Reply to Reviewer 1

Thank you very much for taking the time to review our manuscript and for providing us with constructive feedback. We have carefully addressed all the points raised and will incorporate the necessary changes into the revised manuscript. Please find below the main points to be corrected:

1. **The title, the abstract, and the conclusions are not fully aligned. In the title authors mention to "prediction", in the abstract "interpretation", in the conclusion there are many points mixing the two topics.**

   We agree that the concepts in these sections are somewhat mixed and could be clarified. To address this issue, we have revised the abstract and conclusion to better align with the title, which emphasizes "prediction." We specifically highlighted how the study's results can contribute to improving prediction models.

   Here are the following changes we will make:

   Lines 13-14: *This study demonstrates the potential of causal inference techniques for predicting catchment responses by effectively representing the interconnected processes within hydrological systems in a more interpretable manner.* ~~*This study demonstrates the potential of causal inference techniques in representing the interconnected processes in hydrological systems in a more interpretable and effective manner.*~~

   From line 479:

   Three prediction models, BN, GAM, and RF, in five different settings, namely BN, GAM$\sim$Par, GAM$\sim$All, RF$\sim$Par and RF$\sim$All, are used to predict runoff signatures. These models are executed on the entire dataset as well as 22 clusters, with each configuration undergoing 100 random samplings of training and test sets, resulting in a total of 11,500 model executions.

   From line 485:

   - *The causal parents of the signatures identified by the PC algorithm do not always align with the most influential variables determined by correlation and variable importance analysis. This suggests that strong correlations may result from confounding variables, and causal relationships do not always coincide with high variable importance. This point can impact the robustness of prediction models, especially when the same set of predictor variables is used across diverse environments with varying characteristics.*
   - *BN shows the smallest decrease in accuracy between the training and test samples, demonstrating high transferability. The accuracy of the models is not sensitive to the training sample size and shift in the distribution of predictors. This indicates that $P(Effect \mid Cause)$ remains consistent across environments. Although BN's overall accuracy is lower than that of the nonlinear GAM and RF models, it outperforms RF in predicting mean daily runoff and high flows across different environments (clusters).*
   - *Using causal parents helps mitigate the overfitting problem and improve the robustness in prediction models, particularly in GAM, when the size of the training set is small.* ~~*Using causal parents helps reduce overfitting, particularly for GAM, when the training sample size is small.*~~
   - *The high accuracy of non-causal models, GAM$\sim$All and RF$\sim$All, in the baseline scenarios may be attributed to spurious relationships. This is supported by their reduced accuracy in environments with smaller training sets, highlighting a lack of robustness compared to causal models, which maintain higher reliability under such conditions.* ~~*The high accuracy of non-causal GAM$\sim$All and RF$\sim$All in the baseline models may be due to spurious relationships, as their accuracy decreases in environments with smaller training sets compared to the causal models.*~~
   - *In environments where the target signature is more difficult to predict, such as clusters of the geology category, using causal parents increases prediction accuracy.*

- *Independent variables identified through causal discovery can determine groups of catchments where prediction models exhibit consistent performance. For instance, topographic variables are among the independent variables in this context since all models perform consistently well in clusters 1, 2, and 3, and less effectively in cluster 4. This information helps identify environments where training models achieve higher accuracy, reduced uncertainty, and greater robustness.* ~~*The independent variables identified through causal discovery using DAGs can serve as reliable criteria for catchment classification. This is evident from the models performing consistently well in clusters 1, 2, and 3, while performing less effectively in cluster 4. This information improves model accuracy, reduces prediction uncertainty, and enhances consistency between training and test simulations.*~~
- *Causal inference methods contribute to improving prediction models'* ~~*model*~~ *parsimony, interoperability and robustness in hydrological systems* ~~*modelling*~~*.*

2. **Runoff signature is synonymous of "hydrological response" or "watershed response"? Maybe in the introduction this other common term could be mentioned just to better orient the reader.**

   Runoff signatures represent distinct characteristics of a catchment's response. While "catchment response" and "runoff signature" are sometimes used interchangeably, this could create confusion for readers. In the revised sections, we will carefully use "catchment response" to ensure clarity and avoid any ambiguity.

3. **In the lines 103-113 it should be clarified which is the innovative contribution or the advancement compared to the previous literature accurately listed by the authors**

   We appreciate this reviewer's remark. We have revised the section to clearly highlight the novelty of our work and provide more context. Specifically, we emphasized that the use of causal discovery to identify causal links between catchment attributes, climatic indices, and runoff signatures and integrating these findings into prediction models represents the key innovation of this research.

   To our current knowledge, the proposed study is the first analysis that connects the causal models and catchment attributes. Therefore, we will change this section to:

   This study introduces a novel approach for predicting runoff signatures by integrating causal information into predictive models. To the best of our knowledge, causal inference techniques have not yet been applied for this purpose. Unlike previous studies that primarily rely on correlated-based features for predicting a specific catchment response, we take a step beyond mere correlation by focusing on causally relevant variables, specifically, causal parents. By integrating causal information into predictive models (GAM and RF), we aim to investigate whether it can enhance the prediction models' robustness, interpretability, and parsimony compared to models that do not utilize causal insights. ~~*This study aims to represent the causal relationships between catchment attributes, climate characteristics and runoff signatures. We compare the performance of prediction models (GAM and RF) that incorporate causal information derived from causal discovery methods against models that do not.*~~ *We assume that a specific characteristic of catchment response is directly influenced by a subset of correlated variables, known as causal parents, rather than by all correlated variables.* ~~*runoff signatures are causally influenced by a subset of variables, known as causal parents, rather than by all available variables. We adopt the Peter and Clark (PC) causal discovery method (Spirtes et al., 2001), which is a constrained-based causal discovery algorithm, to identify these causal relationships and to structure the BNs. Our objective is to investigate whether incorporating causal information can provide new insight into hydrological systems modelling, enhance the prediction models' robustness, and improve their parsimony.*~~ *To achieve our objectives, we follow these steps: 1) select potential predictors for each runoff signature among the catchment and climate attributes, 2) identify causal relationships between catchment attributes, climate characteristics, and runoff signatures (network structure) using Peter and Clark (PC) causal discovery method (Spirtes et al., 2001),* ~~*2) identify causal parents and network structure for each signature,*~~ *3) execute models using both the causal parents (causal models) and all selected variables (non-causal models) for entire catchments and subset of catchments, 4) evaluate the robustness of the causal and non-causal models.*

4. **Section 3. Data are crucial for understanding the model application. In the Section 3 there is the attribute list but not the data characterization. A first question that**

could have the reader is "Did they authors select one number for each attribute and for each catchment?" or a time series?

Thank you for raising this point. We will improve Section 3 and include additional information about data characterization to enhance clarity. We will also mention that we do not use the time series data in our analysis. Each catchment in the dataset has five categories of attributes that are outlined in Table 1. We will use this information to categorize catchments (using cluster analysis) based on each category. Therefore, each catchment has been assigned five cluster IDs.

The changes will be as follows:

From line 219:

The clustering classifies the catchments according to the five categories. Time series data is not used for clustering analysis, and only catchment attributes available in the CAMELS dataset, as listed in Table 1, are utilized for this purpose. Table 3 shows the methods used for clustering, the optimum number of clusters according to the elbow and Silhouette scores, and the number of catchments in each cluster.

From line 220:

(a) ***Climate attributes:*** *Climate attributes in the CAMELS dataset are derived from area-weighted averaging of meteorological forcing time series from October 1, 1989, to September 30, 2009. The cluster analysis shows four distinct climate categories, which spread in the east (cluster 1), the Midwest (cluster 2), the west (cluster 3) and the northwest (cluster 4) (Fig. 2a). The largest group of catchments belong to cluster number one, with 334 in the north- and southeast of the US (Table 3). This cluster receives an average of 3.5 mm daily precipitation and has 2.8 mm daily evapotranspiration. Other clusters have the following average precipitation and evapotranspiration levels: Cluster 2 has 2.3 mm of precipitation and 2.7 mm of evapotranspiration, Cluster 3 has 5.5 mm of precipitation and 2.4 mm of evapotranspiration, and Cluster 4 has 2.0 mm of precipitation and 3.3 mm of evapotranspiration.*

(b) ***Soil attributes:*** *The soil properties data, derived from the State Soil Geographic Database (STATSGO), provides information about the top 2.5 meters of soil. Soil texture is represented in 16 classes, of which there are 12 classes based on the United States Department of Agriculture (USDA) and 4 non-soil classes. The saturated hydraulic conductivity and soil porosity are calculated based on the sand and clay fraction using multiple regression analysis. Cluster analysis identifies six groups of catchments. ~~This category is divided into 6 groups.~~ There is no distinctive spatial pattern among soil clusters. However, clusters 2 and 3 are mostly spread across the east and west coastlines (Fig. 2b). The maximum water content and porosity values are influenced by soil texture, which defines the proportion of sand, clay, silt, and other materials. For example, cluster 6 shows the highest soil porosity and maximum water content (Fig. 2b). This cluster has the highest percentage of clay (26%) and silt (47%) fractions among all clusters.*

(c) ***Topographic attributes:*** *The topographic information of catchments, namely catchments' contours, are determined using geospatial fabric (Viger and Bock, 2014) and Geospatial Attributes of Gages for Evaluating Streamflow (GAGES II) methods (Falcone, 2011). These methods are used to determine the area, and the Digital Elevation Model (DEM) is clipped for each catchment. This category is divided into 4 distinctive clusters ~~groups~~. Cluster 1 contains catchments located in the northeast, which are catchments with low elevation and slope (Fig. 2c). Cluster 2 consists of catchments along the west coast spread from the west to the northwest. The catchments with the lowest elevation and slope are in cluster 3, located in the southeast. Cluster 4 contains the highest elevation catchments in the Rocky Mountains (Fig. 2c).*

(d) ***Geological attributes:*** *The geological variables in the CAMELS datasets are derived from the Global Lithological Map (GLiM) (Hartmann and Moosdorf, 2012) and the Global HYdrogeology MaPS (GLHYMAPS) (Gleeson et al., 2014). From the GLiM dataset, sixteen lithological classes are identified, and their proportional areas are calculated for each catchment. The GLHYMAPS dataset is used to estimate subsurface permeability and porosity (Addor et al., 2017). This category is divided into 7 groups. Unlike the climate and topography categories, this category does not show a distinguishable spatial pattern (Fig. 2d). However, the catchments with the highest geological porosity are mainly concentrated in the southeast, and those with the lowest are located in the west (Fig. 2d).*

(e) ***Vegetation attributes:*** *Vegetation is represented using two indicators, vertical density, measured by the Leaf Area Index (LAI), and horizontal density, measured by the Green Vegetation Fraction (GVF). These measurements are derived from a 1-km resolution product of the Moderate Resolution Imaging Spectroradiometer (MODIS). The vegetation or land cover category is divided into 6 different groups (Fig. 2e). The spatial pattern of the vegetation is influenced by climate and topographic categories. According to Fig. 2e, the catchments with the highest forest fractions have the highest maximum leaf area index and are located in the northeast and east of the study area. This area has high precipitation and low evapotranspiration (Fig. 2a). The lowest vegetation cover belongs to the central and southern parts of the US, which are in clusters 4 and 6.*

5. **Figure 1 is not fully clear, Is the cluster analysis necessary? Is it an alternative way to analyze the entire data set? If yes it should be in a different level, like a starting option in the flow chart.**

   Regarding Figure 1, we prefer to use Figure 1 to highlight the core of the analysis, which presents the description of the lists of steps of the proposed comparative analysis on causal models and catchment runoff signatures, rather than a direct flowchart. In this figure, cluster analysis uses the entire dataset and assigns cluster IDs to each catchment. We will modify this figure to make it more understandable as follows:

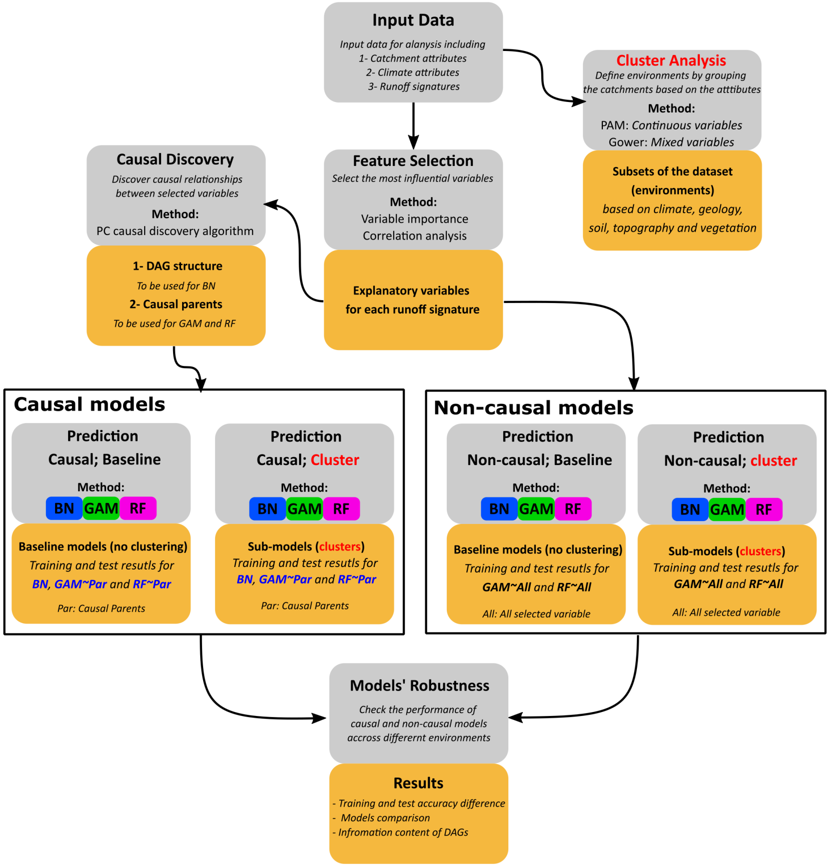

Figure 1: *Flowchart depicting the steps followed in this study. Grey boxes indicate the procedures, orange boxes present the results of these procedures,~~and~~ blue text highlights where information about causality is utilized , and the red text highlights the cluster analysis and indicates where the clustering results are applied.. PC refers to Peter and Clark's causal discovery algorithm, PAM stands for Partition Around the Centroid clustering algorithm, and DAG refers to Directed Acyclic Graph. BN refers to the Bayesian Network, GAM refers to the Generalized Additive Model, and RF refers to the Random Forest. GAMPar and RFPar are causal models (GAM and RF) using only causal parent variables for prediction, while GAMAll and RFAll are non-causal models that use all selected variables as predictors. Baseline models refer to models that use the entire dataset (all 671 catchments) for training and testing, while sub-models use only subsets of the dataset or clusters.*

Regarding the necessity of clustering, When using the entire dataset for training and testing the models, we work with a large, diverse dataset containing a variety of characteristics. However, the question arises: what happens when we focus on regions with fewer catchments that share similar attributes, such as climate, geology, vegetation, topography, or soil properties? To investigate this, we applied clustering analysis to group catchments with similar characteristics, creating categories of varying sizes.

This approach enables us to evaluate the performance of causal and non-causal models across different environments, identifying the situations where the models perform well and where they face challenges. For instance, if the models perform well in a cluster within the topographic category, we can infer that models are effective in regions with specific topographic characteristics and less sensitive to the sample size. Additionally, using causal parents as predictors establishes a causal mechanism for the runoff signature under the assumption that this mechanism remains consistent across different environments. Clustering helps test this assumption and assess the effectiveness of the PC algorithm in identifying the appropriate causal parents using the models' performance. In conclusion, clustering is a complementary rather than an alternative analysis to investigate the causal models further.

6. **The Section 2.2.1 seems incomplete and refers to the Supplementary materials, how-**

ever this step seems important in the whole procedure. More details on how the most important feature are ranked are necessary, indeed the "out-of-bag method" is vague and the sentence "variables are selected based on a combination of correlation analysis, variable importance assessment and consideration of the underlying physics of the runoff signatures." is too general.

We agree that the information provided in this section is not sufficient. Therefore, we have added more details to the feature selection section. We will explain that the out-of-bag method is used as a complementary method to the correlation analysis, which helps identify explanatory variables in categories with low correlation coefficients.

The changes will be as follows:

From line 145:

(a) *Correlation analysis: Pearson, Kendall, and Spearman correlation coefficients are computed to illustrate the potential explanatory variables. The correlation analysis reveals the most influential variables from each category, namely climate, geology, vegetation, topography and soil. In addition, the scatter plot of the data helped visually understand the relationship between variables.*

(b) *Variable importance: Since the results of the correlation analysis are not always consistent, another feature selection procedure is conducted using the random forest method to investigate the feature importance. ~~The same approach as correlation is repeated, using the random forest method to investigate the feature importance. Random forest is implemented using the R package randomForest (Liaw et al., 2015).~~ The variables are ranked using ~~according to~~ the out-of-bag method, which is quantified using the Mean Decreased Accuracy (IncMSE) score. The out-of-bag method ranks variables based on the increase in prediction error caused by removing each variable from the prediction process. Random forest is implemented using the R package randomForest (Liaw et al., 2015).*

With the information provided by the procedures mentioned above, variables are selected based on a combination of correlation analysis, variable importance assessment and consideration of the underlying physics of the runoff signatures. We tried to select the most influential variables from each category, including climate, geology, soil, topography, and vegetation. The number of selected variables varies across categories. Multiple variables are selected from categories where most variables exhibit high correlation. Conversely, only the highest-correlated variable is chosen for categories with a weak correlation to the runoff signature of interest. For example, climatic variables are often highly influential for runoff signatures, leading to the selection of multiple variables. In contrast, geological variables tend to show a weak correlation with some runoff signatures, so only the most influential variable from this category is selected. The results of feature selection are presented in the supplementary materials.

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, HYDROLOGY AND EARTH SYSTEM SCIENCES, 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.

Falcone, J. A.: GAGES-II: Geospatial attributes of gages for evaluating streamflow, Tech. rep., US Geological Survey, 2011.

Gleeson, T., Moosdorf, N., Hartmann, J., and van Beek, L. P. H.: A glimpse beneath earth's surface: GLobal HYdrogeology MaPS (GLHYMPS) of permeability and porosity, GEOPHYSICAL RESEARCH LETTERS, 41, 3891–3898, https://doi.org/10.1002/2014GL059856, 2014.

Hartmann, J. and Moosdorf, N.: The new global lithological map database GLiM: A representation of rock properties at the Earth surface, GEOCHEMISTRY GEOPHYSICS GEOSYSTEMS, 13, https://doi.org/10.1029/2012GC004370, 2012.

Liaw, A., Wiener, M., Breiman, L., and Cutler, A.: Package 'randomforest', 2015.

Spirtes, P., Glymour, C., and Scheines, R.: Causation, prediction, and search, MIT press, 2001.

Viger, R. and Bock, A.: GIS features of the geospatial fabric for national hydrologic modeling, US Geological Survey, 10, F7542KMD, 2014.

Reply to Reviewer 2

Thank you very much for taking the time to review our manuscript and for your deep understanding of this topic by providing constructive feedback. We have responded to all of your points and will incorporate the necessary changes into the revised manuscript.

1 Response to the reviewer's comments

Please find our responses to your points below:

1. **The authors write that they assume that the runoff signatures are sink nodes ("We also assumed that runoff signatures do not cause climate and catchment attributes") and that there are no hidden variables ("It is also assumed that there are no unobserved variables" – see also the assumption that the distribution is Markov and faithful wrt a DAG over the observed variables). But then, identifying the set of causal parents of a runoff signature becomes 'classical' variable selection, a non-causal problem. This implies that a causal analysis is not necessary.**

   Thank you for pointing out this important aspect. Although identifying the causal parents of a runoff signature might seem similar to a variable selection process, our analysis emphasizes investigating the causal mechanisms that drive the runoff signatures. A variable Y and its causal parents $PA(Y)$ define a causal mechanism where Y is conditionally independent of other variables given its parents. This implies that the parents provide a complete explanation for Y regardless of other covariates. With this understanding, we sought to investigate how these causal parents influence the performance of predictive models and to what extent they can explain the target node. If we were only considering the direct causal parents (∼Par) scenario, in principle, it could be considered a variable selection exercise. However, we are also considering and comparing to the full (∼All) DAG - as well as comparing GAM and RF with and without causality considerations.

   We acknowledge that removing variables during the variable selection step may potentially violate the assumptions we made, which is not uncommon when applying the causal inference method to real-world data. However, applying variable selection before running the causal discovery algorithm is crucial in this context. Including all 41 climate and catchment attributes in the PC algorithm increases the dimensionality of the PC algorithm. Increasing the number of covariates in the PC algorithm can reduce the algorithm's detection power (Runge, 2018). This is because higher dimensionality increases the number of conditional independence tests, which can be unreliable in cases of limited sample size or noisy data (Kalisch and Bühlman, 2007; Li and Wang, 2009; Ramsey et al., 2012). Careful variable selection or dimensionality reduction is often necessary to mitigate these issues (Runge et al., 2023).

   Moreover, we attempted to include all continuous variables during the causal discovery process to account for the possibility of causal sufficiency. Despite this, we observed challenges such as the generation of disconnected DAGs with independent nodes or groups of nodes lacking causal relationships with runoff signatures. We also encountered undirected edges in the graphs, indicating potential hidden confounding. These issues further emphasize the limitations of the PC method in real-world applications. It is important to note that our study represents the first attempt to apply causal discovery algorithms in this specific hydrological context, where no established guidelines exist. This required extensive experimentation, including bootstrapping DAGs, refining data processing methods, and testing various conditional independence tests to derive physically meaningful DAGs.

   Causal discovery methods inherently rely on assumptions, such as those concerning hidden variables in the PC algorithm, which are often violated in real-world problems. In our study, which spans numerous catchments across a broad spatial extent, some degree of approximation was inevitable.

However, future research could explore how these assumptions can be relaxed or adjusted without compromising the validity of the results at different scales. We appreciate your feedback, and these points will be better articulated in the manuscript.

We will incorporate these clarifications into Sections 2.2.1 and 2.2.3 to provide a more comprehensive context for our use of causal discovery methods.

2. **This also has some implications on one of the paper's main points regarding the invariance property of the causal parents of Y: If the environments can be modeled as the values of a random variable (if I understand correctly, the environments are created by clustering certain covariates, so we can indeed model them as a child of such covariate(s)), then not only the set of causal parents but also the full set of covariates is invariant – invalidating some of the main points of the paper.**

   This is a good remark that was raised. It is connected to the readability of the manuscript. We do not claim the invariance property of causal parents but the properties of runoff signature (Y) given its causal parents ($PA(Y)$). The prediction of a given response is causally dependent on some (selected) covariates; these covariates constitute/can be seen as descriptors of the environment 'driving' the response. Clustering was used to group the CAMELS catchments into different categories based on specific attributes. Any given catchment will belong to one climate attribute cluster, one soil attribute cluster, one topographic attribute cluster, one geological cluster and one vegetation cluster (i.e. each catchment is 'assigned' 5 cluster values, one for each attribute). The whole process of training and testing the models is now (also) done on separate attribute clusters only, so basically, it is only done on a subset of the available data but using data that share certain characteristics. However, the causal parents/selected variables are the same whether we use clustering or not. The covariates in a DAG include variables from all five categories, and in no environment does a covariate serve as a child of a covariate from another environment.

   We will improve the readability of Section 2.2.2 by clarifying the distinction between clustering and the use of covariates.

3. **(As a side, empirical differences between 'causal' and 'non-causal' methods are then, in my view, 'only' due to different ways of performing variable selection – and worse performance of non-causal methods on test data simply means that the variable selection or regularization can be improved.)**

   Thank you for raising this point. We agree that this point is misleading in our paper, and more explanations need to be added to it. To answer this comment, the variable selection was performed for each runoff signature, and we selected the most influential variables among 41 climate and catchment attributes. The selected variables are used to predict the runoff signature. We called these models non-causal models since they did not use any causal information. Under certain assumptions, we derived the causal relationship between the selected variables. Since, theoretically, the causal parents can explain their child node independent of other covariates, we tested whether using only causal parents affects the prediction model robustness. While this procedure might be similar to variable selection, the main purpose here is not to increase the predictions' accuracy by selecting different sets of predictors but to answer to what extent causal parents can explain their child node compared to using all covariates for prediction. To evaluate this, we used the whole dataset for the prediction, which we called baseline models, and subsets of the dataset, which were called sub-models, with and without utilising causal information. If the causal models performed comparably to or better than non-causal models across different environments, it indicates that causal parents suffice to explain the target variable. In cases where causal models outperformed non-causal ones, it suggests that some covariates in the non-causal models may represent spurious correlations, negatively impacting performance in that specific environment. Overall, the consistency of causal model results highlights the importance of incorporating causal information into prediction frameworks.

   We will add these explanations to Section 2.2.

4. **Even if the above three points were not an issue, the authors do not provide sufficient arguments on why we should trust that the result obtained by PC reflects the causal ground truth. A few points why in my view this is not obvious:**

- **The authors use "expert knowledge (...) to determine the causal direction between two variables with an undirected edge, correct the causally wrong direction between variables and block the spurious edges between variables". But if we know that some of the edges are incorrect, why should we trust the others? (Also, I did not find the description of the process of correcting edges sufficiently clear.)**

  We agree that it is necessary to add further explanation of how expert knowledge is used in developing DAGs. Thank you for letting us know. The way we used expert knowledge in our analysis was specifically to determine the causal direction of undirected edges by considering the underlying physics of the processes. Based on this understanding, first, we added impossible links to a blacklist. We then applied the PC algorithm iteratively, correcting the undirected edges and adding them to a whitelist while blacklisting any spurious links, if identified. This procedure continued until the resulting DAG contained no undirected or spurious links. It is worth mentioning that blacklisting impossible links could reduce the number of iterations to reach a 'stable' DAG.

  It is important to note that we do not claim the resulting DAG to represent the ground-truth DAG. If the ground-truth DAG were known, we could evaluate the resulting DAGs using metrics such as the Bayesian Information Criterion (BIC) or Structural Hamming Distance (SHD). However, in the absence of a known ground-truth DAG, the primary means of evaluation relies on domain knowledge. The structure of the DAG can vary depending on the causal discovery method used and the choice of conditional independence tests. The legitimacy of the graph was assessed using our expert knowledge to judge the plausibility of the inferred causal relationships. We also acknowledge that there are objective methods to evaluate the structure of DAGs. For instance, one approach is to generate synthetic data from the obtained DAG, reapply the causal discovery method to the generated data, and then compare the inferred DAG with the original. This validation procedure is among the considerations we plan to explore in future work.

  To address the reviewer's concerns, we will add the following clarifications to the manuscript in Section 2.2.3:

  – An explanation of how expert knowledge was used to correct undirected and spurious edges.
  – A discussion of the limitations of this approach and its implications for the reliability of the resulting DAG.

  We believe these additions will clarify our methodology and address the concerns raised.

- **The PC algorithm is known to produce results that are not reliable. E.g., relabeling the variables, i.e., simply permuting the columns in the data matrix, or subsampling the data set sometimes change the outcome. Simulation experiments show that even under no model misspecification huge sample sizes can be needed to reliably obtain the ground truth graph.**

  Thank you for raising this important point. Indeed, it is well-documented that the PC algorithm's results can be sensitive to factors such as subsampling the dataset or permuting variable order e.g. (Colombo et al., 2014; Kalisch and Bühlman, 2007). To address these limitations in our study, we implemented a structured approach to mitigate the instability of the algorithm's output.

  Firstly, we utilized the entire dataset (671 catchments) to ensure the largest possible sample size for the causal discovery process. Secondly, we iteratively refined the blacklist and whitelist of edges by running the PC algorithm multiple times. This iterative process allowed us to identify spurious, undirected, or unstable edges, which were then systematically added to the respective lists. By doing so, we ensured that the final execution of the PC algorithm, incorporating the blacklist and whitelist, consistently produced the same DAG.

  While this was a labour-intensive and largely manual task, it represents a step toward addressing the algorithm's inherent variability. We also acknowledge that this process could be further improved and potentially automated in future studies.

  This discussion will be added to section 2.2.3.

- **The paper does not provide any theoretical guarantees. This would probably be too much to ask for an applied paper but I argue below that this question is**

not purely theoretical: the assumptions that are known to be sufficient to obtain theoretical guarantees are most likely violated in this application.

We acknowledge that this study focuses on applying existing methods rather than exploring their theoretical foundations or the assumptions underlying them. However, introducing structural changes to the DAGs across different environments would indeed require theoretical guarantees to ensure that the assumptions used to construct DAGs for the baseline models remain valid for the sub-models with modified DAGs.

In this study, the structure of the DAGs and their variables remains consistent across all environments for each runoff signature. As a result, the assumption of causal sufficiency and the independence of the runoff signature from other covariates, given its parent nodes, holds across all environments. It should be noted that the causal sufficiency assumption is often violated in real-world applications. According to Runge (2018), to satisfy causal sufficiency, we only need to assume that no unobserved variables directly or indirectly influence any pair of our measured variables when working with a limited set of observed data. While we acknowledge that the assumptions underlying causal discovery methods may be strong, combining well-established methods, careful preprocessing, and domain expertise helps mitigate the impact of potential assumption violations.

5. **The paper does not provide sufficient arguments on whether the differences between methods are statistically significant.**

   Thank you for pointing this out. We conducted a random sampling of the training and testing sets 100 times, enabling us to shuffle the catchments between these sets. This approach allowed us to simulate runoff signatures using various combinations of catchment attributes within each environment and to compare the averaged R-squared and RMSE values for each model.

   To address this comment, we performed a nested F-test to calculate the statistical significance of differences between causal (∼Par) and non-causal (∼All) GAM models for train and test results. For random forest, which is a non-parametric and non-linear model, we performed a permutation test to calculate the statistical significance of causal (∼Par) and non-causal (∼All) RF models for training and test results. The details and the results of the tests are presented in the next section. The results of this analysis will be added to the supplementary materials. The tests' results suggest that although the difference between causal and non-causal models is significant during the training phase, it often becomes insignificant during testing.

6. **In my view, the paper is not sufficiently clear about the experimental setup using the different clusters. E.g., how exactly are the training and test sets chosen? The authors mention robustness across environments but then training and test data should be from different clusters?**

   Thank you for letting us know about this issue. In this study, the environments correspond to clusters in the sub-models and the entire dataset in the baseline models. For each environment, we divide the dataset into training and test sets, where 75% of the catchments are randomly selected for training, and the remaining 25% are used for testing. The process of selecting catchments for the training and testing sets is repeated iteratively 100 times, creating different combinations of catchments in these sets for each environment. This approach provides a range of model performances, and their average performance is used for comparison. Importantly, training and testing are conducted within the same environment. Therefore, if a model is trained on catchments from a specific cluster, for example, climate cluster 1, it is also tested on catchments within that same cluster. We will add more explanations on the experimental setup at the end of Section 2.2.

7. **It was unclear to me how the paper accounts for time-dependence of the data points.**

   In this work, we did not use time-dependent variables. Climatic indices and runoff signatures are derived from their respective time series for the whole period from 01/10/1989 to 30/09/2009. These values represent time-aggregated properties of the time series. Additionally, variables related to attributes such as topography, soil, vegetation, and geology can be considered time-independent. The obtained DAGs in this study are static Bayesian Networks.

   If the question is about time dependence in the causal order, time-averaging does not eliminate the underlying causal mechanisms (Gong et al., 2017). The relationships identified in the static DAG still hold because they reflect aggregated causal effects that persist over time. The causal ordering thus remains valid in this aggregated representation. We will add an explanation to Section 2.2 to clarify this point.

8. **(As a side, in general, when considering robustness against a change of environment, using the causal parents as covariates may not be optimal. Instead, one could use what is referred to as the stable blanket.)**

   Thank you for this comment. In our study, the causal structure (DAG) remains consistent, with the target variables (runoff signatures) having no child nodes, and the set of causal parents remaining unchanged across all environments. As a result, the causal parents align with both the Markov blanket and the stable blanket. This is because the causal parents of the target variable form a subset of the Markov blanket, and interventions on non-parent nodes do not alter the functional relationships governing the causal mechanism of the target variable (Pfister et al., 2021). Therefore, in our case, we can consider the causal parents as the stable blanket.

9. **The paper contains several imprecise/incorrect statements. Here are two examples (there are more): "They are the assumptions under which the causal relationship from the observational data can be learned." What precisely does "can be learned" mean? This may sound like a minor point but in my view it is not. One way of making this precise is to write down conditions for uniform consistency. There are few conditions known under which uniform consistency holds. However, such conditions are very restrictive. (E.g., some of such conditions include the assumption that the random variables are jointly Gaussian. If I understand correctly, the authors transform marginals but even this does not suffice.) It is known that, in general, all nonparametric conditional independence tests that are level are trivial (and do not have any non-trivial power), so it may even be impossible to relax such conditions to something reasonable. This is important in that these thoughts may be a reason for why the PC algorithm is usually unreliable in practice (see above). To give another example, "Covariate shift states that if variable Y is to be predicted from X, and X is the cause of Y, the conditional probability P(Y—X) remains the same across all environments if the distribution of X changes" is in my view at least imprecise: covariate shift is usually meant as a non-causal assumption and invariance generally holds only if X is the set of all causal parents of Y.**

   Thank you for your detailed feedback.

   - On the phrase “can be learned”: By “can be learned,” we mean “can be discovered,” as these terms are often used interchangeably in the causal inference literature. Common terminologies include “causal learning,” “causal discovery”, and “learning causal structure” (Peters et al., 2017). To avoid ambiguity, we will standardize these terms throughout the paper to ensure clarity and consistency.
   - On the assumptions for uniform consistency and Gaussian transformation: Gaussianity is one of the assumptions of PC algorithm. Regarding the data transformation to approximate a Gaussian distribution, we followed the approach outlined in the literature. For example, in (Dutta and Maity, 2020), a similar transformation process was employed. We acknowledge that Gaussian assumptions can be restrictive and do not address all the limitations you’ve raised. However, we adopted this approach as it aligns with established practices in the field and enables the use of conditional independence tests that assume Gaussianity.
   - On the description of covariate shift: As you correctly noted, covariate shift is a non-causal assumption. Additionally, the invariance of $P(Y \mid X)$ indeed holds when X includes all causal parents of Y. In our work, we did not claim otherwise. Specifically, we ensured that X contains all causal parents of Y as relevant to our analysis. The environments in our study are constructed using clustering, which creates subsets of catchments containing the same variables but with differing distributions. These environments are designed to test the causal mechanisms (runoff signatures and their causal parents) under varying conditions inspired by the principles of covariate shift.

   We will revise the text to reflect these points clearly and accurately to avoid any potential misunderstandings.

10. **The paper contains several typos, such as "casual" or "Clarck" or "causal models are assum result".**

    We will fix this issue in the entire manuscript.

2 Changes in the manuscript

In this section, the main changes that will be made in the manuscript is shown:

2.2 Methods

The methodology integrates feature selection, clustering, causal discovery and prediction. Fig. 1 shows the methodological procedure used in this study. In Fig. 1, causal models refer to the models that use causal parents, and non-causal models use all selected variables as predictors. Environments are subsets of the dataset obtained by clustering algorithms. Therefore, the words environment, cluster and subset imply the same meaning in this study. Baseline models refer to the models that use the whole dataset, all 671 catchments, for training and testing, and sub-models use subsets of the dataset for this purpose. GAM∼Par and RF∼Par are causal GAM and RF models that employ causal parents for prediction. GAM∼All and RF∼All are non-causal GAM and RF models that use all the selected variables as predictors. A robust model is defined as one that maintains its accuracy across different environments.

Since, theoretically, the causal parents contain all information to explain their child node independent of other covariates, we tested whether using only causal parents affects the prediction model robustness. ~~The goal is to investigate whether the causal discovery can enhance prediction models' robustness, identifiability and parsimony.~~ The primary goal here is not merely to improve prediction accuracy by selecting different sets of predictors. Instead, it is to assess how well causal parents can explain their child node compared to using all covariates for prediction and to determine how integrating causal information can enhance the parsimony and robustness of prediction models.

To evaluate this, we used the whole dataset for the prediction (baseline models) and subsets of the dataset (sub-models) with and without utilising causal information and causal and non-causal models, respectively. If the causal models performed comparably to or better than non-causal models across different environments, it indicates that causal parents suffice to explain the target variable. In cases where causal models outperformed non-causal ones, it suggests that some covariates in the non-causal models may represent spurious correlations, negatively impacting performance in that specific environment. Furthermore, the robustness of the models is assessed by comparing their accuracy in training and test settings.

The steps are explained in the following sections.

Section 2.2; from line 214:

For all models, BN, GAM, and RF, ~~the data is split into 75 % training and 25 % test samples. The models are run 100 times, with training and test sets randomly selected each time.~~ and for each environment, we divide the dataset into training and test sets, where 75% of the catchments are randomly selected for training, and the remaining 25% are used for testing. This process is repeated 100 times using bootstrapping to generate different combinations of training and test sets. This approach provides a range of model performances, and their average performance is used for comparison. Importantly, training and testing are conducted within the same environment. For example, if a model is trained on catchments from a specific climate category cluster, it is also tested on catchments within that same cluster. The models are executed for the whole dataset (baseline models) and each cluster of categories (sub-models). The models' accuracy is evaluated using Root Mean Squared Error (RMSE) and R-squared metrics between prediction and observations. The iteration provides 100 RMSE and R-squared for each run, and the accuracy is reported as their mean value. The following section discusses the obtained results of this study.

2.2.1 Feature Selection

In this section, we conduct the variables selection to 1) identify the most influential factors explaining the target signature and 2) reduce the dimensionality of the causal discovery problem (Runge et al., 2023). Including all 41 climate and catchment attributes in the PC algorithm increases its dimensionality, which can have adverse effects. A higher number of covariates reduces the statistical significance of detected edges and increases the risk of spurious links. This occurs because high dimensionality requires more conditional independence tests, which can become unreliable in cases of limited sample sizes (Kalisch and Bühlman, 2007; Li and Wang, 2009; Le et al., 2016; Ramsey et al., 2012). It is worth mentioning that we attempted to include all continuous variables in the causal discovery process without applying variable selection. This approach was tested to address the causal sufficiency assumption in the PC algorithm. Despite this, we observed challenges such as the generation of disconnected DAGs with independent nodes or groups of nodes lacking causal relationships with runoff signatures.

The explanatory variables for each signature are selected based on 1) ranked correlation coefficients and 2) variable importance. It should be noted that to develop the BN, which is a probabilistic graphical model, the selected variables (nodes) shouldn't be the deterministic functions of each other; otherwise, the conditional dependency structure of DAGs will change. Therefore, the aridity index, a function of precipitation and potential evapotranspiration, is removed from the selection procedures. Additionally, it is assumed that the selected variables satisfy causal Markov and faithfulness assumptions (Spirtes et al., 2001) when used for the PC causal discovery algorithm. They are the assumptions under which the causal relationship from the observational data can be learned. These assumptions relate the d-separation in the graph to conditional dependencies in the joint distribution (Pearl, 2009). These assumptions are explained in the following sections. The methods used for correlation analysis and variable importance are as follows:

1. Correlation analysis: Pearson, Kendall, and Spearman correlation coefficients are computed to illustrate the potential explanatory variables. The correlation analysis reveals the most influential variables from each category, namely climate, geology, vegetation, topography and soil. In addition, the scatter plot of the data helped visually understand the relationship between variables.

2. Variable importance: Since the results of the correlation analysis are not always consistent, another feature selection procedure is conducted using the random forest method to investigate the feature importance. ~~The same approach as correlation is repeated, using the random forest method to investigate the feature importance. Random forest is implemented using the R package randomForest (Liaw et al., 2015).~~ The variables are ranked using ~~according to~~ the out-of-bag method, which is quantified using the Mean Decreased Accuracy (IncMSE) score. The out-of-bag method ranks variables based on the increase in prediction error caused by removing each variable from the prediction process. Random forest is implemented using the R package randomForest (Liaw et al., 2015).

With the information provided by the procedures mentioned above, variables are selected based on a combination of correlation analysis, variable importance assessment and consideration of the underlying physics of the runoff signatures. We tried to select the most influential variables from each category, including climate, geology, soil, topography, and vegetation. The number of selected variables varies across categories. Multiple variables are selected from categories where most variables exhibit high correlation. Conversely, only the highest-correlated variable is chosen for categories with a weak correlation to the runoff signature of interest. For example, climatic variables are often highly influential for runoff signatures, leading to the selection of multiple variables. In contrast, geological variables tend to show a weak correlation with some runoff signatures, so only the most influential variable from this category is selected. The results of feature selection are presented in the supplementary materials.

2.2.2 Clustering

The CAMEL dataset provides five categories of catchment and climate attributes for each catchment. Clustering catchments based on each category of attributes is assumed to provide groups of catchments with homogeneous characteristics (Blöschl et al., 2013). Clustering is used to group the CAMELS catchments into different categories based on specific attributes. Any given catchment will belong to one climate attribute cluster, one soil attribute cluster, one topographic attribute cluster, one geological cluster and one vegetation cluster (i.e. each catchment is 'assigned' 5 cluster values, one for each attribute). The whole process of training and testing the models is now (also) done on separate attribute clusters only, so basically, it is only done on a subset of the available data but using data that share certain characteristics. However, the causal parents/selected variables are the same whether we use clustering or not.

We investigate the performance of the sub-models within each cluster of catchments. Each cluster is considered a new environment with certain properties to investigate the robustness of models with and without causal parents. The selected covariates are the same in all environments for each runoff signature. The properties of covariates in each cluster/environment are assumed to be homogeneous with respect to the specific attributes. The models are trained and tested for each cluster with homogeneous properties. Clusters are considered subsets of data where the distribution of covariates shifts from one cluster to another. This idea is inspired by Peters et al. (2016), where subsets of data are considered as different environments. The causal mechanism (the target variable and its parents) for each signature remains unchanged if there is a change in the distribution of parents (Woodward, 2008). Therefore, causal models (models with causal parents as explanatory variables) are expected to perform with consistent accuracy across different environments. This concept is influenced by the covariate shift assumption (Quionero-Candela et al., 2009). Covariate shift states that if variable Y is to be predicted from X, and X is the cause of Y, the conditional probability $P(Y|X)$ remains the same across all environments if the distribution of X changes. The assumption is tested by measuring the change in the accuracy of models when using causal parents as predictors across different environments. This information will help investigate the performance of the causal compared to non-causal models.

Two clustering methods are employed to group the catchment attributes in the CAMEL dataset. The K-medoids or Partitioning Around Mediods (PAM) clustering algorithm (Rdusseeun and Kaufman, 1987) is used for categories of attributes with continuous variables. PAM is a more robust method for handling outliers and noises than the K-mean method. The Gower distance (Gower, 1971) is used for mixed variables. This method is developed for datasets containing continuous, binary or multiattribute variables (Hennig and Liao, 2013). The elbow and silhouette methods are used to find the optimum number of clusters.

2.2.3 Causal Discovery

Causal discovery is used to partially or fully infer the causal structure, Directed Acyclic Graph (DAG), from observational data or distribution under certain assumptions (Heinze-Deml et al., 2018). Here, we try to find causal structures from the observational data without specifying the underlying physical equations using a causal discovery method. The causal discovery method is applied to the selected variables for the whole dataset and each runoff signature.

This study uses the constrained-based PC algorithm (Spirtes et al., 2001), named after its authors Peter and Clark ~~Clarck~~. This method identifies the DAG under faithfulness and Markov assumptions. Markov's assumption states that DAG represents all the conditional independencies in the dataset, and faithfulness states that conditional dependencies in the joint distribution of the data reflect the d-separation in DAG; in other words, the distribution is faithful to DAG (Peters et al., 2017). It is also assumed that there are no unobserved variables. We also assumed that runoff signatures do not cause climate and catchment attributes s and it is a sink node. PC algorithm assumes that the variables have a normal distribution. Therefore, the Box-Cox transformation is applied to the data (Dutta and Maity, 2020). The bnlearn R package (Scutari, 2009) is used to apply the PC algorithm. Mutual information with the Mont Carlo permutation test is chosen as the conditional independence test.

Since it is well-documented that the PC algorithm's results can be sensitive to factors such as sample size or permuting variable order, e.g. (Colombo et al., 2014; Kalisch and Bühlman, 2007), we applied an interactive process based on the expert knowledge to make sure that our results are reproducible. Therefore, first, a blacklist of edges is created to specify all impossible links prior to running the PC algorithm. The algorithm is then executed to derive the initial structure of the graph. Expert knowledge is applied to correct the causally incorrect edge directions by blacklisting the specific incorrect direction and to remove spurious links by blacklisting both directions. Additionally, corrected causal links are added to a separate list called the whitelist. We then iteratively applied the PC algorithm until the resulting DAG contained no undirected or spurious links. It is worth mentioning that blacklisting impossible links is important to reduce the number of iterations to reach a stable DAG.

We do not claim the resulting DAG to be the ground-truth DAG using this procedure. If the ground-truth or reference DAGs were known, we could evaluate the resulting DAGs using metrics such as the Bayesian Information Criterion (BIC) or Structural Hamming Distance (SHD). However, in the absence of a known ground-truth DAG, the primary means of evaluation relies on domain knowledge. The structure of the DAG can vary depending on the causal discovery method used and the choice of conditional independence tests. The legitimacy of the graph was assessed using our expert knowledge to judge the plausibility of the inferred causal relationships.

The obtained DAG structures are used to predict runoff signatures using Bayesian Network methods. Additionally, Generalized Additive Models and Random Forests are applied to predict runoff signatures: once using all variables in the DAGs (non-causal models) and once using only the causal parents of the target nodes (causal models). Since the target variable (runoff signature) has no child nodes, its causal parents provide an optimal blanket for the regression models. This is because the causal parents form a subset of the Markov blanket, and interventions on non-parent nodes do not affect the functional relationships underlying the causal mechanism of the target variable (Pfister et al., 2021).

Supplementary

12 Significance of Differences Between Causal and Non-Causal Models

12.1 Difference between GAM∼Par and GAM∼All

To determine whether the differences between causal and non-causal GAM models are statistically significant, we perform a nested F-test. This test is chosen because the causal model is nested within the non-causal model. We did this test to compare the train and test results of the causal (GAM∼Par) and non-causal (GAM∼All) models. If the result of this test is not significant (P-value >0.05) it means that GAM∼All is not significantly better than GAM∼Par. The steps to perform the nested F test are as follows:

1. Hypotheses:
   - H_0: The simpler (nested/causal) model provides an adequate fit to the data.
   - H_1: The more complex model (non-causal) provides a significantly better fit.
2. F-statistic: The F-statistic is calculated as:

$$F = \frac{\frac{\mathrm{RSS}_{\mathrm{nested}} - \mathrm{RSS}_{\mathrm{full}}}{df_{\mathrm{full}} - df_{\mathrm{nested}}}}{\frac{\mathrm{RSS}_{\mathrm{full}}}{n - df_{\mathrm{full}}}}$$

   where:

   - $\mathrm{RSS}_{\mathrm{nested}}$: Residual sum of squares of the nested (simpler/causal) model
   - $\mathrm{RSS}_{\mathrm{full}}$: Residual sum of squares of the full (complex/non-causal) model
   - df_{nested}: Residual degree of freedom of the nested model
   - df_{full}: Residual degree of freedom the full model
   - n: Number of observations
3. Derive P-value: The F-statistic follows an F-distribution with degrees of freedom:

$$\mathrm{df}_1 = df_{\mathrm{full}} - df_{\mathrm{nested}}, \quad \mathrm{df}_2 = df_{\mathrm{full}}$$

4. Decision Rule: Compare the calculated F-statistic to the critical value from the $F(\mathrm{df}_1, \mathrm{df}_2)$-distribution or use the corresponding P-value. Reject H_0 if:

$$P < \alpha$$

   where α is the significance level and is equal to 0.05.

Table 1: Statistical significance of differences between causal (GAM∼Par) and non-causal (GAM∼All) models. The significance level (a) is set to 0.05, and P-values are calculated using the nested F-test for both train and test results. The stars, *****, indicate statistically significant differences between causal and non-causal models, and **NS** stands for Not Significant within the respective environment, indicating that the null hypothesis has not been rejected. In the ”Environment” column, ”Clim” refers to climate, ”Geol” to geology, ”Topo” to topography, and ”Vege” to vegetation.

Statistical significance of difference between GAM∼Par and GAM∼All												
Environment	**Baseflow Index**		**High Q Dur**		**High Q Freq**		**Low Q Dur**		**Low Q Freq**		**Q mean**	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Baseline	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Clim 1	*	NS	*	NS	*	NS	*	NS	*	NS	*	*
Clim 2	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Clim 3	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Clim 4	NS	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Geol 1	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Geol 2	NS	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Geol 3	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Geol 4	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Geol 5	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Geol 6	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Geol 7	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
Soil 1	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Soil 2	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Soil 3	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Soil 4	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Soil 5	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Soil 6	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Topo 1	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Topo 2	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Topo 3	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Topo 4	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Vege 1	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Vege 2	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Vege 3	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Vege 4	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Vege 5	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Vege 6	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS

Table 1: (continued) Statistical significance of differences between causal (GAM∼Par) and non-causal (GAM∼All) models. The significance level (a) is set to 0.05, and P-values are calculated using the nested F-test for both train and test results. The stars, *****, indicate statistically significant differences between causal and non-causal models, and **NS** stands for Not Significant within the respective environment, indicating that the null hypothesis has not been rejected. In the "Environment" column, "Clim" refers to climate, "Geol" to geology, "Topo" to topography, and "Vege" to vegetation.

Statistical significance of difference between GAM∼Par and GAM∼All										
Environment	**Q5**		**Q95**		**Runoff Ratio**		**Slope of FDC**		**Stream Elas**	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Baseline	*	NS	*	NS	*	*	*	NS	*	NS
Clim 1	*	NS	*	*	*	*	*	NS	*	NS
Clim 2	*	NS	*	NS	*	NS	*	NS	*	NS
Clim 3	*	NS	*	NS	*	NS	*	NS	*	NS
Clim 4	*	NS	*	NS	*	NS	*	NS	*	NS
Geol 1	*	NS	*	NS	*	NS	*	NS	*	NS
Geol 2	*	NS	*	NS	*	NS	*	NS	*	NS
Geol 3	*	NS	*	NS	*	NS	*	NS	*	NS
Geol 4	*	NS	*	NS	*	NS	*	NS	*	NS
Geol 5	*	NS	*	NS	*	NS	*	NS	*	NS
Geol 6	*	NS	*	NS	*	NS	*	NS	*	NS
Geol 7	NS	NS	NS	NS	NS	NS	NS	NS	*	NS
Soil 1	*	NS	*	NS	*	NS	*	NS	*	NS
Soil 2	*	NS	*	NS	*	NS	*	NS	*	NS
Soil 3	NS	NS	*	NS	*	NS	*	NS	*	NS
Soil 4	NS	NS	*	NS	*	NS	*	NS	*	NS
Soil 5	*	NS	*	NS	*	NS	*	NS	*	NS
Soil 6	*	NS	*	NS	*	NS	*	NS	NS	NS
Topo 1	*	NS	*	NS	*	NS	*	NS	*	NS
Topo 2	*	NS	*	NS	*	NS	*	NS	*	NS
Topo 3	*	NS	*	NS	*	NS	*	NS	*	NS
Topo 4	*	NS	*	NS	*	NS	*	NS	*	NS
Vege 1	*	NS	*	NS	*	NS	*	NS	NS	NS
Vege 2	*	NS	*	NS	*	NS	*	NS	*	NS
Vege 3	*	NS	*	NS	*	NS	*	NS	*	NS
Vege 4	NS	NS	*	NS	*	NS	*	NS	*	NS
Vege 5	*	NS	*	NS	*	NS	*	NS	*	NS
Vege 6	*	NS	*	NS	*	NS	*	NS	*	NS

12.2 Difference between RF∼Par and RF∼All

To assess the significance of the difference between the causal (∼Par) and non-causal (∼All) random forest (RF) models, we employ a non-parametric permutation test. This method is appropriate because random forests are non-parametric and nonlinear, making methods like the nested F-test unsuitable. In this test, the R^2 and $RMSE$ values of training obtained from 100 runs of each model are resampled (with replacement) to construct a null hypothesis distribution of performance differences. This is achieved by randomly shuffling the labels of RF∼All and RF∼Par, recalculating the performance difference for each shuffle. A total of 10,000 permutations are performed to ensure the robustness of the null distribution for both train and test results. The P-value is then determined as the proportion of permuted differences that are as large or larger than the observed difference. A significance threshold (a) of 0.05 is used to evaluate the results. If the obtained P-value is greater than 0.05, the difference between RF∼All and RF∼Par is not statistically significant.

Table 2: Statistical significance of differences between causal (RF∼Par) and non-causal (RF∼All) models. The significance level (a) is set to 0.05, and P-values are calculated using the permutation test for both train and test results. The stars, *****, indicate statistically significant differences between causal and non-causal models, and **NS** stands for Not Significant within the respective environment. In the "Environment" column, "Clim" refers to climate, "Geol" to geology, "Topo" to topography, and "Vege" to vegetation.

Statistical significance of difference between RF∼Par and RF∼All												
Environment	**Baseflow Index**		**High Q Dur**		**High Q Freq**		**Low Q Dur**		**Low Q Freq**		**Q mean**	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Baseline	*	*	*	*	*	*	*	*	*	*	*	*
Clim 1	*	*	*	NS	*	*	*	NS	*	*	*	*
Clim 2	*	*	*	*	*	*	*	*	*	*	*	NS
Clim 3	*	NS	*	NS	*	*	*	*	*	*	*	*
Clim 4	*	*	*	*	*	*	*	*	*	*	*	NS
Geol 1	*	NS	*	*	*	*	*	NS	*	NS	NS	NS
Geol 2	*	*	*	NS	*	*	*	NS	*	NS	NS	NS
Geol 3	*	*	*	NS	*	*	*	NS	*	NS	NS	NS
Geol 4	*	*	*	NS	*	NS	*	NS	*	*	*	NS
Geol 5	*	*	NS	NS	*	NS	*	NS	*	*	NS	NS
Geol 6	*	NS	*	NS	*	*	*	*	*	*	*	NS
Geol 7	*	NS	*	*	*	NS	*	NS	*	NS	*	NS
Soil 1	*	NS	*	*	*	*	*	NS	*	NS	*	NS
Soil 2	*	*	*	NS	*	*	*	NS	*	*	NS	NS
Soil 3	*	NS	*	*	*	*	*	NS	*	NS	*	NS
Soil 4	*	NS	*	NS	*	*	*	NS	*	NS	*	NS
Soil 5	*	*	*	*	*	NS	*	*	*	NS	*	NS
Soil 6	*	*	*	NS	*	*	*	NS	*	*	NS	NS
Topo 1	*	*	*	NS	*	*	*	*	*	*	NS	NS
Topo 2	*	*	*	NS	*	*	*	NS	*	*	*	NS
Topo 3	*	*	*	NS	*	*	*		*	*	*	*
Topo 4	*	NS	*	*	*	*	*	*	*		*	NS
Vege 1	*	*	*	NS	*	*	*	*	*	*	*	NS
Vege 2	*	*	*	NS	*	*	*	NS	*	*	*	*
Vege 3	*	*	*	NS	*	*	*	NS	*	*	*	NS
Vege 4	*	NS	*	NS	*	NS	*	NS	*	NS	*	NS
Vege 5	*	*	*	NS	*	*	*	*	*	*	*	NS
Vege 6	*	NS	*	*	*	*	*	*	*	*	*	NS

Table 2: (continued) Statistical significance of differences between causal (RF∼Par) and non-causal (RF∼All) models. The significance level (a) is set to 0.05, and P-values are calculated using the permutation test for both train and test results. The stars, *****, indicate statistically significant differences between causal and non-causal models, and **NS** stands for Not Significant within the respective environment. In the "Environment" column, "Clim" refers to climate, "Geol" to geology, "Topo" to topography, and "Vege" to vegetation.

Statistical significance of difference between RF∼Par and RF∼All										
Environment	**Q5**		**Q95**		**Runoff Ratio**		**Slope of FDC**		**Stream Elas**	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Baseline	*	*	*	*	*	*	*	*	*	*
Clim 1	*	*	*	*	*	*	*	*	*	*
Clim 2	*	NS	*	NS	*	*	*	*	*	*
Clim 3	*	NS	*	*	*	*	*	NS	*	*
Clim 4	*	NS	*	NS	*	*	*	*	*	NS
Geol 1	NS	NS	*	NS	*	*	*	NS	*	*
Geol 2	NS	NS	*	NS	*	*	*	*	*	NS
Geol 3	*	NS	*	NS	*	*	*	*	*	NS
Geol 4	NS	NS	*	NS	*	*	*	NS	*	*
Geol 5	*	*	*	NS	*	NS	*	NS	*	NS
Geol 6	*	*	*	NS	*	*	*	*	*	*
Geol 7	*	NS	*	NS	*	*	*	NS	*	NS
Soil 1	*	NS	*	NS	*	*	*	NS	*	NS
Soil 2	*	NS	*	NS	*	NS	*	NS	*	*
Soil 3	*	NS	*	NS	*	*	*	NS	*	*
Soil 4	*	NS	*	NS	*	*	*	*	*	NS
Soil 5	*	*	*	NS	*	*	*	*	*	*
Soil 6	NS	NS	*	NS	*	*	*	*	*	NS
Topo 1	*	*	*	NS	*	*	*	*	*	*
Topo 2	*	NS	*	NS	*	*	*	*	*	NS
Topo 3	*	NS	*	*	*	*	*	*	*	*
Topo 4	*	NS	*	NS	*	*	*	*	*	NS
Vege 1	*	*	*	NS	*	*	*	*	*	NS
Vege 2	*	*	*	*	*	*	*	*	*	*
Vege 3	*	*	*	NS	*	NS	*	*	*	NS
Vege 4	*	NS	*	*	*	NS	*	*	*	NS
Vege 5	*	NS	*	*	*	*	*	*	*	NS
Vege 6	*		*	*	*	*	*	*	*	*

References

Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., and Savenije, H.: Runoff prediction in ungauged basins: synthesis across processes, places and scales, Cambridge University Press, 2013.

Colombo, D., Maathuis, M. H., et al.: Order-independent constraint-based causal structure learning., J. Mach. Learn. Res., 15, 3741–3782, 2014.

Dutta, R. and Maity, R.: Temporal networks-based approach for nonstationary hydroclimatic modeling and its demonstration with streamflow prediction, Water Resources Research, 56, e2020WR027 086, 2020.

Gong, M., Zhang, K., Schölkopf, B., Glymour, C., and Tao, D.: Causal discovery from temporally aggregated time series, in: Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence, vol. 2017, NIH Public Access, 2017.

Gower, J.: GENERAL COEFFICIENT OF SIMILARITY AND SOME OF ITS PROPERTIES, BIOMETRICS, 27, 857–&, https://doi.org/10.2307/2528823, 1971.

Heinze-Deml, C., Maathuis, M. H., and Meinshausen, N.: Causal Structure Learning, in: ANNUAL REVIEW OF STATISTICS AND ITS APPLICATION, VOL 5, edited by Reid, N., vol. 5 of *Annual Review of Statistics and Its Application*, pp. 371–391, https://doi.org/10.1146/annurev-statistics-031017-100630, 2018.

Hennig, C. and Liao, T. F.: How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification, JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES C-APPLIED STATISTICS, 62, 309–369, https://doi.org/10.1111/j.1467-9876.2012.01066.x, 2013.

Kalisch, M. and Bühlman, P.: Estimating high-dimensional directed acyclic graphs with the PC-algorithm., Journal of Machine Learning Research, 8, 2007.

Le, T. D., Hoang, T., Li, J., Liu, L., Liu, H., and Hu, S.: A fast PC algorithm for high dimensional causal discovery with multi-core PCs, IEEE/ACM transactions on computational biology and bioinformatics, 16, 1483–1495, 2016.

Li, J. and Wang, Z. J.: Controlling the false discovery rate of the association/causality structure learned with the PC algorithm., Journal of Machine Learning Research, 10, 2009.

Liaw, A., Wiener, M., Breiman, L., and Cutler, A.: Package 'randomforest', 2015.

Pearl, J.: Causality, Cambridge university press, 2009.

Peters, J., Buhlmann, P., and Meinshausen, N.: Causal inference by using invariant prediction: identification and confidence intervals, JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY, 78, 947–1012, https://doi.org/10.1111/rssb.12167, 2016.

Peters, J., Janzing, D., and Schölkopf, B.: Elements of causal inference: foundations and learning algorithms, The MIT Press, 2017.

Pfister, N., Williams, E. G., Peters, J., Aebersold, R., and Bühlmann, P.: Stabilizing variable selection and regression, The Annals of Applied Statistics, 15, 1220–1246, 2021.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and D. Lawrence, N.: Dataset Shift in Machine Learning, MIT Press, Cambridge, MA, 2009.

Ramsey, J., Zhang, J., and Spirtes, P. L.: Adjacency-faithfulness and conservative causal inference, arXiv preprint arXiv:1206.6843, 2012.

Rdusseeun, L. and Kaufman, P.: Clustering by means of medoids, in: Proceedings of the statistical data analysis based on the L1 norm conference, neuchatel, switzerland, vol. 31, 1987.

Runge, J.: Causal network reconstruction from time series: From theoretical assumptions to practical estimation, Chaos: An Interdisciplinary Journal of Nonlinear Science, 28, 2018.

Runge, J., Gerhardus, A., Varando, G., Eyring, V., and Camps-Valls, G.: Causal inference for time series, NATURE REVIEWS EARTH & ENVIRONMENT, 4, 487–505, https://doi.org/10.1038/s43017-023-00431-y, 2023.

Scutari, M.: Learning Bayesian networks with the bnlearn R package, arXiv preprint arXiv:0908.3817, 2009.

Spirtes, P., Glymour, C., and Scheines, R.: Causation, prediction, and search, MIT press, 2001.

Woodward, J.: Invariance, modularity, and all that: Cartwright on causation, in: Nancy Cartwright's philosophy of science, pp. 210–249, Routledge, 2008.

List of Changes in the Manuscripts

The list of changes in the manuscript is as follows:

Table 1: Changes in the manuscript.

Section	**Changes**
Abstract	Changes made to make it more relevant to the title and main objectives of the paper.
Section 1 (Introduction)	The novelty of the paper is emphasized.
Section 2.1 (Data)	Additional explanation has been added to the section and Table 1 to clarify the data used in the study.
Section 2.2 (Methods)	Additional explanation on the independent causal mechanism is added to the section. Fig. 1 is modified. The data splitting into train and test sets is further explained at the end of the section.
Section 2.2.1 (Methods)	The explanation of the necessity of feature selection, such as reducing dimensionality, is added. Further explanation of correlation analysis and variable importance is added.
Section 2.2.2 (Methods)	The reasons for applying cluster analysis and how the resulting environments are used are further explained.
Section 2.2.3 (Methods)	The steps for applying causal discovery methods to data and the evaluation of the resulting DAGs are further explained.
Section 3.1 (Results)	More explanations about the properties of data within each category (e.g., climate, geology, and so on) are added to the clustering results. The caption of Table 3 is modified.
Section 3.2 (Results)	An explanation of the independent causal mechanism has been added to the section. Fig. 3 is modified to specify the nodes of the causal mechanism of the runoff signature. The caption of Table 4 is modified.
Section 3.4 (Results)	The results from the statistical significance test are added to the explanations of the models' comparisons.
Section 4 (Discussion)	Complementary explanations are to the section to clarify the results.
Section 5 (Conclusion)	The paper's outcomes are explained in a way that aligns with the topic, abstract, and main objectives of the paper.
Supplementary	The details of the statistical significance tests is added to the supplementary materials.