Response to comments from reviewer #1 Uwe Ehret

We thank Dr. Uwe Ehret for their detailed and precise comments. Their suggestions in various sections have helped bring more clarity and good structure to the manuscript. Also, their suggestion of a modified analysis of the timeseries prediction has made the results more robust and increased our confidence in causal methods. Below we provide our response to their comments and some initial suggested changes in the manuscript.

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.

Major points

Points related to the definition of causality

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.

   Suggested changes in manuscript from authors: Paragraph to be added on Line 52:

   "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 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 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.

   Author response:

We agree that using multivariate equations for extracting causal interactions can induce different DAG structures due to the ambiguity introduced by 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.

Suggested changes in manuscript from authors: Towards the above comment we suggest adding the following text in Line 228.

"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.

    Author response:

    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 et al., 2018) 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.

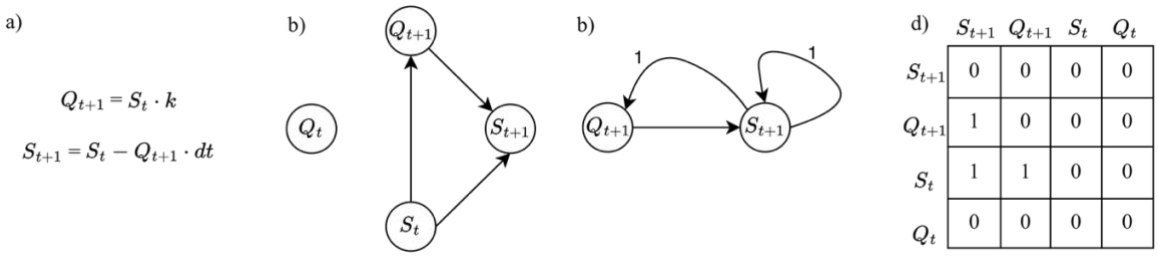

Figure 1. Representation of time indexed causal relations via DAGs and Adjacency matrix.

Suggested changes in manuscript from authors: None

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.

   Author response: 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 be 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.

   Suggested changes in manuscript from authors: Paragraph to be added in Discussions section:

   We evaluated the efficacy of CD methods in a simulated environment, with many variables. However, in many real world systems, such rich data are unavailable. Understanding the causal structure in such cases is an interesting problem. Thus, a future study can explore a scenario, where in a simulated environment, the available data are restricted to only those commonly available (e.g. precipitation, discharge and potential evaporation). The results can then be explored to see what can be inferred about missing (restricted) variables. In a recent study Abbasizadeh et al., 2025, used CD for finding the causal parent (drivers) of runoff signatures from

catchment and climate characteristics 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.

   Author response: 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.

Points related to the manuscript structure

1. 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 shall add this description of a climate zone wise discussion in the introduction.

   Suggested changes in manuscript from authors: We shall add description stating the analysis and interpretation of results, climate zone-wise in the revised manuscript.

2. 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 rephrasing RQ 4 to something like "Can CD methods help building parsimonious and robust hydrological models?"

   Author response: Thank you for suggesting title for RQ4. We shall adopt the same.

   Suggested changes in manuscript from authors: Change RQ4 to "Can CD methods help building parsimonious and robust hydrological models?"

3. 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 shall make changes accordingly.

Suggested changes in manuscript from authors: In the revised manuscript, we shall move the Sect 2.5.4 (Granger causality based: Temporal Causal Discovery Framework) to top of Sect 2.5, so it is in the same order as Table 1 and the rest of the paper.

4. 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 shall make changes accordingly.

   Suggested changes in manuscript from authors: We move Sect 2.5.7 into a new section, Sect 2.6.

5. 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 keep the Results and Discussions section separate as we think being a newer topic, keeping the results and discussion is necessary to avoid overwhelming information at once.

   Suggested changes in manuscript from authors: Rename subsections in Sect. 3 to reflect the associated RQ.

6. 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.

   Suggested changes in manuscript from authors: Make changes accordingly.

7. 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 would be beneficial for clarity and avoid information overload.

   Suggested changes in manuscript from authors: Make changes accordingly.

Points related to manuscript content

1. 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 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 Dr. Uwe Ehret.

Also, you may see the results of the analysis as suggested by Reviewer #2 in the other response document.

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:

- 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.
- Causality based models suffer smaller drop in performance compared to PCC based models.
- Compared to PCC, Causality based models don't need training periods to be as long to achieve stability in accuracy.

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 Uwe Ehret.

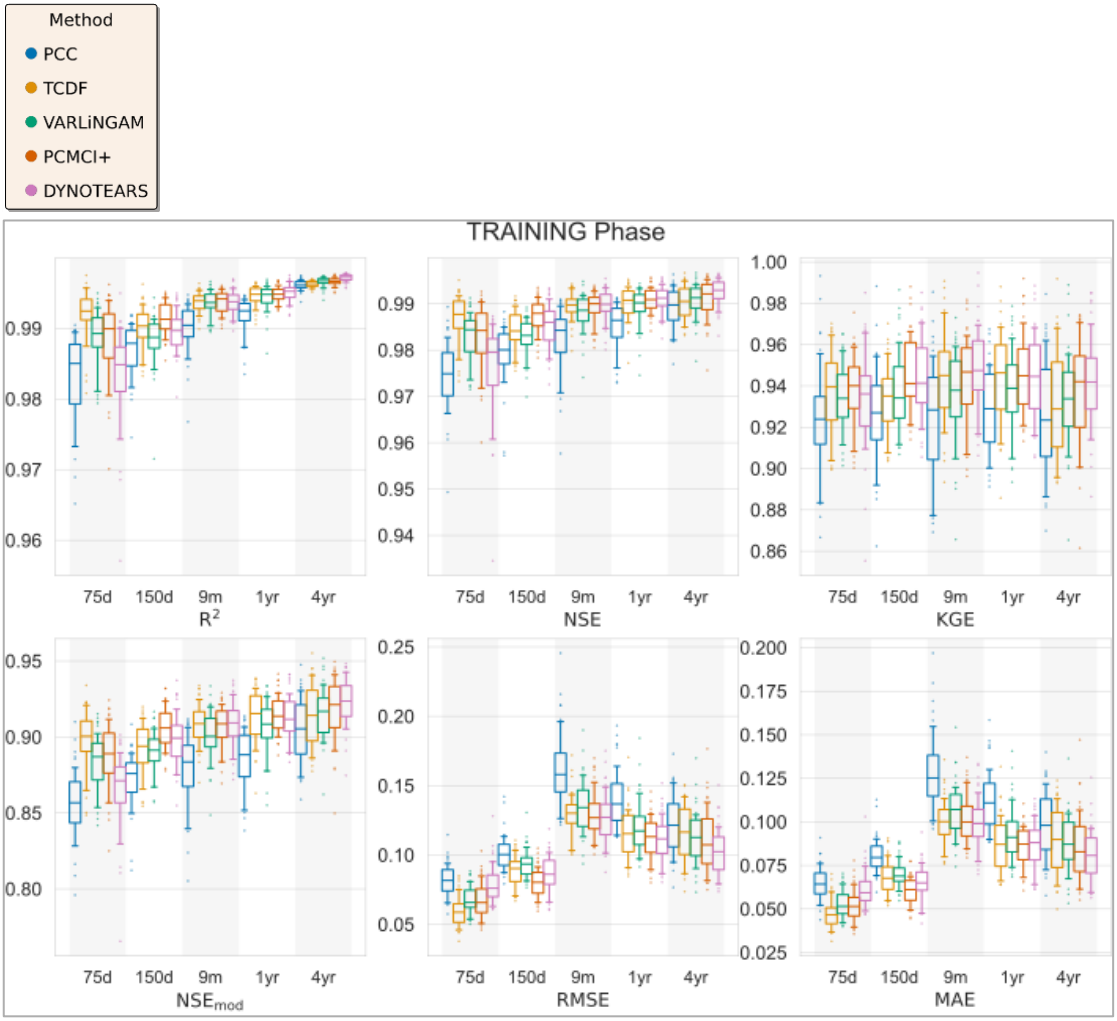

Figure 2a. Training period performance (and error) metrics of experiment suggested by Uwe Ehret. Results shown for increasing periods of training length.

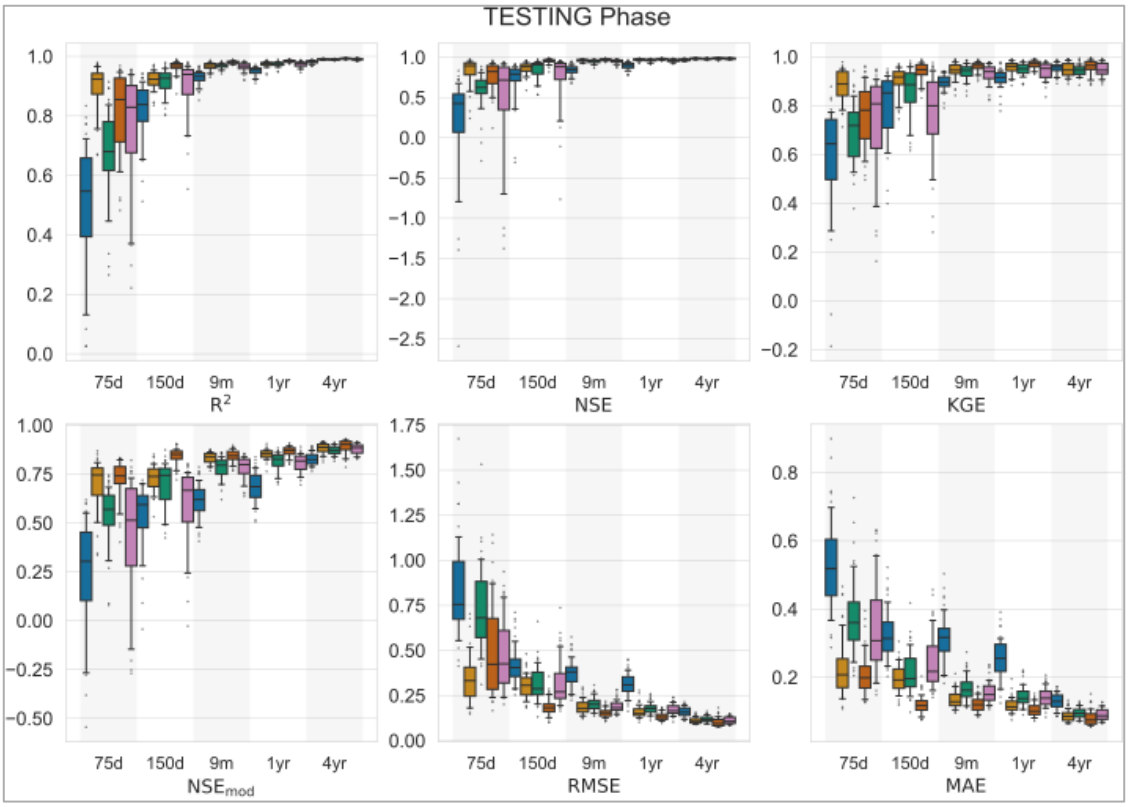

Figure 2b. Same as figure 2a but for testing period.

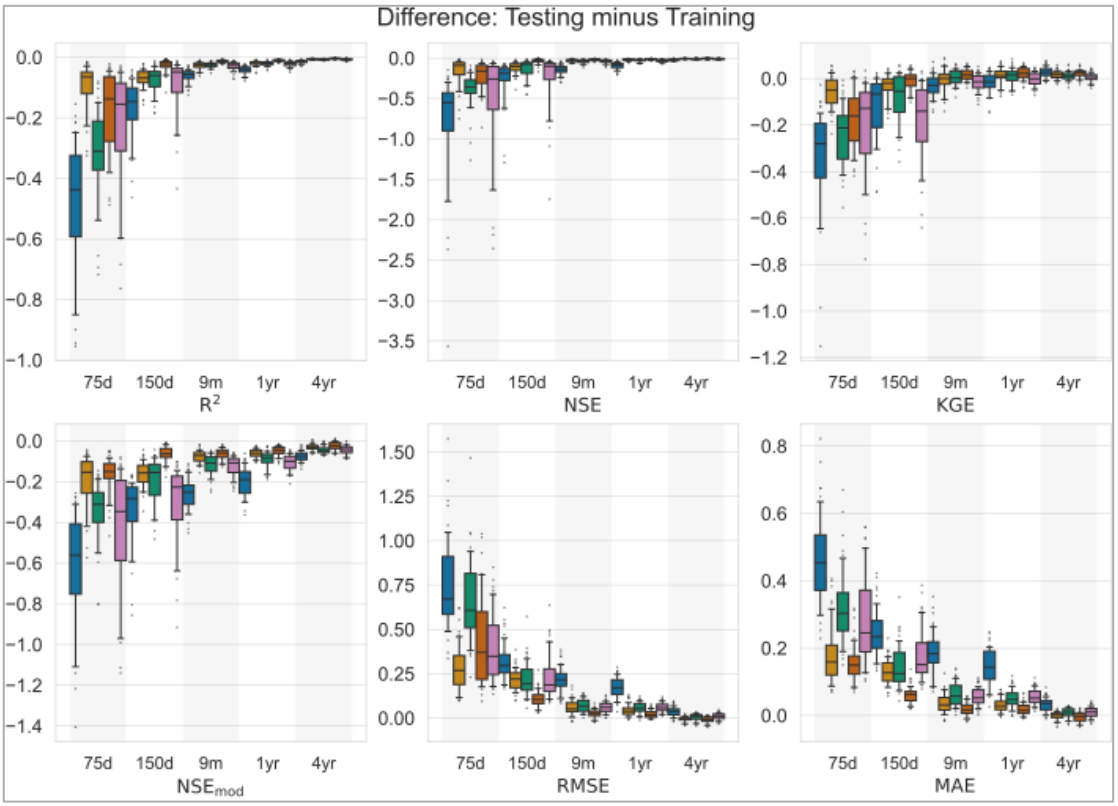

Figure 2c. Difference in performance (and error) metrics between testing and training periods (figure 2b minus figure 2a).

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 2: 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.

2. 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 shall move the description how PCC was used from Appendix B to section 2.5.

   Suggested changes in manuscript from authors: Move PCC description from Appendix B to Sect 2.5

3. In Sect. 3.5, results are reported for the ANN approach, but in the Appendix C SVR results are also shown. Consider removing them if not relevant, or also discuss them.

   Author response: Thank you for pointing out the omission. Since we show the main figure (Fig 6) in the results section for only a single level of noise, we wanted to show the results at 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.

Suggested changes in manuscript from authors: We shall mention the results obtained with SVR model in Sect. 3.5, after the main results in Line 619.

Minor points

1. Line 81: I assume you mean "Time series produced by hydrological systems are …"

   Author response: Thank you for pointing out the incorrect wording.

   Suggested changes in manuscript from authors: Modify Line 81 as "Time series produced by hydrological systems are …"

2. Line 176: causes in rows and effect in columns: This is opposite to what's shown in Fig. 1.

   Author response: Indeed, we shall correct the same.

   Suggested changes in manuscript from authors: Correction of Fig. 1, we shall replace it with the figure below, showing causes in rows and effects in columns.

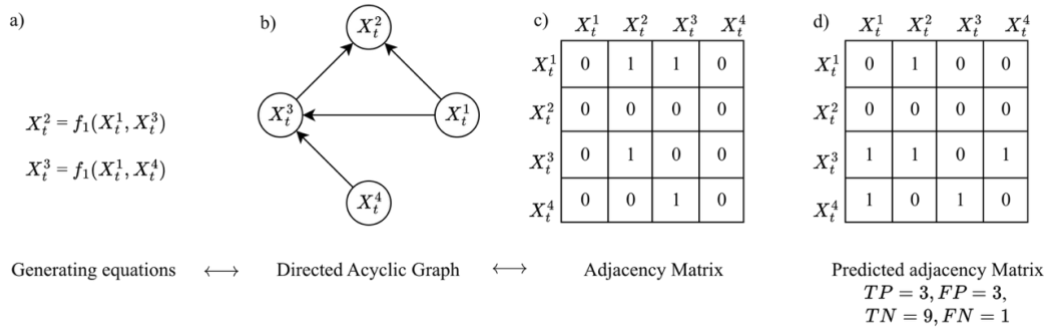

Figure 3. Figure shows the corrected representation of causal relations (a and b) in the adjacency matrix (c and d), causes in rows and effects in columns.

3. Line 211: Slightly misleading. I suggest rephrasing to: "We surveyed various models and their outputs with the …"
4. 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.

   Suggested changes in manuscript from authors: Modify Line 211 as "We surveyed various global models and their outputs with the …"

5. Fig. 2:
   - In a) ,please make clear what are the causes and what are the effects

     Author response: Agreed, please see below.

     Suggested changes in manuscript from authors: Modify 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"

- 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.

  Suggested changes in manuscript from authors: Replace description of Fig 2b-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 location we selected 5 grid points, thus a total of 45 grid points were selected for analysis."

- 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

  Suggested changes in manuscript from authors: We shall clarify this in the caption of Fig. 2.

6. Line 243: what's k in the equation?

   Author response: *k* represents any variable in the timeseries data **X**. We shall correct the omission.

   Suggested changes in manuscript from authors: Include the definition of k in Line 243 as:

$$X = \{x_k^t\}_{t\in(0,1,\dots,T)} \quad \text{for} \quad k \in (1,\dots,d) \quad \text{for} \quad \{x_k^t\} \in \mathbb{R}^d$$

7. 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.

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

8. Fig. 5: Unclear which subplot is for which method. Please add labels

   Author response: We shall make bigger labels of the figure in the revised manuscript.

   Suggested changes in manuscript from authors: Make bigger labels for Fig 5.

9. 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.

$$\mathrm{NSE}_{modified} = 1 - \frac{\sum_{i=1}^{n}|S_i - O_i|}{\sum_{i=1}^{n}|S_i - \bar{O}_i|}$$

where S_i is the simulated timeseries and O_i is the observed timeseries.

Suggested changes in manuscript from authors: We shall add description of the metrics used in Fig. 7.

- 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 shall mention this in the text.

  Suggested changes in manuscript from authors: Add 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.

10. Line 729: remove "not"

    Author response: Thank for pointing out the error, we shall correct.

    Suggested changes in manuscript from authors: Correction in Line 729, remove 'not'.

11. 788: 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.

    Suggested changes in manuscript from authors: We shall add the following description in Line 788.

    "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. 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.

3. Runge, Jakob, et al. "Detecting and quantifying causal associations in large nonlinear time series datasets." Science advances 5.11 (2019): eaau4996.
4. 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.