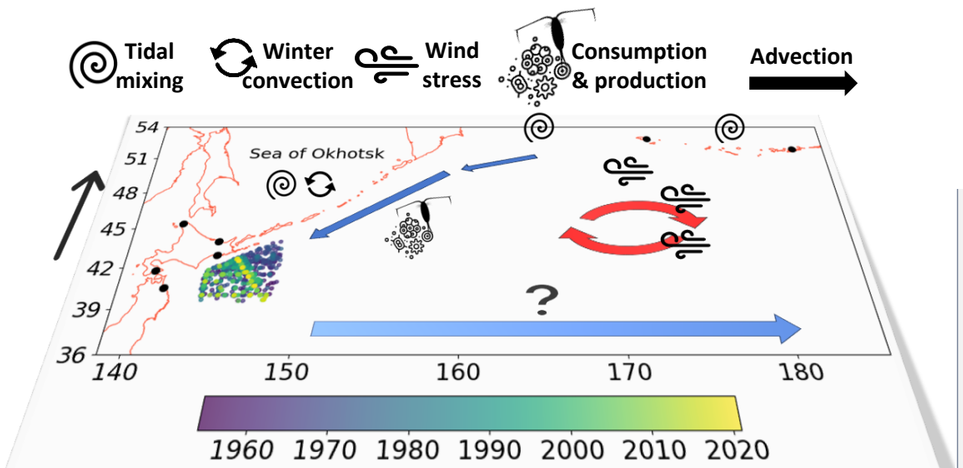
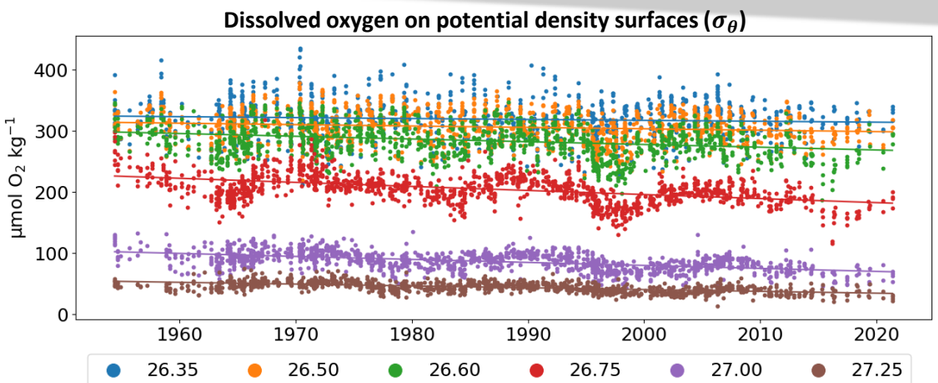
Causal inference in oceanography - new tools for tackling physical and biogeochemical questions with observational data

Benedict Blackledge[1], Oliver Andrews[1,2], Urmi Ninad[3], Mattias Green[4], Rory Bingham[1]

The problem: "Does the nodal tide cause the bi-decadal oscillations in dissolved oxygen observed in the Oyashio region of the North-West Pacific?"

The phrase 'correlation is not causation' is familiar to many. While true, the researcher may be left wondering what causation even is…
Methods to attribute remote causes are often based on lagged regression. These are associative measures, which often perform poorly due to confounding variables and non-linear causal links [1,2]. In fact, it can be shown that such measures do not necessarily detect causality even in principle [3].
We apply a causal inference approach to the long-standing question of whether variations in remote tidal mixing drive the high oxygen variability in the Oyahsio region [4], off Hokkaido, Japan.





What can we assume about the system?

- **Cold, oxygenated, relatively fresh and low PV water** leaves the Sea of Okhotsk (SO) [4]
- **Tidal mixing, particularly in the diurnal band, takes place within SO and its outflow channels**.
- Brine rejection by sea-ice formation permits oxygenated water to reach denser isopycnals than convection alone**, but mixing is most likely required to reach even denser levels**
- **The transport of the oxygenation signal is linear**

The data!

- **The quality controlled dissolved oxygen, temperature, salinity and isopycnal depth time series were kindly provided by Daisuke Sasano (Japan Meteorological Agency)**
- **The nodal tidal index is derived from tide gauges (black dots, map above).**
- **The index of the Pacific decadal oscillation (PDO) is the US NCEI PDO index**

Our approach: PCMCI applied to marine oxygen time series

PCMCI allows us to estimate the lagged dependencies within a dataset, with a minimal number of pre-specified parameters. It is a combination of the PC algorithm (for the identification of causal graph networks) [5] and the momentary conditional independence test (MCI) [1] – The methods are implemented using the **Tigramite package**.

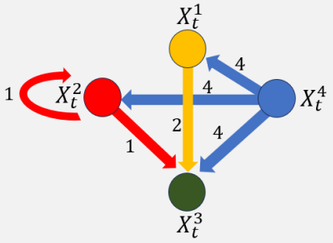
PCMCI applies a conditional independence test (CI) to two variables at a time, one of which is lagged and a conditioning set $\boldsymbol{S}$, which may contain the causal 'parent' variables: $\boldsymbol{CI}\left(X^i_{t-\tau},\ X^j_t,\ \boldsymbol{S}\right)$ to evaluate $X^i_{t-\tau}X^j_t \mid \boldsymbol{S}$. Here we use the linear test, partial correlation (ParCorr). Refer to [1] for the details, but **ParCorr removes the effect of $\boldsymbol{S}$ on $X^i$and X^j using OLS regression, then tests for independence of the residuals**.

But to efficiently evaluate $\boldsymbol{CI}\left(X^i_{t-\tau},\ X^j_t,\ \boldsymbol{S}\right)$, $\boldsymbol{S}$ should be as small as possible – ideally, we remove irrelevant variables. This is done via the PC selection step (see below). For a detailed Earth sciences example, see [6].

Example 'process' graph..

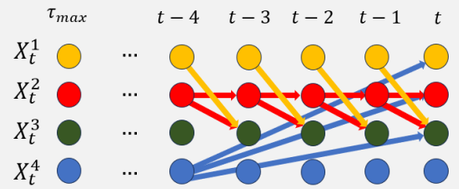
The system of interest, such as the Oyashio time series, can be expressed as a graph network. The causal effects occur at some lag (numbers)

The causal links can be linear or non-linear. **We make a linear assumption**



.. and as a time series graph

The true causal links between nodes (variables) exist at some lag, t. **These links and their lags are what we aim to identify**

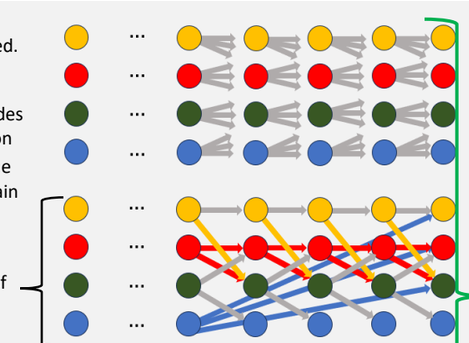


Part 1: apply the PC method

A fully connected graph is generated.

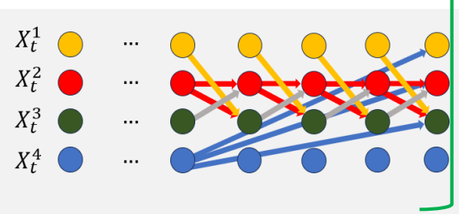
Iteratively remove nodes. Only nodes that are conditionally dependent on X^j_tat the required significance value are kept. The new set should contain the true causal links, but probably spurious ones as well

This is now the reduced superset of parents (variables at some lag, for all variables), $\boldsymbol{S}$

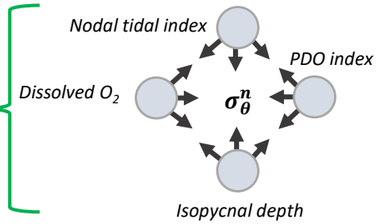


Part 2: apply the MCI test – the causal discovery step

Run the conditional independence test for every pair of nodes in $\boldsymbol{S}$.



Example graph from the start of Part 2 (PC step)

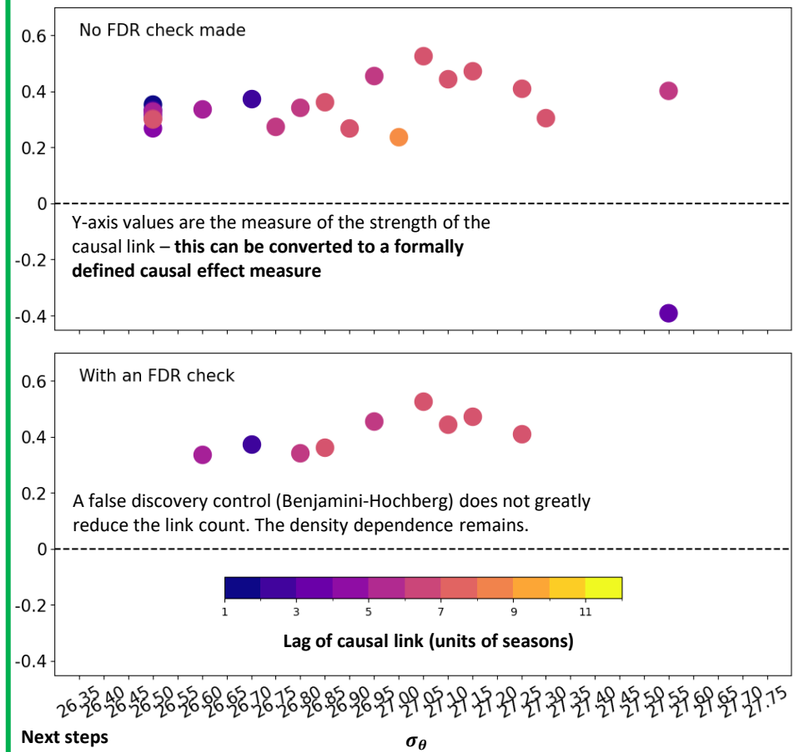


Any questions or thoughts?
uv20102@bristol.ac.uk

Further CI info (+ refs)!

Results: where can we identify a direct link from the nodal tide to dissolved oxygen?

Nodal tidal causal links can be identified within a density range of 26.5 to 27.5. The lags increase with density, with a maximum of 7 (~1.75 years). This is comparable to the estimated transport time from the SO outflows to the Oyashio region [4] found with lagged correlation.



Next steps

- Establish whether the nodal tidal signal in the density range above propagates to the Eastern Pacific – the same analysis applied to the Ocean station Papa record.
- With a causal network established, additional experiments can be performed. Notably, counter-factual analysis, i.e. **If the nodal tide is set to 0 while a strongly positive PDO exists how does the oxygen concentration on σ^n_θ respond?**
- However large uncertainties remain. The networks presented here are simple, with few mediating variables (i.e. Tide → Oxygen, rather than, Tide → Latent variable → Oxygen). These forms can change the outcome of a counter factual analysis considerably, and are the subject of ongoing work

Affiliations
1) University of Bristol, School of Geographical sciences, 2) University of York, Department of Environment and Geography, 3) Technische Universität Berlin, 4) Bangor University, School of Ocean sciences