INFERRING CAUSATION FROM TIME SERIES IN EARTH SYSTEM SCIENCES

Causal Network exam 2021-2022

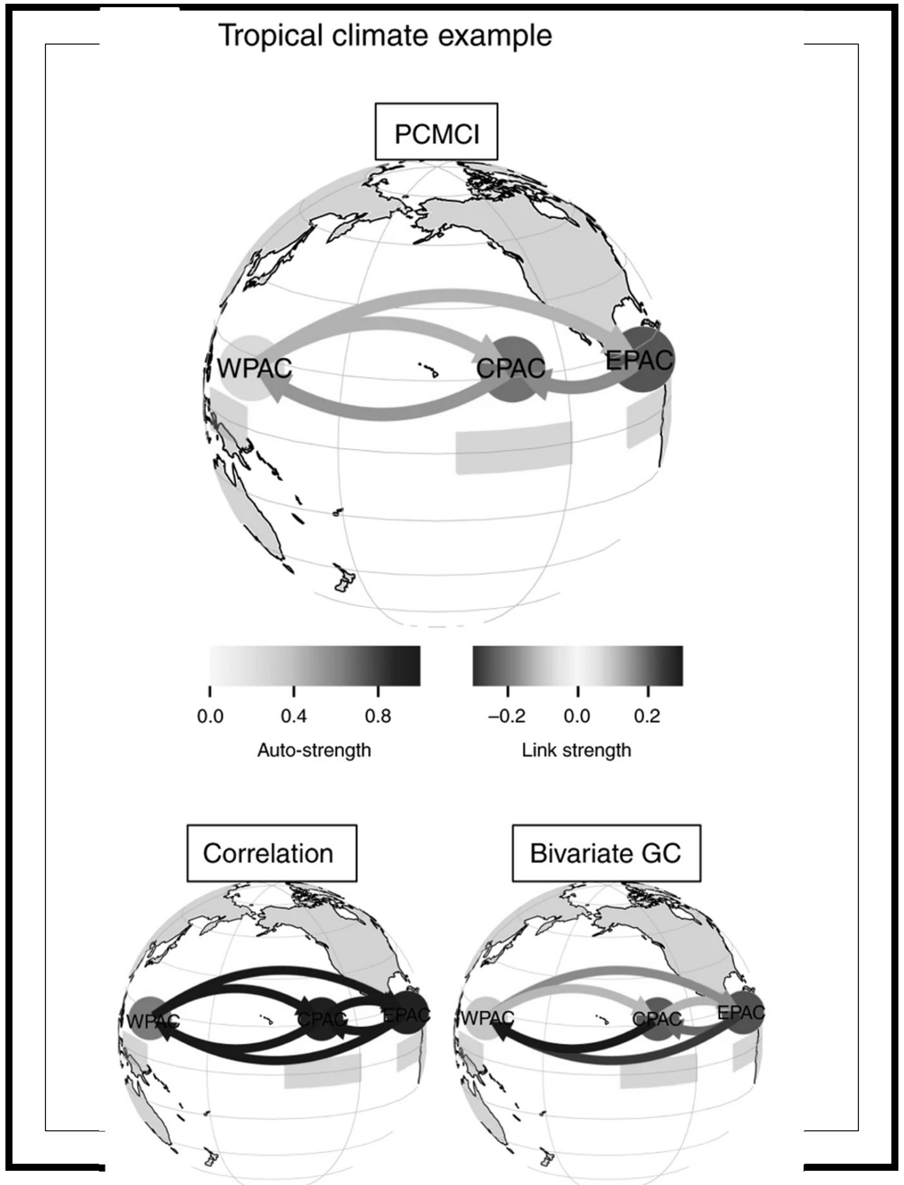
Ilaria Erba 2° year PhD

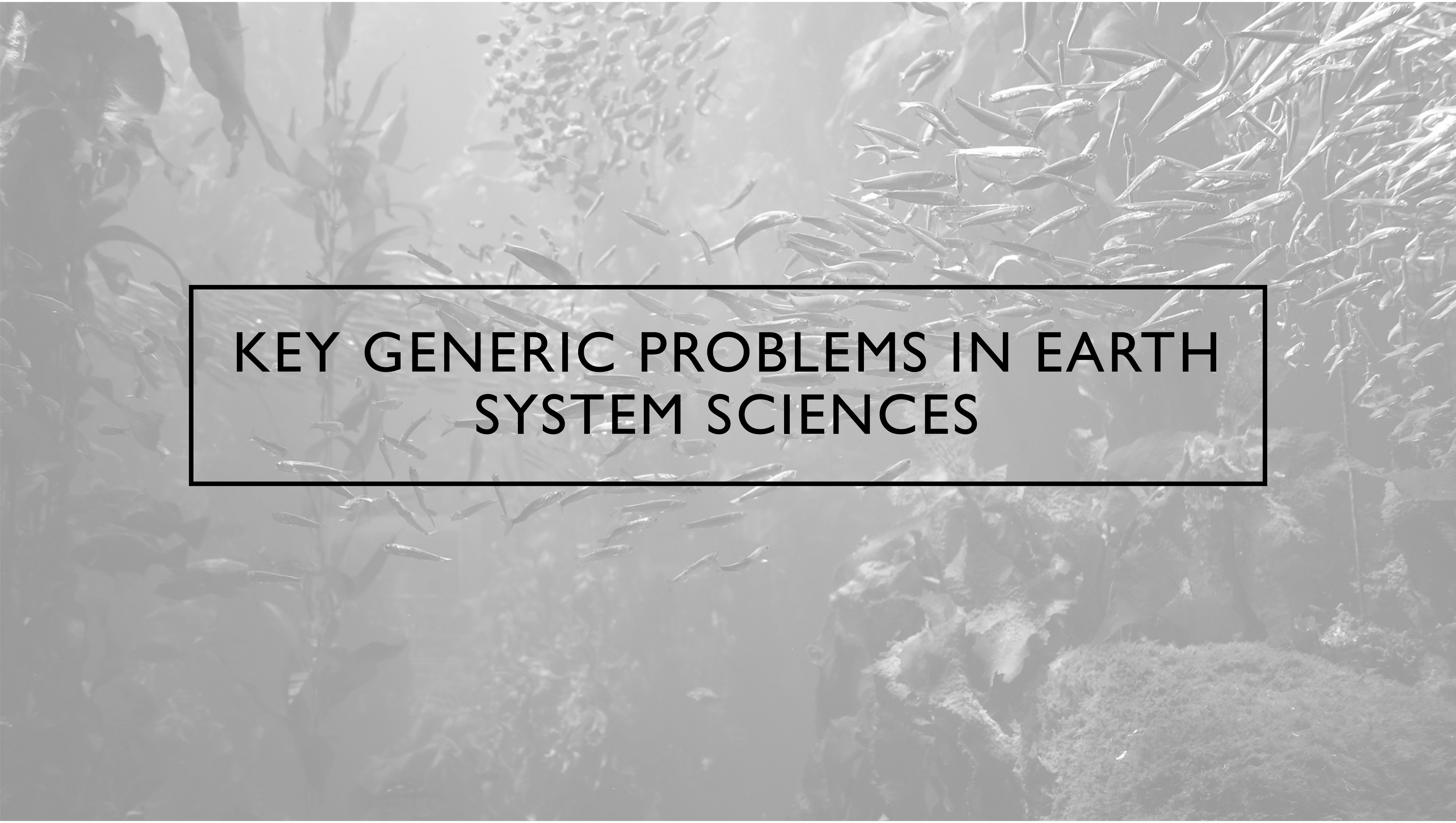
Mirko Paolo Barbato 2° year PhD

INTRODUCTION

- Earth system is a large-scale complex dynamical system
  - interventional experiments are either <u>infeasible</u> or <u>ethically problematic</u>
- Commonly used tools:
  - Simulations
  - Correlations
  - Regression
- Correlation does not imply causation
- Reichenbach's cause principle
- Data-driven machine learning methods contrast
- Causal inference methods have the potential to advance the

state-of-the-art

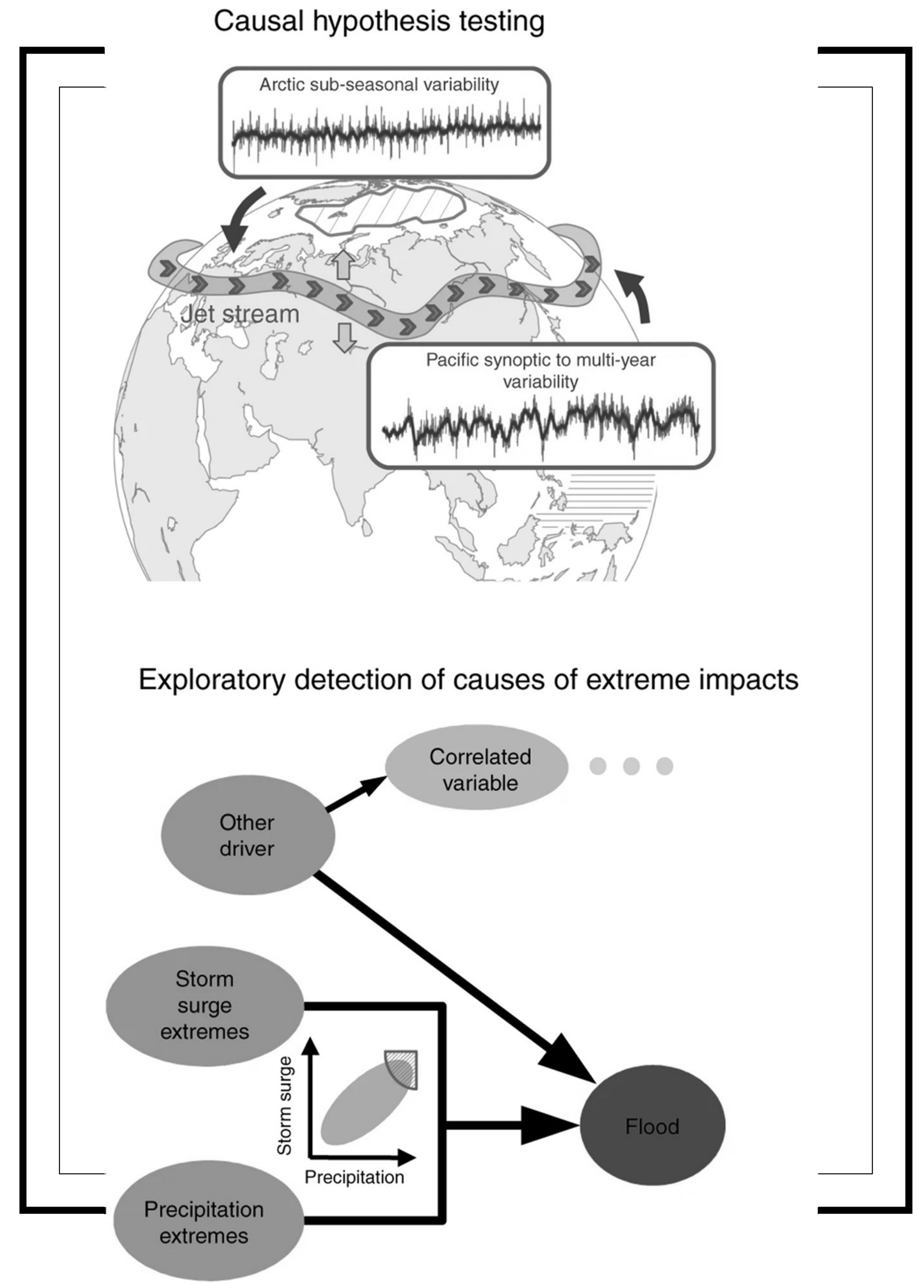




KEY GENERIC PROBLEMS IN EARTH SYSTEM SCIENCES

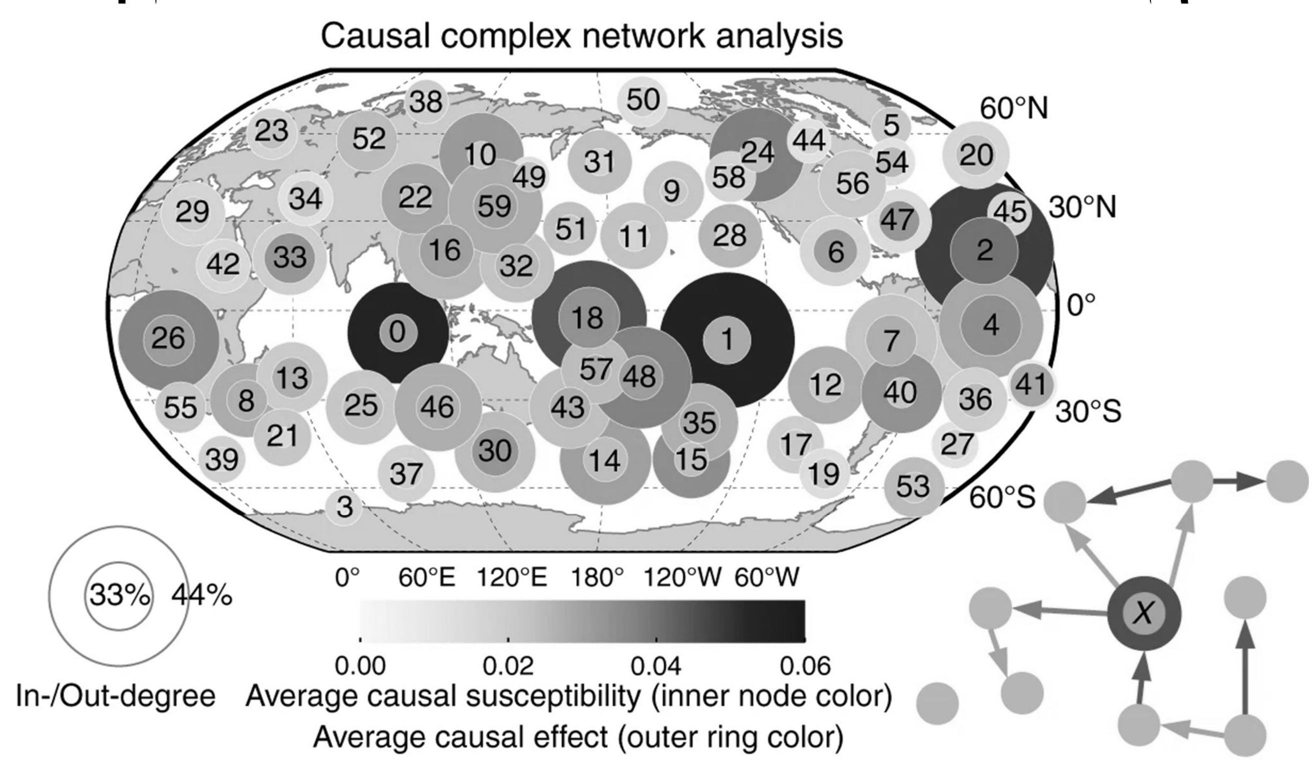
IDENTIFYING CAUSAL RELATIONSHIP

- From observational data we want to understand the cause and effect
- Extraction and definition of data → usually extracted from gridded spatiotemporal datasets
- Reconstructing the causal relations between these extracted variables
    - Different time scales between processes
    - Distributions of climate variables, for example precipitation, are often non-Gaussian
- Detection of causes of extreme impacts
    - Small sample size of observed impacts
    - Synergistic effects



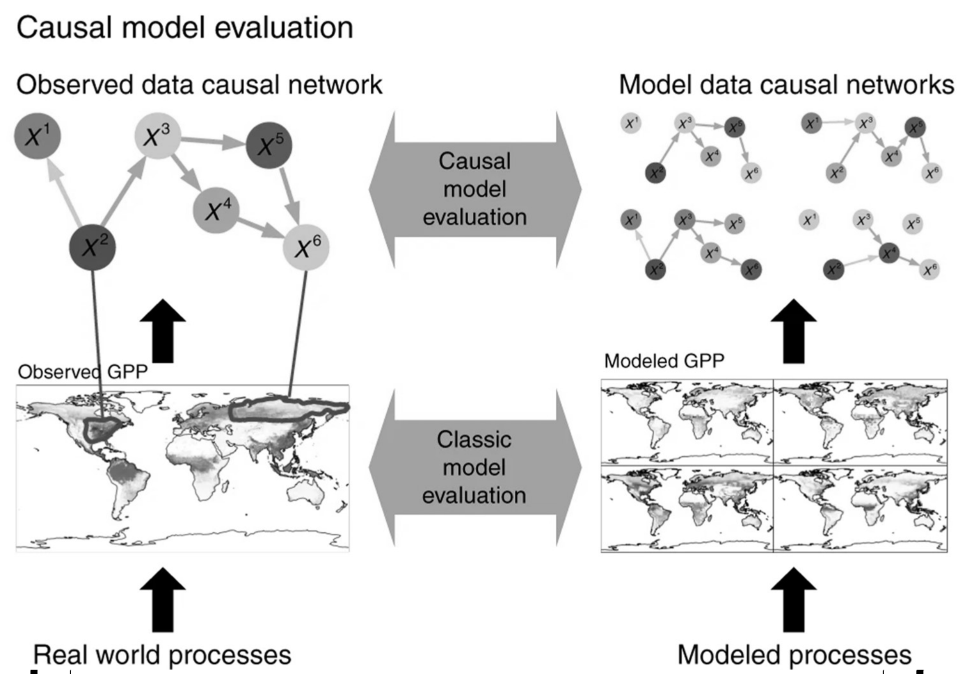
NETWORK ANALYSIS

- Causal complex network analysis
  - Nodes are defined as time series at different grid locations
  - Links are based on correlations between the grid point time series
  - Node degree quantifies the number of processes linked to a node → do not allow for a causal interpretation
  - Proposal: network measure that takes causality into account



EVALUATION OF PHYSICAL MODELS

- Causal evaluation of physical model
  - Models are partly based on known process and partly on approximating processes
  - Small differences in parametrization can lead to different models
  - Underdetermination of equifinality
  - Proposal: comparison of reconstructed causal dependencies
    - causal dependencies are more directly linked to the physical processes
    - more robust against overfitting than simple statistics
    - yield more reliable future projections

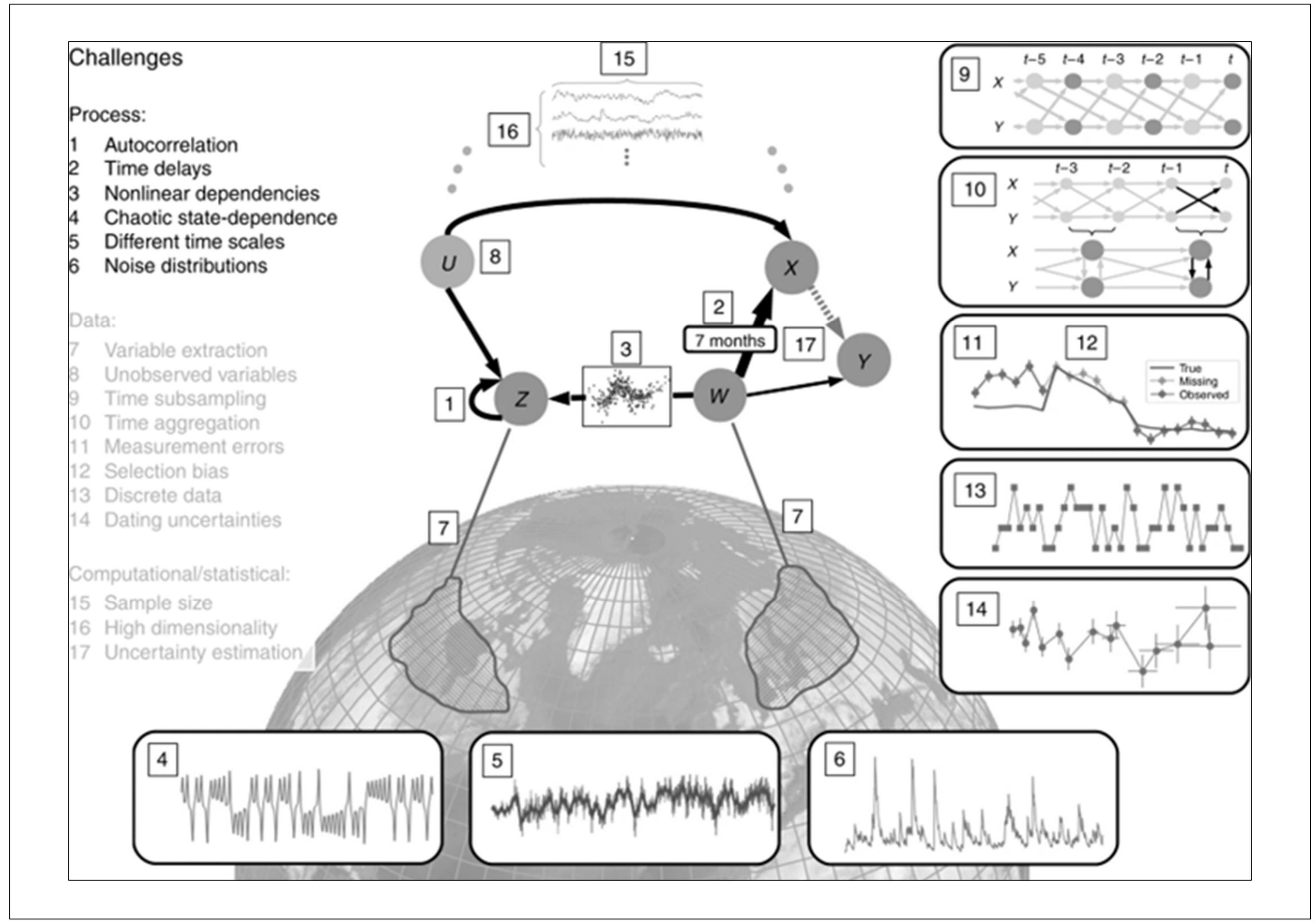


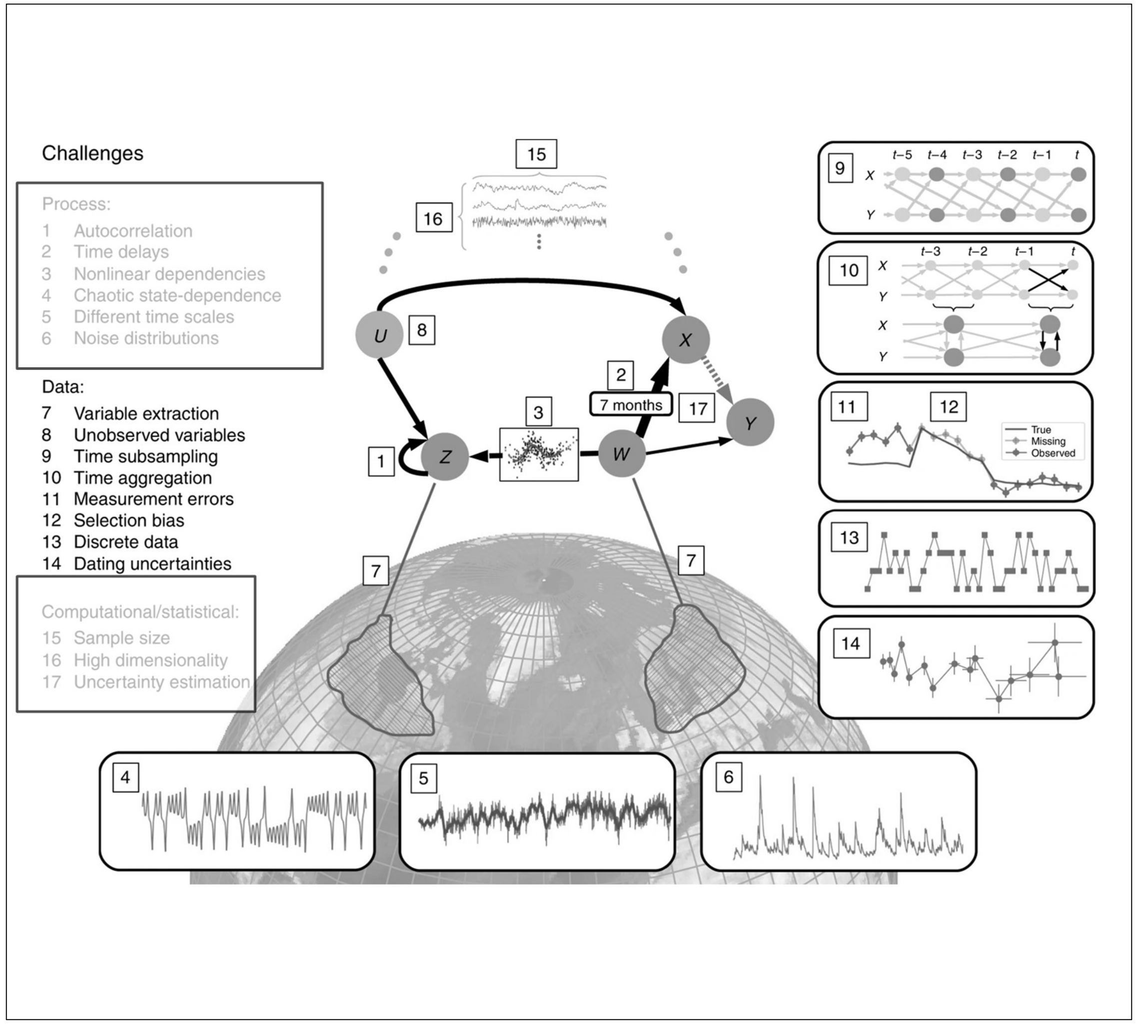
CHALLENGES FROM METHODOLOGICAL PERSPECTIVE

PROCESS CHALLENGES

Methodological challenges for causal discovery in complex spatio-temporal systems such as the Earth system. At the **process level**:

- autocorrelation
- time delays
- nonlinearity
- synergistic behavior
- noise distribution

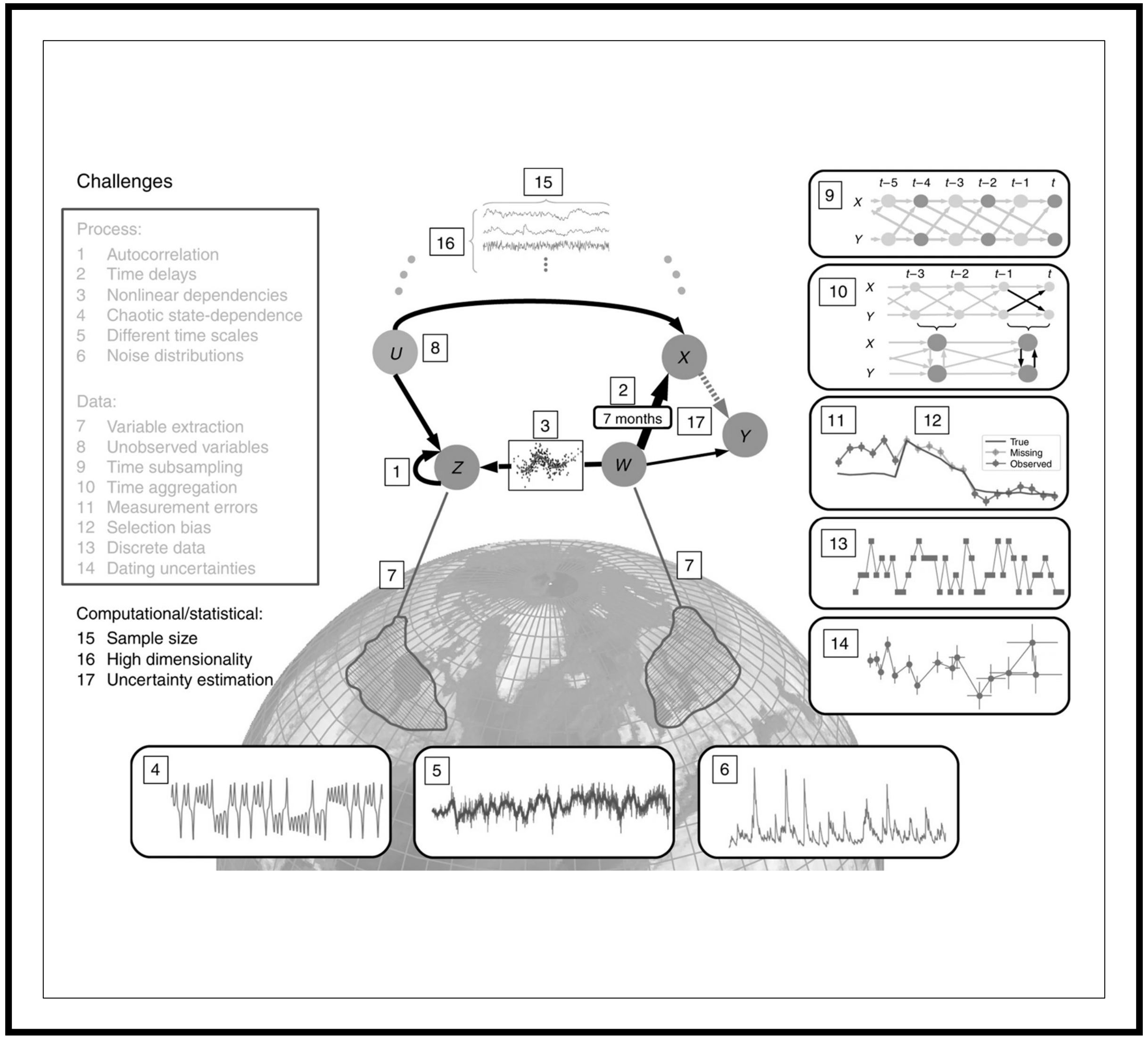




DATA CHALLENGES

Data has to be rapresentative and interpretative of the sub-processes of the system

- Unobserved variables → No causal sufficiency
- Time sub-sampling → causal dependencies appear contemporaneous or cyclic
- Data quality
  - Measurement errors
    - Observational noise
    - Systematic biases
    - Missing values
- Data type and class imbalance



COMPUTATIONAL AND STATISTICAL CHALLENGES

- Scalability
  - Sample size
  - High dimensionality
- Interpretation of the causal conclusions are based on the assumptions underlying the different methods which may alter conclusions for a particular application
- Most of the challenges discussed in this section are the same for correlation or regression methods which are, in addition, ambiguous to interpret and often lead to incorrect conclusions

OVERVIEW OF CAUSAL INFERENCE METHODS

PRELIMINARY STATE OF ART

General assumptions of many causal inference methods for **time series**:

Time-order: causes precede effects

Causal Sufficiency: direct common drivers are observed

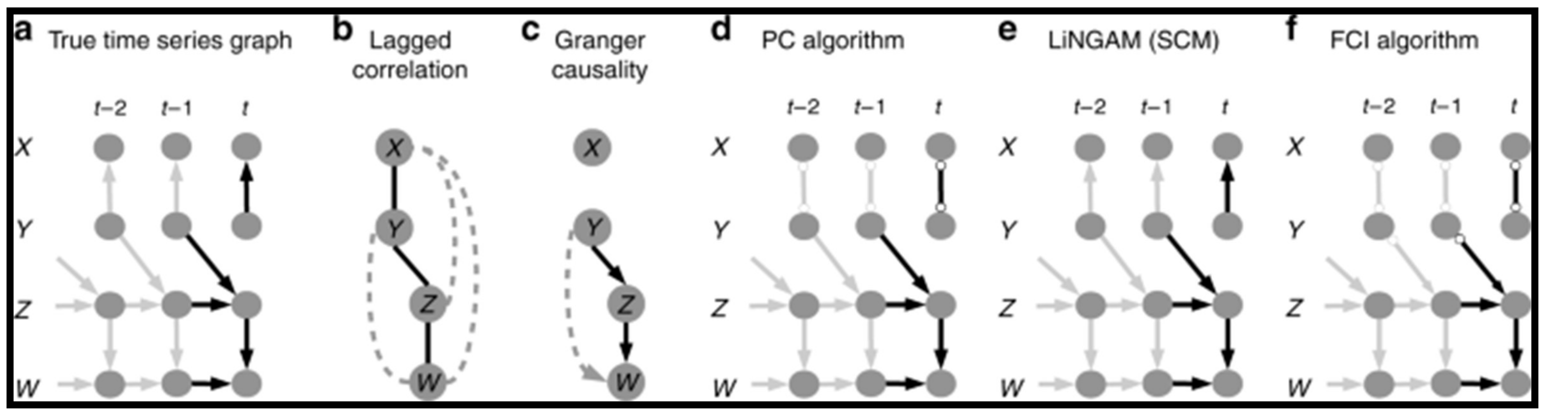
Causal Markov Condition stating that in a graphical model a variable Y is independent of every other variable (that is not affected by Y) conditional on Y's direct cause

Recent work shows that **some of these assumptions can be relaxed**.

GRANGER CAUSALITY (GC)

Test if omitting the past of a time series X in a time series model including Y's own and other covariates' past increases the prediction error of the next time step of Y

- Different kind of time series models:
  - The Granger causality test is based on **linear** autoregressive modeling
  - Nonlinear dependencies can be modeled with more complex time series models
    - Multivariate extensions of GC fail if too many variables are considered, or dependencies are contemporaneous due to time-sampling
- Limitations:
  - To lagged causal dependencies
  - Deficiencies in the presence of subsampled time series

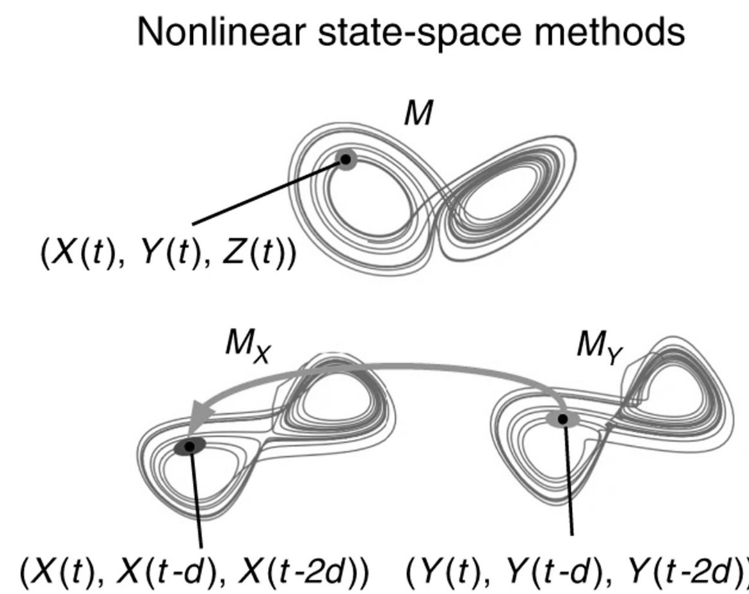


NONLINEAR STATE-SPACE METHODS (CCM)

While GC view systems as having interactions that arise from an underlying stochastic process, convergent cross-mapping take a different dynamical systems perspective

Interactions occur in an **underlying dynamical system** and attempt to uncover causal relationships based on Takens' theorem and nonlinear state-space reconstruction

- A causal relationship between two dynamical variables X and Y can be established if the variable X can be predicted using the reconstructed system based on the time-delay embedding of variable Y, then we know that X had a causal effect on Y

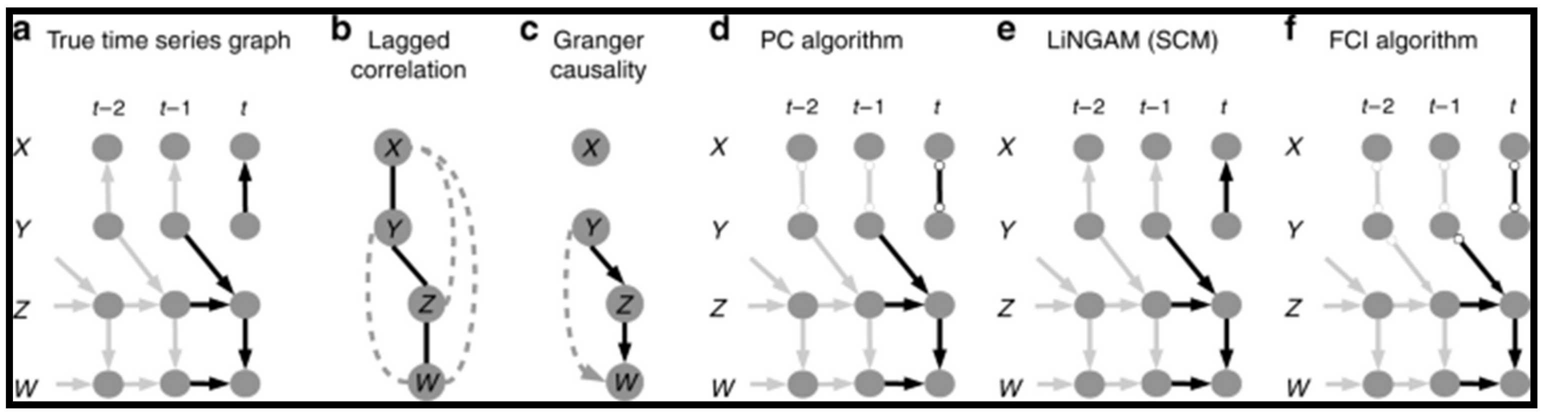


CAUSAL NETWORK LEARNING ALGORITHMS

CCM is less well suited for time series that are of a stochastic nature

Multivariate extensions of GC fail if too many variables are considered, or dependencies are contemporaneous due to time-sampling

- The common assumptions for the causal network learning algorithms are **Markov condition** and **Faithfulness**
- Search architecture classification
  - Empty or fully connected graph
  - Statistical criterion for removing or adding an edge
- Search architecture examples
  - Greedy equivalence search starts with an empty graph and use Conditional independencies tests or Score function that quantify the likelihood of a graph structure



STRUCTURAL CAUSAL MODEL FRAMEWORK (SCMS)

GC requires a time delay between cause and effect to identify causal directionality

Many causal network learning algorithms account for contemporaneous dependencies, but they can only identify causal graphs up to a Markov-equivalence class

- It gives origin to ambiguity. E.g.: measuring that X is conditionally independent of Y given Z, while all other (conditional) relationships are dependent results in X←Z→Y X→Z→Y X←Z←Y
- Structural causal models (SCMs) can identify causal directions in such cases because they permit assumptions about the functional class of models
- SCMs have not yet been applied in Earth system sciences except for one work in remote sensing

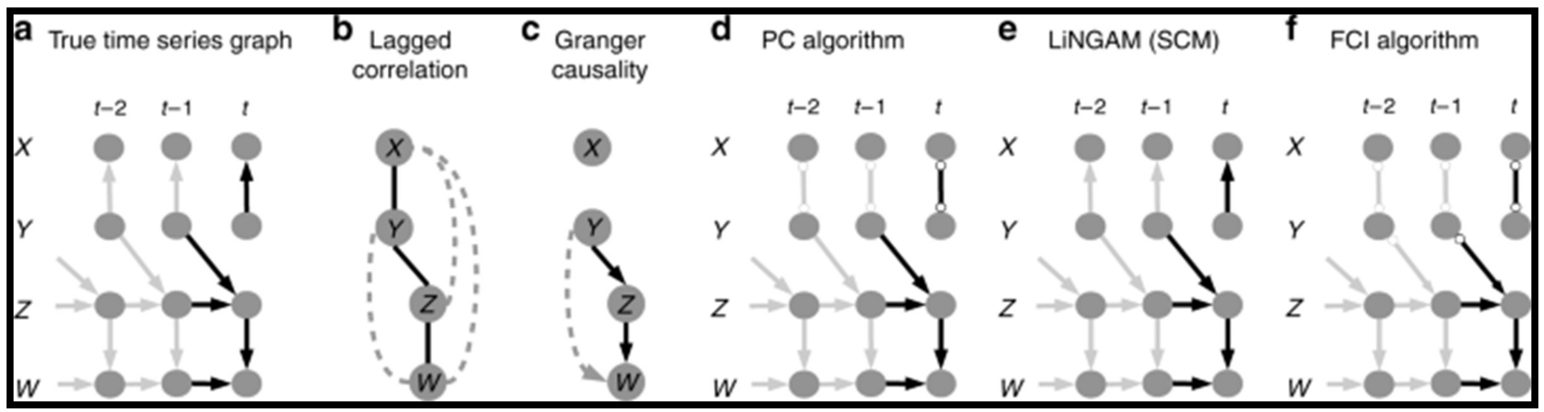
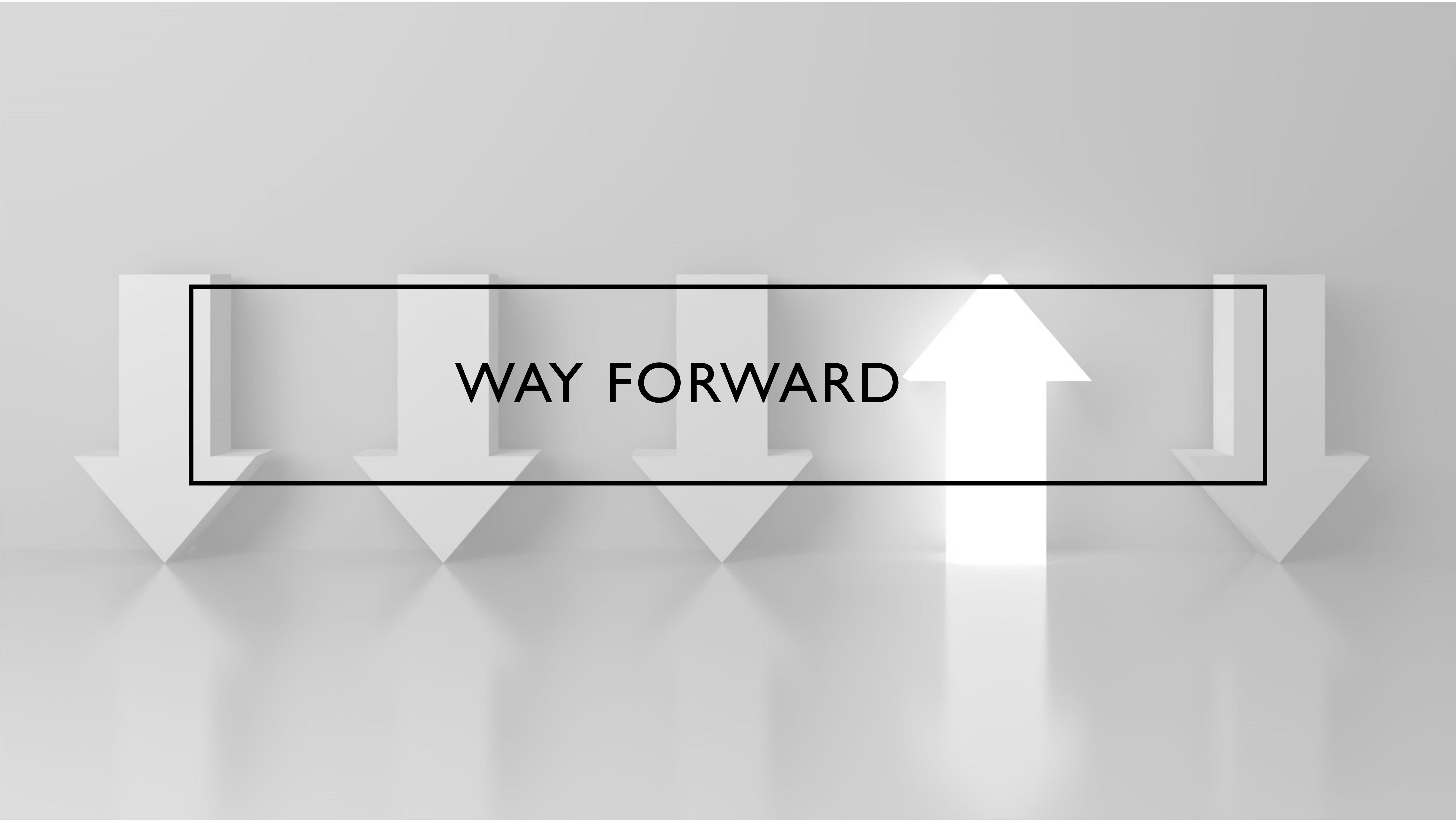


Table 1 List of methods, key strengths, and further research directions addressing current limitations

Method	Key strengths	Further research directions
Granger causality and nonparametric extensions[9,37,99]	Significance assessment; nonparametric versions	Dealing with contemporaneous effects and feedback cycles; high-dimensionality; deterministic dependencies; synergistic effects; time scales; unobserved variables
Nonlinear state-space methods[10,11]	State-dependent nonlinear systems; contemporaneous effects	Significance assessment; high-dimensionality; highly synchronous dynamics; high stochasticity; time scales; unobserved variables
Conditional independence-based algorithms[12]	High-dimensionality; unobserved variables; nonparametric tests	Significance assessment; deterministic effects; synergistic effects; time scales; contemporaneous feedback cycles
PCMCI[23,24]	High-dimensionality; time delays; strong autocorrelation; nonparametric tests	Unobserved variables; deterministic effects; synergistic effects; time scales; contemporaneous feedback cycles
Information-theoretic algorithms[23,24,51]	High-dimensionality; nonparametric; time delays; information-theoretic interpretation	Significance assessment; unobserved variables; deterministic effects; synergistic effects; time scales; contemporaneous feedback cycles; efficient entropy estimation
Structural causal models[13,38]	Contemporaneous effects; nonparametric versions	High-dimensionality; synergistic effects; time scales; unobserved variables; time delays
Invariance-based methods[4,13,57,58,60,61]	Utilizes heterogeneity in space and time	Causality in stationary regimes; same as for SCMs
Bayesian score-based approaches[48]	Bayesian uncertainty assessment; inclusion of expert knowledge	High-dimensionality; nonlinearity; deterministic effects; synergistic effects; time scales; contemporaneous feedback cycles; unobserved variables; combine with cond. independence-based methods[100]

This table is intended to be a rough method guide. A detailed overview is beyond the scope of this Perspective and hardly possible because comparison studies are currently largely lacking. Spurring research to overcome this lack is a goal of this Perspective and the accompanying platform causeme.net. The terms used in this table are explained in the challenges section and illustrated in Fig. 4

METHODS COMPARISON


WAY FORWARD

AVENUES OF FURTHER METHODOLOGICAL RESEARCH

In the short-term existing methods already address some of the mentioned challenges

- Combining different conceptual approaches
- Filtering methods as pre-processing steps

In the mid-term it is worth exploring methods that have not been applied to Earth system data

- Method development and comparison require benchmark datasets with known causal ground truth for validation
- The lack for datasets is compensated by **physical simulation models** and generation of **synthetic data**
- causality benchmark platform causeme.net with synthetic models mimicking real data challenges
- Combining observational causal inference and physical modelling


THANK FOR THE ATTENTION!