The authors attempt to use “real-world data” (GLDAS) to systematically evaluate the effectiveness of different causal inference methods. I believe this is an important step for the development of the causal discovery field, which has long lacked benchmark testbeds based on realistic and complex systems for a rigorous evaluation of algorithmic performance. This absence of realistic benchmarks has made me skeptical about the practical usefulness of different causal discovery algorithms—especially those built on different assumptions—when applied under real-world, highly complex conditions. From this perspective, I think this work represents a valuable and meaningful attempt in this direction. However, there are several MAJOR issues in the study that, in my view, require further clarification and discussion.

Concerns with “true” causality matrix

When constructing the “true” causal adjacency matrix, the authors appear to consider only a one-day lagged causal effect. While this assumption may be reasonable for some fast-response flux variables (e.g., surface energy fluxes), many land-surface variables are known to exhibit pronounced long-term memory, such as soil moisture, groundwater storage, and snowpack. These memory effects typically influence the current state through multi-step Markov processes, rather than solely through the immediately preceding time step. It is worth noting that nearly all causal discovery methods evaluated in this study are, in principle, capable of explicitly accounting for multi-lag causal relationships. Therefore, assuming a uniform one-day lag in the reference “true” adjacency matrix may limit the realism of the benchmark and potentially affect the fairness of the evaluation.

Concerns with the selection of causality discovery methods

The authors compare four causal discovery algorithms that originate from distinct methodological paradigms. However, the manuscript does not sufficiently justify why these four specific algorithms were selected over other causal inference methods that are more commonly used in the hydrometeorological community. While PCMCI and its variants have seen increasing adoption in atmospheric and hydrological studies, widely used approaches such as CCM and Granger causality have not been directly included in the comparison. I therefore recommend that the authors provide a more explicit and systematic rationale for their choice of algorithms.

Concerns with the comparison of different methods

The manuscript presents an extensive set of quantitative analyses, using multiple statistical metrics to demonstrate the overall performance of different causal discovery methods. However, the connection between the discovered causal structures and real hydrometeorological processes is not sufficiently explored. This will bring some misleading result. For example, given that causal graphs are subject to Markov equivalence, different DAG structures may yield similar performance.

In this regard, the manuscript would benefit from the inclusion of more concrete case studies. For example, what specific driving variables are identified by CD methods in different basins or climate regimes, and how do these compare with those selected by PCC? Are the identified drivers consistent with known physical mechanisms governing land–atmosphere interactions, hydrological processes, or energy balance? Conversely, which suspicious or physically implausible links are removed by CD methods relative to correlation-based approaches? Moreover, the authors may consider presenting spatial patterns of the inferred causal drivers, for example by mapping the dominant drivers across grid points within a given basin or region. Such spatially explicit analyses would offer a more intuitive and diagnostic perspective on method performance.

Concerns with the result of predicting time-series

The authors attempt to evaluate different causal discovery methods by using the variables selected by each method to drive machine learning models, and then comparing predictive performance as a proxy for causal effectiveness. However, I believe the resulting conclusions require more careful interpretation.

First, the number of features selected by PCC and by the CD methods differs substantially. Under identical training data, training protocols, and hyperparameter settings, models with a larger number of input features generally have a higher risk of overfitting and poorer generalization performance. As such, the reported differences in predictive skill may primarily reflect differences in feature dimensionality rather than the intrinsic quality or causal relevance of the selected predictors.

Second, soil moisture is a state variable with strong temporal memory. Its current value is typically highly and approximately linearly correlated with its lag-1 state, which, in practice, already contains most of the predictive information about the system. From my experience, introducing complex models or large sets of external predictors can sometimes degrade this physically consistent memory structure, leading to unstable or nonphysical mappings.

If the authors wish to retain a prediction-based comparison, I strongly recommend adopting a more controlled and interpretable experimental design. For example, this could include: (i) enforcing the same number of input features across different methods (e.g., using only the top-k ranked predictors), (ii) ensuring comparable model capacity or parameter counts across experiments, and (iii) explicitly accounting for the role of lag-1 soil moisture as a baseline or control predictor.

Specific comments

Line 201: The use of the Matthews Correlation Coefficient (MCC) requires further explanation, as it is not a commonly used metric in this context. In particular, the authors should clarify its interpretation and why it is more appropriate than standard metrics under the highly imbalanced adjacency matrix setting.

Line 250: Incorporating acyclicity as a soft constraint in the loss function does not strictly guarantee a DAG solution, and such soft constraints can fail in practice, especially under finite-sample conditions or suboptimal hyperparameter choices.

Line 266: Eq.(9) appear before Eq.(8).

Line 569-470: The manuscript does not clearly justify why these specific years were selected for model training and validation.

Line 475: I do not consider the addition of random noise alone to constitute a more realistic scenario, nor does it meaningfully test robustness, since many error assumptions are already based on Gaussian noise.