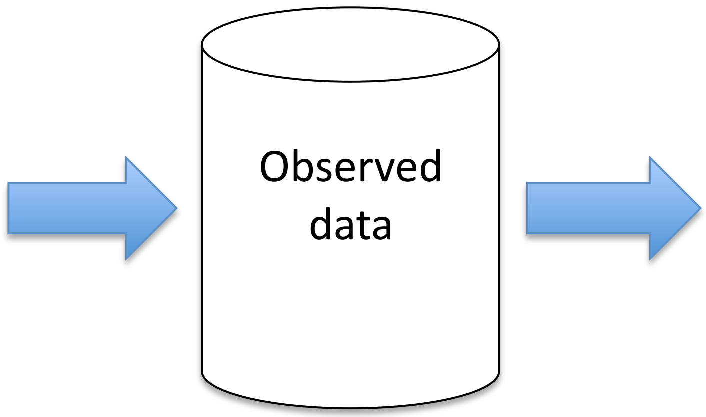
Applying Causal Discovery Methods in the Geosciences

Imme Ebert-Uphoff

Research Faculty, Electrical and Computer Engineering,

Colorado State University

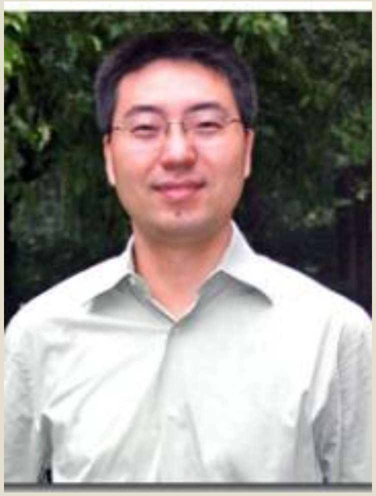


What/how can we learn about physical processes from data?

BIRS workshop – Mar 12, 2016.

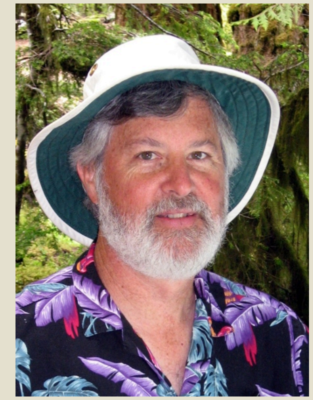
Big Data Tsunami at the Interface of Statistics, Environmental Sciences and Beyond

Collaborators



Yi Deng
Earth and Atmospheric
Sciences, Georgia Tech

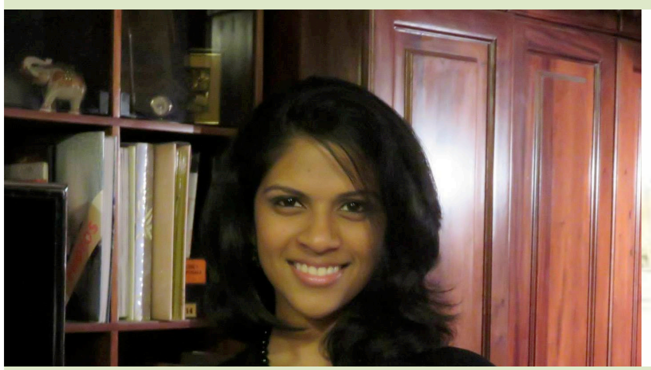
Chuck Anderson
Computer Science
Colorado State



Dorit Hammerling
NCAR

Allison Baker
NCAR

Students at Colorado State:



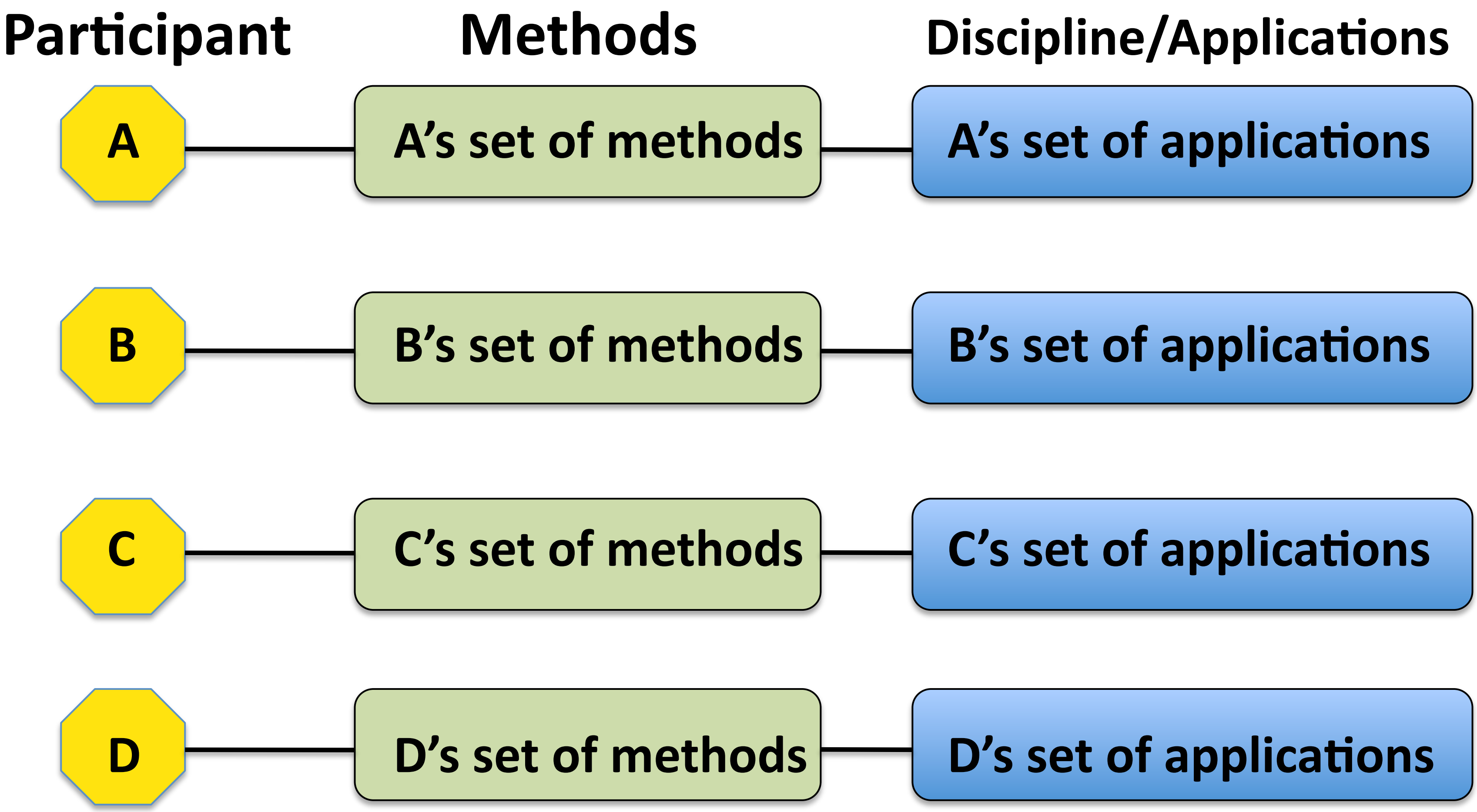
Savini Samarasinghe
Electr. & Comp. Eng.

Christian Rodriguez
Electr. & Comp. Eng.

Melinda Ryan
Computer science

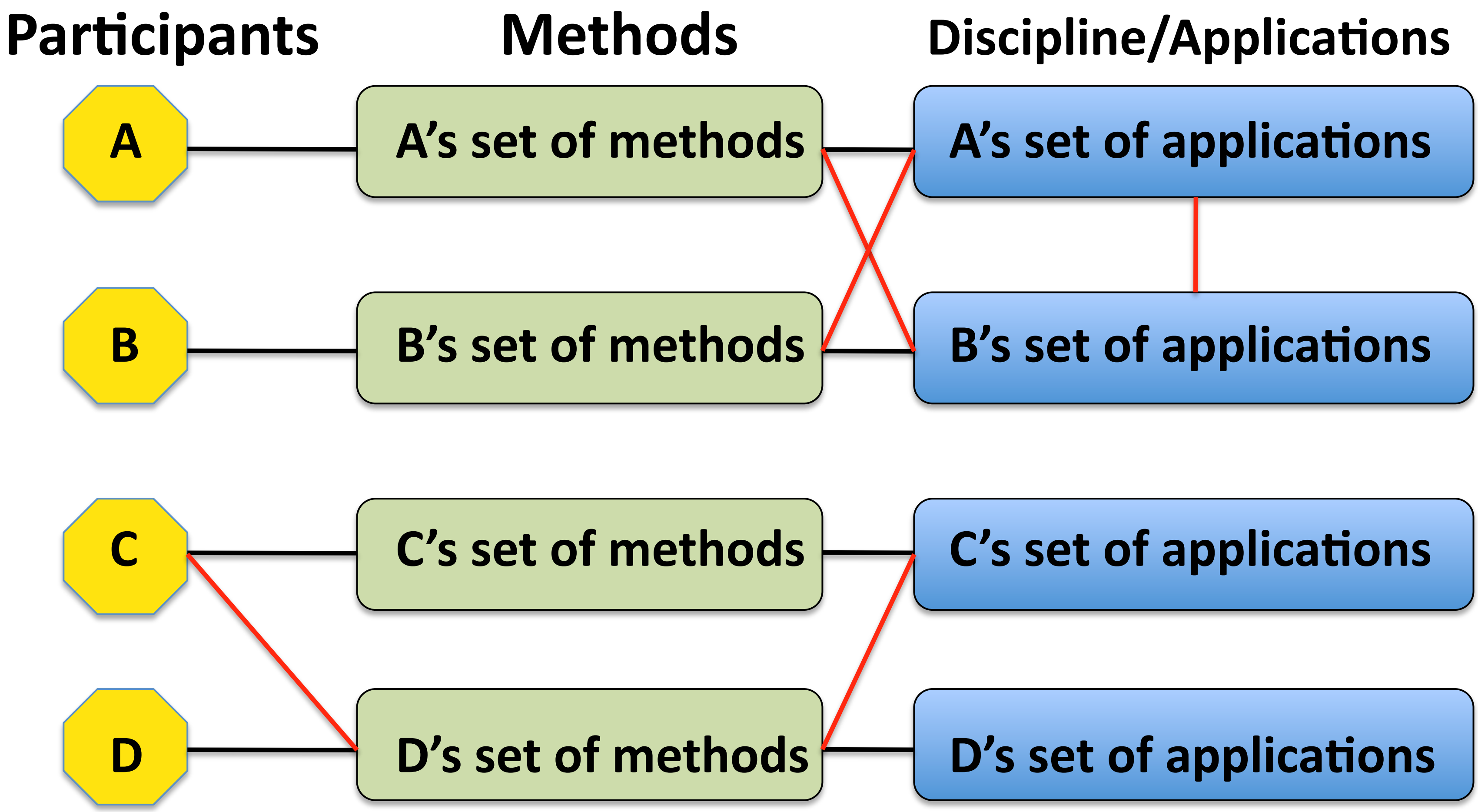
My naïve view of this workshop

Each participant has special knowledge of certain methods and applications.



At the end of the workshop

Cross fertilization: participants learn about new matches between methods and applications.



Applications / methods

In my case:

Method = Causal discovery

Applications = Geosciences, especially atmospheric/climate science, large-scale dynamic processes

Purpose = Scientific discovery (<u>not</u> prediction, <u>not</u> forecasting, etc.)

Typical geoscience applications

- Complex systems; many variables.
- Often spatially distributed → spatio-temporal data
- Data sets are large in size, but that is because
  - Dimensionality is high,
  - While sample size is actually small (often 60 years of daily/monthly/yearly data).
- Properties of many underlying mechanisms not yet fully understood

  → opportunities for **scientific discovery from data**

Reading suggestion

- **Report of "*2015 Workshop on Intelligent and Information Systems for Geosciences*".**
- Yolanda Gil and Suzanne Pierce (+ 32 participants)
- 59 pages.
- **Includes discussion of geoscience applications in need of new analysis methods.**
- **Available at is-geo.org.**

Causal Discovery Theory - 101

Goal: **Learn potential cause-effect relationships from observed data.**

Causal discovery theory

- Provides algorithms for that purpose.
- Based on **Probabilistic Graphical Models**.
- **Input: Observed data.**
- **Output: Graph structure (diagram) showing potential causal connections.**

Terminology:

- If final model is **directed graph**: called "**Bayesian network**"
- If final model is **undirected graph**: called "**Markov network**"

Causal Discovery – quick history

Development:

- Path diagrams (Wright 1921), Granger “causality” (1969)
- Causal calculus: late 1980s (Pearl, Rebane)
- Hidden common causes: Spirtes, Glymour, Scheines (1990s)
- More algorithms: 1980s to now
- Computationally feasible since 1990s
- Constantly pushing boundaries for # of variables.

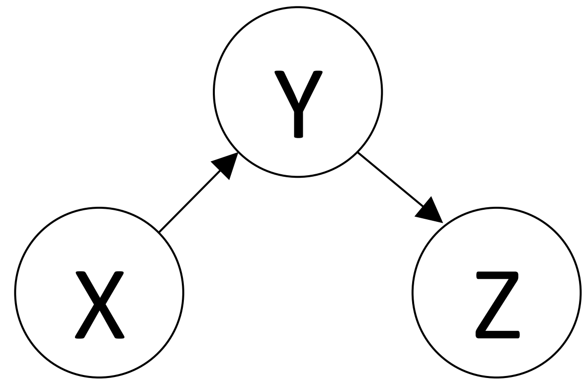
Applications:

- Used extensively in social science and economics (since 1980s)
- 2011: Turing award (=Nobel prize in computer science) to Judea Pearl
- Many **recent success stories in bioinformatics**:
  - identifying gene regulatory networks,
  - identifying protein interactions,
  - discovering neural connections in the brain.
- Emerging in geosciences.

Concept 1: Direct vs. **indirect** connections

Example: See system on right.

Arrows indicate: **cause → effect**.



In this plot:

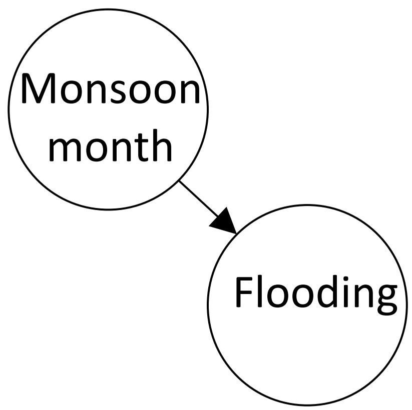
- X is a **direct** cause of Y,
- Y is a **direct** cause of Z,
- X is only an **indirect** cause of Z.

Goal of causal discovery: we want to identify only direct connections. Eliminate all others.

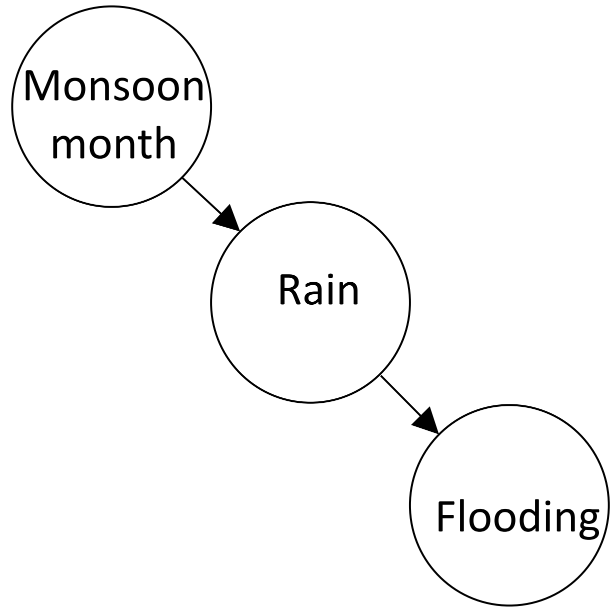
Caution: Directness is **relative** property

One can always transform a direct connection into an indirect one by including an intermediate cause!

Toy example:



Monsoon month is **direct** cause of flooding in this model.



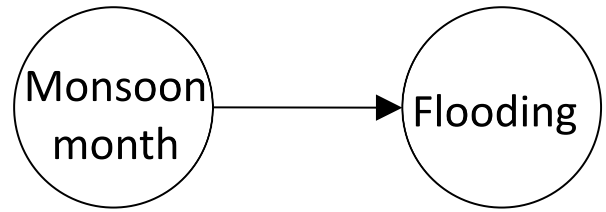
Monsoon month is only **indirect** cause of flooding in this model.

Both models are correct!

Directness is only defined relative to variables included in model.

Concept 2: Causality is probabilistic relationship

Example:



This graph implies:

1) Flooding is *more likely* in monsoon months, but *not* certain.
2) Flooding can also happen outside of monsoon months.

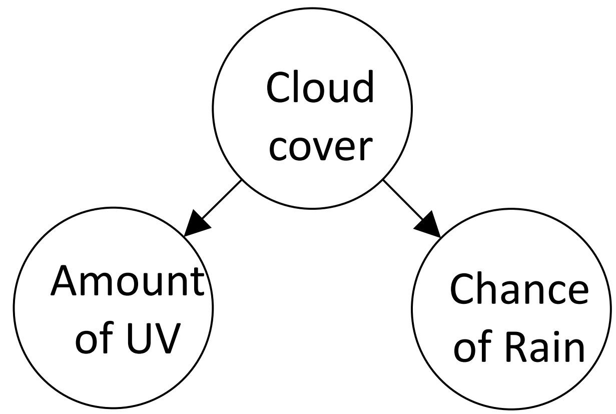
→ Supplement graph with probabilities.
→ **Use framework of "Probabilistic graphical models"**

But:

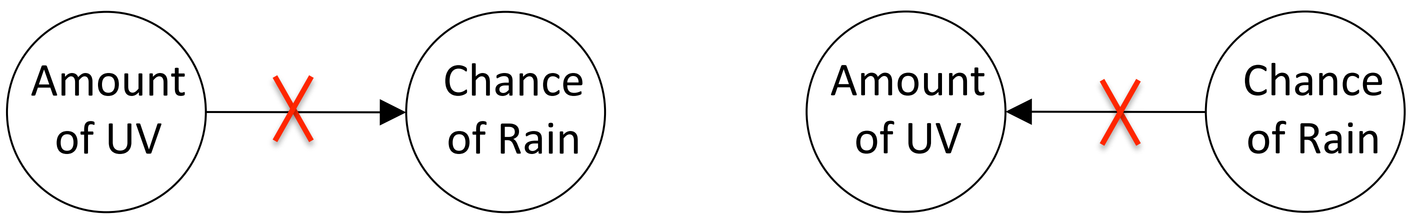
- For our applications we so far do **not** care about the *exact* probabilities.
- Just want to identify **graph** showing *strongest potential* causal connections.

Concept 3: Hidden common causes (latent variables)

Ex.: Cloud cover is **common cause** of UV and rain variables.



If we do not include the common cause in model, results are no longer causal:



Conclusion:

1) We can never prove causal connections.

2) But we can disprove causal connections.

→ Tool for that: Conditional independence tests.

A basic algorithm to find the graph

Use classic statistical tests (e.g. Fisher's Z-test) to detect and eliminate *indirect* connections.

Basic algorithm for learning <u>independence graph</u> from data:

1. Nodes of graph = observed variables.
2. Start with **fully connected graph** = assume that every variable is a cause of every other variable.
3. Eliminate as many edges as possible using conditional independence tests.
4. Establish arrow directions (using more statistical tests and/or temporal constraints).

Whatever is left at end: **potential causal connections.**

(Elimination procedure.)

Assumptions for causal interpretation

A) From <u>data</u> (probability distribution) to <u>independence graph</u>:

Faithfulness: graph model actually models the underlying data well.

1) Probability distributions are i.i.d.
2) No selection bias.
3) If developing directed model, no loops allowed.
4) Causal signals strong enough to be picked up by statistical tests.

B) From <u>independence graph</u> to <u>causal interpretation</u>:

Causal sufficiency: "no hidden common causes"

If any two nodes, X, Y, of the graph have a common cause Z, then Z must also be included in the graph.

Causal sufficiency usually NOT satisfied in geoscience

- There may always be a **hidden common cause** we are not aware of, that cannot be measured, or including them all may make model too complex.
- Need to keep that possibility in mind when interpreting results
  → results are only causal *hypotheses*.
- **Each hypothesis could be direct connection, due to hidden common cause, or combination of both.**

How do we deal with that? Add "evaluation step".

- In results, every link (or group of links) must be checked by domain expert.
- Can we find physical mechanism that explains it?

  If Yes → confirmed.

  If No → **new hypothesis** to be investigated by domain expert

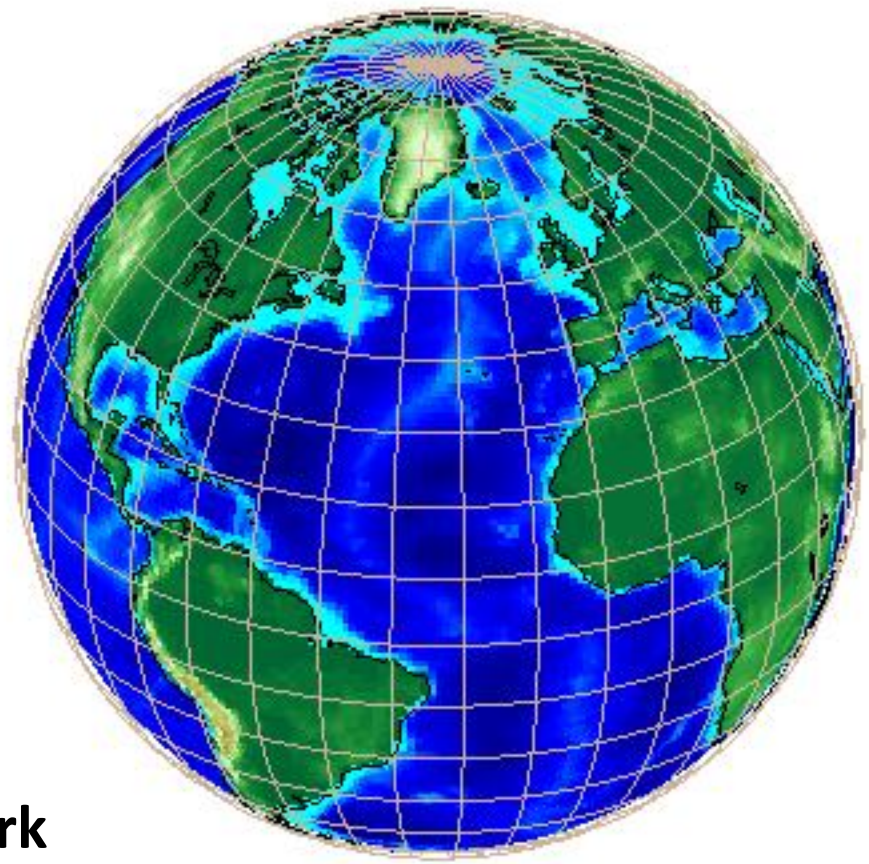
Application 1: Climate Networks

Tsonis and Roebber (2004) introduced "climate networks"

1) Define grid around globe.

2) Evaluate an atmospheric field at all grid points.
→ Time-series data at grid points.

3) Identify all **pairs** of grid points with high correlation
→ **correlation-based climate network**



Existing Climate Networks

Correlation-based climate networks:

- Yield undirected graph, static model.
- Focus on **similarities** between geographical regions
- **Great for identifying tele-connections** (= regions that are far apart, but behave similarly)

Two additional (less common) types:

1. Mutual information network
2. Phase synchronization network

All existing climate networks:

Use **only pair-wise tests** involving data for

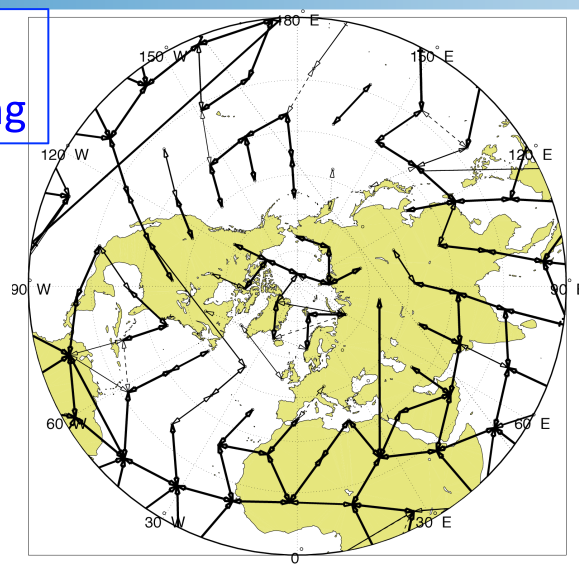
nodes X,Y to decide whether X-Y should be connected.

Example: Interaction maps from geopotential height

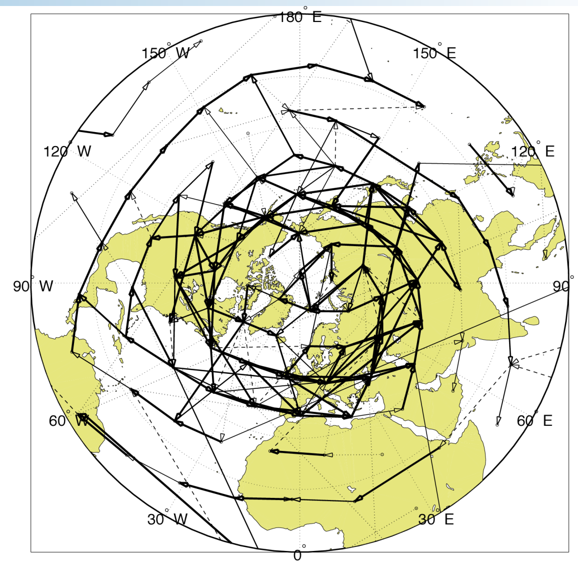
Joint work with Yi Deng

Data:

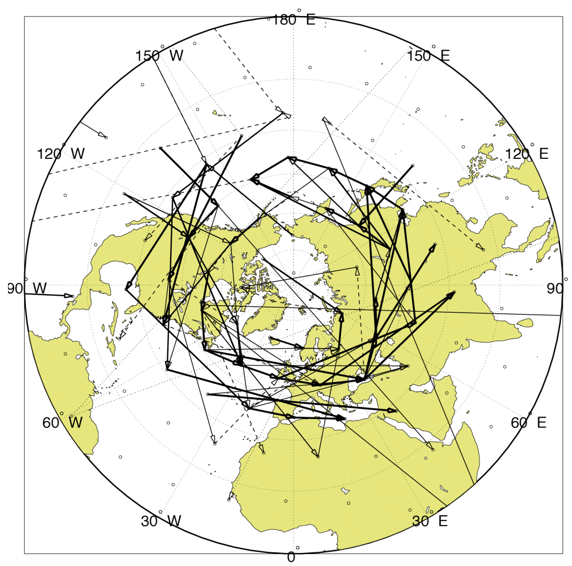
- **500 mb geopotential height**
- NCEP/NCAR Reanalysis
- 1948-2011
- Results for winter (DJF months)
- Fekete grid



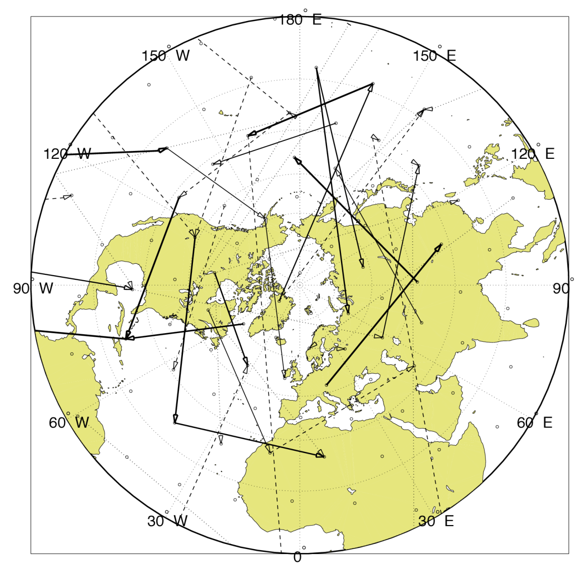
(a) 0-day-delay



(b) 1-day-delay



(c) 2-day-delay



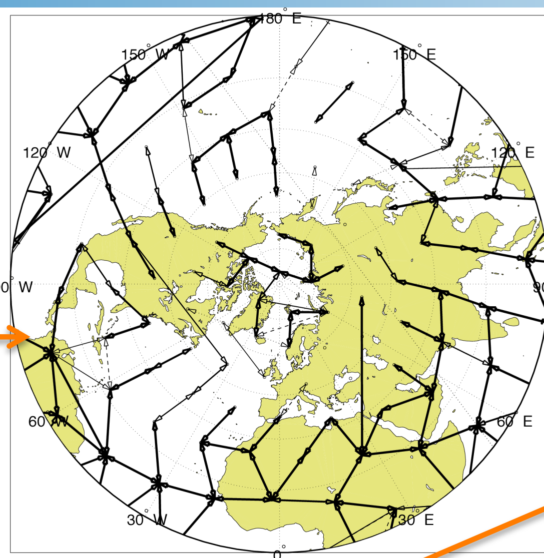
(d) 3-day-delay

Shown here:

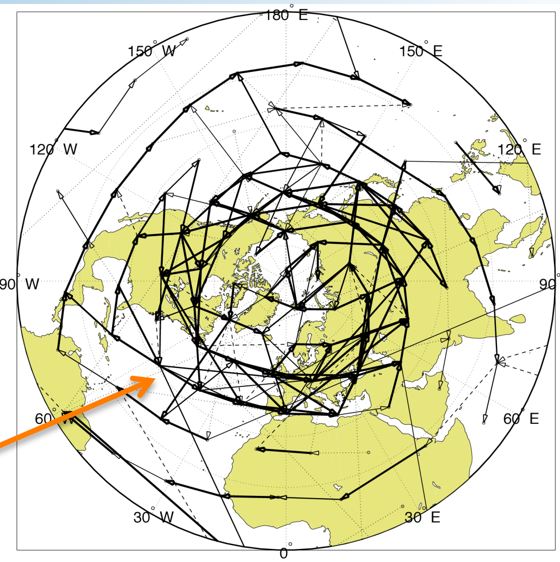
- Stereo-graphic projection (North)
- **Strongest direct connections for 0, 1, 2, 3 days.**

What we learned later from synthetic experiments

But what
Is this ?!?



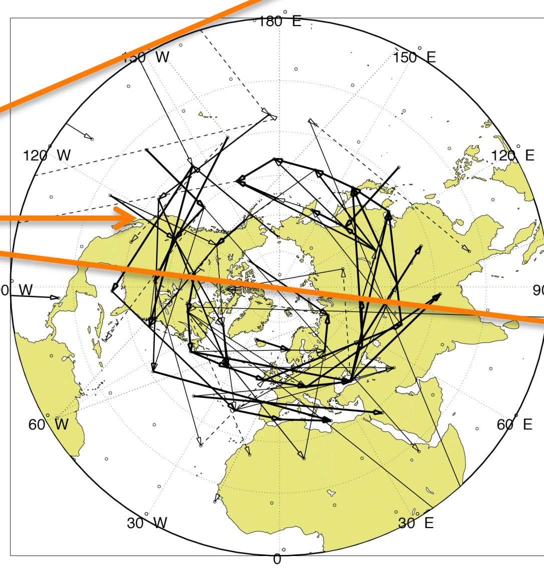
(a) 0-day-delay



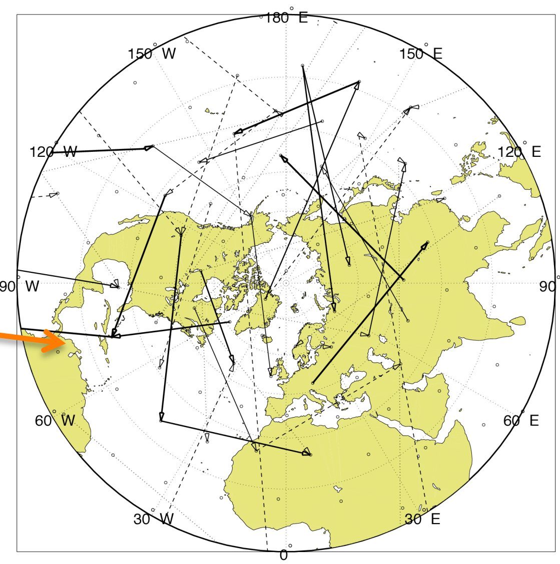
(b) 1-day-delay

This one was clear right away:

This is what **advection** looks like



(c) 2-day-delay



(d) 3-day-delay

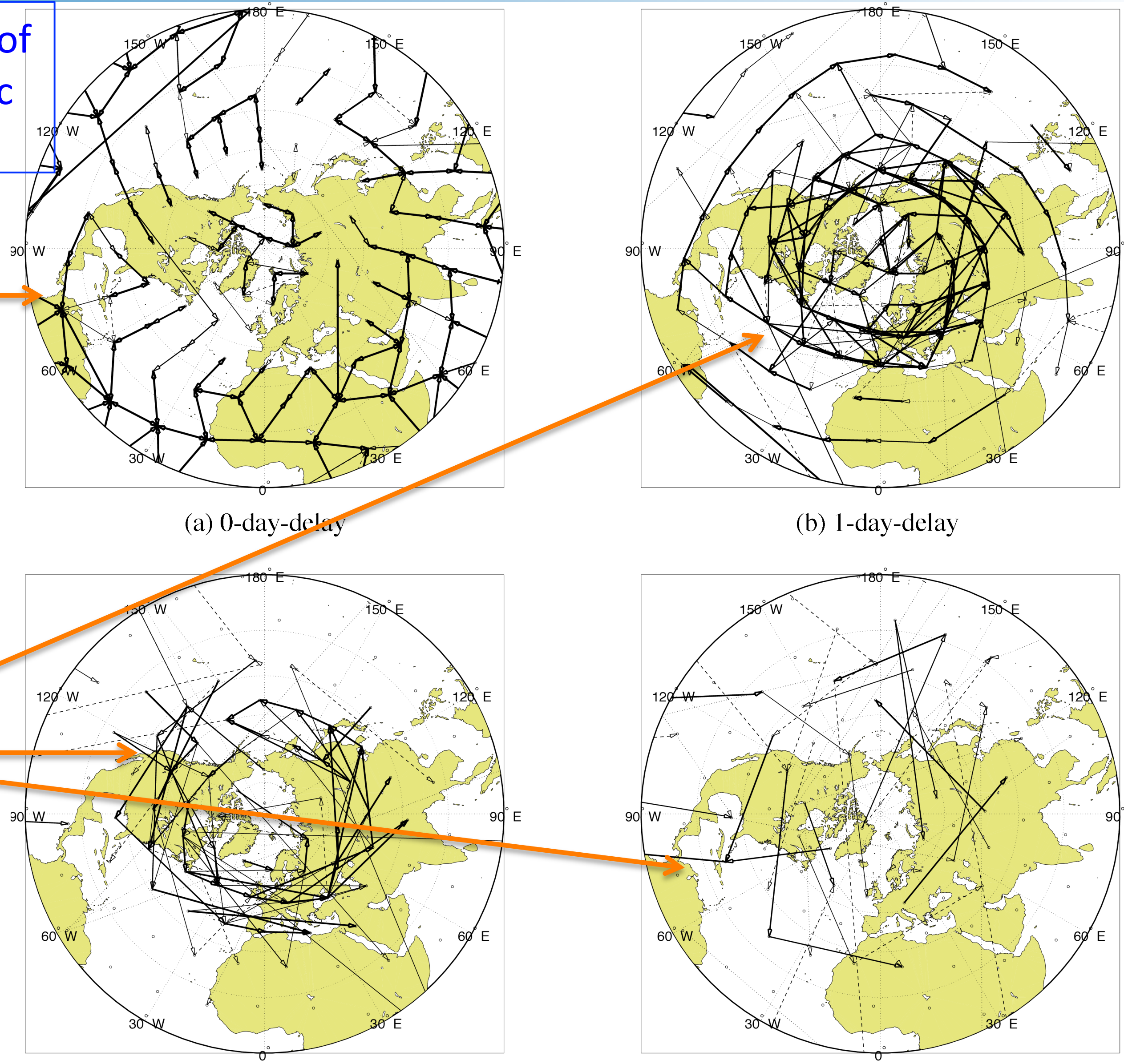
What we learned later from synthetic experiments

It took us 3 years and lots of experiments with synthetic data to find this out:

This is what **diffusion** looks like

This one was clear right away:

This is what **advection** looks like



(a) 0-day-delay

(b) 1-day-delay

(c) 2-day-delay

(d) 3-day-delay

We can now do this in 3D, too!

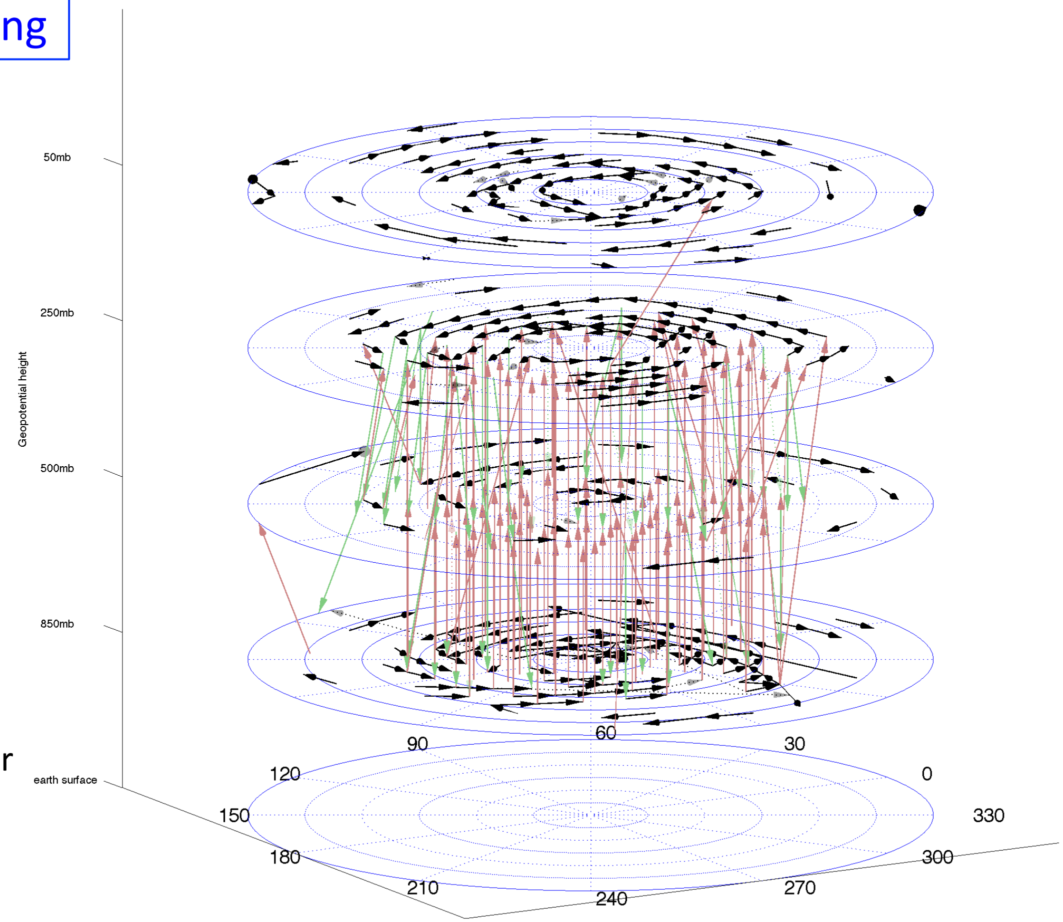
Joint work with Yi Deng

Input = observed daily geopotential height data.

Causal Discovery →
Track physical interactions around the globe to study specific effects (QBO, etc).

“Selective reverse engineering”

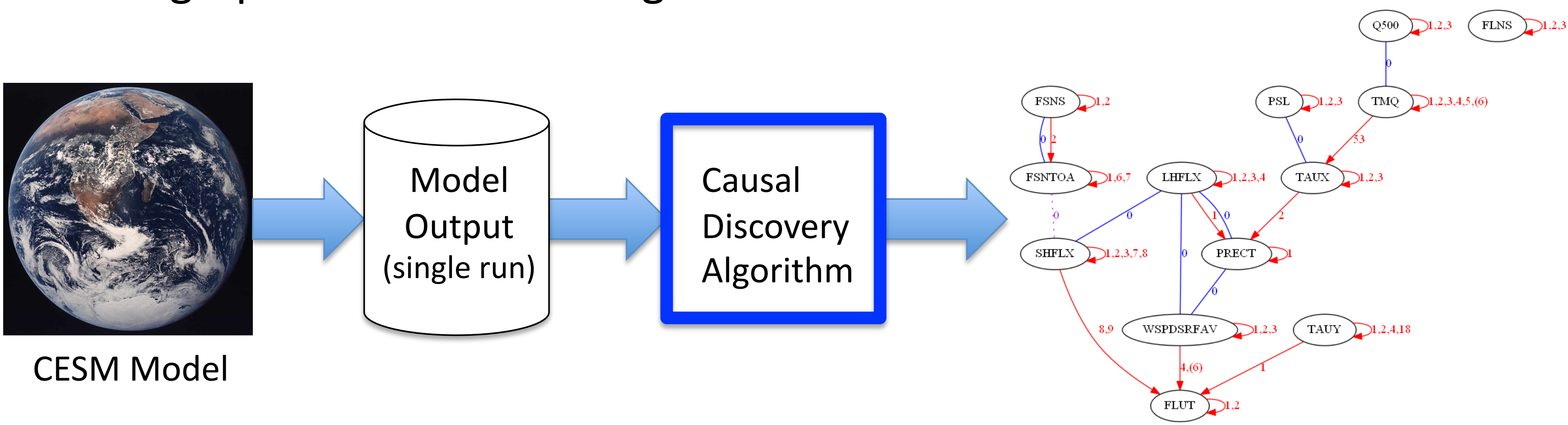
Data: NCEP/NCAR Reanalysis data, 1948-2011.
Daily geopotential height for 850, 500, 250, 50mb.
Data during QBO up transition.
Northern hemisphere, stereo-graphic projections for 850, 500, 250, 50mb.
400 point grid. Timescale: D=1 day.



Here: Observed data → Causal discovery → Interaction Maps

Application 2: Apply to Climate Model Runs

Idea by Dorit Hammerling: Use interaction maps as "dynamic fingerprints" or "causal signatures" of climate model runs.



CESM Model

- Calculate "causal signature" for individual model outputs (e.g. different initial conditions), then compare their "signature".
- First experiments: use only 15 variables, use global averages.

Here: Model data → Causal discovery → Interaction Maps

Sample Results: Effect of compression

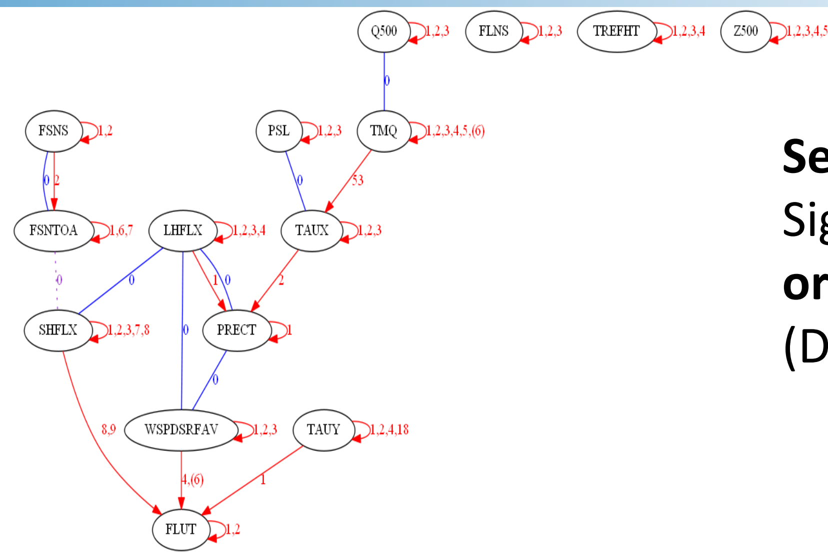
How to read the plots:

1) **Every connection is only a potential cause-effect relationship** (could be due to common cause).

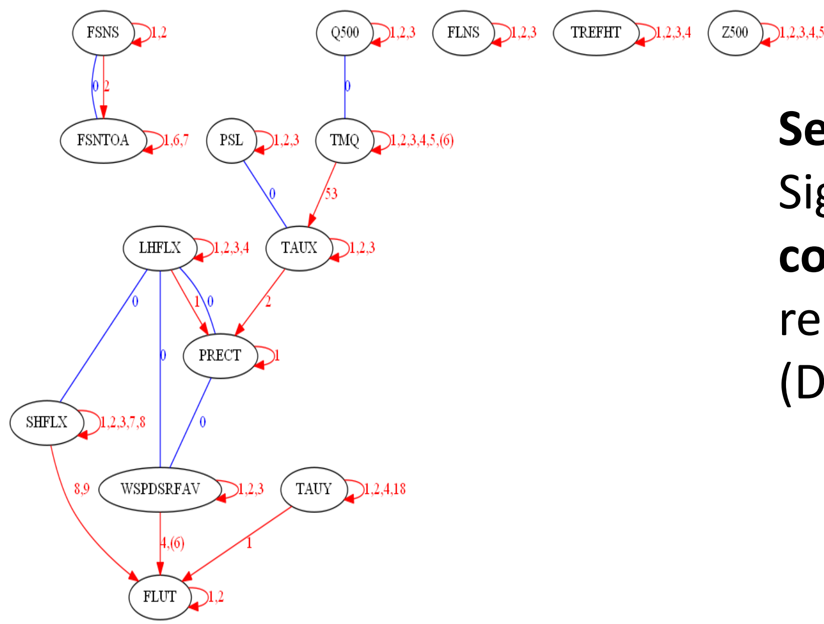
2) Connections can be directed or undirected.

3) Number(s) next to line = delay from potential cause to potential effect.
Here: daily time scale.

Observation: compression is causing only *tiny* differences.



Set 31:
Signature from **original** data
(D=1 day)



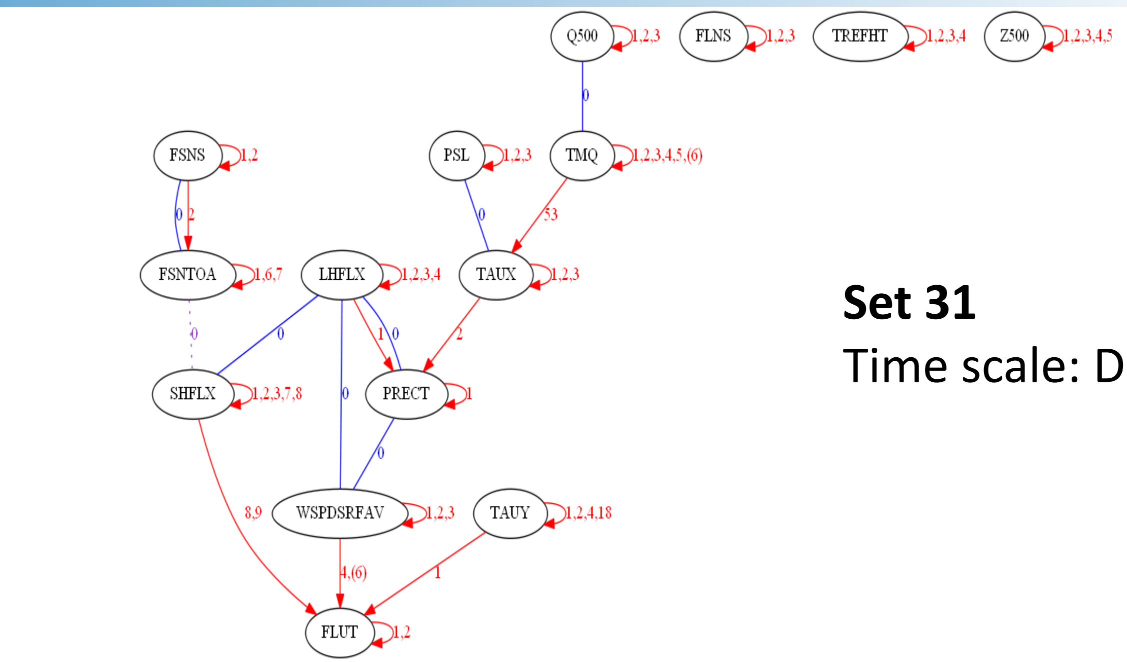
Set 31C:
Signature after **compression** and reconstruction
(D=1 day)

Sample Results: Effect of initial conditions

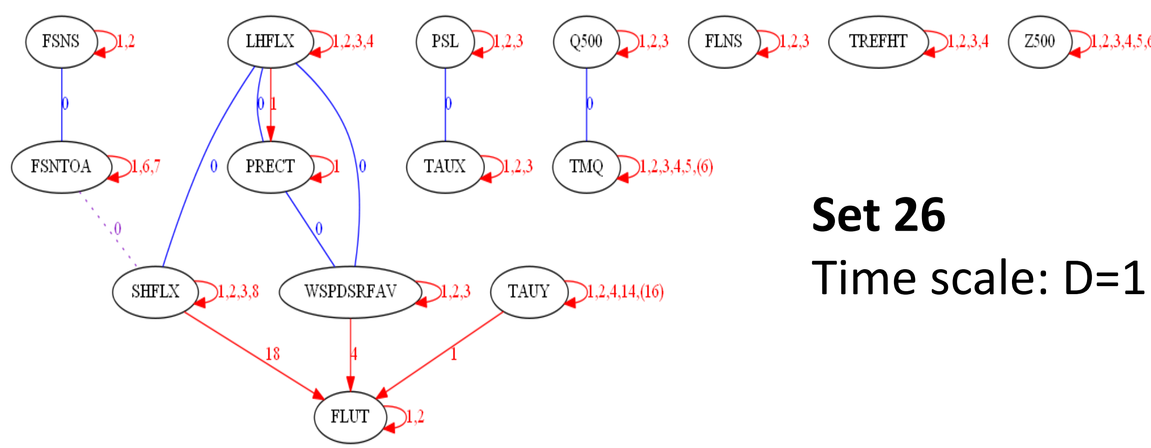
Shown here: Interactions on daily time scale.

Observation: Different initial conditions do yield some differences.

But there is always a "basic minimal pattern" that stays the same. Needs more study …



Set 31
Time scale: D=1 day



Set 26
Time scale: D=1 day

Opportunities of Causal Discovery in Geosciences

- Apply to observed data or model data for reverse engineering
  → extract big picture of interactions from observed/model data.
- **Interaction maps are intuitive** → great communication tool.
- Interaction maps are useful for **scientific discovery**:
  - Learn *details* about (physical) mechanisms that are not yet fully understood.
  - Details can be: Location / direction / magnitude of effect, causal pathway.
  - Study *trends* for different conditions.

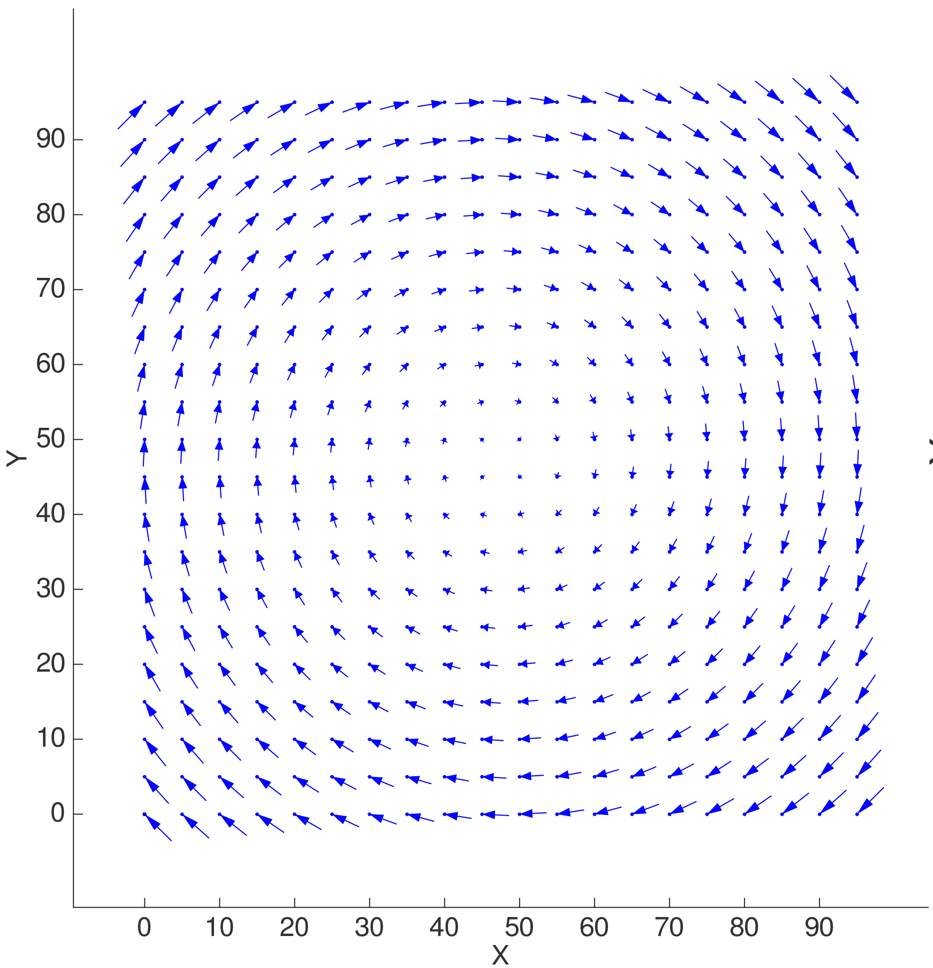
    Example: How do mechanisms change in a warming climate?

Limitations + Challenges of Causal Discovery

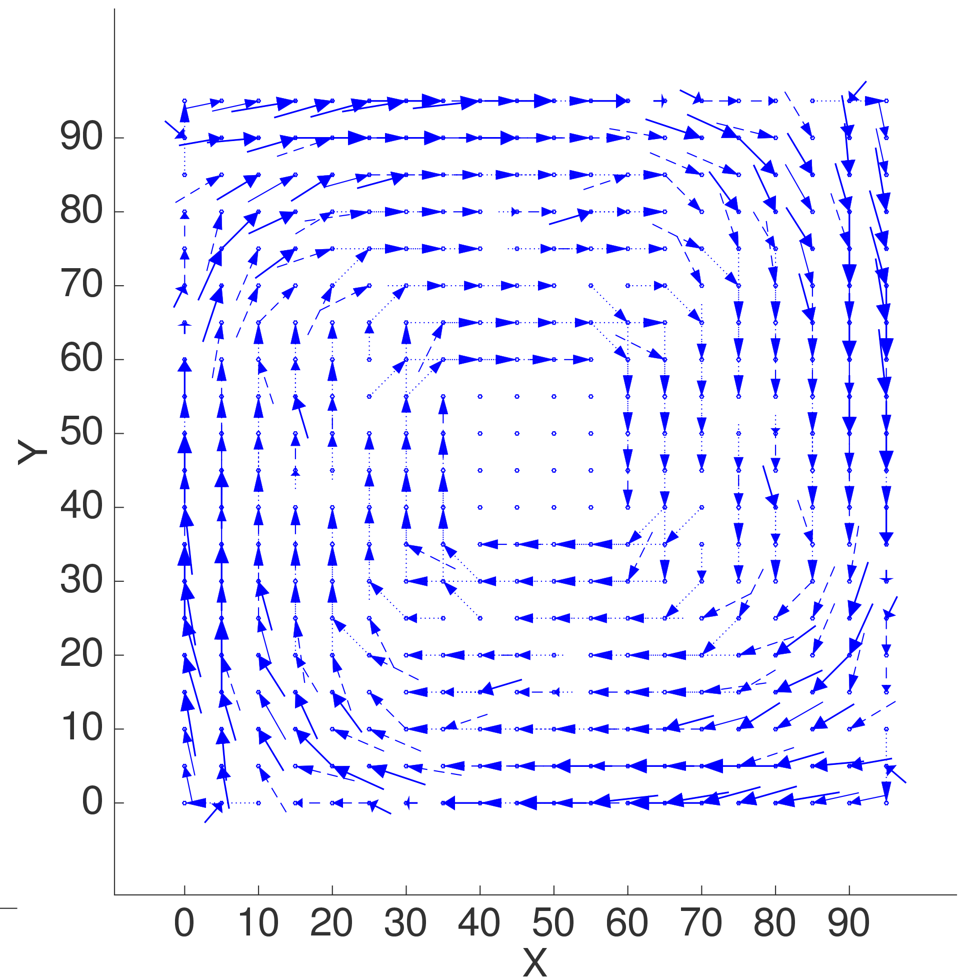
1) **Large sample size required** for statistical tests (robustness).
2) **Computational complexity** – can limit spatial resolution.
3) **Grid bias** → signals along grid symmetry are picked up best.
4) **Signal speed bias: signals with speeds** around (Δx/Δt) get picked up best.
5) **Ground truth rarely available to test and calibrate** methods → need to generate and test on synthetic data.
6) In practice, method catches **only the strongest interactions** for any variable/location. (If there are strong + weak interactions at one location, do *not* expect to pick up the weak one.)

Experiments with synthetic data: advection + diffusion

Original advection velocity field (input)

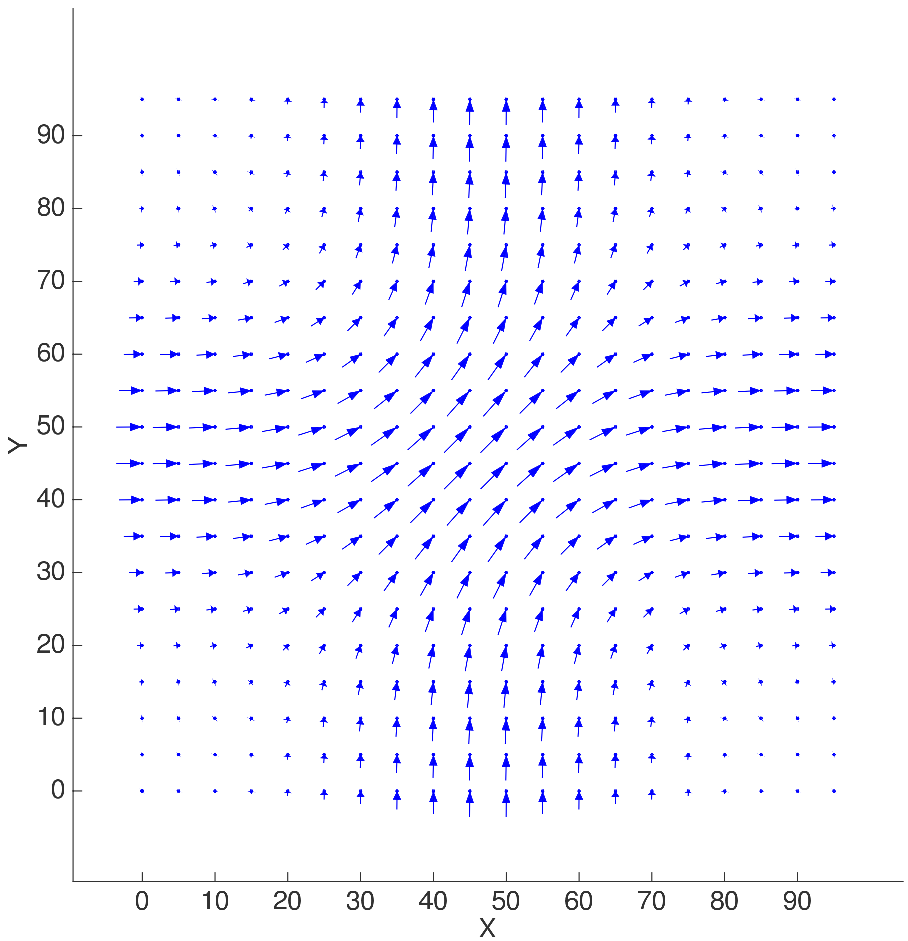


Result: Estimated velocity field

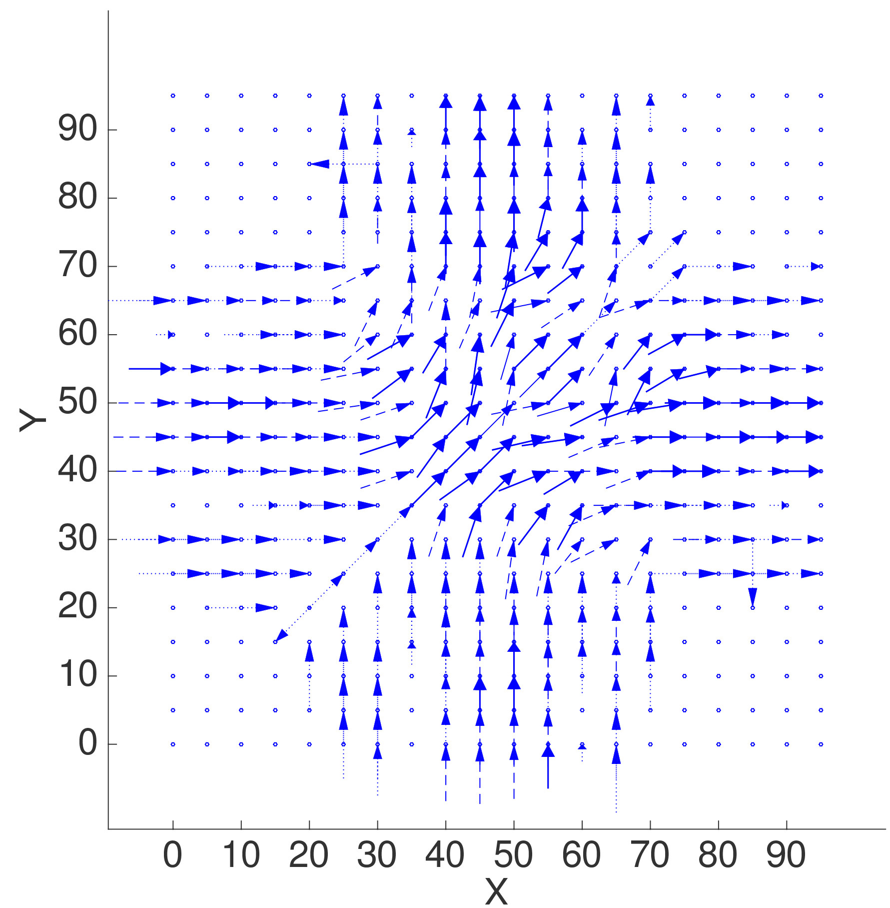


Experiments with synthetic data

Original advection velocity field (input)



Result: Estimated velocity field



Interpretation of interaction maps is <u>hard work</u> !

1. **Identify physical mechanism for each interaction found**:
    - Many different mechanisms can be at work simultaneously.
    - Only domain scientist can determine what each connection represents.
    - Some may be due to hidden common causes.
2. **Determine effect of grid bias, signal speed bias** ($\Delta x / \Delta t$), etc..

   Ex: use several different grids/resolutions and compare results.
3. Conduct experiments with **synthetic data to learn typical causal signature**s of different physical mechanisms.

Conclusions

- **Causal discovery is emerging** in many new applications.
- **Causal interpretation requires caution**: we can only identify *potential* cause-effect relationships.
- **Knowledge discovery – of *any* kind – has much to contribute to geosciences and similar disciplines.**
- There are still *so many processes* of this earth that are ***not yet fully understood***. → Lots of potential.

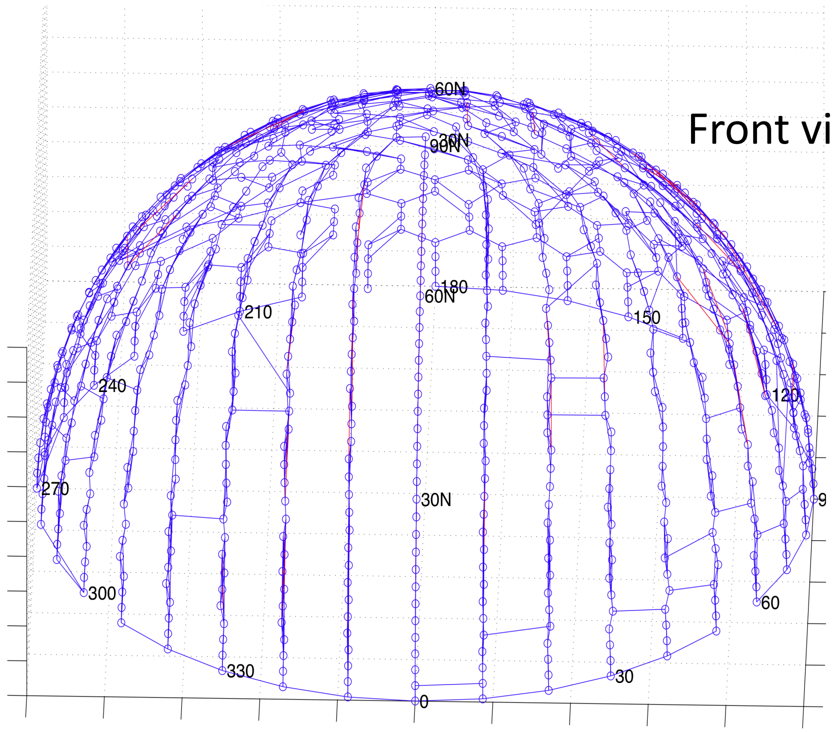
The End.

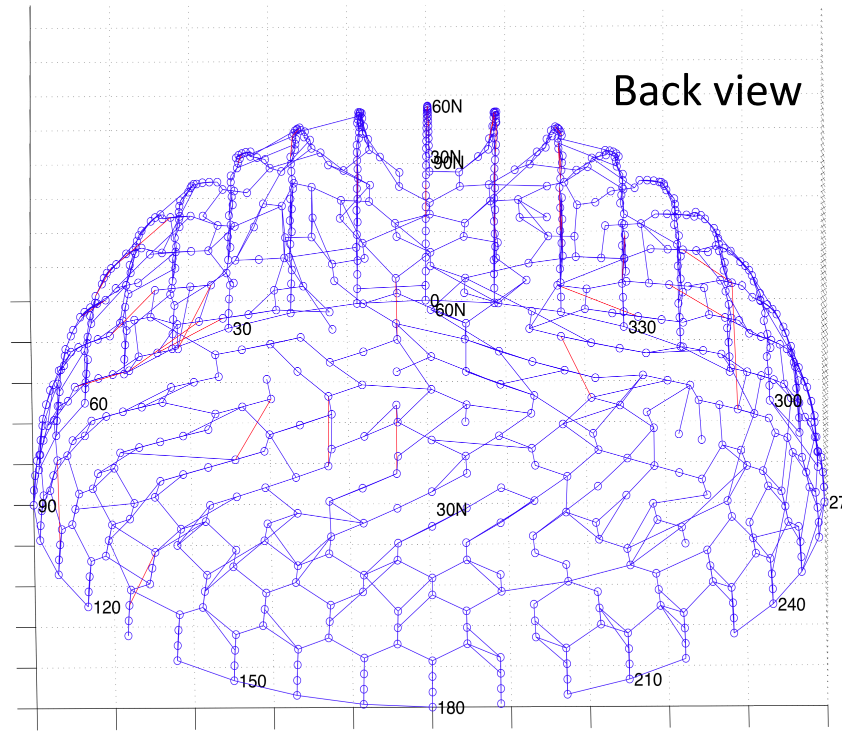
Questions or Suggestions?

Motto: *To boldly go, where no causal discovery algorithm has gone before*.

One of our first experiments – what's going on?

Static model, simple equal-area grid, Northern hemisphere shown

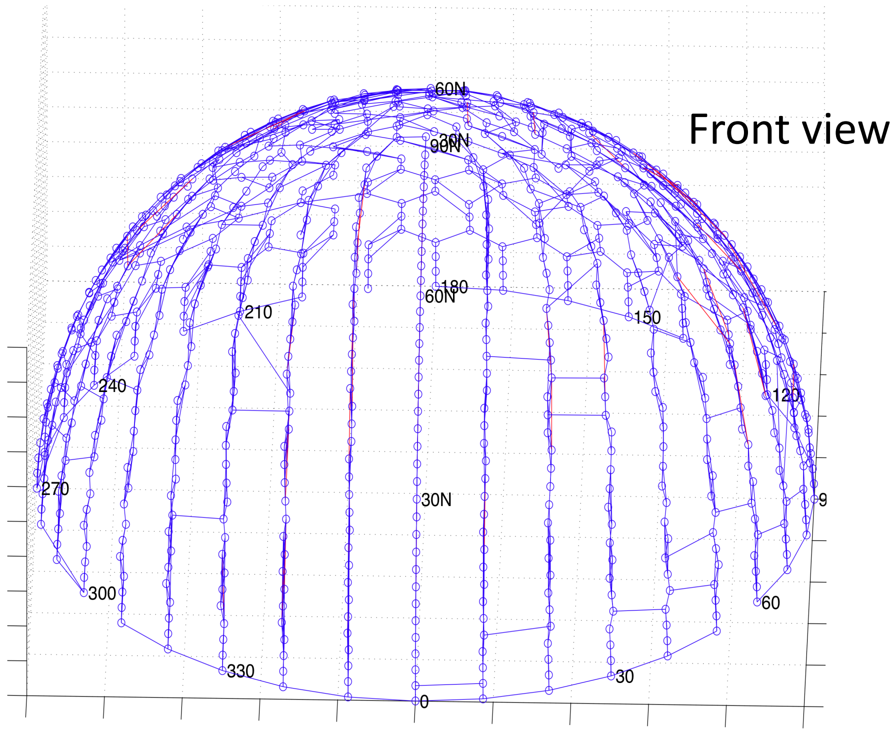


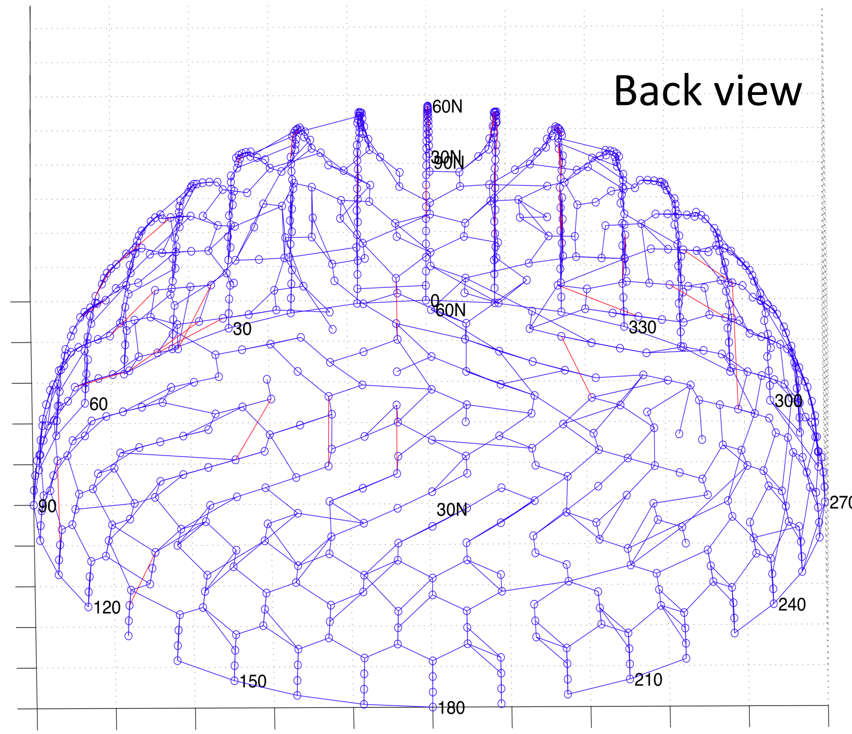


- Straight connections in Africa & hexagons in Pacific ?!?
- Does this make any sense ?!?
- What do you think happened here?

One of our first experiments: showing grid bias

Static model, simple equal-area grid, Northern hemisphere shown

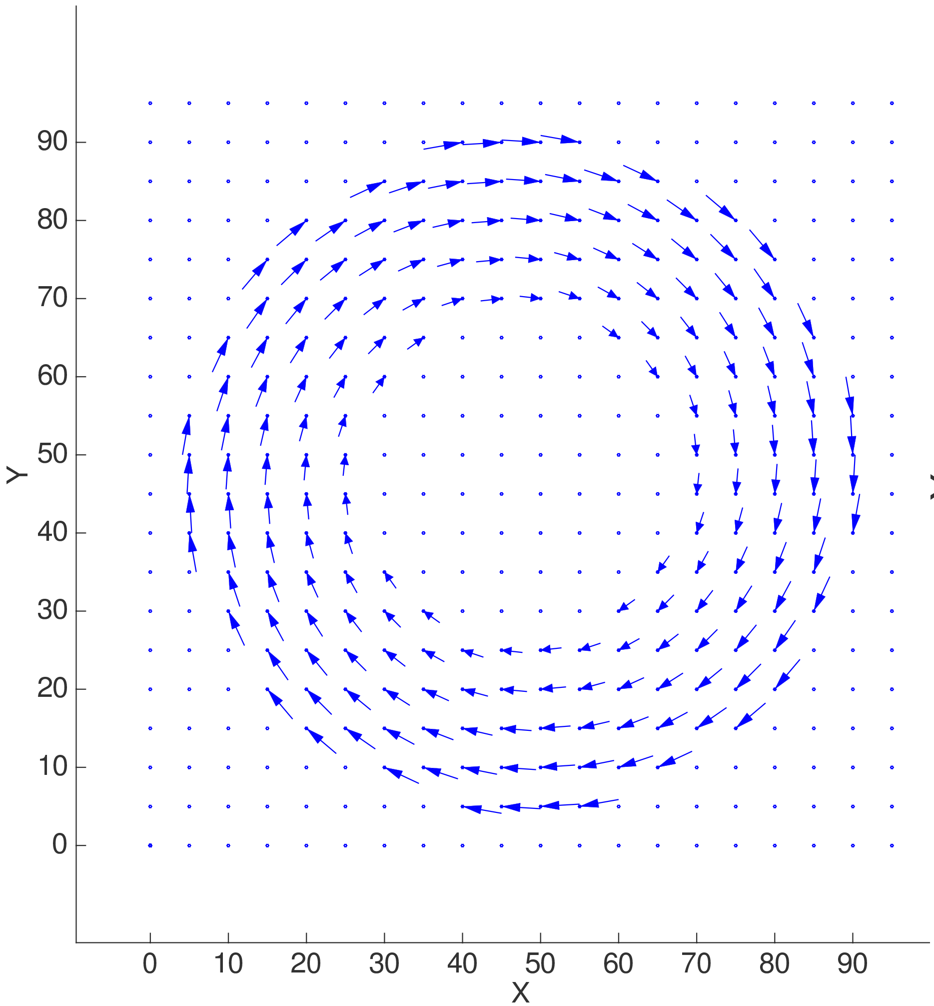




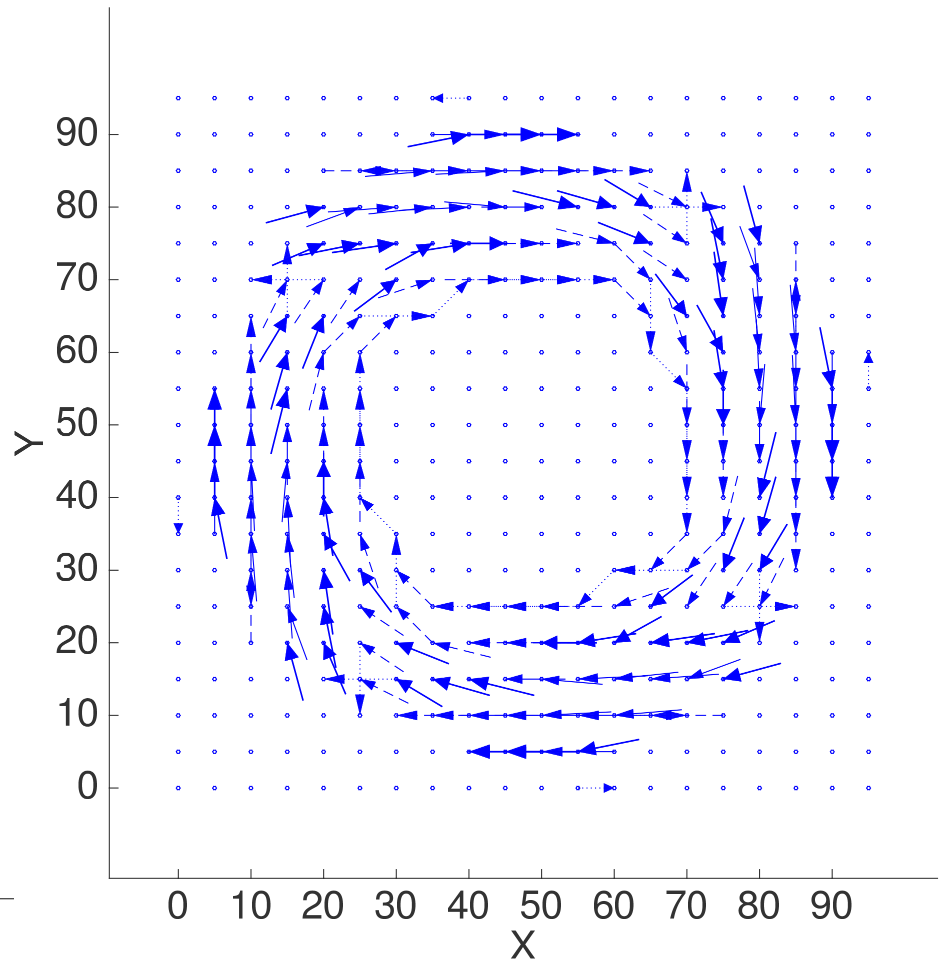
- Straight connections in Africa & hexagons in Pacific ?!?
- **Unequal proximity in grid is stronger signal than "causal" signal.**
- **Direction bias because of uneven proximity of some neighbors.**
- → Any two points close to each other are connected! Not what we intended!
- **Solution: Isotropic grid (Fekete grid) → reduces bias for direction.**

Experiments with synthetic data

Original advection velocity field (input)



Result: Estimated velocity field



Application to Climate Models

Goals:

1. **Study effect of lossy compression** in output data – *Does fingerprint look very different after compression and reconstruction?*
2. **Detect errors in individual runs** (e.g. maybe one software component not linked in properly). *Do we pick up such errors in the fingerprint?*
3. Can we **classify ensemble members** based on their causal signatures?

First experiments: Focus on only 15 variables of climate model output, use global averages.