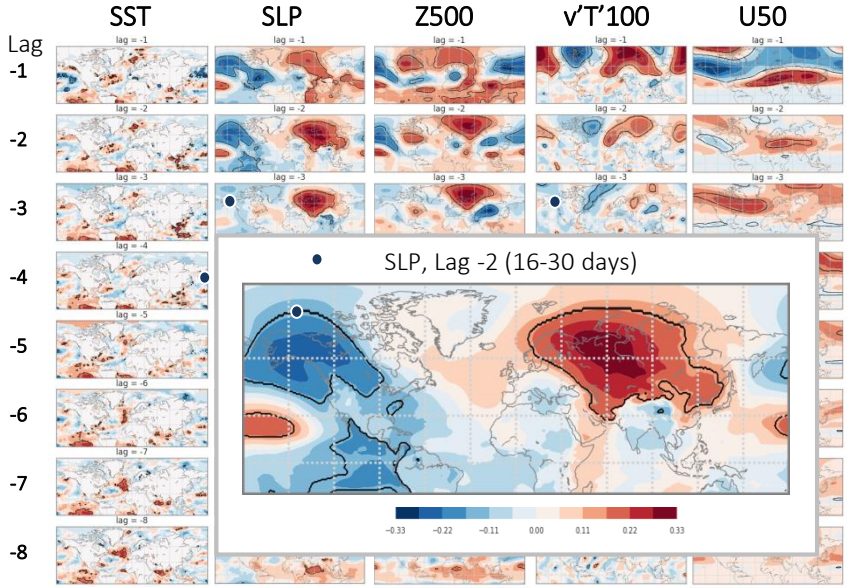
Issue: Overfitting

Prediction Model of the Polar Vortex

Index of interest: Winter PoV (15day mean)

Step 1: Find all precursor regions

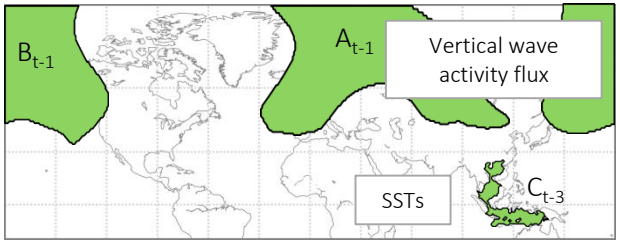
471 precursor regions



Step 2: Detect *Causal* Precursors

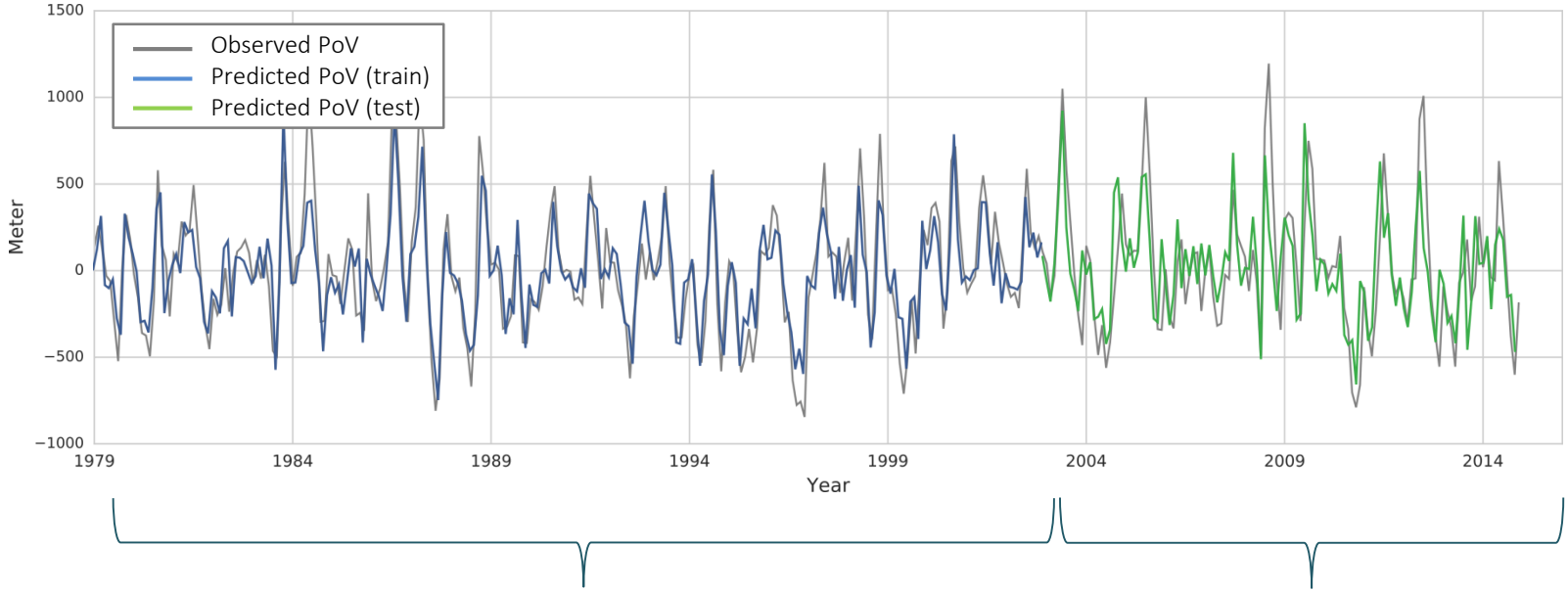
4 *causal* precursors

$$\underbrace{PoV_t = \beta_0 PoV_{t-1} + \beta_1 A_{t-1} + \beta_2 B_{t-1} + \beta_3 C_{t-3}}_{\text{Prediction model}}$$



Kretschmer et al. *GRL* (2017)

Skillful Prediction of the Polar Vortex

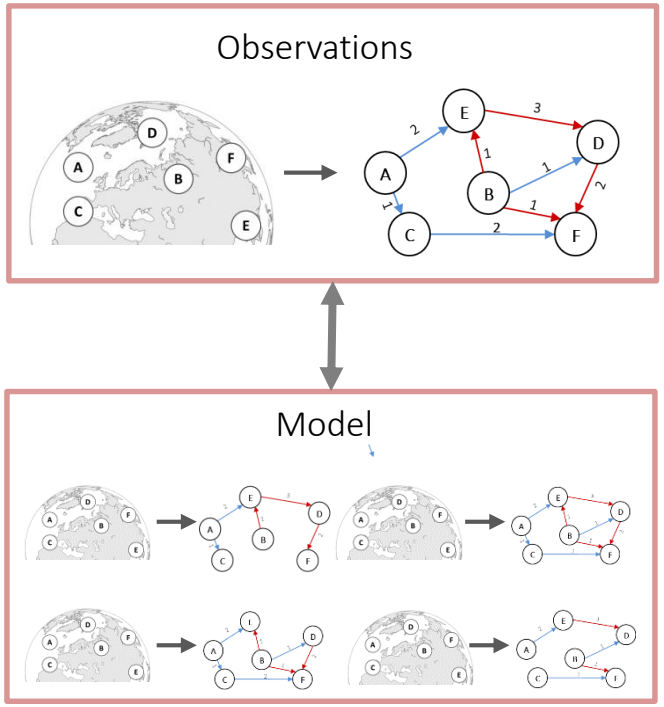


Kretschmer et al. *GRL* (2017), Di Capua et al. *WAF* (2019)

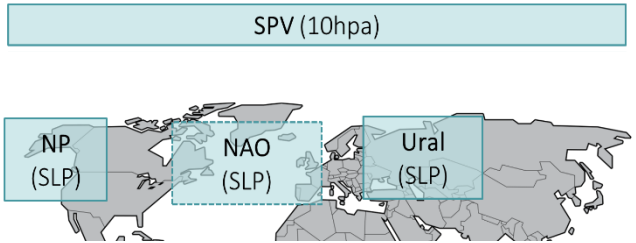
EXAMPLE 3: PROCESS-BASED MODEL EVALUATION

Issue:

How well are teleconnections represented in models?



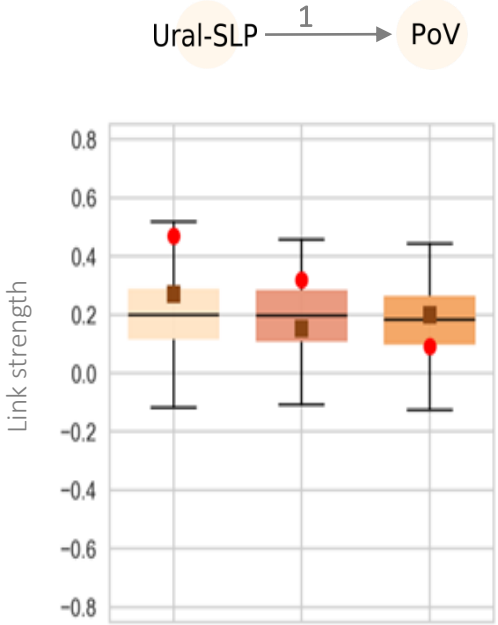
TROPOSPHERE-STRATOSPHERE COUPLING IN SEAS5



Data sets	SeaS5 (ERA-I)
Initialization	1st of November
Months	November-March
Period	1981– 2016
Time-scale	Monthly means

- **Bootstrap approach:**
  For each variable 1000 time-series are created by each year picking a random ensemble member
- **Run PCMCI for:**
  early winter (DJ)
  middle winter (JF)
  late winter (FM)

Comparison of Link Strength



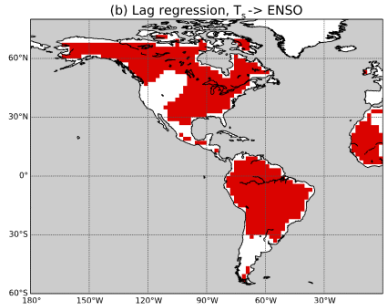
To be continued…

SUMMARY

THANK YOU!

Why?

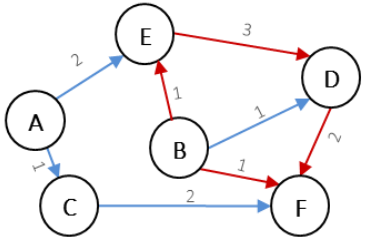
Understanding teleconnections is key to improve forecasting



Correlation analysis is too limited to achieve this

What?

Causal discovery algorithms (e.g. PCMCI) can overcome some of these problems



Already successfully applied to different data sets

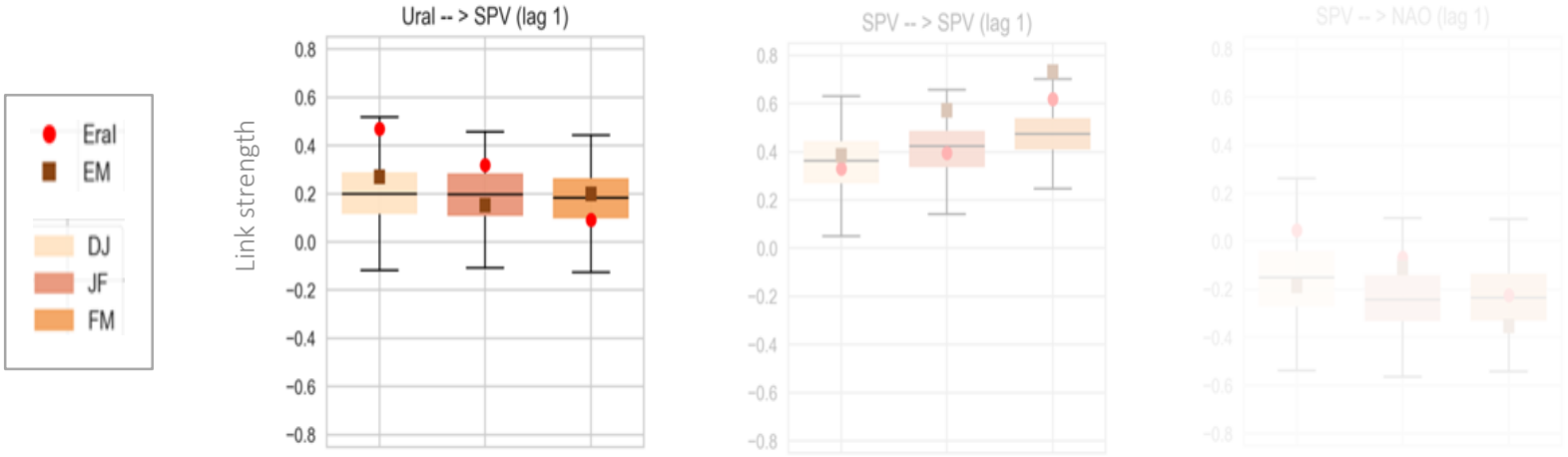
How?

Useful for hypothesis testing, statistical forecasting, model evaluation…

Expert knowledge on data is needed!

SI

Comparison of Link Strength



ASSUMPTIONS (PITFALLS)

A. Causal Sufficiency
(Unobserved variables, Sub-sampled time-series)

B. Causal Markov Condition
(Non-Markovian process, Time-aggregation)

C. Faithfulness
(Counteracting mechanisms, determinism, non-pairwise dependencies)

D. Assumption of no instantaneous effects

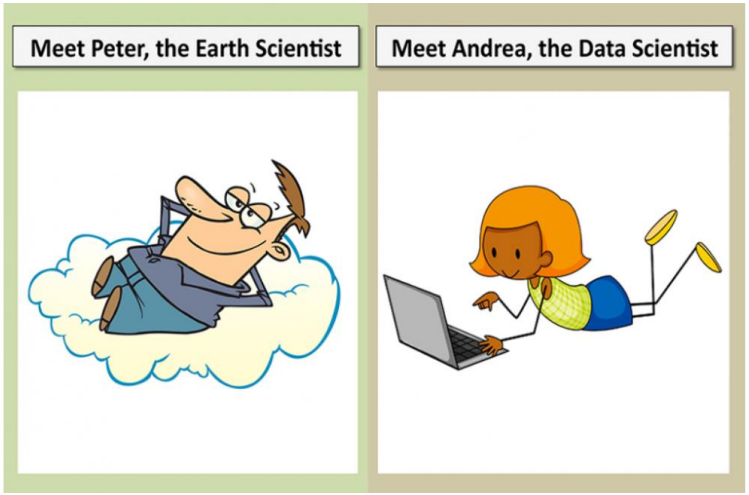
E. Causal Stationarity
(non-stationarity due to confounding)

F. Dependence type assumptions
(Nonlinearity)

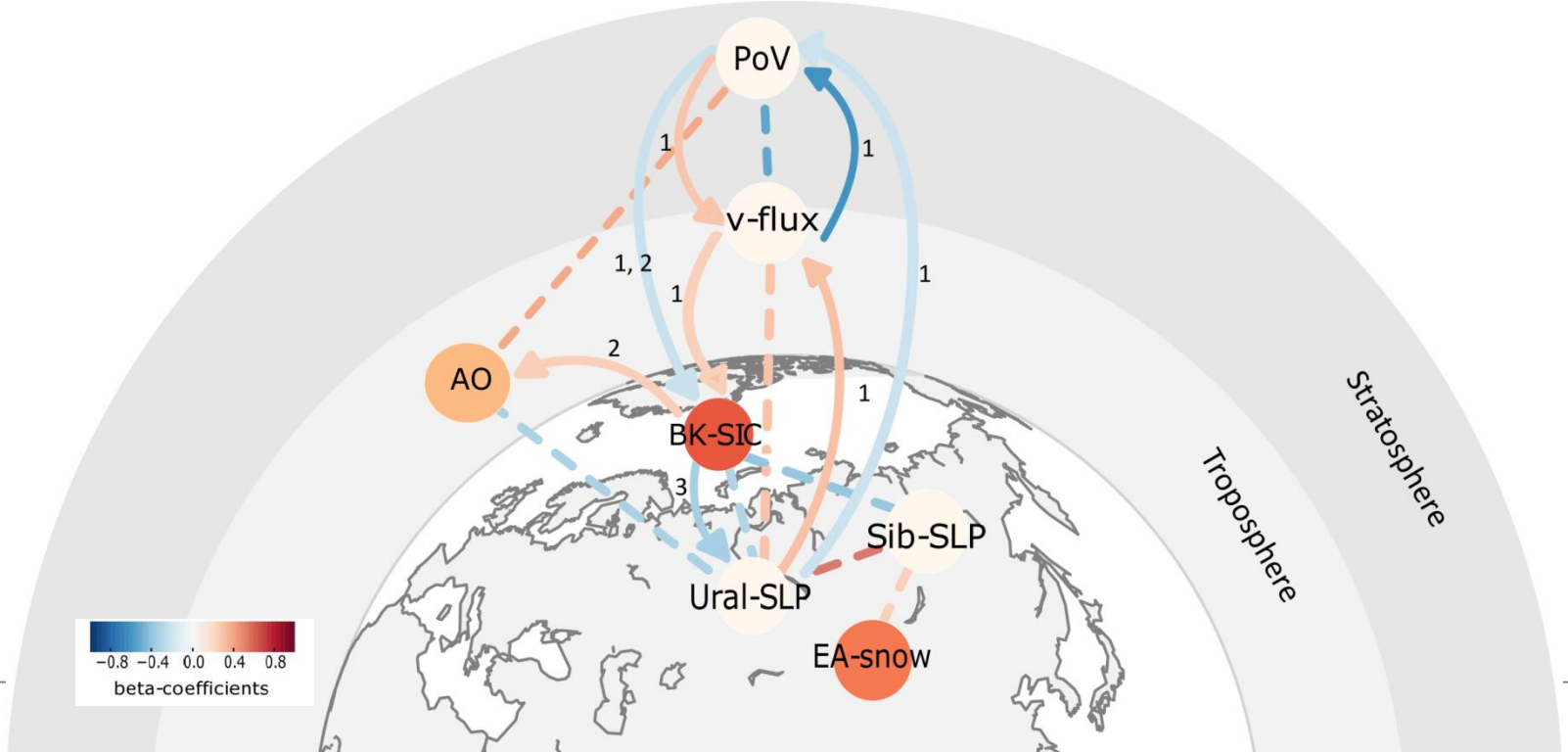
G. Data Quality
(measurement error)

Expert knowledge required!

Three Steps to Successful Collaboration with Data Scientists



Runge et *al.* (Chaos, 2018), Ebert-Uphoff & Deng (EOS, 2017)


PoV
v-flux
AO
BK-SIC
Sib-SLP
Ural-SLP
EA-snow
1
1
1, 2
1
1
2
1
3
Stratosphere
Troposphere
−0.8 −0.4 0.0 0.4 0.8
beta-coefficients