Dear Editor and Reviewers,

We thank the editor and reviewers for their detailed comments which have brought more structure and clarity to the paper. We also thank the reviewers for their suggestions which have made the results more robust and increased our confidence in Causal methods. Please read below for our detailed response to the comments raised in the review of our manuscript.

For each comment, under the "Author response" heading we provide our broad agree/disagree/clarification to the comment. Then we explain our position. Under the "Actual changes in manuscript from authors", we provide the exact changes in the manuscript, as a result of the response. These changes are implemented in the revised manuscript.

To Reviewer #R1, Dr Uwe Ehret.

We have merely added the changes originally proposed, or in some cases updated those changes in a way that is consistent with the original response document.

To Reviewer #R2,

For one case, we have changed our mind relative to the response document; in all other cases we have merely added the changes originally proposed, or in some cases updated those changes in a way that is consistent with the original response document.

Further, the table (Table 1) below lists the instances where our response has a major update or change compared to our original *response to reviewers* posted on 10th Feb 2026. Please note that the comments where we have not made major or any changes are described as "-". While places where we have updated our response or added the response (in-line with the first response) are stated as "Updated" and "Added" respectively. While "Changed" marks the instance where we have changed our response compared to the first response.

Finally, we have updated the description of the VARLiNGAM algorithm to make it more explicit and lucid, and corrected an incorrect mathematical formulation (Sect 2.5.2).

Table No.	Comment number (Row No.)	What changed	
		Author Response	Actual changes
Updates/changes in response to Reviewer #R1			
4	4	-	Updated
4	6	-	Updated
4	11	-	Added
4	12	-	Updated
4	13	-	Added
4	14	-	Added
4	23	Updated	Updated
4	25	-	Added
4	26	-	Added
Updates/changes in response to Reviewer #R2			
5	1	**Changed**	**Changed**
5	3	Updated	Updated
5	4	Updated	Added
5	8	-	Added

Table 1: Description of changes implemented in contrast to our first response posted 10th Feb 2026.

Table No.	Table name	Colour code
3	Response to comment by Editor	
4	Response to comments by Dr Uwe Ehret (Reviewer #1)	
5	Response to comments by Reviewer #2	

Table 2: Table describing the response tables.

<u>Response to comment by Editor (Table 3)</u>

1.	**Comment by Editor (Yonggen Zhang) towards Reviewer #2 comment and our response** Regarding the restriction to one-day lagged causal effects raised by the 2nd reviewer, I have read your rebuttal, but I agree with the reviewer that this remains a significant concern. While I understand your argument, the physical variables being simulated, such as soil moisture, snowpack, and groundwater, inherently possess long-term memory in the real world. Limiting the benchmark strictly to a lag-1 relationship may fail to thoroughly test the methods' ability to handle the multi-step dependencies they will inevitably encounter. To adequately address this point, please implement one of the following approaches: either include a robustness check or sensitivity test, or provide a substantially stronger justification. You need to provide a compelling theoretical justification for why a lag-1 restricted evaluation remains a fair, robust, and comprehensive benchmark.	(Our response is a unified response to comments raised by both Reviewer #2 and the Editor.) We thank the Editor and Reviewer #2 for their insightful comments regarding the inclusion of long-term memory of hydrological variables and the potential need to account for multi-step Markov processes. We agree that real-world hydrological systems often exhibit pronounced memory effects. However, we wish to clarify that the restriction of our “true” causality adjacency matrix to contemporaneous and one-day lagged relationship is not an arbitrary assumption, but a strict mathematical reflection of the physically based environment used in our analysis. We have carefully considered the comment and respectfully offer the following detailed clarification, which has also been incorporated into the revised Discussion. As the Editor and Reviewer #2 pointed out, the true adjacency matrix constructed in this study contains causal relations from contemporaneous flux variables alongside contemporaneous and one-day lagged state variables. Our method for deriving these “true” causal drivers is based on the mathematical governing equations of the CLSM F-2.5 model (Appendix 2). Within this physically based simulated environment, the governing partial differential equations are essentially formulated as first-order Markov processes. A new state variable is computed using only the state from the immediately preceding time step (as the initial condition) combined with the current meteorological forcing. 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. If we were to expand the true causality matrix beyond one-day lag relations, the rows for the multi-step lag variables would simply be zeros and redundant. However, despite the first-order nature of the governing physics, we completely agree with the editor that many real-world hydrologic variables, such as soil moisture, groundwater storage, and snowpack, exhibit pronounced long-term memory when observed empirically. In observational datasets, we rarely capture the complete state vector of the environment. This partial observability means that the unmeasured physical delays (e.g., percolation time, routing) manifest statistically as long memory. Therefore, if a purely data-driven modelling framework is to be used for prediction based directly on observations, relying solely on lag-1 variables is often insufficient, and higher-order (multi-step) Markov processes or autoregressive models are required to account for the system's integrated memory. This distinction highlights that the suitable choice between including only lag-1 variables versus variables with longer lags depends entirely on the purpose of the causality analysis. If the analysis aims to construct a purely empirical, predictive data-driven model from observations, incorporating longer lags is practically essential to capture the apparent memory effects caused by unobserved intermediate states. However, if the analysis aims to identify causal drivers to inform, validate, or parameterize a physically based model (as is the basis of our CLSM F-2.5 benchmark), the inclusion of only lag-1 variables is both necessary and sufficient, as it perfectly reflects the governing differential equations. We agree with the editor that a more comprehensive analysis would test the ability of CD methods to accurately identify multi-step causal relations. However, as explained above, the simulated environment in which we conducted the analysis natively contains only contemporaneous and one-day lag relations. Thus, our benchmark is strictly limited to these transitions. Furthermore, this limitation implies that our use of a multivariate regression model to compare the efficacy of

		predictors—selected via PCC vs CD methods—may not sufficiently highlight the inherent merits of the true predictors determined by the causality analysis. To ensure this important context is clear to the readers, we have added a dedicated explanation of these dynamics and limitations to the revised Discussion section (Sect 5.3, Line 915) as follows: “Fifth, we applied CD methods in a simulated environment that represents physical processes with through a contemporaneous and one-day lag relations. In contrast, many real-world hydrological variables, such as soil moisture, groundwater storage, and snowpack, exhibit pronounced long-term memory when observed empirically. In observational datasets, we rarely capture the complete state vector of the environment. This partial observability means that the unmeasured physical delays (e.g., percolation time, routing) manifest statistically as long memory. Therefore, if a purely data-driven modelling framework is to be used for prediction based directly on observations, relying solely on lag-1 variables is often insufficient, and higher-order (multi-step) Markov processes or autoregressive models are required to account for the system's integrated memory. While the CD methods evaluated in the manuscript can be adapted to allow detection of multi-step causal relations in a system, derivation of true drivers in this study is based on the mathematical governing equations of the CLSM F-2.5 model (Appendix 2). Because this simulated environment is governed contemporaneous and one-day lag transitions, the resulting true causality matrix only contains contemporaneous and one-day lag relations. Thus, the CD methods were evaluated on their ability to accurately identify the correct causal variable and its correct lag, up-to one timestep. This constraint was an intended design to robustly contrast the key drivers selected from CD against those selected by empirical correlation-based methods, and to evaluate their predictive skills assuming that an exhaustive set of hydrological variables (both forcing and state) is accessible. However, the limited availability of complete observation data indicates that variables with longer lag need to be considered when applying CD to real-world datasets. Investigating causal structures under these conditions is an interesting topic of future

		research, as some apparent long-lag causal relationships may actually represent pseudo-causality induced by hidden variables and limited observability. Further studies are warranted to explore scenarios within simulated environments where data availability is artificially restricted to commonly measured variables (e.g. precipitation, discharge and potential evaporation), allowing researchers to determine what can be causally inferred about the missing states. For example, Delforge et al. (2022) utilised CD to reveal hydrological connectivity in a Karst aquifer system using time-series data for rainfall, potential evapotranspiration, resistivity, and percolation rates. By applying CD in both a synthetic case study and real-world observations, they incorporated lag relations of up to five time-steps to accurately capture the time span of preferential flow peaks."

Response to comments by Dr Uwe Ehret (Reviewer #1, Table 4)

Major points		
1.	A cause-effect relation in the sense of this manuscript only exists between directly coupled nodes in a DAG. This differs from the colloquial interpretation, where indirect relations, e.g. between precipitation and streamflow, would also qualify as one. To help the reader, a clear definition of cause-effect relation (and how it differs from correlation) should be placed at the beginning of the manuscript, e.g. in the paragraph starting at Line 52.	Author response: We agree with reviewer, see the paragraph below. Actual changes in manuscript from authors: Text added on Line 55: "Causal relations are defined as direct physical and dynamical influences from the causal drivers (causes) onto a variable (effect). For a given variable, its causal drivers conditionally isolate it from the remaining system and represent only direct interactions. This is fundamentally different from correlation-based approaches like Pearson's correlation coefficient, which aim to identify a statistical dependence between variables, accounting for both, direct and indirect relations. Thus, for example a correlation may exist between rainfall and transpiration, however a causal relationship may not be found between them, given the causal drivers of transpiration are accounted for."
2.	There is a fundamental ambiguity in the way how the "true" causal linkages are defined as those that can be extracted from model equations, as is done in the manuscript: Any	Author response: We agree that using multivariate model equations for extracting causal interactions can induce different DAG structures due to the ambiguity introduced by

	multivariate equation of the form y = f (x1,x2,x3) can be re-expressed as a sequence of nested equations, e.g. y = f(x1, g(x2,x3)), or y = f(g(x1,x2),x3), etc. The choice of the nesting and sequential execution is more or less left to the preferences of the programmer. It will not change results, but it will change the resulting DAG, and with it what qualifies as a cause-effect relation, and what not, according to the definition in the manuscript. Any CD performed on a virtual reality derived from a set of process equations will suffer from this ambiguity, and I wonder to which degree this makes results useless.	nesting different variables together. Taking the example provided by the reviewer, where y = f(x1, x2, x3) can be re-written such that y = f(x1, Z) where Z = g(x2, x3) and y = f(W, x3) where W = h(x1, x2). Such a case would yield two different DAGs as: x1 –> y <– Z and W –> y <–x3. However, we define causality as the interaction of various state and flux variables of the model, as defined by their structural generating equations. Thus, the adjacency matrix captures the interactions between meaningful physical (simulated) variables and represents causal interactions of actual physical processes, rather than different intermediaries possible from nesting permutations of variables. With this definition, in the above example if Z represents an actual physical variable of the earth system, we consider it to be a causal driver of y. This eliminates the other two formulations and reduces the ambiguity, while grounding its causal interaction with y as physical process. <u>Actual changes in manuscript from authors:</u> Text added in Line 247. "Thus, the true adjacency matrix represents only the causal interactions of various state and flux variables, as represented by the model's generating equations, thereby rooting the causality in physical processes only."
3.	Most hydrological models use non-iterative numerical schemes, where a flux equation is followed by a state-updating equation. E.g. outflow from a linear reservoir is calculated as ○ Q(t+1) = S(t)*k ○ S(t+1) = S(t) - Q(t+1) * dt If I understood correctly from the manuscript, such a process equation structure cannot be represented by a DAG, because Q=f(S) and S = f(Q) and hence DAG-based CD methods cannot be applied. If correct, this would be a hindrance for the adoption of CD methods in hydrology, and should be mentioned as a limitation in the discussion or conclusion.	<u>Author response:</u> We clarify that such relations can be represented with DAGs, if the time indexing of variables is considered, which ensures the acyclicity of the resultant DAG. In the example provided by the reviewer, the causal relations can be represented via a DAG after considering the time indexing. The explicit time ordering can be shown by creating a new node for the time lagged variable as shown in figure 1b. Alternatively (in favour of brevity), the lag information is usually annotated on the arrows of a regular DAG, as shown in figure 1c. Similarly, the adjacency matrix must be expanded to accommodate the lagged variable as a causal parent, as shown in figure 1d.

		We note that such notations are standard in timeseries causal graphs (Runge 2019) and thus do not possess any hindrance to their application. We make use of time indexed variables for representing time lagged relations such as equation A6 (in manuscript) in the creation of our true adjacency matrix, see the bottom half of figure 2a in manuscript. <u>Actual changes in manuscript from authors:</u> None

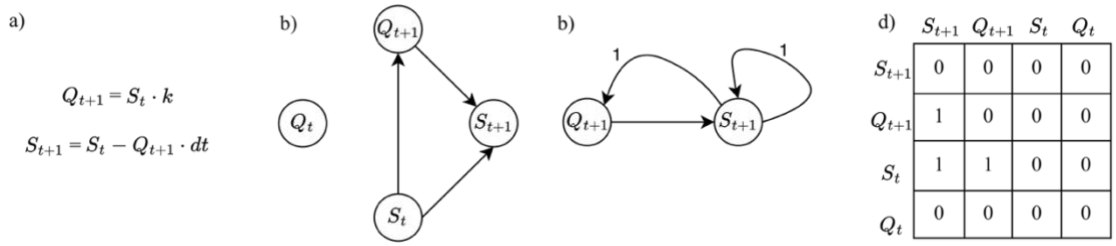

Figure 1. Representation of time indexed causal relations via DAGs and Adjacency matrix.

4.	In their study, the authors use an ideal situation where the full causal structure is perfectly known (see Fig. 2a) to evaluate CD methods. This is fine, but it will also be interesting to hydrologists what the potential of such methods is for system architecture identification. I.e. when only a few observables (usually forcing and target variables) are available, and the size and structure of the underlying system should be learned. I recommend adding a few words (and references) about this matter, e.g. in Sect. 4.4 or 5.	<u>Author response:</u> The reviewer has pointed out that hydrologists would be interested to use CD methods for understanding the size and structure of say a catchment system. By this we understand that they would typically be interested in understanding the various connections between processes with limited observed variables, say precipitation, discharge and potential evaporation. We conducted our experiment in a large system with many variables, to evaluate the efficacy of CD methods, to identify various (simulated) causal processes and structures. However, we agree in many systems such rich data are unavailable. Understanding the causal structure in such cases is an interesting problem. Thus, a future study (again in a synthetic environment) can look into this problem by restricting the available variables to only those which are commonly available (precipitation, discharge and potential evaporation). The results can then be explored to what can inferred about “missing” (restricted) variables. Also, such data scarce scenario may violate causal sufficiency, however designing the

		scope of the system is a choice of the researcher and the results should be interpreted accordingly. In this regard we suggest adding the text below. <u>Actual changes in manuscript from authors:</u> Text added in Sect 4.4 Limitations, Line 930: "However, the limited availability of complete observation data indicates that variables with longer lag need to be considered when applying CD to real-world datasets. Investigating causal structures under these conditions is an interesting topic of future research, as some apparent long-lag causal relationships may actually represent pseudo-causality induced by hidden variables and limited observability. Further studies are warranted to explore scenarios within simulated environments where data availability is artificially restricted to commonly measured variables (e.g. precipitation, discharge and potential evaporation), allowing researchers to determine what can be causally inferred about the missing states. For example, Delforge et al. (2022) utilised CD to reveal hydrological connectivity in a Karst aquifer system using time-series data for rainfall, potential evapotranspiration, resistivity, and percolation rates. By applying CD in both a synthetic case study and real-world observations, they incorporated lag relations of up to five time-steps to accurately capture the time span of preferential flow peaks. Similarly, Abbasizadeh et al., 2025, although with more comprehensive real-world data of climate and catchment attributes, used CD to identify the causal parents (drivers) of runoff signatures and report them to align with existing knowledge of the physical processes generating runoff."
5.	Out of curiosity: In causality analysis, does the concept of an inhibitor exist? I.e. a variable that would effectively mask an existing causal relationship? For example, assume z = x + y. If x=1 and y=0, z=1. Also, for x=0 and y=1, z=1. So y effectively masks the causal dependency of z and x. This is not something to be addressed in the manuscript, but I would appreciate a reply.	<u>Author response:</u> Indeed, the concept of inhibitors, where two or more variables interact to mask each other's effects, exists in Causality and is generally studied under multivariate causality and joint interactions (Runge et al., 2019, Goodwell et al., 2020). In the example provided by the reviewer, a causal analysis would yield an adjacency matrix whose coefficient for causal effects of X and Y on Z, would be negative to each other. Looking at this information, a researcher can conclude the inhibiting

		action of X on Y, when impacting Z. Similarly, the inhibiting action of Y on X when impacting Z. Note that the example provided is a case of a perfect deterministic system where two variables X and Y are negatively related and interact in a manner such that their combined effect is null on Z. Such a case of perfect determinism violates the causal faithfulness assumption and constraint-based CD methods fail in such scenarios (Runge et al., 2019). However, other methods like TCDF, VARLiNGAM and DYNOTEARS do not assume faithfulness (Table 2 in manuscript) and should yield the correct causal structure.
6.	The results are often discussed separately for the different climate zones, or compared among them (e.g. Fig. 3, or Lines 510 pp). This is not reflected in the abstract and in the formulation of the research questions at the end of section 1. I recommend doing so.	Author response: Agreed. We have added the description of a climate zone wise discussion in the introduction. See below. Actual changes in manuscript from authors: We shall add description stating the analysis and interpretation of results, climate zone-wise in the revised manuscript. Modify line 114 as: a) Can CD methods identify the true drivers in a complex simulated hydrometeorological system, across different climate types? Modify line 116 as: b) What is their overall performance, in terms of identifying causal relations and eliminating non-causal co-relations, across different climate types? Modify line 122 as: “By reviewing the causal discovery literature, we select methods better suited for hydrometeorological systems. We apply the methods in diverse climate types of a large and complex simulated environment to recover the process drivers.”
7.	Line 109: Research Question (RQ) 4 is ambiguous: At this point in the manuscript, it is unclear what it means, and later in the manuscript it is used at two places: In Sect. 3.4 and Sect. 3.5. Sect 3.4 essentially addresses RQ 2 for a subsystem, so it should be labelled otherwise. Sect 3.5 addresses RQ 4. I suggest	Author response: Thank you for suggesting title for RQ4. We have adopted the same. Actual changes in manuscript from authors:

	rephrasing RQ 4 to something like "Can CD methods help building parsimonious and robust hydrological models?"	Line 119 changed RQ4 to "Can CD methods help building parsimonious and robust hydrological models?"
8.	Sect. 2.5.1 - 2.5.4: Here the order of models differs from that in Table 1 and Sect 3.4. Please harmonize (I suggest keeping the order as in Table 1).	Author response: We agree with the reviewer and make changes accordingly. Actual changes in manuscript from authors: Line 266, In the revised manuscript, we have reordered the Sect 2.5, so it is in the same order as Table 1 and the rest of the paper. TCDF > VARLiNGAM > PCMCI+ > DYNOTEARS.
9.	Sect 2.5.7 should be a separate section 2.6, as it is topically separate from 2.5.1-2.5.6, which are all about CD methods.	Author response: We agree with the reviewer and make changes accordingly. Actual changes in manuscript from authors: Line 503, We have moved Sect 2.5.7 into a new section, Sect 2.7.
10.	In Sect. 3, results are not only reported but also discussed. I suggest renaming it to "Results and Discussion". Also, I suggest mentioning at the beginning that the main substructure in this section is by the research questions RQ1-RQ4 and also reflecting this in the subsection headers. E.g. "3.1 RQ1: Can CD methods …"	Author response: We agree to rename the subsections with their associated RQs. However, we have decided to keep the Results and the Discussions sections separate as we think that would make description of results and following conceptual and technical discussions more concise without making them overwhelming to follow through. Actual changes in manuscript from authors: We have renamed subsections in Sect. 3 to reflect the associated RQ. Line 546, 560, 578, 591.
11.	Sect. 3.4.6 is a summary statement, and would be better placed later in the manuscript	Author response: We agree with the reviewer and move it to the Discussion section. Actual changes in manuscript from authors: Text added in Line 794: “A striking feature of CD methods is their ability to correctly eliminate lagged variables as false

		positives. Whereas PCC classifies each lagged relation of the causal parent as a causal driver. Hence the CD methods are able to handle auto-correlated variables.”
12.	Sect. 4: Here the main structure differs from that in Sect. 3. I recommend merging the two, structuring them along the RQs, and moving any parts that go beyond the immediate results and discussion of the experiment to the last section, which could then be named "Summary, Conclusions and Outlook"	Author response: We agree to move the Sect. 4.2, 4.3 and 4.4 (Caveats, Perspectives and Limitations) to Sect. 5 Conclusion and renaming it as "Conclusions and Outlook". However, as mentioned above, we think keeping Results and Discussion separate would be beneficial for clarity and avoid information overload. Actual changes in manuscript from authors: Line 872, moved sections 4.2, 4.3 and 4.4 into Sections 5.1, 5.2 and 5.3.
Points related to manuscript content		
13.	I really like the experiments and analyses related to RQs A-C, but the experiment for RQ D is not convincing. Why should a random error imposed on the identified drivers help distinguishing robust models with good generalization from non-robust models with poor generalization? From the information inequality we know that "information does not hurt", i.e. adding predictors will never worsen predictions. This is always true, and it shows in the superior performance of the PCC-based model in training, but the catch is that with increasing number of predictors, the curse of dimensionality kicks in, the available sample quickly becomes non-representative, overfitting occurs, and out-of-sample performance will drop. So a convincing demonstration of "CD returns fewer drivers than PCC, therefore the training data are more representative, therefore out-of-sample prediction is better" must include sample size. I recommend doing the following: Learn the different ML models (with input as selected by PCC and the CD models) on differently sized training data (from very small to the entire period 2000-2003) and apply on 2004. The CD-based models should do much better for small training sample sizes than the PCC-based.	Author response: We thank the reviewer for these comments, which have been beneficial to consider in depth. In summary, we have undertaken additional analysis as described below in response to this point. We propose that the additional analysis supplement, rather than replace, the analysis already in the manuscript. Please read below for our detailed response. 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 tables). 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 Dr. Uwe Ehret. Also, you may see the results of the analysis as suggested by Reviewer #2 in the other response table. We partially agree to the comment and thank the reviewer for their suggestion. We emulated the analysis as suggested by them and report the following conclusions based on the results: o With shorter periods of training length, models predict less accurately, but this drop in accuracy stabilises if the training period is longer than a year (Fig 3.). o Causality based models suffer smaller drop in performance compared to PCC based models (Fig 3, 4). o Compared to PCC, Causality based models don't need training periods to be as long to achieve stability in accuracy (Fig 3, 4). We show the details of the analysis and results below:

Predictors	Same as manuscript
Noise added	**None, as suggested by reviewer**
Training period	**Varying: 75 days, 6 months, 9 months, 1 year and 4 years; starting from 01-01-2000**
Testing period	01-01-2004 to 31-12-2004
Location	Ganga Basin (Same as manuscript)
Target variable	Surface soil moisture (Same as manuscript)

Table 1: Experiment details for analysis as suggested by Reviewer #1 Uwe Ehret.

Figure 2. Training period performance (and error) metrics of experiment suggested by Uwe Ehret. Results shown for increasing periods of training length.

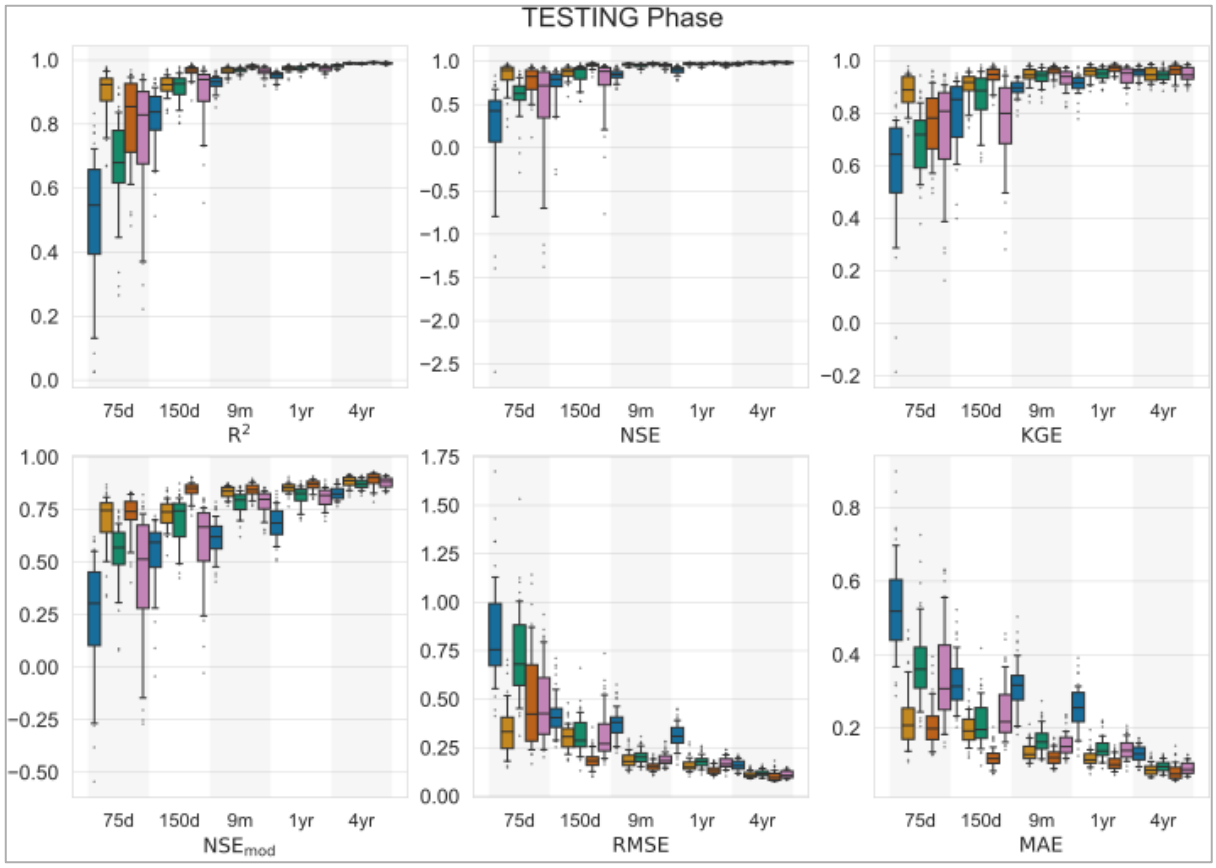

Figure 3. Same as figure 2 but for testing period.

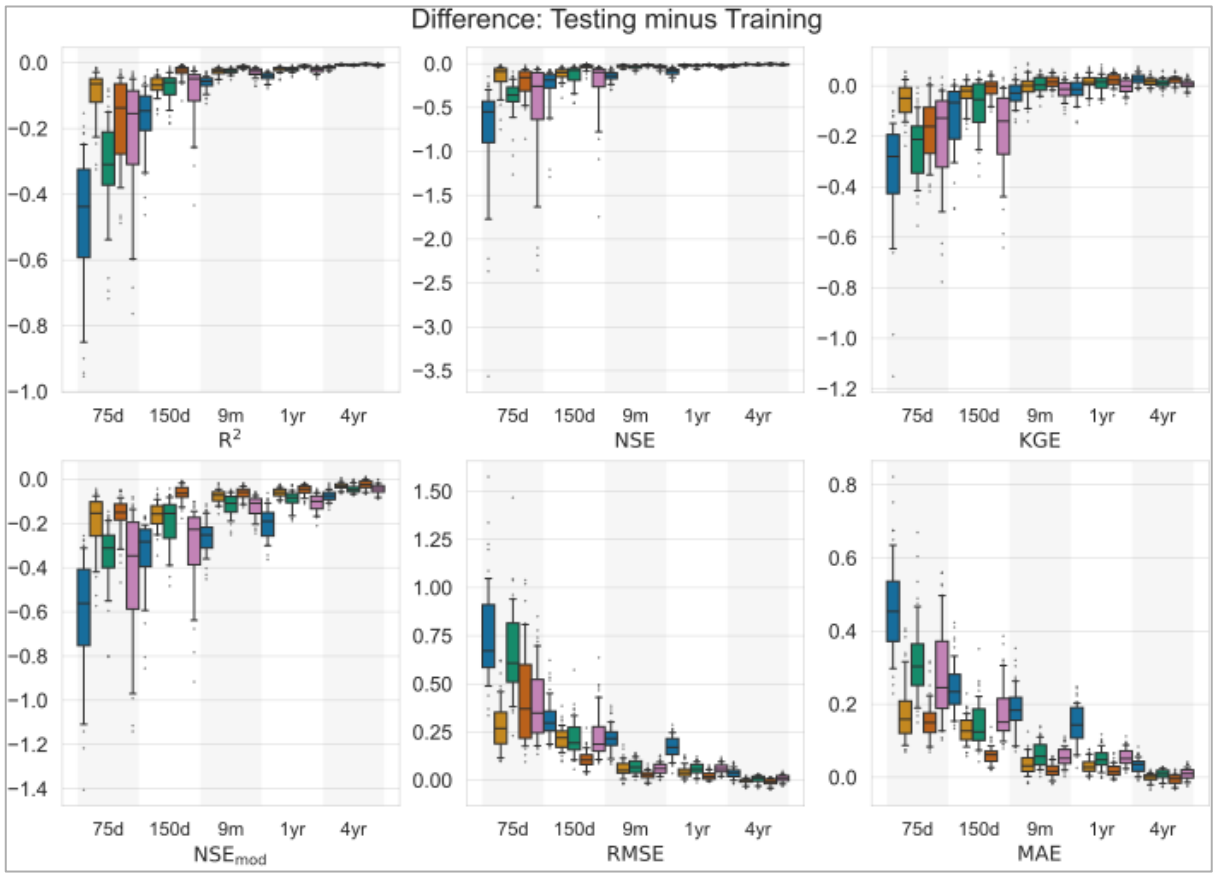

Figure 4. Difference in performance (and error) metrics between testing and training periods (figure 3 minus figure 2).

		Actual changes in manuscript from authors: We intend to keep the results from all the three sets of analyses, and we would improve the discussion on the efficacy of various methods based on them. 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. The actual text is not described here in favour of readability. Please read the revised manuscript instead. To describe the methodology, analysis and results of the two new analyses, we add text in the appropriate sections of the manuscript. Overall, we add the objective of both new analyses with new text in Line 534. Then describe the results with new text in Line 761. While the analysis details are provided in two new Appendices, Appendix D Line 1149 and Appendix E Line 1156. The associated tables are Tables A2, A3, A4 and A5 show the analysis details. While the figures are Fig A7, A8, A9, A10 and A11 show the results.

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 2: Summary of experiments for causality and PCC based time-series prediction models.

14.	Sect 2.5 A description of how PCC was used in the study is missing. Please add, comparable to the descriptions of the CD methods in Sects 2.5.1-2.5.4	Author response: We have moved the description how PCC was used from Appendix B to section 2.5. Actual changes in manuscript from authors: New section in Line 492: "2.6 Methods: Non-causal methods Pearson's correlation coefficient is a widely used method to measure the co-relation between two variables. It quantifies the strength of the co-relation as the ratio of their covariance to the product of their standard deviations. To test the statistical significance of the obtained PCC value, a hypothesis test can be performed. To do this, a null hypothesis of zero correlation, i.e. no linear dependence between the data is assumed, a significance level α is selected and the p-value associated to the PCC value is calculated. α denotes the probability of rejecting the null hypothesis when in fact it is true. The p-value is the probability of obtaining a PCC value equal to that obtained, under the assumption that the null hypothesis is true. Thus, if the p-value is less than α, the hypothesis is rejected and the obtained PCC value is considered statistically significant at significance level α. For identifying the drivers of a target variable, we found its Pearson's correlation coefficient with all the remaining variables in the system, both at contemporaneous time step and by creating their one-step-lagged time-series." Modify Line 1087 as: "For PCC, we selected only those variables as drivers where the p-value was smaller than 0.05 and absolute correlation coefficient greater than 0.2 (Wu and Chau, 2011)"
15.	In Sect. 3.5, results are reported for the ANN approach, but in the Appendix C SVR results	Author response:

	are also shown. Consider removing them if not relevant, or also discuss them.	Thank you for pointing out the omission. Since we show the main figure (Fig 7) in the results section for only a single level of noise, we wanted to show the results with increasing levels of noise for completeness and robustness (Appendix figure A2). Similarly, we repeated the experiment with an SVR model as well. Thus, the appendix figure A1 shows the surface soil moisture prediction scores, across increasing levels of noise, with an SVR model. Actual changes in manuscript from authors: We have added text describing the results obtained with SVR model in Sect. 3.5, after the main results. Text added after Line 755: “Further, we repeated the above analysis for different levels of added noise, Figure A2 shows the results. We observe that with increasing levels of noise in the data, the performance of PCC based ML models degrades significantly. While CD-based models show smaller reductions in performance. Figure A1 shows results similar to Figure A2 but with a different machine learning model, support vector regression. The figure shows similar conclusions, where with increasing levels of noise PCC based models suffer larger reductions in performance compared to CD-based models. Interestingly, at the noise level of one standard deviation, PCC based ML models perform only slightly worse than CD-based models.”
Minor points		
16.	Line 81: I assume you mean "Time series produced by hydrological systems are …"	Author response: Thank you for correction. Actual changes in manuscript from authors: Modified Line 87 as "Time series produced by hydrological systems are …"
17.	Line 176: causes in rows and effect in columns: This is opposite to what's shown in Fig. 1.	Author response: The reviewer is correct. It is a mistake from authors. We have replaced Fig 1., with the correct figure. Actual changes in manuscript from authors:

		Line 176, Figure R1, Correction of Fig. 1, we have replaced it with the figure below (Fig 5.), showing causes in rows and effects in columns and corrected the figure caption.

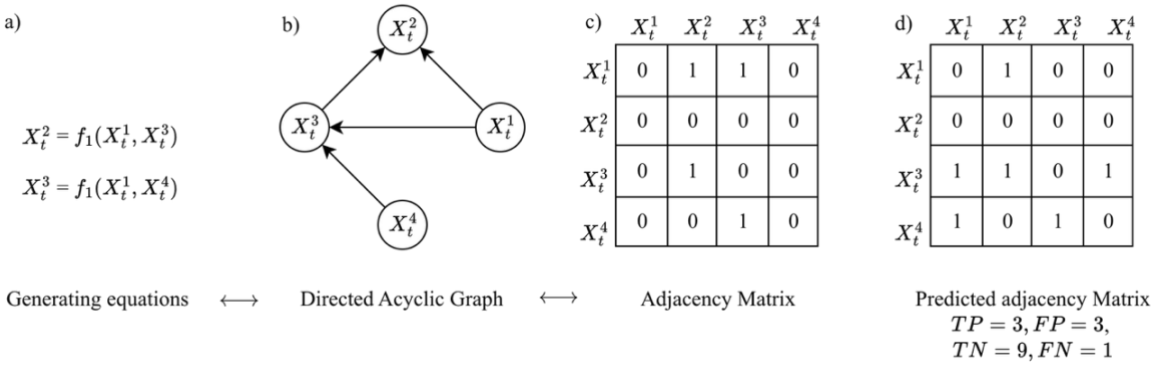

Figure 5. Figure shows the corrected representation of causal relations (a and b) in the adjacency matrix (c and d), causes in rows and effects in columns.

18.	Line 211: Slightly misleading. I suggest rephrasing to: "We surveyed various models and their outputs with the …" Line 211 pp: Was global coverage also a criterion? If yes please mention.	Author response: Accepted the rephrasing. Indeed, the global coverage was a criterion, we shall mention in the changes. Actual changes in manuscript from authors: Modified Line 229 as "We surveyed various models and their outputs with the following criteria in mind: a) all data generated by the model are available for use, b) all model forcing variables are available, c) all the time-series are available at the same resolution at which they were generated or used, and d) the model provides a global coverage of land area."
19.	Fig. 2: ○ In a) ,please make clear what are the causes and what are the effects	Author response: Agreed, please see below. Actual changes in manuscript from authors: Modified the description of Fig 2a with: “The True adjacency matrix representing the causal relationships between the simulated and forcing variables of CLSM-F2.5 model. Similar to Fig. 1c the matrix shows the relationship of causes (row

		variables) to their effects (column variables). The matrix is created based on the generating equations of the model (Appendix A2) and the definition of adjacency matrix adopted in Section 2.2"
20.	○ b) - f): Pictures and legend to not match: It is 5 regions, in the text it says six major river basins. Also, the stars in the maps, which I assume depict the grid points, are not nine per map (as stated in the legend).	Author response: Thank for pointing out the error. We replace the description with below. Actual changes in manuscript from authors: Modified the description of Fig 3 b)-f) with: "We extracted data from locations in nine Köppen-Geiger Climate Classes (eight unique classes), these locations are spread across 5 river major river basins. For each Köppen-Geiger Climate Class we selected 5 grid points, thus a total of 45 grid points were selected for analysis."
21.	Later in the manuscript, in Fig. 3, are shown for 9 Köppen-Geiger classes and 9 river basins, which does not match the 5 plots in Fig. 2. Please harmonize.	Author response: Agreed. We wanted to show the different basin names from where the data were extracted alongside their Köppen-Geiger class. In four of the five basins (Amazon, Murray, Mississippi, Danube) used, two different Köppen-Geiger classes were dominant. Thus, we selected both those classes, in the four basins, for analysis. Hence, we have 9 Köppen-Geiger class data from 5 basins as: 4x2 +1 Actual changes in manuscript from authors: We have clarified this in the caption of Fig. 2.

22.	Line 243: what's k in the equation?	Author response: *k* represents any variable in the timeseries data X. We have corrected the omission. Actual changes in manuscript from authors: Included the definition of k in Line 264 as: $X = \{x_k^t\}_{t\in(0,1,\dots,T)}$ where $k \in (1,\dots,d)$ for $\{x_k^t\} \in \mathbb{R}^d$

23.	Line 468: Why did you select the Ganga basin? Please justify	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. Actual changes in manuscript from authors: We have modified the description for the same in Line 513. "Similar to Zou et al., 2023, we use PCC and CD methods to identify the predictors of surface soil moisture. Then, we train machine learning models, based on these sets of predictors. To evaluate the performance of these ML models under contrasting conditions, we selected a location and period which underwent a significant change in climatic conditions. Thus, we choose a grid location in the Ganga basin which exhibited normal conditions between 2000 and 2003 but suffered drought during the 2004-05 period. Hence, we trained the model with CLSM data from 01 January 2000 to 31 December 2003. While we evaluate their performance during the drought period from 01 January 2004 to 31 December 2005."
24.	Fig. 5: Unclear which subplot is for which method. Please add labels	Author response: We have made bigger labels of the figure in the revised manuscript. Actual changes in manuscript from authors: Updated Fig 5. with bigger labels, moved to the top of the sub-figures.
25.	Fig. 7 ○ The metrics are not explained. E.g. what is NSEmod?	Author response: We shall add a brief description for the metrics used. NSE-modified was used as it is more robust to outliers. $$\text{NSE}_{modified} = 1 - \frac{\sum_{i=1}^{n}\lvert S_i - O_i\rvert}{\sum_{i=1}^{n}\lvert S_i - \bar{O}_i\rvert}$$ where S_i is the simulated timeseries and O_i is the observed timeseries. Actual changes in manuscript from authors: Text added in caption of Fig 7:

		"The metrics adopted are commonly used metrics in hydrology. R2-Coefficient of Determination, NSE-Nash-Sutcliffe efficiency, NSEmod-modified Nash-Sutcliffe efficiency, KGE-Kling-Gupta efficiency, RMSE-Root Mean Square Error and MSE-Mean Squared Error, (Jackson et al., 2019)".
26.	○ d): If I interpret correctly, testing performance for the CD models expressed by RMSE and MAE is better than for the training period. Is this correct? It could be because testing is in a dry year, where soil moisture is generally lower, therefore absolute errors are also lower. In any case, please add an explanation to the text. ○ The requested changes might be obsolete if the figure is completely changed (see my comment on RQ D)	Author response: Indeed, you are correct. We have added text to mention this. Actual changes in manuscript from authors: Added description for the lower error metrics seen in the testing period as a result of the lower absolute values of soil moisture in the drought years. Text added in Line 750: "We note that absolute values of soil moisture during the dry years of 2004-05 are lower compared to the values during the normal years. This resulted in smaller RMSE and MSE values for the CD-based models in the testing period."
27.	Line 729: remove "not"	Author response: Thank for pointing out the error, we have corrected it. Actual changes in manuscript from authors: Correction in Line 903, remove 'not'.
28.	Why are the snow-related variables ignored? Please explain.	Author response: We ignored the snow-related variables since the data extracted did not have snow related dynamics, thus the timeseries where zeros throughout the period. Moreover, we tried to include some climate classes from snow regions but found it difficult to analyse since the number of snow-related variables with valid data were highly varying. This would result in inconsistent number of true positives and true negatives to be identified, compared to other climate classes and even within the snow-related grid points. Actual changes in manuscript from authors:

		We have added the following description in Line 966. “We avoided snow regions in our analysis due to highly varying snow-related variables where valid data was available.”

References:

1. Runge, Jakob. "Causal network reconstruction from time series: From theoretical assumptions to practical estimation." Chaos: An Interdisciplinary Journal of Nonlinear Science 28.7 (2018).
2. Delforge, Damien, et al. "Detecting hydrological connectivity using causal inference from time series: synthetic and real karstic case studies." Hydrology and Earth System Sciences 26.8 (2022): 2181-2199.
3. Abbasizadeh, Hossein, et al. "Can causal discovery lead to a more robust prediction model for runoff signatures?." Hydrology and Earth System Sciences 29.19 (2025): 4761-4790.
4. Runge, Jakob, et al. "Detecting and quantifying causal associations in large nonlinear time series datasets." Science advances 5.11 (2019): eaau4996.
5. Goodwell, Allison E., et al. "Debates—Does information theory provide a new paradigm for Earth science? Causality, interaction, and feedback." Water Resources Research 56.2 (2020): e2019WR024940.
6. Jackson, Elise K., et al. "Introductory overview: Error metrics for hydrologic modelling–A review of common practices and an open source library to facilitate use and adoption." Environmental Modelling & Software 119 (2019): 32-48.

Response to comments by Reviewer #2 (Table 5)

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	(Our response is a unified response to comments raised by both Reviewer #2 and the Editor.) We thank the Editor and Reviewer #2 for their insightful comments regarding the inclusion of long-term memory of hydrological variables and the potential need to account for multi-step Markov processes. We agree that real-world hydrological systems often exhibit pronounced memory effects. However, we wish to clarify that the restriction of our “true” causality adjacency matrix to contemporaneous and one-day lagged relationship is not an arbitrary assumption, but a strict mathematical reflection of the physically based environment used in our analysis. We have carefully considered the comment and respectfully offer the following detailed clarification, which has also been incorporated into the revised Discussion.

	the realism of the benchmark and potentially affect the fairness of the evaluation.	As the Editor and Reviewer #2 pointed out, the true adjacency matrix constructed in this study contains causal relations from contemporaneous flux variables alongside contemporaneous and one-day lagged state variables. Our method for deriving these "true" causal drivers is based on the mathematical governing equations of the CLSM F-2.5 model (Appendix 2). Within this physically based simulated environment, the governing partial differential equations are essentially formulated as first-order Markov processes. A new state variable is computed using only the state from the immediately preceding time step (as the initial condition) combined with the current meteorological forcing. 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. If we were to expand the true causality matrix beyond one-day lag relations, the rows for the multi-step lag variables would simply be zeros and redundant. However, despite the first-order nature of the governing physics, we completely agree with the editor that many real-world hydrologic variables, such as soil moisture, groundwater storage, and snowpack, exhibit pronounced long-term memory when observed empirically. In observational datasets, we rarely capture the complete state vector of the environment. This partial observability means that the unmeasured physical delays (e.g., percolation time, routing) manifest statistically as long memory. Therefore, if a purely data-driven modelling framework is to be used for prediction based directly on observations, relying solely on lag-1 variables is often insufficient, and higher-order (multi-step) Markov processes or autoregressive models are required to account for the system's integrated memory. This distinction highlights that the suitable choice between including only lag-1 variables versus variables with longer lags depends entirely on the purpose of the causality analysis. If the analysis aims to construct a purely empirical, predictive data-driven model from observations, incorporating longer lags is practically essential to capture the

		apparent memory effects caused by unobserved intermediate states. However, if the analysis aims to identify causal drivers to inform, validate, or parameterize a physically based model (as is the basis of our CLSM F-2.5 benchmark), the inclusion of only lag-1 variables is both necessary and sufficient, as it perfectly reflects the governing differential equations. We agree with the editor that a more comprehensive analysis would test the ability of CD methods to accurately identify multi-step causal relations. However, as explained above, the simulated environment in which we conducted the analysis natively contains only contemporaneous and one-day lag relations. Thus, our benchmark is strictly limited to these transitions. Furthermore, this limitation implies that our use of a multivariate regression model to compare the efficacy of predictors—selected via PCC vs CD methods—may not sufficiently highlight the inherent merits of the true predictors determined by the causality analysis. To ensure this important context is clear to the readers, we have added a dedicated explanation of these dynamics and limitations to the revised Discussion section (Sect 5.3, Line 915) as follows: “Fifth, we applied CD methods in a simulated environment that represents physical processes with through a contemporaneous and one-day lag relations. In contrast, many real-world hydrological variables, such as soil moisture, groundwater storage, and snowpack, exhibit pronounced long-term memory when observed empirically. In observational datasets, we rarely capture the complete state vector of the environment. This partial observability means that the unmeasured physical delays (e.g., percolation time, routing) manifest statistically as long memory. Therefore, if a purely data-driven modelling framework is to be used for prediction based directly on observations, relying solely on lag-1 variables is often insufficient, and higher-order (multi-step) Markov processes or autoregressive models are required to account for the system's integrated memory. While the CD methods evaluated in the manuscript can be adapted to allow detection of multi-step causal relations in a system, derivation of true drivers in this study is based on the mathematical governing equations of the CLSM F-2.5 model

		(Appendix 2). Because this simulated environment is governed contemporaneous and one-day lag transitions, the resulting true causality matrix only contains contemporaneous and one-day lag relations. Thus, the CD methods were evaluated on their ability to accurately identify the correct causal variable and its correct lag, up-to one timestep. This constraint was an intended design to robustly contrast the key drivers selected from CD against those selected by empirical correlation-based methods, and to evaluate their predictive skills assuming that an exhaustive set of hydrological variables (both forcing and state) is accessible. However, the limited availability of complete observation data indicates that variables with longer lag need to be considered when applying CD to real-world datasets. Investigating causal structures under these conditions is an interesting topic of future research, as some apparent long-lag causal relationships may actually represent pseudo-causality induced by hidden variables and limited observability. Further studies are warranted to explore scenarios within simulated environments where data availability is artificially restricted to commonly measured variables (e.g. precipitation, discharge and potential evaporation), allowing researchers to determine what can be causally inferred about the missing states. For example, Delforge et al. (2022) utilised CD to reveal hydrological connectivity in a Karst aquifer system using time-series data for rainfall, potential evapotranspiration, resistivity, and percolation rates. By applying CD in both a synthetic case study and real-world observations, they incorporated lag relations of up to five time-steps to accurately capture the time span of preferential flow peaks."
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	<u>Author response:</u> 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 89-103, these methods are bivariate and cannot find the correct causation under confounding variables. Similarly, many variables show strong state dependence (self-

	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.	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 104-111). Towards this, we shall add text to describe the rationale of not selecting PCMCI. <u>Actual changes in manuscript from authors:</u> We have added 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	<u>Author response:</u> We partially agree with the reviewer and modify Sect 3.4 with more explicit analysis of the results. 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: • 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? - The number of analysed basins, causal drivers and number of methods together make such a

	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.	discussion very large. Thus, we avoided explicitly discussing the results for such an analysis. However, the results are explicitly shown in Figures 5 and 6, interested readers can analyse for themselves. However, we have added a general overview discussion of these results. • 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. <u>Actual changes in manuscript from authors:</u> Based on Fig. 5 and 6, we modify Sect. 3.4 to include a more explicit discussion of the following:

		• Classify the identified causal drivers according to known hydrometeorological processes, water and energy budgets. • Analyse the discovered variables into causal and non-causal drivers • Analyse the spatial (climate/basin-wise) patterns of discovered causal drivers. The revised text is very long and in favour of readability not provided here. Please read from the revised manuscript instead, Line 608-738.
4.	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)	<u>Author response:</u> We thank the reviewer for these comments, which have motivated us to consider an additional utility of CD-based methods from the perspective of the 'Paradox of Overparameterization'. Based on the comment, we have undertaken additional analysis as described below in response to this point. We propose that the additional analysis supplement, rather than replace, the analysis already in the manuscript. Please read below for our detailed response. 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 table). 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.

	explicitly accounting for the role of lag-1 soil moisture as a baseline or control predictor.	Also, you may see the results of the analysis as suggested by Dr. Uwe Ehret in the other response table. 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 (Figure 6 & 7). 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 (Table 3): We choose K=8, as it is the equal to the actual number of true causal drivers of surface soil moisture. This yielded the following set of drivers for each method (Table 4). 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 (Table 5).

Selection criteria of TOP-K predictors for different methods

Method	Selection criteria
PCC	Sort the variables selected by PCC according to their absolute 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 absolute matrix coefficient and select the top 8 variables.

PCMCIplus	Same as VARLiNGAM
DYNOTEARS	Same as VARLiNGAM

Table 3. Selection criteria to implement the TOP-K approach for various methods.

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

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 5. Details of experimental setup as suggested by Reviewer #2.

<u>Results for TOP-K approach</u>:

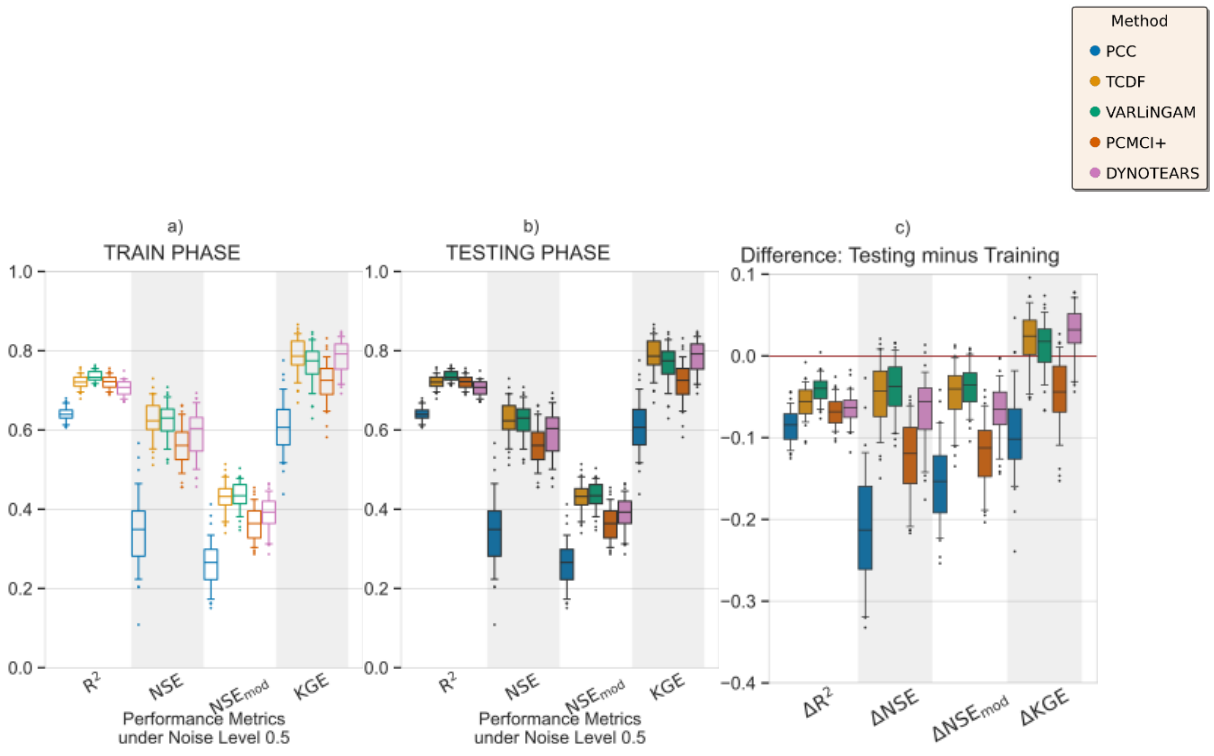

Figure 6. Performance metrics for machine learning models based on predictors in Table 4.

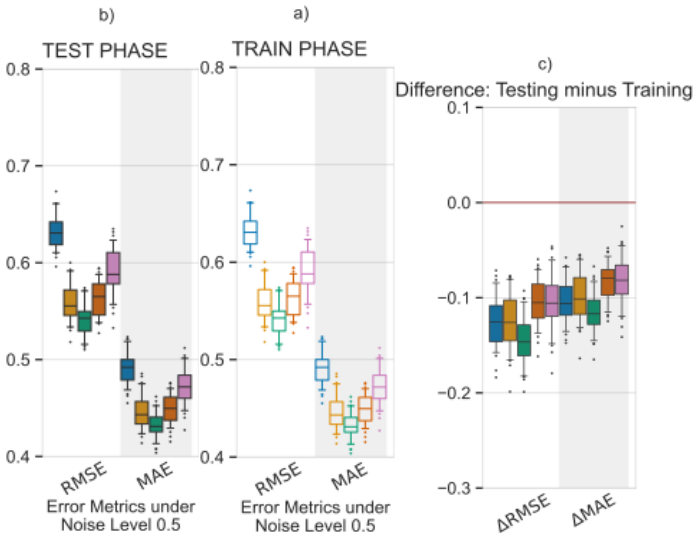

Figure 7. Error metrics for machine learning models based on predictors in Table 4.

		<u>Actual changes in manuscript from authors:</u> We intend to keep the results from all the three sets of analyses, and we would improve the discussion on the efficacy of various methods based on them. After considering the suggestion of both the reviewers, we now have three sets of timeseries prediction results, each having different methodologies. The table (Table 6) below summarises them. The actual text is not described here in favour of readability. Please read the revised manuscript instead. To describe the methodology, analysis and results of the two new analyses, we add text in the appropriate sections of the manuscript. Overall, we add the objective of both new analyses with new text in Line 534. Then describe the results with new text in Line 761. While the analysis details are provided in two new Appendices, Appendix D Line 1149 and Appendix E Line 1156. The associated tables are Tables A2, A3, A4 and A5 show the analysis details. While the figures are Fig A7, A8, A9, A10 and A11 show the results.

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 6: Summary of experiments for causality and PCC based time-series prediction models.

Specific comments		
5.	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.	<u>Author response:</u> 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 214), 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 560. <u>Actual changes in manuscript from authors:</u> We have added the following text in Line 217. "…classes of the confusion matrix. Such metrics can show a biased picture in cases 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 any 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."
6.	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. Actual changes in manuscript from authors: Modify the text in Line 415 to remove ensuring its acyclicity with "penalizing its cyclicity".
7.	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. Actual changes in manuscript from authors: Line 432 moved the loss equation to appear before the acyclicity equation, now Eq. 17 and 18 respectively.
8.	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 period where the climatic conditions underwent significant change. The Ganga basin faced drought during 2004-05 period; thus, we selected it. Actual changes in manuscript from authors: We have added a description for the same in Line 511. Text modified in Line 513: "Similar to Zou et al., 2023, we use PCC and CD methods to identify the predictors of surface soil moisture. Then, we train machine learning models, based on these sets of predictors. To evaluate the performance of these ML models, we selected a location and period which went under a significant change in climatic conditions. Thus, we choose a grid location in the Ganga basin which suffered drought during the 2004-05 period."

9.	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 and revise our text accordingly. 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 an error-free 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. Actual changes in manuscript from authors: We have revised the text in Line 526 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.
2. Zou, Liangfeng, et al. "Coupling the causal inference and informer networks for short-term forecasting in irrigation water usage." Water Resources Management 37.1 (2023): 427-449.