Beyond correlation: Cause-effect relations in hydrometeorological systems

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

[1]Interdisciplinary Centre for Water Research, Indian Institute of Science Bengaluru
[2]Faculty of Engineering and Information Technology, The University of Melbourne
Email: viveky@iisc.ac.in

Abstract: Hydrometeorological systems are complex systems with multiple closely related variables. For predicting a variable of interest, data driven methods like statistical regression and, in recent times, big data machine learning have been increasingly used as an alternative to physically-based modelling. However, these techniques rely on finding a co-relation between variables, rather than cause-effect relations. While embedding physical laws into machine learning models has been suggested to make them more reliable, identifying parsimonious and causal (cause-effect) related drivers remains important for any method to make it more interpretable and reliable.

This work explores causal discovery (CD) algorithms (Peters et al., 2017) to recover the drivers of a system, in a simulated environment where the true causal relations are known—namely, the Global Land Data Assimilation System (GLDAS, v2.0). By applying four theoretically distinct CD algorithms– *(i)* TCDF *(ii)* VARLiNGAM, *(iii)* PCMCI+, and *(iv)* DYNOTEARS, we cover the broad spectrum of CD approaches and evaluate their efficacy in identifying the known causal drivers (Assaad et al., 2022). Further, we contrast these CD results with a non-causal method of simple correlation, Pearson's correlation coefficient (PCC). While identifying a causal link is important to understand cause-effect relations between variables, eliminating spurious correlation as false causality is important to remove non-causal relations and obtain a parsimonious set of predictors. Accordingly, we evaluate the performance of CD methods and PCC on both these aspects.

The results show that CD methods identify fewer false drivers compared to PCC, which is prone to spurious associations from cross-correlations and lagged correlations, typically present in hydrometeorological systems. In contrast, 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 a high number of false drivers. Overall, CD methods perform similar to or better than PCC, with PCMCI+ and DYNOTEARS performing the best. These findings highlight the value of CD for eliminating spurious relations and deriving a robust, parsimonious set of predictors for process understanding and predictions under diverse climate conditions.

This study overviews, demonstrates and tests the efficacy of CD methods in identifying 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.

REFERENCES

Assaad, C.K., Devijver, E. and Gaussier, E., 2022. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, *73*, pp.767-819.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

Keywords: *Causality, hydrology, machine learning*