Beyond correlation: Cause-effect relations in hydrometeorological systems

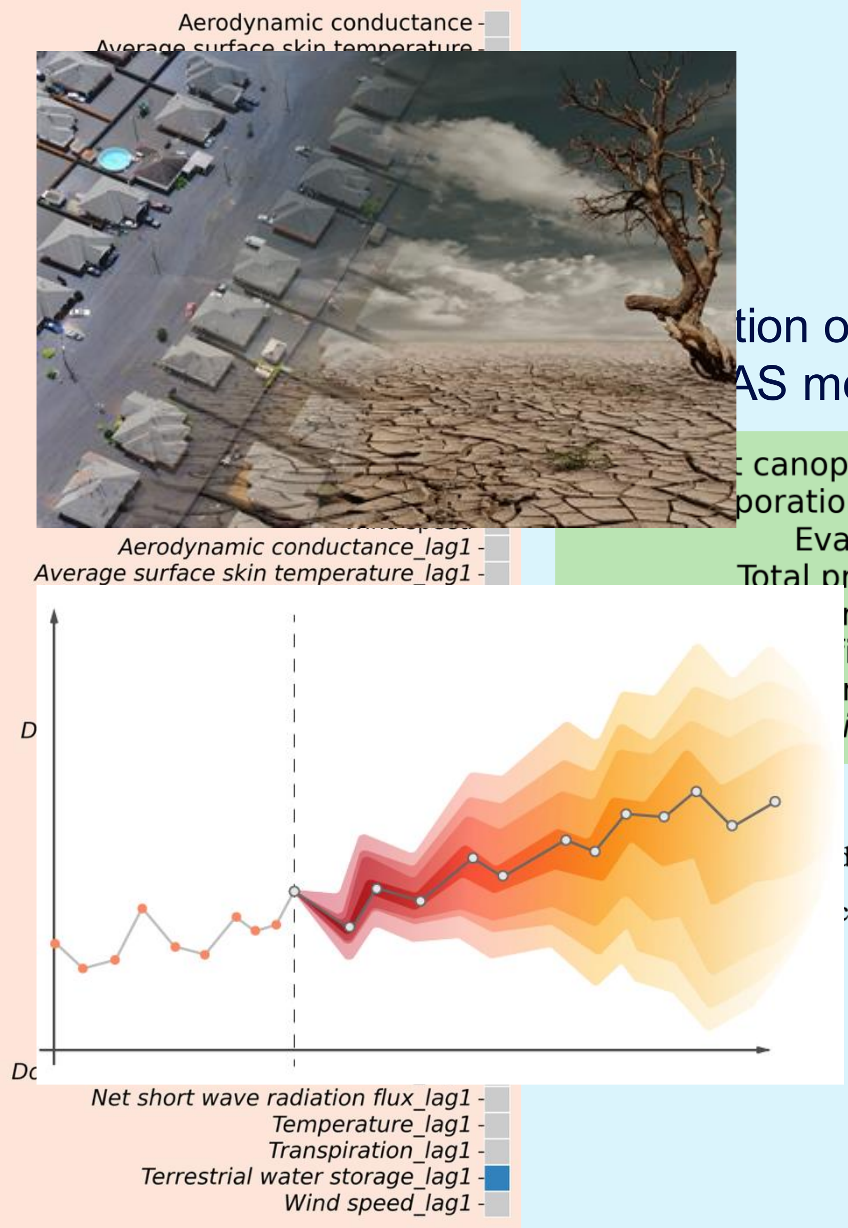
Vivek Kumar Yadav[1,2], Murray Peel[2],
Keirnan Fowler[2], Dongryeol Ryu[2], Bramha Dutt Vishwakarma[1]

K2: Statistical, post-processing and verification methods for environmental prediction, MODSIM 2025, 04 Dec 2025

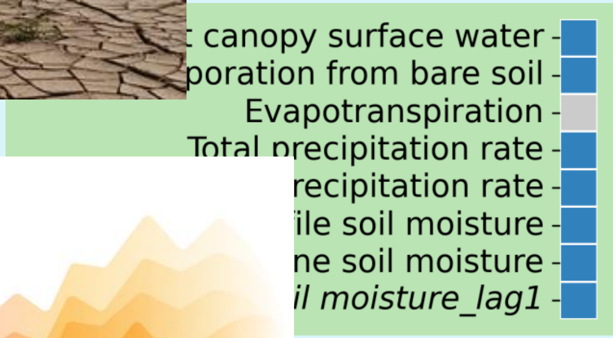
[1]Interdisciplinary Centre for Water Research, Indian Institute of Science Bengaluru
[2]Faculty of Engineering and Information Technology, The University of Melbourne

Drivers of surface soil moisture

- Hydrometeorological systems are highly correlated:
  - multiple cross and lagged correlations, typical to hydrology.
- Correlation does not imply causation.
- *getting right answers for wrong reasons.*



tion of surface soil moisture
AS model environment



Need to identify *robust* and *direct* drivers of variables

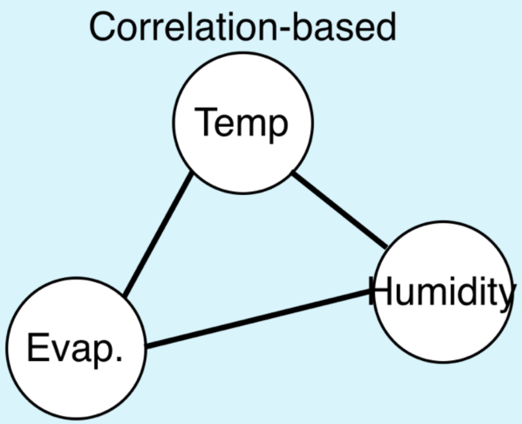
Causality: Science of cause-effect

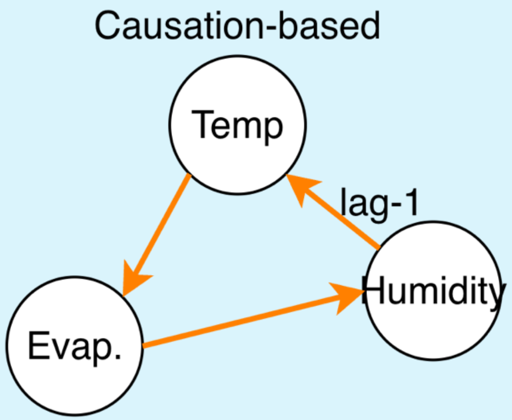
Causal Discovery (CD)

Discovering a direct cause-effect relation between entities (variables or events).

Causal Inference

Quantifying the influence of an entity (variable or event) onto another.





- Structure
- Directionality

Objective: Evaluate CD methods in a simulated hydrometeorological system to recover drivers of the system

Methodology

GLDAS Land surface model

Literature review of generating equations

8 forcing variables
19 simulated variables

Reference truth
80 causal links,
1378 links absent

Timeseries data over 45 grids.
9 Climates, 5 grids

Application of 4 different CD methods and PCC

Literature review of CD methods:
- Stochastic
- Multivariate
- Theoretically distinct

Efficacy of CD methods: Comparison of estimated drivers and reference truth

Extraction of drivers of surface SM: split sample timeseries model

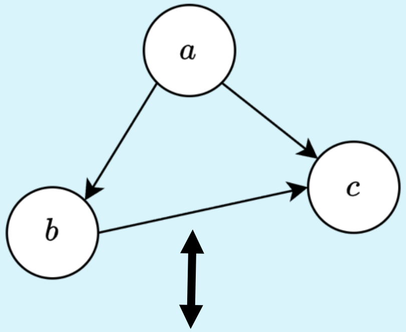
How to represent causal relations? And interesting properties

Node: Variable (timeseries) or event

Link/Edge/Adjacency: Relationship b/w nodes

Directed edge: Cause effect in direction of edge

Directed-Acyclic Graph (DAG)

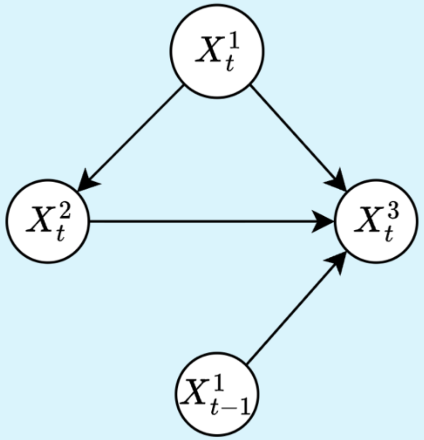


Adjacency Matrix

	a	b	c
a	0	1	1
b	0	0	1
c	0	0	0

$$X_t^1 = \eta_t^1$$
$$X_t^2 = a \cdot X_t^1$$
$$X_t^3 = b \cdot X_t^1 - c \cdot X_t^2 + d \cdot X_{t-1}^1$$

Acyclicity



Lower triangular ordering

	X_t^1	X_t^2	X_t^3	X_{t-1}^1
X_t^1	0	1	1	0
X_t^2	0	0	1	0
X_t^3	0	0	0	0
X_{t-1}^1	0	0	1	0

	X_t^3	X_t^2	X_t^1	X_{t-1}^1
X_t^3	0	0	0	0
X_t^2	1	0	0	0
X_t^1	1	1	0	0
X_{t-1}^1	1	0	0	0

Structural vector autoregressive model (SVAR)

$$\begin{bmatrix} X_t^1 \\ X_t^2 \\ X_t^3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ a & 0 & 0 & 0 \\ b & -c & 0 & d \end{bmatrix} \begin{bmatrix} X_t^1 \\ X_t^2 \\ X_t^3 \\ X_{t-1}^1 \end{bmatrix}$$

$$X = AB$$

How is Casual Discovery done?

- Granger Causality
- Noise-based
- Constraint-based
- Score-based
- State-space reconstruction based: Convergent Cross Mapping
- Others

TCDF (Nauta et al., 2020)

- Find DAG

- Method: CNN to model time-series, Attention scores to unravel causality

VARLiNGAM (Hyvärinen et al., 2008)

- Classic SVAR model

- Method: Exploits lower triangular ordering of adjacency matrix

PCMCI+ (Runge et al., 2022)

- Find DAG

- Method: Conditional independence tests

DYNOTEARS (Pamfil et al., 2020)

- Classic SVAR model

- Method: Numerical optimization to minimize the model error and acyclicity score

Comparing two adjacency matrices

- One-to-one comparison of adjacency matrices yields a confusion matrix.

Confusion matrix

	Actual	
Predicted	True Positive	True Negative
	False Positive	False Negative

Recall (*TPR*): ability to accurately identify the correct links in the adjacency matrix

$$= \frac{True\ Positive}{True\ Positive + False\ Negative}, \in [0,1]$$

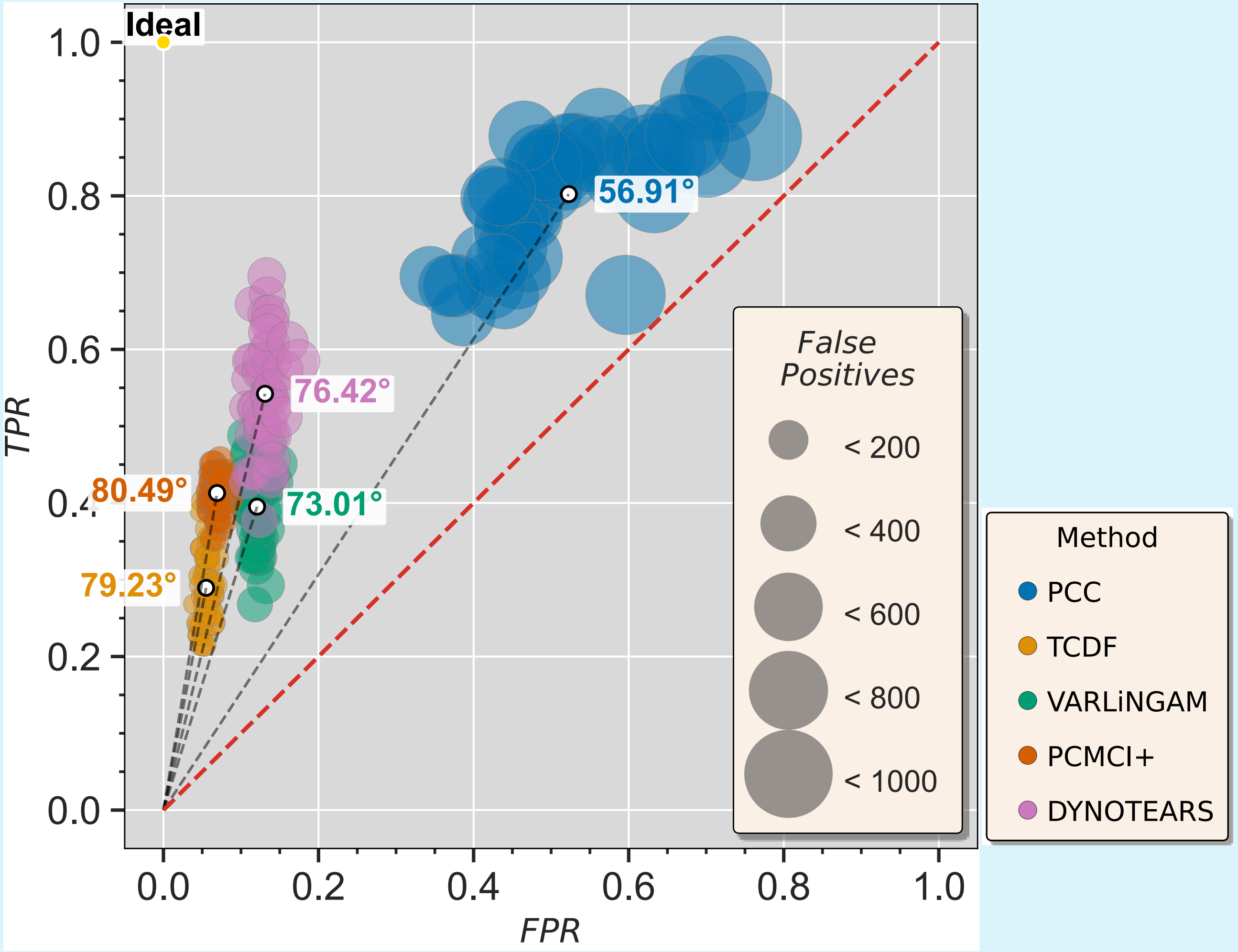
False positive ratio (*FPR*): ability to accurately identify the incorrect links in the adjacency matrix

$$= \frac{False\ Positive}{False\ Positive + True\ Negative}, \in [0,1]$$

Results

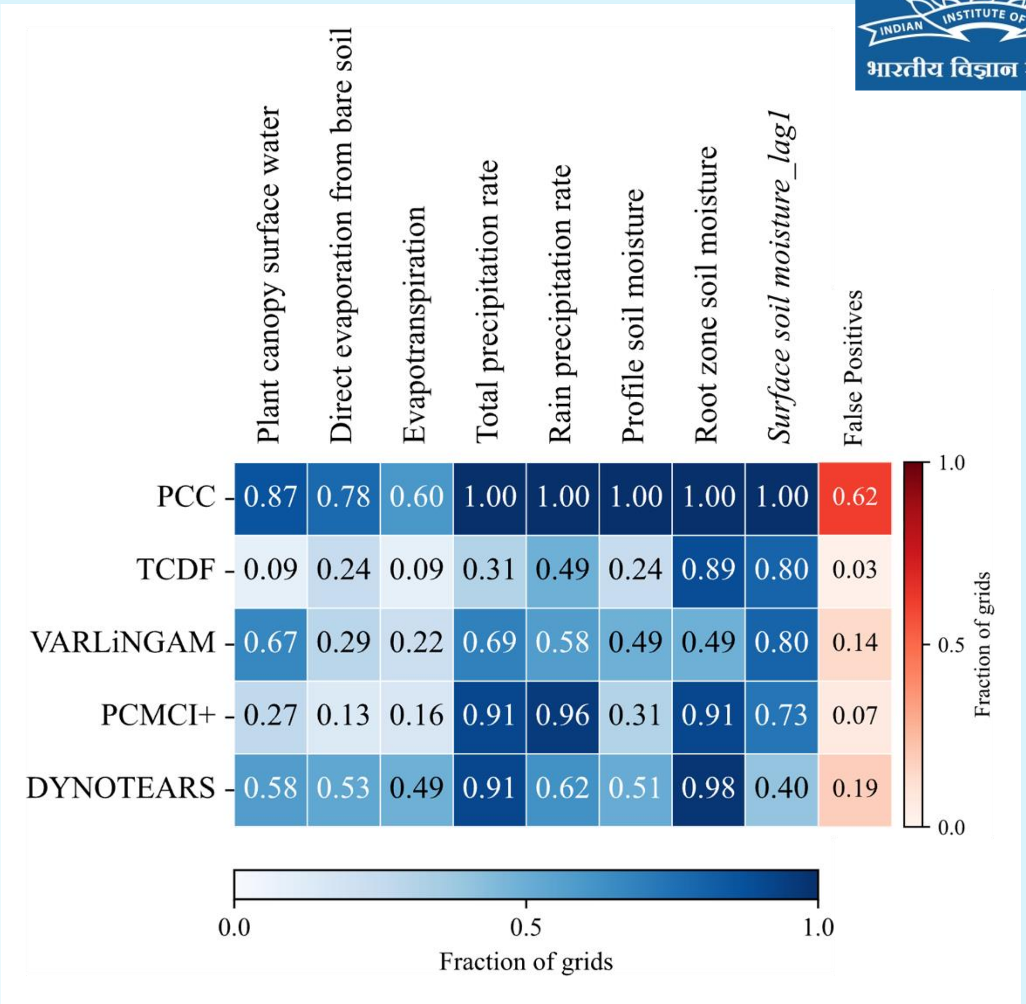
Identifying causal drivers of the complete system

- Cost of identifying causal links is the accumulation of false positives.
- PCC shows high variance along TPR and FPR, indicating lack of robustness.
- Comparatively, CD methods show less variance along FPR, demonstrating robustness towards eliminating false positives.
- For a unit gain in FPR, CD methods have a higher TPR.



Causal drivers of surface soil moisture

- Compared to PCC, CD methods identify fewer causal drivers.
- However, CD methods eliminate large number of correlated but *causally unrelated* variables.
- This allows a parsimonious surface soil moisture representation and model.

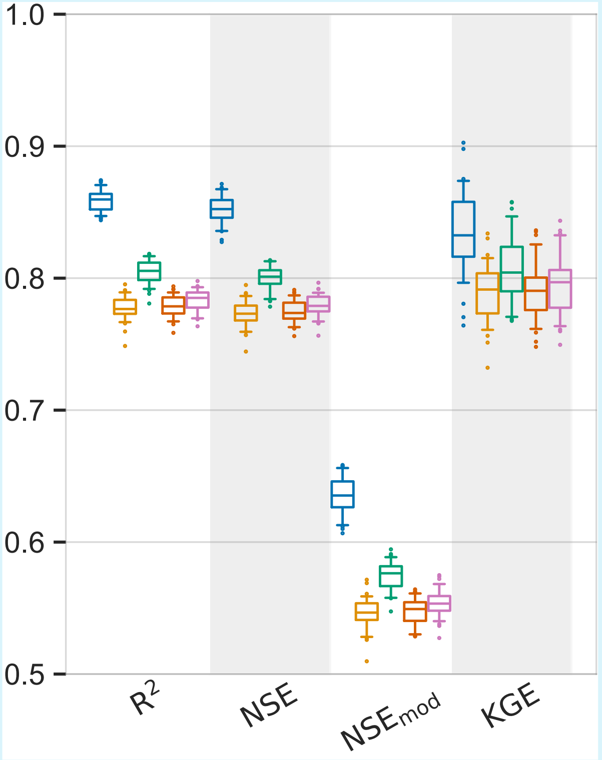


Time-series prediction with causal drivers

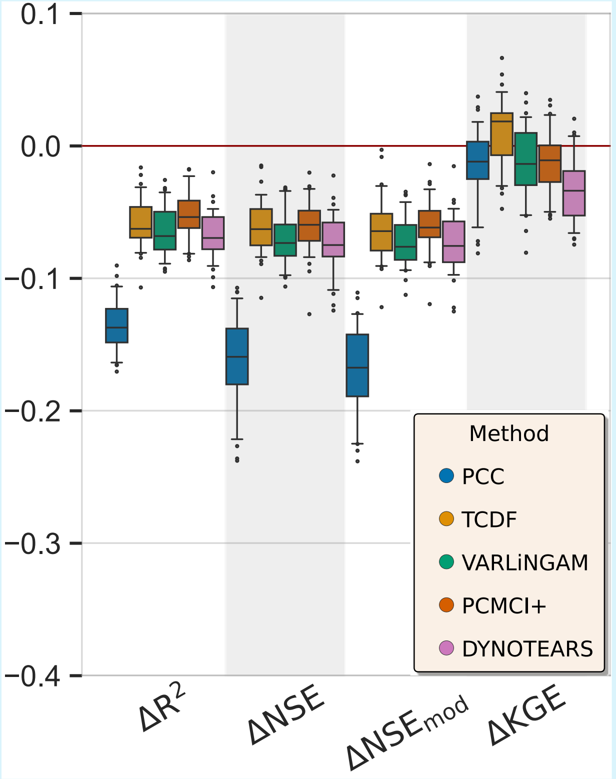
Location	Ganga basin, India
Model	Support Vector Regression
Training	Normal period (01 Jan 2000 to 31 December 2003)
Testing	Drought period (01 Jan 2004 to 31 December 2005)
Noise added	Random Gaussian (100 Monte Carlo simulations)

Performance metrics

Training period



Difference: Testing – Training



Conclusions

- Hydrometeorological systems are highly interconnected, with many cross and lagged correlations.
- Causal Discovery can help identify direct process drivers of variables in such systems.
- Robust identification of drivers is important for predictions, especially under changing conditions like drought, climate change.

References

1. Ombadi, Mohammed, et al. "Evaluation of methods for causal discovery in hydrometeorological systems." Water Resources Research 56.7 (2020): e2020WR027251.
2. Pearl, Judea. Causality. Cambridge university press, 2009.
3. Runge, Jakob, et al. "Inferring causation from time series in Earth system sciences." Nature communications 10.1 (2019): 2553.
4. Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. Elements of causal inference: foundations and learning algorithms. The MIT press, 2017.
5. Nauta, Meike, Doina Bucur, and Christin Seifert. "Causal discovery with attention-based convolutional neural networks." Machine Learning and Knowledge Extraction 1.1 (2019): 19.
6. Pamfil, Roxana, et al. "Dynotears: Structure learning from time-series data." *International Conference on Artificial Intelligence and Statistics*. Pmlr, 2020.
7. Hyvärinen, Aapo, Shohei Shimizu, and Patrik O. Hoyer. "Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-Gaussianity." *Proceedings of the 25th international conference on Machine learning*. 2008.
8. Runge, Jakob. "Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets." *Conference on uncertainty in artificial intelligence*. Pmlr, 2020.

https://doi.org/10.5194/egusphere-2025-4650

Abstract | Discussion | Metrics

05 Nov 2025

Status: this preprint is open for discussion and under review for Hydrology and Earth System Sciences (HESS).

Cause-effect discovery in Hydrometeorological Systems: Evaluation of Causal Discovery methods

Vivek Kumar Yadav, Murray Peel, Keirnan Fowler, Dongryeol Ryu, and Bramha Dutt Vishwakarma

Abstract. Identifying the driver(s) of a process or phenomenon is central to understanding and predicting its future state. In complex hydrometeorological systems, a process can have multiple drivers dynamically coupled to the system across timescales. Thus, a robust method to identify drivers is imperative. In hydrological sciences, methods like multivariate regression and, more recently, Big Data machine-learning approaches rely on finding a *co*-relation between variables, rather than identifying cause-effect relations. This study evaluates cause-effect discovery (Causal Discovery or CD) algorithms in hydrometeorological systems. Although earlier studies have made important contributions to exploring CD methods, they have primarily focused on bivariate methods in simple synthetic environments. Specifically, we evaluate the following four theoretically distinct multivariate CD algorithms, (i) TCDF (ii) VARLiNGAM, (iii) PCMCI+, and (iv) DYNOTEARS. We evaluate these algorithms within a large, complex simulated environment of the Global Land Data Assimilation System (GLDAS) where the drivers, reference truth, are known perfectly. We evaluate the drivers identified by CD methods against this reference truth and also contrast its results with the widely used method of *co*-relation identification, Pearson's Correlation Coefficient (PCC). The results show that CD methods identify fewer false drivers compared to PCC, across a range of Köppen-Geiger climate types. For example, PCC failed to distinguish true drivers from instantaneous and lagged cross-correlations, typically present in hydrometeorological systems. Whereas, CD methods eliminate a higher number of false instantaneous and lagged drivers. Thus, though PCC identifies the highest number of true drivers, it suffers from high false drivers. Overall, CD methods perform similar to or better than PCC, while PCMCI+ and DYNOTEARS performed the best. Further, we test whether time-series prediction models perform better when predictors are limited to those identified as causal by CD methods. Evaluation of surface soil moisture predictions during drought shows that CD-based models outperform PCC-based models and are more parsimonious. Thus, we demonstrate the effectiveness of using causal discovery to eliminate spurious relations and obtain a robust set of drivers for prediction and process understanding across different climate conditions. This study overviews, demonstrates and tests efficacy of CD methods in studying cause-effect relations in hydrometeorological systems. By exposing their capabilities and differences in a simulated environment, we hope to encourage their use in the real world and move beyond *co*-relation.

Preprint paper

Thank you!