Response to comments from reviewer #2

We thank the reviewer for their comments and suggestions. We tried to address their comments to the best of our ability and understanding. Further their suggestion to improve the timeseries prediction model has increased our confidence in causal methods. Please read below our response to their comments and suggested changes to address them.

In the document below, we have used ink colour as:

Authors response	Agreeing/disagreeing/clarifying our position on the comment. Then explain our position.
Suggested changes in manuscript from authors	We detail (if applicable) the changes we make in response to the comment and our response to it.
Text in green	Text common in the response documents, towards both the reviewers.

1. **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.

   Author response: Only one-day lagged causal effects are considered in the construction of the true adjacency matrix. We clarify that this is based on the generating equations of the CLSM F-2.5 model and not taken as an assumption.

   Indeed, many processes can have muti-step Markov formulation. In the case of the CLSM F-2.5 model, these relations were restricted to just one-step Markov processes. For example, equation A6 shows groundwater storage as a function of contemporaneous root zone soil moisture, baseflow and groundwater storage at the previous timestep.

   Consequently, the constructed true adjacency matrix reflects this and restricts causal relations for one-step lagged relations only.

   Suggested changes in manuscript from authors: None.

2. **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.

Author response: We partially agree with the reviewer, that the rationale for selecting the CD methods evaluated in this work may not be explicit.

The commonly used methods of CD, Granger causality (GC), transfer entropy (TE), CCM and PC-alg, were considered when selecting algorithms.

However as argued in Lines 83-95, these methods are bivariate and cannot find the correct causation under confounding variables. Similarly, many variables show strong state dependence (self-causation via autocorrelation) which cannot be handled by these methods. Further GC, TE and PC-alg cannot find the correct causal structure in contemporaneous causal interactions. CCM is based on a deterministic system assumption, where hydro-meteorological systems are typically stochastic. Finally, PCMCI (a precursor of selected PCMCI+) cannot discover contemporaneous relations.

The four CD methods evaluated in this work were chosen to represent theoretically distinct methodologies of finding causal relations in time-series data. These methods are suitable to multivariate systems, non-deterministic settings and can find self-causal and contemporaneous relations. By selecting methods based on, noise-based assumptions, score-based, constraint-based and GC-inspired, we cover the broad spectrum of causal discovery approaches (Lines 89-103).

Towards this, we shall add text to describe the rationale of not selecting PCMCI.

Suggested changes in manuscript from authors: We shall add the text below in Line 95.

“Finally, PCMCI, a method gaining rapid adoption in hydrological and atmospheric science, was not selected as it cannot discover contemporaneous relations.”

3. **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.

Author response: We partially agree to the reviewer.

The reviewer points out that our quantitative analysis is extensive but lacks an explicit analysis of the discovered causal drivers with respect to known physical mechanisms and exploring the climate/basin wise analysis.

They suggest exploring the following:

- Are the causal drivers identified by CD methods, consistent with existing physical mechanisms of land-atmosphere, hydrological process and energy balance equation? And how they are different from those identified by PCC?
  - In Sect. 3.4 we explored exactly this aspect. We focused on the causal drivers of surface soil moisture, which is jointly driven by water budget and energy budget equations. However, we did not explicitly classify them under different process types in the ensuing section.
- Present a spatially explicit/basin-wise or climate-wise analysis of identified causal drivers.
  - Fig. 5 and 6 show the identified causal drivers by PCC and CD methods across the different climate classes. Though we describe some climate-wise patterns observed in them, we did not explicitly discuss them.
- Finally, the reviewer points to a limitation of DAGs which can potentially lead to similar performance of different DAGs.
  - We clarify that in our case such ambiguity does not occur. The metrics used, Recall, MCC and FPR are calculated accounting for the directionality of a causal relation in the adjacency matrix. Thus, if two graphs belong to the same Markov equivalence class, they would not obtain the same scores, because they have different directionality.

Suggested changes in manuscript from authors: Based on Fig. 5 and 6, we shall modify Sect. 3.4 to include a more explicit discussion of the following:

- Classify the identified causal drivers according to known hydrometeorological processes.
- Analyse the spatial (climate/basin-wise) patterns of discovered causal drivers.

4. **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.

Author response: We agree with the reviewer, implemented their suggestions. The results support the hypothesis that focusing on CD drivers for timeseries prediction yields superior results.

The analysis for timeseries prediction model based on causality and PCC shown in Sect 3.5 was flagged as unsatisfactory by both reviewers. Where they raised issues regarding the methodology adopted and both reviewers gave suggestions for a more thorough analysis. We thank them for their suggestions and report that their suggestions have increased our confidence on causal methods.

In summary, we implemented the suggestions from both the reviewers (please find full details in either response document). The results support the hypothesis that focusing on CD drivers for timeseries prediction yields superior results.

Further, in response to a comment raised by Reviewer #2 (last comment), we have modified the description of adding random noise. Thus, we do not claim it to be realistic scenario, rather a non-idealized setting where random noise represents observational noise, typically present in hydrometeorological systems.

Please read below for the conclusions and details of the analysis suggested by Reviewer #2.

Also, you may see the results of the analysis as suggested by Dr. Uwe Ehret in the other response document.

We thank the reviewer for their comment. They raise the point that the conclusions of our study in Sect 3.5 may be due to:

a. A result of high dimensionality of the predictors identified by PCC compared to the CD methods. Towards this they suggest restricting the number of predictors to a common number, across the different methods. And a common model architecture (suggestion i and ii). For this they suggest using a TOP-K approach.

    - Towards this we report that we adopted their suggestions and emulated the experiment, called TOP-K hereafter, and report the results below.

b. Soil moisture being a state variable shows high temporal memory and thus be explicitly included in the timeseries prediction of all models (suggestion iii).

    - Towards this we report that the predictors identified by all the methods (PCC and CD methods) in the manuscript did in-fact identify the lag-1 soil moisture as a causal driver. Hence all the results shown in Figure 7 of manuscript are based on timeseries models having lag-1 soil moisture as an explicit predictor, for all methods (PCC and CD).

- Further, as can be seen in Table 2 below, this was ensured in this experimental setup (TOP-K analysis) as well.

Please read below for the conclusions and details of the analysis suggested by Reviewer #2. Also, you may see the results of the analysis as suggested by Reviewer #1 (Uwe Ehret) in the other response document.

We partially agree to the comment and thank the reviewer for their suggestion. We emulated the TOP-K approach suggested by them and report the following conclusions based on the results:

- ○ CD based models show higher performance compared to PCC based models, both, in the training and testing periods.
- ○ CD based models show higher robustness compared to PCC based models across the testing and training periods.
- ○ Similarly, CD based models show lower errors compared to PCC based models, both in the training and testing periods.
- ○ However, PCC and CD based models show similar robustness in error metrics.

Below we show the details of the experiment and the results.

TOP-K approach: To select a common number of drivers for all methods we took the predictors selected by different methods (i.e. the same as the manuscript) and filtered them with the following approach:

We choose K=8, as it is the equal to the actual number of true causal drivers of surface soil moisture.

Selection criteria of TOP-K predictors for different methods	
Method	**Selection criteria**
PCC	Sort the variables selected by PCC according to their correlation coefficient and select the top 8 variables.
TCDF	None. Difficult to extract the Adjacency matrix from the python code. Thus, we keep predictors same as the manuscript, which incidentally were 8.
VARLiNGAM	Sort the variables of the adjacency matrix according to their matrix coefficient and select the top 8 variables.
PCMCIplus	Same as VARLiNGAM
DYNOTEARS	Same as VARLiNGAM

Table 1. Selection criteria to implement the TOP-K approach for various methods.

This yielded the following set of drivers for each method:

	PCC	**TCDF**	**VARLiNGAM**	**PCMCIplus**	**DYNOTEARS**
1.	SoilMoist_RZ	Evap	Qs	Qg_tavg	AvgSurfT
2.	SoilMoist_S_lag1	Rainf	Rainf_f	Qs	Lwnet
3.	SoilMoist_RZ_lag1	SoilMoist_P	Rainf	Rainf_f	Qair_f
4.	Tws	CanopInt_lag1	SoilMoist_RZ	Rainf	Rainf_f

5.	SoilMoist_P	GWS_lag1	Tws	SoilMoist_RZ	SoilMoist_RZ
6.	Tws_lag1	SoilMoist_RZ_lag1	SoilMoist_RZ_lag1	Swdown_f	Tair_f
7.	SoilMoist_P_lag1	SoilMoist_S_lag1	SoilMoist_S_lag1	Tws	AvgSurfT_lag1
8.	GWS	Tws_lag1	Tveg_lag1	SoilMoist_S_lag1	SoilMoist_S_lag1

Table 2. Table showing the set of predictors for each method, after filtering the respective set identified by each method respectively.

Details of machine learning models: Based on the set of predictors in Table 2 machine learning models (feedforward neural network) were trained for surface soil moisture prediction. The model architecture was ensured to be the same for all the methods.

The other details, i.e. the train and test period, location, target variable and noise level is same as the manuscript.

Predictors	**Filtered via TOP-K approach**
Noise added	0.5 standard deviation (Same as manuscript)
Training period	01-01-2000 to 31-12-2003 (Same as manuscript)
Testing period	01-01-2004 to 31-12-2004 (Same as manuscript)
Location	Ganga Basin (Same as manuscript)
Target variable	Surface soil moisture (Same as manuscript)

Table 3. Details of experimental setup as suggested by Reviewer #2.

Results for TOP-K approach:

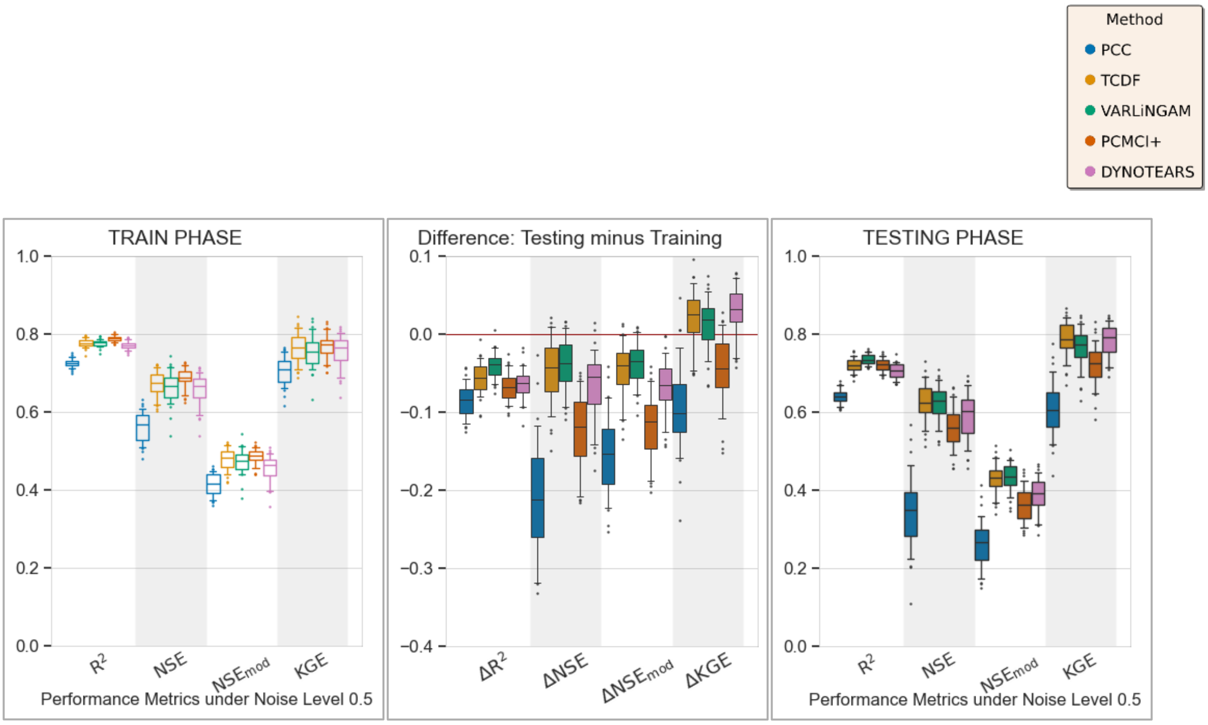

Figure 1a. Performance metrics for machine learning models based on predictors in Table 4.

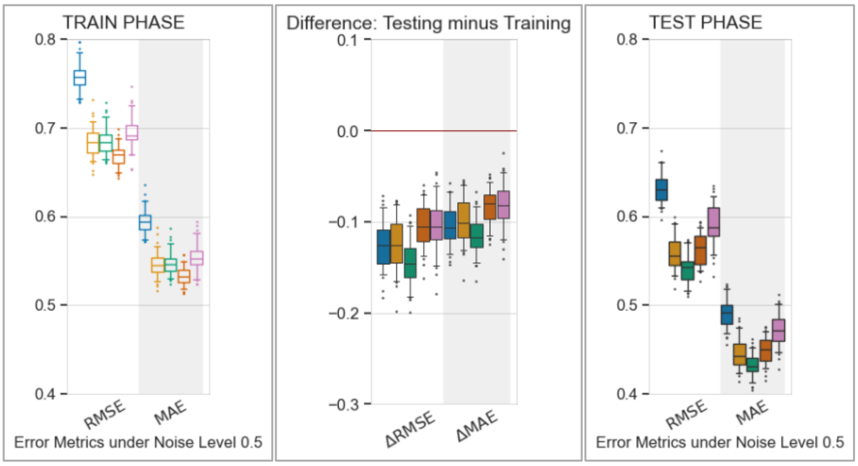

Figure 1b. Error metrics for machine learning models based on predictors in Table 4.

Suggested changes in manuscript from authors: We intend to keep all the (three) set of results in the manuscript.

After considering the suggestion of both the reviewers, we now have three sets of timeseries prediction results, each having different methodologies. The table (Table 2) below summarises them:

Experimental setup	**Main conclusions**
Noise added to predictors selected by PCC and CD methods – from the manuscript.	• High performance of PCC based models in training but significant reduction under testing phase. • Decent performance of CD based models in training while robust results under testing.
No noise added to variables, rather variation of training period – as suggested by Uwe Ehret.	• With shorter periods of training length, models fail to stabilize across the training and testing periods, where stability is achieved after approximately a year of training. • Causality based models suffer smaller drop in performance compared to PCC based models • Causality based models stabilize earlier than PCC based models.
TOP-K (8) predictors filtered from the predictors selected by PCC and CD methods – as suggested by Reviewer 2	• CD based models show higher performance compared to PCC based models in the training and testing periods. • CD based models are more robust across the periods compared to PCC based models.

Table 4: Summary of experiments for Section 3.5

We adopted the methodology as shown in the manuscript as representative of a more *traditional* approach towards timeseries prediction modelling. Based on the experimental setup and results of the two suggestions, we would improve the discussion on the efficacy of various methods. We shall add a paragraph in addition to existing text describing the results shown above and the corresponding plots in the supplementary.

Specific comments

1. 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.

   Author response: We agree that MCC is not a common metric used in the context of comparing DAGs. However, in our case it is the suitable metric, as in RQ2 we evaluate how different methods perform when considering both the ability to find true causal relations and eliminating the false relations. Thus, MCC acts as a metric balancing the ability to find true causal drivers and retaining a parsimonious representation.

   Moreover, metrics ignoring certain classes of the confusion matrix are susceptible to show a biased picture in case of heavy class imbalance (Chicco and Jurman, 2020). This can be seen in our results as well, where PCC shows the highest Recall scores, however it shows lowest MCC scores.

   Since we had not described our true adjacency matrix at this point in the manuscript (Line 201), we avoided mentioning the class imbalance present in it. Thus, we described its effect later for showing the MCC based results in Sect 3.2 Line 499.

   Suggested changes in manuscript from authors: Towards this we suggest adding the following text in Line 204.

   "…classes of the confusion matrix. Such metrics can show a biased picture in case of high class imbalance (TP>>TN or TP<<TN). MCC considers all the four classes (TP, TN, FP and FN) and thus is unaffected by imbalance in the dataset (Chicco and Jurman, 2020). Moreover, by considering both the ability to find true causal relations and eliminating false relations, MCC acts as a metric balancing the ability to find causal drivers and retaining a parsimonious representation."

2. 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.

   Author response: We agree that the acyclicity as a soft constraint does not necessarily ensure a DAG solution. We shall modify the text accordingly.

   Suggested changes in manuscript from authors: Modify the text in Line 250 to remove *ensuring its acyclicity* with "penalizing its cyclicity".

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

   Author response: Agreed, we shall modify the text to maintain the chronological order of appearance of equations.

   Suggested changes in manuscript from authors: Move loss function equation 9 in text before acyclicity equation 8.

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

   Author response: We selected a location where the climatic conditions underwent significant change. The Ganga basin faced drought during 2004-05 period; thus, we selected it.

   Suggested changes in manuscript from authors: We shall add a description for the same in Line 468.

5. 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.

   Author response: We agree with the reviewer. The idea behind introducing additive noise was to remove the perfect determinism of the model environment. Thus, we adopted this methodology to test the performance of different methods in non-ideal settings to prevent trivial model fit and test models under a representation of observational noise, rather than within a perfectly deterministic environment.

   Further we clarify that we did Monte Carlo simulations to test the robustness against different levels of synthetically added noise. As a single realization of random noise can have very high or low levels of randomness added. We did not intend to claim the overall robustness of the models. We shall revise the text accordingly.

   Suggested changes in manuscript from authors: We shall revise the text in Line 475 as below:

   “Thus, we introduced random additive Gaussian noise to the data to relax this idealized environment. This prevented trivial model fits within a deterministic environment and allowed us to test the models under a representation of observational noise, typically present in hydrometeorological systems.

   To evaluate the sensitivity of results to the magnitude and realization of the added noise, we conducted Monte Carlo simulations across multiple noise levels (Appendix C).”

   References:

1. Chicco, Davide, and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." BMC genomics 21.1 (2020): 6.