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 is not the answer to everything but might help to get an additional perspective on climate data.
“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

Outline
Why? What? How?
WHY SHOULD WE USE CAUSAL DISCOVERY ALGORITHMS IN CLIMATE SCIENCE?
Motivating Example

(a) Lag regression, ENSO -> $T_s$

(b) Lag regression, $T_s$ -> ENSO

**AUTO-CORRELATION AFFECTS THE CORRELATION STRENGTH**

\[
X_t = a \ X_{t-1} + \varepsilon_t^X \\
Y_t = b \ Y_{t-1} + c \ X_{t-1} + \varepsilon_t^Y
\]

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

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

Climate data is often strongly auto-correlated!

---

Runge et al., *Journal of Climate* (2014)
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?
A and B are (unconditionally) independent

\[ \iff \]

\[ P(A, B) = P(A) \cdot P(B) \]

A and B are conditionally independent given C

\[ \iff \]

\[ P(A, B \mid C) = P(A \mid C) \cdot P(B \mid C) \]
**Measuring Conditional (In)Dependence**

**Correlation**

\[ r(A, B) \]

**Partial Correlation**

\[ r(A, B \mid C) \]

\[ A = a C + res_A \]
\[ B = b C + res_B \]

- Dependent
- Conditionally independent

\[ res_B \]
\[ res_A \]
RECONSTRUCTING CAUSAL RELATIONSHIPS FROM DATA

Input

Time-series
A  B  C
D  E  F

PCMCI

PCMCI: Adapted version of PC Algorithm
Spirites & Glymour (1991), Runge et al. (2014, 2018)

\[ r(A_{t-\nu}, B_t) \]

Iterate through combinations of conditions

Identify Spurious Correlations

Auto-Correlation
A  B

Indirect Links
A  C  B

Common Driver
A

Output

Causal Effect Network

A  B  C
D  E  F

\[ \rightarrow \]
Step 1: Estimate Parents

- Number of conditions = 0
  - \( r(B_{t-1}, A_t), \ldots, r(F_{t-5}, A_t) \)
  - \( P_0 = \{A_{t-1}, B_{t-1}, B_{t-3}, A_{t-2}, D_{t-1}, E_{t-4}, F_{t-2}\} \)

- Number of conditions = 1
  - \( r(B_{t-1}, A_t \mid A_{t-1}), \ldots, r(F_{t-2}, A_t \mid A_{t-1}) \)
  - 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: \( P_A = P_2 \)
Step 2: Estimate Link Strength

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

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

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
**EXAMPLE (WITH KNOWN GROUND TRUTH)**

\[ Z_t = 2 \cdot \text{Nino}_{t-1} + \eta_t^Z \]

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

- Kim et al. (2014)
  - Low Barents Kara sea ice (BK-SIC)
  - Ural Blocking (Ural-SLP)
  - Enhanced vertical wave activity flux (v-flux)
  - Weak Polar Vortex (PoV)
  - Negative Arctic Oscillation (AO)

- Cohen et al. (2014)
  - High Eurasian snow cover (EA-snow)
  - Siberian Blocking (Sib- SLP)

TIME-SERIES OF INVOLVED PROCESSES

Monthly anomalies 1979-2014

**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)
Example 2: Statistical Forecasting

Causal Precursors = Predictors

Issue: Overfitting
**Prediction Model of the Polar Vortex**

**Step 1: Find all precursor regions**

<table>
<thead>
<tr>
<th>Lag</th>
<th>SST</th>
<th>SLP</th>
<th>Z500</th>
<th>v’T’100</th>
<th>U50</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 471 precursor regions
- SLP, Lag -2 (16-30 days)

**Step 2: Detect Causal Precursors**

- 4 causal precursors

\[
\text{PoV}_t = \beta_0 + \beta_1 \text{PoV}_{t-1} + \beta_2 A_{t-1} + \beta_3 B_{t-1} + \beta_4 C_{t-3}
\]

Prediction model

- Vertical wave activity flux
- SSTs

Index of interest: Winter PoV (15day mean)

Kretschmer et al. *GRL* (2017)
SKILLFUL PREDICTION OF THE POLAR VORTEX

Observed PoV
Predicted PoV (train)
Predicted PoV (test)

Train regression model
Corr = .79

1-step (1-15 days) ahead predictions
Corr = .73

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 PCMCl for:** early winter (DJ) middle winter (JF) late winter (FM)
COMPARISON OF LINK STRENGTH

To be continued...
### Summary

**Why?**
Understanding teleconnections is key to improve forecasting

![Map showing teleconnections](image)

Correlation analysis is too limited to achieve this

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

![Causal diagram](image)

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

![Comparison of Link Strength Diagrams](image-url)
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)

Three Steps to Successful Collaboration with Data Scientists

Expert knowledge required!

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