

A causal approach to circulation variability in the Southern Hemisphere

Master Research Project

Elena Saggioro, PhD candidate Mathematics of Planet Earth CDT ¹
Prof Ted Shepherd, main supervisor ²

¹Dept. of Mathematics
University of Reading and Imperial College London

²Dept. of Meteorology
University of Reading

SFB-1294 Seminar, 16 November 2018



The stratosphere - troposphere coupling

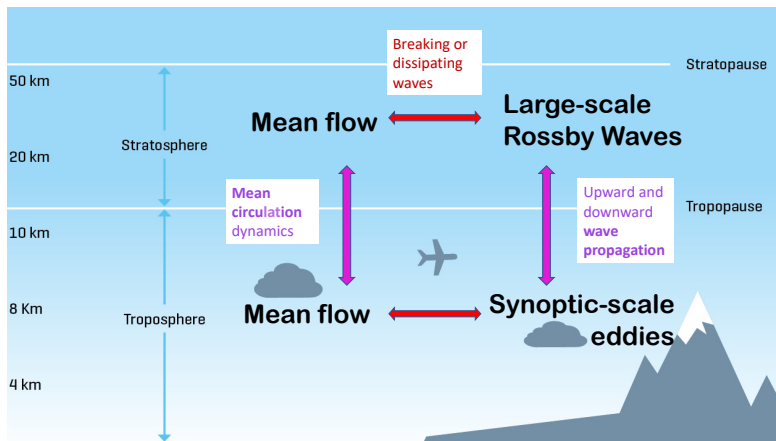


Figure: Schematic of dynamical mechanism adapted from Haynes (2005)

Downward influence

How does stratosphere influence the tropospheric dynamics?
Theory still does not provide good understanding... however

- Downward influence is **hinted at by observations**.
- Might be missing key to explain some tropospheric seasonal patterns
- Numerical simulations show **improved seasonal forecast** skills for troposphere when **better stratosphere implemented**

Coupling in the Southern Hemisphere (SH)

Vertical **coupling in the SH** is detectable by **looking at circulation variability**.

(Variability = statistical behaviour of difference between weather and climatology, i.e. expected average climate.)

Coupling in the Southern Hemisphere (SH)

Vertical **coupling in the SH** is detectable by **looking at circulation variability**.

(Variability = statistical behaviour of difference between weather and climatology, i.e. expected average climate.)

Features of SH variability:

- Maximises between October to December (beginning of austral summer)
- Especially at high-latitudes $[90 : 50]^{\circ}S$
- Large at both stratosphere and troposphere
- Why relevant?

Prominent seasonal phenomena affected (Southern Annular Mode, El Niño Southern Oscillation).

Tropospheric variability

Variable : **Eddy-driven Jet**(Jet), west-to-east wind current confined at mid/high latitude. This jet is generated by momentum convergence of small scale turbulence.

Climatology: the Jet **shifts its center of mass towards the equator**

Variability: **very variable timing of shift**

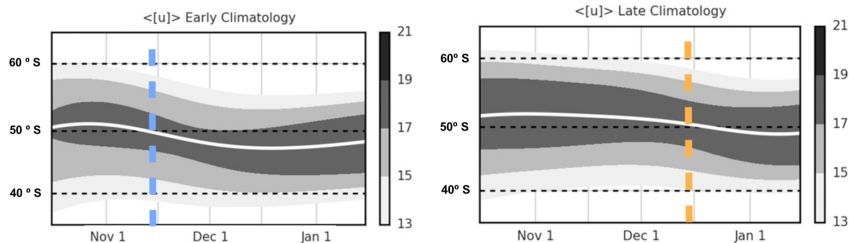


Figure: High-integrated zonal-mean zonal wind $\langle [u] \rangle(t, lat)$ in ms^{-1} (grey) and its maximum as a Jet proxy (white). Early and late timing for the shift. From Byrne (2017).

In the mean time...Stratospheric variability

Variable : **Polar Vortex** (PoV), large-scale region of air contained by a strong west-to-east jet stream circling the polar region. Exists only in winter.

Climatology: **PoV experiences its springtime breakdown** (loses strength due to increased solar absorption).

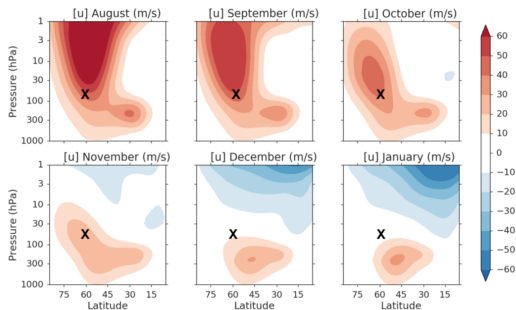


Figure: Climatology of zonal circulation from August to December. 'X' = approximate PoV location. Adapted from Byrne et al. (2018).

In the mean time...Stratospheric variability

Variable : **Polar Vortex** (PoV), large-scale region of air contained by a strong west-to-east jet stream circling the polar region. Exists only in winter.

Climatology: **PoV experiences its springtime breakdown** (loses strength due to increased solar absorption).

Variability: Notably, its **timing is highly variable too**

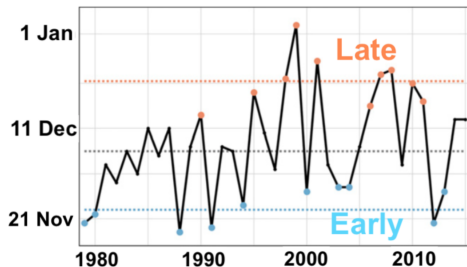


Figure: Breakdown dates of each year on record. Byrne et al. (2017)

PoV associated with Jet shift

- Studies on **reanalysis data** (Black 2007; Byrne 2017,2018) show PoV breakdown **strongly associated** with Jet shift.
Qualitative analysis: composite and dripping paint plots (not shown);
quantitative: correlations (Figure).
- Physical arguments (backed by some numerical experiments, Sun 2009) suggest **association is likely to be a downward influence**.

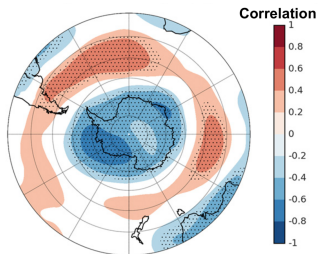


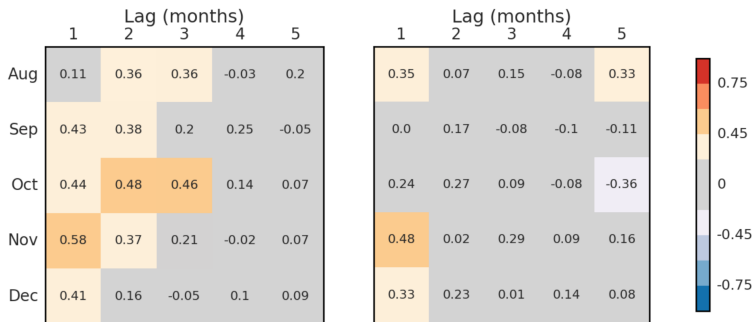
Figure: Correlation between PoV breakdown dates and tropospheric DJF circulation anomalies: significant at high latitudes. From Byrne et al.(2018)

Lagged correlation to quantify predictability

Byrne et al. (2018) want to **quantify potential predictability of troposphere given the state of the stratosphere**.

They perform a statistical analysis of **observations**(reanalysis) using **lagged correlations**:

$$\text{entry} = \rho(X_m, Y_{m+\text{lag}})$$



(a) $\rho(PoV_m, Jet_{m+lag})$

(b) $\rho(Jet_m, Jet_{m+lag})$

Problematic aspects of the works presented so far:

- **qualitative** analysis (Baldwin 2001; Black 2007; Byrne 2017) : only show qualitative association between PoV and Jet. No influence, no directionality.
- use of **lagged correlation** to infer predictability / causation (Byrne 2018): is unfit for purpose as prone to biases!

Goal:

Use **reanalysis data** to infer direction, strength and time-scale of the coupling **dynamics** using a **quantitative** and **casual** approach.

Misuse of lagged cross-correlation

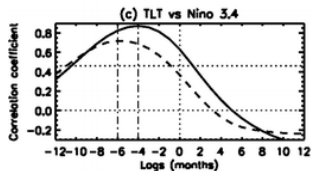


Figure: Gu et al. (2011)

In many **meteorological studies**, **max of the lagged cross correlation** used to **detect time scale , direction and strength of influence** between pairs of variables.

However...

Misuse of lagged cross-correlation

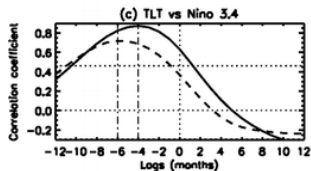


Figure: Gu et al. (2011)

In many meteorological studies, **max of the lagged cross correlation** used to **detect time scale , direction and strength of influence** between pairs of variables.

However...

Lagged cross correlation

Given a pair of jointly wide sense stationary processes (X_t, Y_t) :

$$\rho(X_t, Y_{t+\tau}) \doteq \frac{\text{Cov}[X_t, Y_{t+\tau}]}{\sigma_X \sigma_Y} \quad \forall t$$

- Measure of linear relation \Rightarrow **Max of ρ = delayed signals best aligned**
- Effective in context of signal processing (GPS, radar echolocation)
- **No causation** (depends on joint probability)

Causal Discovery via Bayesian networks

Big data + computational power available + Earth's climate is a high dimensional complex systems ... perfect for

- **Causal discovery**: given observations for N variables evolving in time and interacting via some unknown relations, can we infer dependencies?

Causal Discovery via Bayesian networks

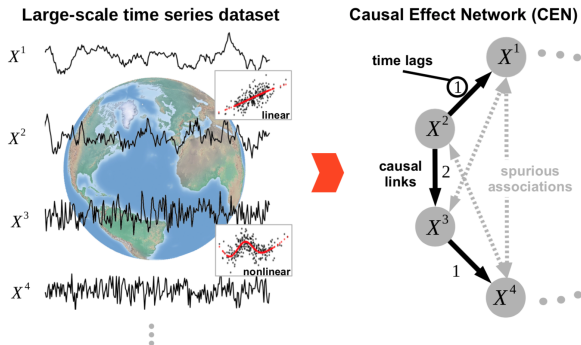
Big data + computational power available + Earth's climate is a high dimensional complex systems ... perfect for

- **Causal discovery**: given observations for N variables evolving in time and interacting via some unknown relations, can we infer dependencies?
- Cannot use lagged cross-correlation: many links would be spurious!

Causal Discovery via Bayesian networks

Big data + computational power available + Earth's climate is a high dimensional complex systems ... perfect for

- **Causal discovery:** given observations for N variables evolving in time and interacting via some unknown relations, can we infer dependencies?
- Cannot use lagged cross-correlation: many links would be spurious!
- **Use Information Theory measures on Bayesian Causal Networks**



Definition of Time-series Causal graph

Be a N-variate process \mathbf{X} with set of components V .

\mathbf{X} is associated to its **time-series causal graph** $\mathcal{G} = (V \times \mathbb{Z}, E)$ where

- node (v, t) :
each individual variable $v \in V$ at a specific time $t \in \mathbb{Z}$
- link $e \in E$:
lag-specific causal link between variables $X_{t-\tau}$ and Y_t if and only if

$$\tau > 0 \text{ and } X_{t-\tau} \not\perp\!\!\!\perp Y_t \mid \mathbf{X}_t^- \setminus \{X_{t-\tau}\}$$

where $\not\perp\!\!\!\perp$ means not independent and $\mathbf{X}_t^- := (\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \dots)$ is the past of the whole process.

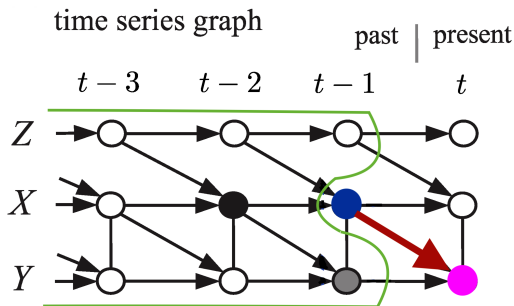


Figure: Representation of time-series causal graph, under the hypothesis of causal stationarity.

- In words, a link (\rightarrow) exists if there is some **information flowing** from $X_{t-\tau}$ to Y_t that is **not already contained in any of the nodes** in the **past of the process**.
- If each link exist for all t i.e. causal graph is invariant under time translation : **causal stationarity**.

How to measure $X \perp Y | Z$?

- **Conditional Mutual Information (CMI)**

$$I(X, Y | Z) \doteq \int \int \int p(x, y, z) \log \left[\frac{p(x, y | z)}{p(x | z)p(y | z)} \right] dx dy dz$$

Good because $X \perp Y | Z \iff I(X, Y | Z) = 0$

How to measure $X \perp Y | Z$?

- **Conditional Mutual Information (CMI)**

$$I(X, Y | Z) \doteq \int \int \int p(x, y, z) \log \left[\frac{p(x, y | z)}{p(x | z)p(y | z)} \right] dx dy dz$$

Good because $X \perp Y | Z \iff I(X, Y | Z) = 0$

- Therefore $I_{X \rightarrow Y}^{LINK}(\tau) \doteq I(X_{t-\tau}, Y_t | \mathbf{X}_t^- \setminus \{X_{t-\tau}\})$ measure for link existence

How to measure $X \perp Y | Z$?

- **Conditional Mutual Information (CMI)**

$$I(X, Y | Z) \doteq \int \int \int p(x, y, z) \log \left[\frac{p(x, y | z)}{p(x | z)p(y | z)} \right] dx dy dz$$

Good because $X \perp Y | Z \iff I(X, Y | Z) = 0$

- Therefore $I_{X \rightarrow Y}^{LINK}(\tau) \doteq I(X_{t-\tau}, Y_t | \mathbf{X}_t^- \setminus \{X_{t-\tau}\})$ measure for link existence
- To remove spurious links it is sufficient to evaluate the **Momentary Information Transfer (MIT)**(Runge et al. (2012))

$$I_{X \rightarrow Y}^{MIT}(\tau) \doteq I(X_{t-\tau}, Y_t | \mathcal{P}_{Y_t} \setminus \{X_{t-\tau}\} \cup \mathcal{P}_{X_{t-\tau}})$$

where parents of a variable are $\mathcal{P}_{Y_t} \doteq \{Z_{t-\tau} : Z \in \mathbf{X}, \tau > 0, Z_{t-\tau} \rightarrow Y_t\}$.

How to measure $X \perp Y | Z$?

- **Conditional Mutual Information (CMI)**

$$I(X, Y | Z) \doteq \int \int \int p(x, y, z) \log \left[\frac{p(x, y | z)}{p(x | z)p(y | z)} \right] dx dy dz$$

Good because $X \perp Y | Z \iff I(X, Y | Z) = 0$

- Therefore $I_{X \rightarrow Y}^{LINK}(\tau) \doteq I(X_{t-\tau}, Y_t | \mathbf{X}_t^- \setminus \{X_{t-\tau}\})$ measure for link existence
- To remove spurious links it is sufficient to evaluate the **Momentary Information Transfer (MIT)**(Runge et al. (2012))

$$I_{X \rightarrow Y}^{MIT}(\tau) \doteq I(X_{t-\tau}, Y_t | \mathcal{P}_{Y_t} \setminus \{X_{t-\tau}\} \cup \mathcal{P}_{X_{t-\tau}})$$

where parents of a variable are $\mathcal{P}_{Y_t} \doteq \{Z_{t-\tau} : Z \in \mathbf{X}, \tau > 0, Z_{t-\tau} \rightarrow Y_t\}$.

1. $I^{MIT} = 0 \iff I^{LINK} = 0$
2. $I^{MIT} \geq I^{LINK}$ so MIT more detectable
3. Computing MIT doable for large networks, as dimensionality of conditions is reduced
4. MIT is reliable estimate for link strength (see theorems on coupling-strength autonomy)

Causal Discovery algorithm, Runge et al.(2014,2018)

Idea of PCMCI: for each pair of lagged X_t and $Y_{t-\tau}$ compute $I(X, Y | Z)$.
Choose Z a low-dimensional set made of X and Y 's "potential parents" .

1. PC-step:

Start from a fully connected t.s. graph

for each variable $\{X^i\}_{i=1}^N$:

reduce full past process $Z^i \rightarrow Z^{i*}$ =potential parents
(via few low dimensional cond. independence tests)

2. MCI-step:

for each pair $(X^i_{t-\tau}, X^j_t)$ and $\tau = 0, 1, ..\tau_{max}$:

compute $I(X^i_{t-\tau}, X^j_t | Z_t^{j*} \cup Z_{t-\tau}^{i*}) \rightarrow (value, p\text{-val})$

if $p\text{-val} < \alpha$:

assign link $X_{t-\tau} \rightarrow Y_t$

Parameters:

- implementation for CI test (linear, non-linear, nearest-neighbour)
- significance level α , max lag τ_{max}

Interpretation and caveats

A link $X_{t-\tau} \rightarrow Y_t$ stands for **conditional independence between $X_{t-\tau}$ and Y_t is unlikely within the dataset**:

- Not necessarily physical causation: no experiment available
- Ok for hypothesis testing
- Ok for identification of optimal predictors within dataset

Causal discovery is well posed only if :

- Separation in the graph equivalent to independence in the process
- All common drivers are included in data

And the **PCMCI implementation is suitable if:**

- Causal stationarity, adequate choice of CI test, no measurement errors

Vertical coupling in the SH: data

1. stratospheric polar vortex index:

$$\text{PoV}(\text{day}) = [u](\text{day}, \phi = -60^\circ, \text{press} = 50\text{hPa}) \text{ (ms}^{-1}\text{)} \quad (1)$$

2. tropospheric eddy-driven jet index:

$$\text{Jet}(\text{day}) = \sum_{\phi=-50^\circ}^{-65^\circ} \sin(\phi) [u](\text{day}, \phi, \text{press} = 850\text{hPa}) \text{ (ms}^{-1}\text{)}. \quad (2)$$

where $\text{day} \in [01/01/1979 : 31/12/2016]$ and $[u]$ is the daily measure of the zonal-mean ($[\]$) zonal-wind (u) field.

ERA-Interim reanalysis data for u .

Yearly time-series (grey) and climatologies (black):

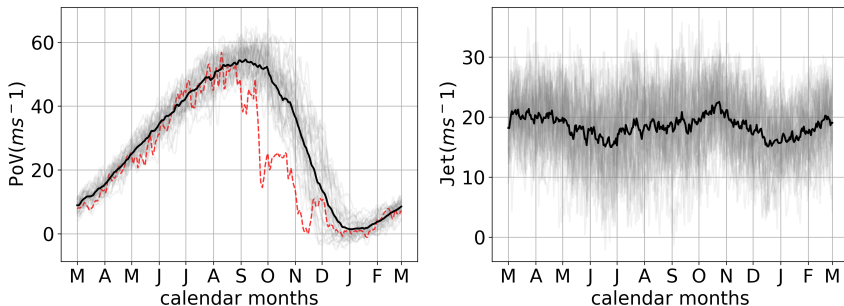


Figure: Red: year 2002, outlier because is the only SH stratospheric sudden warming on record.

Causal analysis 1: month-to-month time selection

For $X, Y \in \{PoV, Jet\}$ and $m \in \{A, S, O, N, D\}$, $lag \in [+1 : +5]$ compare:

1 **Correlation** : $\rho(X_m, Y_{m+lag})$

Causal analysis 1: month-to-month time selection

For $X, Y \in \{PoV, Jet\}$ and $m \in \{A, S, O, N, D\}$, $lag \in [+1 : +5]$ compare:

1 **Correlation** : $\rho(X_m, Y_{m+lag})$

2 **Causation via ParCorr-MIT** : $\rho(X_m, Y_{m+lag} | Z)$

- Z found with PCMCI algorithm
- CI test chosen is Partial Correlation
 $\rho(X, Y | Z) \doteq \rho(X_Z, Y_Z)$ where X_Z, Y_Z residual of linear regression of X, Y onto Z.

Causal analysis 1: month-to-month time selection

For $X, Y \in \{PoV, Jet\}$ and $m \in \{A, S, O, N, D\}$, $lag \in [+1 : +5]$ compare:

1 **Correlation** : $\rho(X_m, Y_{m+lag})$

2 **Causation via ParCorr-MIT** : $\rho(X_m, Y_{m+lag} | Z)$

- Z found with PCMCI algorithm
- CI test chosen is Partial Correlation
 $\rho(X, Y | Z) \doteq \rho(X_Z, Y_Z)$ where X_Z, Y_Z residual of linear regression of X, Y onto Z.

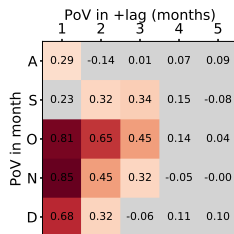
Why ParCorr-MIT?

Being consistent with Corr \Rightarrow test for linear dependency

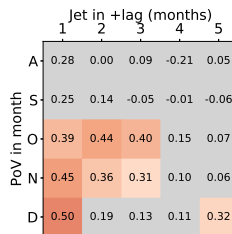
$$\Rightarrow I(X, Y | Z) = -\frac{1}{2} \log(1 - \rho(X, Y | Z)^2).$$

Because linear CMI = 0 \Leftrightarrow ParCorr = 0, then ParCorr-MIT equivalent to MIT to detect link.

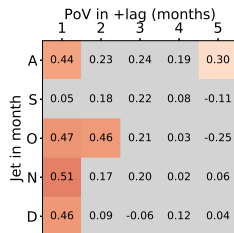
Correlation matrices



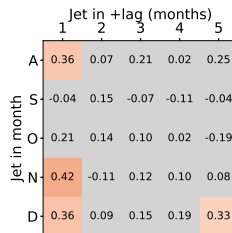
(a) PoV



(b) PoV before Jet

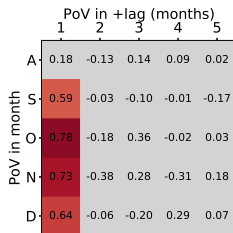


(c) Jet before PoV

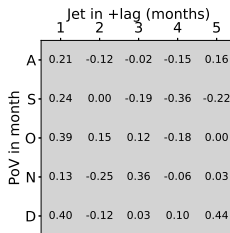


(d) Jet

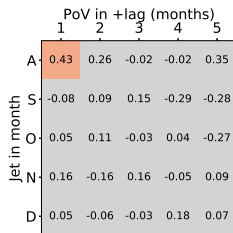
Partial Correlations matrices



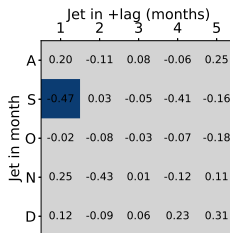
(e) PoV \rightarrow PoV



(f) PoV \rightarrow Jet !



(g) Jet \leadsto PoV



(h) Jet \leadsto Jet

Why Cross-correlations are biased

- **Large PoV auto-dependency strongly inflates the values of both lagged cross-correlations:**¹

Strong auto-dependence detected is on PoV, so is the main confounding effect removed by ParCorr-MIT:

$$\rho(\text{PoV}_m, \text{Jet}_{m+\text{lag}}) \gg 0 \text{ becomes } \rho(\text{PoV}_m, \text{Jet}_{m+\text{lag}} \mid \text{PoV}_{m-1}) \sim 0$$

and same $\rho(\text{Jet}_m, \text{PoV}_{m+\text{lag}} \mid \text{PoV}_{m+\text{lag}-1}) \sim 0$.

- **Are PoV and Jet decoupled on this time scale? Not necessarily!**

Can be due to low sample size used in month-to-month matrices (37 data points for each entry) .

...With larger sample-size might be able to detect other links.

¹ Inflation of cross-corr due to strong auto-corr is proved in linear theory on MIT (Runge 2013).

Causal analysis 2: breakdown-dependent time-window

Except timing, assume physical mechanisms explaining spring-to-summer variability should be the same each year (causal stationarity).

The transition starts about 2-3 months before the breakdown .

⇒ **In each year j select the days $[t_j - 70 : t_j + 20]$ with t_j = Polar Vortex breakdown date.**

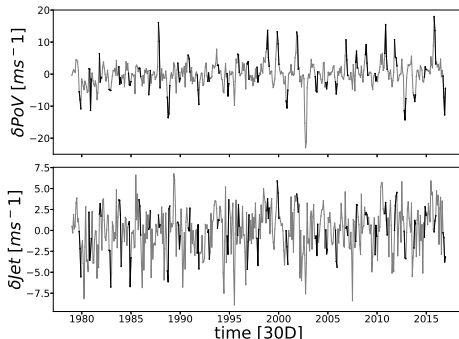
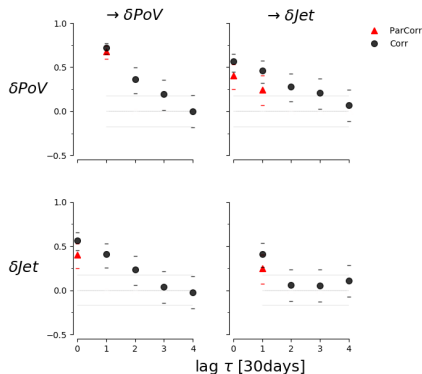


Figure: Total sample size 111 days (black).

Intra-seasonal Causal Network

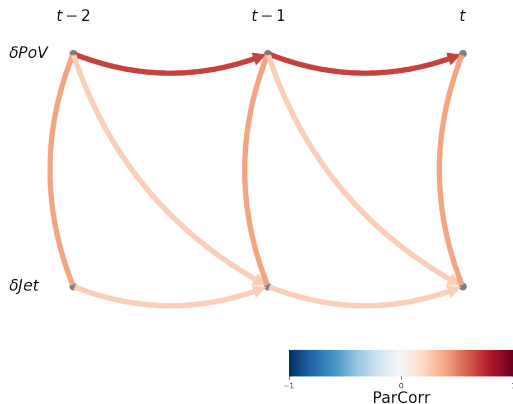
PCMCI gives **4 ParCorr-MIT values significant** at 0.025 confidence level (red). Compared with **correlation values** (black).

NB: Because assumed causal stationarity, measures are only function of lag.



Intra-seasonal Causal Network

Which translates into graph:



Intra-seasonal shift modelled by VAR(1)

CI = ParCorr \Rightarrow the Causal Network approximates a VAR(1) :

$$\delta\text{PoV}_t = a \delta\text{PoV}_{t-1} + \epsilon_t^P$$

$$\delta\text{Jet}_t = b \delta\text{Jet}_{t-1} + c \delta\text{PoV}_{t-1} + \epsilon_t^J$$

with linear coefficients and covariance matrix of 2d noise term ²:

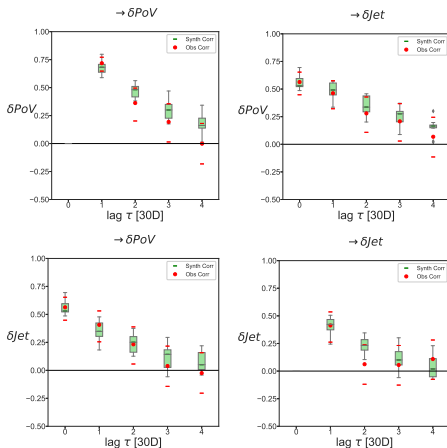
$$\mathbf{a} = \mathbf{0.68} \pm \mathbf{0.06} ; \mathbf{b} = \mathbf{0.19} \pm \mathbf{0.09} ; \mathbf{c} = \mathbf{0.37} \pm \mathbf{0.09};$$

$$\Sigma = \begin{bmatrix} \sigma_P^2 & \sigma_{PJ} \\ \sigma_{PJ} & \sigma_J^2 \end{bmatrix} = \begin{bmatrix} 0.53 & 0.24 \\ 0.24 & 0.75 \end{bmatrix}$$

²Coeffs. form linear fit of each variable onto its “parents” + 2d Gaussian fit on residuals.

A first validation using lagged correlations

Lagged correlations of VAR(1) synthetic time series vs data?








Conclusions

- 1 **PoV strong auto-correlation inflates all cross-correlations .**
- 2 Found intra-seasonal **causal link from PoV to Jet** when enough statistics and adequate time-window selection (PoV-breakdown dependent) .
- 3 PoV linear **statistical predictor** of Jet ? It does reproduce well correlations...

Future work

- Short term : explore implications of VAR(1) in the SH context.
- Long term 1 : how can we improve confidence in results from **small sample size data**? (eg Knock-off (Barber2015))
- Long term 2 : apply CEN techniques to **climate models' outputs**.
Do models reproduce observed causal paths (rather than correlations)?
Can we use it to constraint inter-model spread in long term projection?
How to account for model error?

References I

-  R. X. Black et al. “Interannual Variability in the Southern Hemisphere Circulation Organized by Stratospheric Final Warming Events”. In: *Journal of the Atmospheric Sciences* 64.8 (Aug. 2007), pp. 2968–2974.
-  N. J. Byrne et al. “Nonstationarity in Southern Hemisphere Climate Variability Associated with the Seasonal Breakdown of the Stratospheric Polar Vortex”. In: *Journal of Climate* 30.18 (Sept. 2017), pp. 7125–7139.
-  N. J. Byrne et al. “Seasonal Persistence of Circulation Anomalies in the Southern Hemisphere Stratosphere and Its Implications for the Troposphere”. In: *Journal of Climate* 31.9 (May 2018), pp. 3467–3483.
-  P. H. Haynes. “Stratosphere-Troposphere coupling”. In: *SPARC Newsletter* 25 (2005), pp. 27–30.
-  J. Runge et al. “Detecting causal associations in large nonlinear time series datasets”. <https://arxiv.org/pdf/1702.07007.pdf>. June 2018.

References II



J. Runge et al. “Escaping the Curse of Dimensionality in Estimating Multivariate Transfer Entropy”. In: *Physical Review Letters* 108.25 (June 2012).



J. Runge et al. “Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy”. In: *Physical Review E* 86.6 (Dec. 2012).



J. Runge et al. “Quantifying the Strength and Delay of Climatic Interactions: The Ambiguities of Cross Correlation and a Novel Measure Based on Graphical Models”. In: *Journal of Climate* 27.2 (Jan. 2014), pp. 720–739.



Elena Saggioro. “A causal approach to climate variability in the Southern Hemisphere”. MA thesis. University of Reading and Imperial College London, 2018.



M. Sigmond et al. “Enhanced seasonal forecast skill following stratospheric sudden warmings”. In: *Nature Geoscience* 6.2 (Jan. 2013), pp. 98–102.