



Cause-effect based modelling for reliable results under changing climatic conditions

Vivek Kumar Yadav^{1,2}, Murray Peel², Keirnan Fowler², Dongryeol Ryu², and Bramha Dutt Vishwakarma¹

¹Interdisciplinary Centre for Water Research, Indian Institute of Science, Bengaluru, 560012, India (viveky@iisc.ac.in)

²The University of Melbourne, Faculty of Engineering and Information and Technology, Melbourne, Australia

Identifying the drivers of a process is imperative to its understanding and forecasting, especially under changing climate. Hydrometeorological systems are complex with multiple closely related variables. In such systems a process can have multiple drivers, coupled to the system, across timescales. Thus, identifying the drivers of a process becomes challenging. In Hydrology, multivariate regression and recently Big Data machine learning methods have gained popularity. However, these methods rely on finding correlation between variables and fall short of identifying causal (cause-effect) relations.

This work explores causal discovery (CD) algorithms to identify the drivers in a hydrological system. 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 and 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 contrast its results with the widely used method of co-relation identification, 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 also important to obtain a parsimonious set of predictors. Accordingly, we evaluate the performance of CD methods and PCC for 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, although 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.

Further, we evaluate the effect of focusing on causal drivers by training machine learning models for surface soil moisture prediction. We evaluate their performance under changing climate conditions of drought. PCC-based models show higher performance in the training period (median $R^2=0.85$ & $NSE=0.84$); however, they suffer a sharp drop in performance during the test period. In

contrast CD-based models show decent performance in training (median $R^2 \sim 0.8$ & $NSE \sim 0.78$) and are more robust in the testing period. Together, these findings highlight the value of CD for eliminating spurious relations and retrieving 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 correlation.