Can causal discovery lead to a more robust prediction model for runoff signatures?

Hossein Abbasizadeh[1], Petr Maca[1], Martin Hanel[1], Mads Troldborg[2], and Amir AghaKouchak[3,4,5]

[1]Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Czech Republic
[2]The James Hutton Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK
[3]Department of Civil and Environmental Engineering, University of California, Irvine, CA, USA
[4]United Nations University Institute for Water, Environment and Health, Hamilton, ON, Canada
[5]Department of Earth System Science, University of California, Irvine, CA, USA

Correspondence: Hossein Abbasizadeh (abbasizadeh@fzp.czu.cz)

Abstract. Runoff signatures characterize a catchment's response and provide insight into the hydrological processes. These signatures are governed by the co-evolution of catchment properties and climate processes, making them useful for understanding and explaining hydrological responses. However, catchment behaviours can vary significantly across different spatial scales, which complicates the identification of key drivers of hydrologic response. This study represents catchments as networks of variables linked by cause-and-effect relationships. We examine whether the direct causes of runoff signatures, representing independent causal mechanisms, can explain these catchment responses across different environments. To achieve this goal, we train the models using the causal parents of the runoff signatures and investigate whether it results in more robust, parsimonious, and physically interpretable predictions compared to models that do not use causal information. Peter and Clark (PC) causal discovery algorithm is applied separately for 11 runoff signatures to derive causal relationships between catchment attributes, climate indices, and the corresponding runoff signature. Three prediction models including Bayesian Network (BN), Generalized Additive Model (GAM), and Random Forest (RF) are used for predictions. The results indicate that among models, BN, a linear model with a structure based on the causal network, exhibits the smallest decline in accuracy between training and test simulations compared to the other models. Across nearly all environments and runoff signatures, using causal parents enhances robustness and parsimony while maintaining the accuracy of GAMs. While RF achieves the highest overall performance, it also demonstrates the most significant drop in accuracy between the training and test phases. When the sample size for training is small, the accuracy of the causal RF model, which uses causal parents as predictors, is comparable to that of the non-causal RF model, which uses all selected variables as predictors. This study demonstrates the potential of causal inference techniques for interpreting and enhancing the prediction of catchment responses by effectively representing the interconnected processes within hydrological systems in a cause-and-effect manner.

1 Introduction

Hydrological processes result from complex interactions between climate inputs and catchment characteristics (Sivapalan, 2006). These processes manifest in the catchment response at the catchment outlet. Therefore, catchments can be conceptu-

alized as a unit in which the cumulative effect of all interacting processes defines their runoff behaviour, commonly referred to as "runoff signatures." Runoff signatures encapsulate key characteristics of the runoff process in a catchment, including stream flow magnitude, frequency, and timing. These signatures are essential for a wide range of engineering and scientific applications (Blöschl et al., 2013), especially when causal interpretation or assessment is not possible due to insufficient data. McMillan (2020) outlined a wide range of applications for runoff signatures, such as assessing the performance of hydrological models (Clark et al., 2011; Todorovic et al., 2024), selecting appropriate model structures (Hrachowitz et al., 2014; Spieler and Schuetze, 2024) and estimating parameters (Pokhrel et al., 2012; Pizarro and Jorquera, 2024). They are also instrumental in streamflow prediction in ungauged basins (Yadav et al., 2007; Zhang et al., 2014; Matos and Oliveira e Silva, 2024), and in understanding catchment runoff responses at different spatial and temporal scales (Ficchi et al., 2019).

Although all processes in a catchment contribute to its runoff response, each runoff property (or signature) is directly influenced by a distinct set of climatic and catchment-specific characteristics. As an example, Chagas et al. (2024) studied the regional patterns of low flows across 1400 river gauges in Brazil. They showed that catchment characteristics, especially geological properties, have a significantly greater influence on low flows than climate attributes. Guzha et al. (2018) investigated the effects of changes in forest cover on annual mean flow, high flow, and low flow in 37 catchments of different climatic and physiographic properties in East Africa, concluding that not all catchments exhibit a significant response to forest loss. Therefore, it is necessary to identify a set of variables or covariates that are causally associated with a specific runoff signature and can reliably explain it under various environmental conditions. Understanding these variables allows for explaining the signature of interest across environments with different climatic and physiological conditions.

The main drivers of runoff signatures are commonly investigated using classification and regression methods. These techniques are applied to identify the main drivers influencing catchment response and assess their spatial dependencies. Classification criteria often include runoff properties (Ley et al., 2011; Sawicz et al., 2011; Kuentz et al., 2017), climate, and catchment similarities (Olden et al., 2012; Singh et al., 2016; Yang and Olivera, 2023; Ciulla and Varadharajan, 2024). Additionally, machine learning and statistical methods are widely used for the same purpose. For example, Addor et al. (2018) used random forest to predict 15 runoff signatures across 600 catchments in the US. They showed that climatic attributes are among the most influential predictors of the runoff signatures. McMillan et al. (2022) investigated the dominant process by linking climate and catchment attributes to hydrological signatures over large sets of catchments in the US, UK, and Brazil. They found that although some signatures, such as runoff ratio and baseflow index, were among the most robust metrics for characterizing processes, in some cases, the correlation found among variables and signatures in a country may not always generalize to others. They noted that these diverging correlations could result from statistical associations rather than true causal relationships.

We postulate that investigating the relationship between hydrological variables from causal-and-effect perspectives might solve the problem of diverging correlations reported by McMillan et al. (2022). A variable X is considered the cause of a variable Y if the value of Y depends on or is influenced by X in any given circumstances (Pearl et al., 2016; Pearl, 2009). Therefore, the probability of a target variable, such as a runoff signature, given its causes, should be the same under different conditions or across different environments. Broadly, there are two widely used frameworks for discovering causal relationships and estimating causal effects from observational data, including structural causal modelling (Pearl, 2009) and potential outcome

framework (Rubin, 1974). The methods used to discover causality and quantify causal effects and their strength are broadly referred to as Causal Inference Methods (CIMs).

One application of CIMs is to complement machine learning approaches by addressing the problems of transfer and generalization (Schölkopf et al., 2021; Ombadi, 2021), by identifying dependencies and confounding factors using multivariate analyses (Runge et al., 2019a). In an under-investigation cause-and-effect relationship, a confounding variable is an unknown or unmeasured variable that influences both the supposed effect and supposed cause (Pearl et al., 2016). Identifying confounders and unravelling causal effects make CIMs a valuable tool for enhancing the interpretability of Earth system models (Reichstein et al., 2019). CIMs are established based on a robust mathematical framework that identifies conditional dependencies in observational data (Pearl, 2009). This process often involves deriving a causal graph based on our understanding of relevant processes using methods such as the Bayesian Network (BN) or Bayesian Belief Network (Verma and Pearl, 1990).

In the last decade, significant efforts have been made to investigate and develop applications for CIMs in the field of Earth system modelling. These studies, primarily focused on uncovering causal relationships from time series, cover a broad range of topics including climate science (Runge et al., 2019b; Kretschmer et al., 2016), remote sensing (Perez-Suay and Camps-Valls, 2019), soil moisture-precipitation feedback detection (Wang et al., 2018), runoff behaviour (Zazo et al., 2020), the causal discovery of summer and winter evapotranspiration drivers (Ombadi et al., 2020), and study of hydrological connectivity (Sendrowski and Passalacqua, 2017; Rinderera et al., 2018; Delforge et al., 2022). However, the causal relationships between catchment attributes, climate characteristics, and runoff signatures have yet to be thoroughly investigated. A catchment can be represented as a probabilistic network of interconnected processes leading to a runoff signature. To achieve this, catchments can be conceptualized as Bayesian Networks (BNs), where variables are causally linked. BNs, part of the family of probabilistic graphical models, consist of nodes representing variables and directed edges indicating causal directions (Koller and Friedman, 2009). The structure of BNs is usually identified through causal discovery methods and expert knowledge (Runge et al., 2019a). Methods for causal discovery, also known as structural learning or causal search, can be categorized into constrained-based, score-based, and asymmetry-based approaches (Runge et al., 2023). Constrained-based methods use conditional independence tests to identify the causal graph, while score-based methods evaluate multiple causal graphs using a scoring function, selecting the highest-scoring one. Asymmetry-based methods are used to infer causal direction in the bivariate relationships (Runge et al., 2023).

The information about the causal relationships between catchment variables can be incorporated into prediction models to predict runoff signatures. Predictions using BNs are primarily designed for discrete datasets that can model complex interactions between variables. The rigorous probabilistic theories involved in BN make them popular for environmental modelling (Aguilera et al., 2011). However, Nojavan et al. (2017) and Qian and Miltner (2015) showed that the results of BNs are influenced by the discretization choice of continuous variables. Inference with BN for continuous variables is still a challenging task (Li and Mahadevan, 2018). Gaussian BN is a widely used method for modelling continuous variables. It assumes that the relationships between variables are linear and variables follow a Gaussian distribution (Marcot and Penman, 2019). To relax these assumptions, non-parametric continuous BNs have been developed (e.g. Qian and Miltner (2015)). However, Gaussian BNs remain a robust and widely-used framework, supported by various software packages (Geiger and Heckerman, 1994). Gaussian

BNs have been successfully applied in environmental modelling, particularly for water-quality studies e.g. Jackson-Blake et al. (2022) and Deng et al. (2023).

Given the success of Gaussian BNs in other fields, in this study, we adopt Gaussian BNs to predict runoff signatures. The links between variables of BN are derived from Peter Spirtes and Clark Glymour's (PC) causal discovery algorithm. Additionally, two non-linear models, the Generalized Additive Model (GAM) and Random Forest (RF), are used in this study. GAM is an extension of the Generalized Linear Model (GLM) that models non-linear relationships between explanatory and response variables using sums of arbitrary functions of the explanatory variables (Hastie et al., 2009). GAMs have been widely used for hydrological studies including flood frequency analysis (Ouali et al., 2017), low flow frequency analysis (Ouarda et al., 2018), flood peak prediction (Dubos et al., 2022), analysis of nuisance flooding (Vandenberg-Rodes et al., 2016), spatial analysis of extremes (Love et al., 2020), climate-crop yield relationships (Zachariah et al., 2021). RFs, first developed by Breiman (2001), are non-linear, non-parametric models used extensively for regression, classification, prediction, and variable selection. RF-based models have also been used in the field of environmental modelling, including for flow frequency analysis (Desai and Ouarda, 2021), runoff signature prediction (Addor et al., 2018), water level forecasting (Nguyen et al., 2015), downscaling (Arshad et al., 2024), and understanding of drivers of hazards (Seydi et al., 2024).

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. 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. These causal parents, together with the runoff signature, form a causal mechanism that is theoretically independent of other variables and can explain the variations in the signature. In this context, our objective is to test whether this fundamental concept applies to complex, real-world hydrological systems. 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), 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.

2 Data and methods

2.1 Data

The Catchment Attributes and MEteorology for Large-sample Studies dataset (CAMELS) is used in this study (Newman et al., 2015; Addor et al., 2017). It includes time series of streamflow and hydrometeorological variables, climatic indices (derived

from hydrometeorological time series), catchment attributes, and runoff signatures (derived from streamflow time series) for 671 catchments spanning the contiguous United States. The attributes in the CAMELS dataset are divided into 5 categories, including climate, geology, soil, topography, and vegetation (land cover) categories. CAMELS also includes comprehensive explanations of the techniques employed to derive catchment attributes and a discussion of potential limitations in the data sources. The variables used in this study include catchment characteristics, climate attributes, and runoff signatures, which are outlined in Table 1 and Table 2. The streamflow and hydrometeorological time series are not included in this study.

Table 1. Catchment and climate attributes, calculated over the period from 01/10/1989 to 30/09/2009 (Table 2 in Addor et al. (2018)).

Category	No	Attribute	Description	Unit
Climate	1	p_mean	Mean daily precipitation	mm/day
	2	pet_mean	Mean daily PET (Priestley-Taylor)	mm/day
	3	p_seasonality	Seasonality and timing of precipitation	-
	4	frac_snow	Fraction of precipitation as snow	-
	5	aridity	pet_mean/p_mean	-
	6	high_prec_freq	Frequency of high precipitation days	days /year
	7	high_prec_dur	Average duration of high precipitation events (precipitation > 5×p_mean)	days
	8	low_prec_freq	Frequency of high precipitation events (precipitation > 5×p_mean)	days /year
	9	low_prec_dur	Average duration of dry periods (precipitation < 1mm)	days
	10	high_prec_timing	Season during which most high precipitation days occur (precipitation > 5×p_mean)	season
	11	low_prec_timing	Season during which most dry days occurs (precipitation < 1mm)	season
Topography	1	gauge_lat	Gauge latitude	degrees north
	2	gauge_lon	Gauge longitude	degrees east
	3	elev_mean	Mean elevation of catchment	m
	4	slope_mean	Mean slope of catchment	m/km
	5	area_gages2	Area of catchment	km^2
Geology	1	geol_1st_class	Most common geological class in the catchment	-
	2	geol_2nd_class	Second most common geological class in the catchment	-
	3	glim_1st_class_frac	Fraction of most common geological class	-
	4	glim_2nd_class_frac	Fraction of second most common geological class	-
	5	carbonate_rocks_frac	Fraction of carbonate rock	-
	6	geol_porosity	Subsurface porosity	-
	7	geol_permeability	Subsurface permeability	m^2

Table 1. (continued) Catchment and climate attributes, calculated over the period from 01/10/1989 to 30/09/2009, (Table 2 in Addor et al. (2018)).

Category	No	Attribute	Description	Unit
	1	soil_depth_pelletier	Depth to bedrock (< 50m)	m
	2	soil_depth_statsgo	Soil depth (< 1.5m)	m
	3	soil_porosity	Volumetric soil porosity (averaged over the top 1.5m of soil)	-
	4	soil_conductivity	Saturated hydraulic conductivity (harmonic mean over the top 1.5m of soil)	cm/hr
Soil	5	max_water_content	Maximum water content (averaged over the top 1.5m of soil)	m
	6	sand_frac	Sand fraction (averaged over the top 1.5m of soil)	%
	7	silt_frac	Silt fraction (averaged over the top 1.5m of soil)	%
	8	clay_frac	Clay fraction (averaged over the top 1.5m of soil)	%
	9	water_frac	Fraction of water in 1.5m of topsoil	%
	10	organic_frac	Fraction of the soil depth marked as organic matterial (fraction of soil_depth_statsgo)	%
	11	other_frac	Fraction of other components (fraction of soil_depth_statsgo)	%
	1	frac_forest	Forest fraction of catchment	-
	2	lai_max	Maximum monthly leaf area index	-
	3	lai_diff	Difference between maximum and minimum leaf area index	-
Vegetation	4	gvf_max	Maximum monthly green vegetation fraction	-
	5	gvf_diff	Difference between maximum and minimum green vegetation fraction	-
	6	dom_land_cover	dominant land cover type	-
	7	dom_land_cover_frac	fraction of dominant land cover	-

Table 2. Runoff Signatures in CAMELS dataset, calculated over the period from 01/10/1989 to 30/09/2009

No	Signature	Description	Unit	Reference
1	baseflow_index	The ratio of mean daily baseflow to mean daily discharge	-	(Ladson et al., 2013), Table 2 in Addor et al. (2018)
2	high_q_dur	The average duration of high-flow events (successive days of flow events > 9 × median daily flow)	days	(Clausen and Biggs, 2000), Table 2 in Addor et al. (2018)
3	high_q_freq	Frequency of high-flow days (flow events > 9 × median daily flow)	days/year	(Clausen and Biggs, 2000), Table 2 in Addor et al. (2018)
4	low_q_dur	The average duration of low-flow events (successive days of flow events < 0.2 × mean daily discharge (q_mean))	days	(Olden and Poff, 2003), Table 2 in Addor et al. (2018)
5	low_q_freq	Frequency of low-flow days (flow events < 0.2 × mean daily discharge (q_mean))	days/year	(Olden and Poff, 2003), Table 2 in Addor et al. (2018)
6	q_mean	Mean daily discharge	mm/day	Table 2 in Addor et al. (2018)
7	Q5	Low flow: 5% flow quantile (95% exceedance probability)	mm/day	Table 2 in Addor et al. (2018)
8	Q95	High flow: 95% flow quantile (5% exceedance probability)	mm/day	Table 2 in Addor et al. (2018)
9	runoff_ratio	Mean daily discharge to mean daily precipitation	-	(Sawicz et al., 2011), Table 2 in Addor et al. (2018)
10	slope_FDC	The slope of flow duration curve	-	(Sawicz et al., 2011), Table 2 in Addor et al. (2018)
11	stream_elast	Steam flow elasticity (sensitivity of annual streamflow to variations in precipitation)	-	(Sankarasubramanian et al., 2001), Table 2 in Addor et al. (2018)

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 defined as subsets of the dataset obtained through clustering algorithms. Therefore, the word "environment" refers to the clusters or subsets of data. The whole dataset itself can also be considered an environment; however, in this study, we primarily refer to clusters when discussing environments. Baseline models refer to the models that use the whole dataset (i.e., 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.

In this study, we explore the concept of independent mechanisms in the context of modelling runoff signatures. The independent mechanisms assumption suggests that the causal generative process of a system's variables is made up of self-contained modules that operate independently, without influencing or providing information to one another, and these mechanisms stay stable even when the data distribution changes (Schölkopf et al., 2012; Peters et al., 2017). Using the Directed Acyclic Graph (DAG) obtained from causal discovery, we identified the causal parents of the target runoff signature, which represent the independent causal mechanism generating this variable. Independent mechanisms, as modular components, can be trained separately across different environments and tend to be more adaptable and reusable, a quality we refer to as robustness in this study (Parascandolo et al., 2018). They may also be easier to interpret and provide more insight since these causal mechanisms correspond to physical mechanisms. To evaluate the real-world applicability of this mechanism, we used the identified causal parents as predictors to train RF and GAM. This approach tests whether the independent mechanism derived from the DAG can effectively explain and predict the target variable, supporting the idea that these causal conditionals serve as robust and interpretable modules in the prediction of runoff signatures.

To achieve this, we use the whole dataset for the prediction in baseline models and subsets of the dataset in sub-models, both with and without utilizing causal information, corresponding to 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 and checking whether the difference between causal and non-causal models is statistically significant in both settings. The methods used to calculate statistical significance tests comparing causal and non-causal models are presented in the supplementary material.

The steps are explained in the following sections.

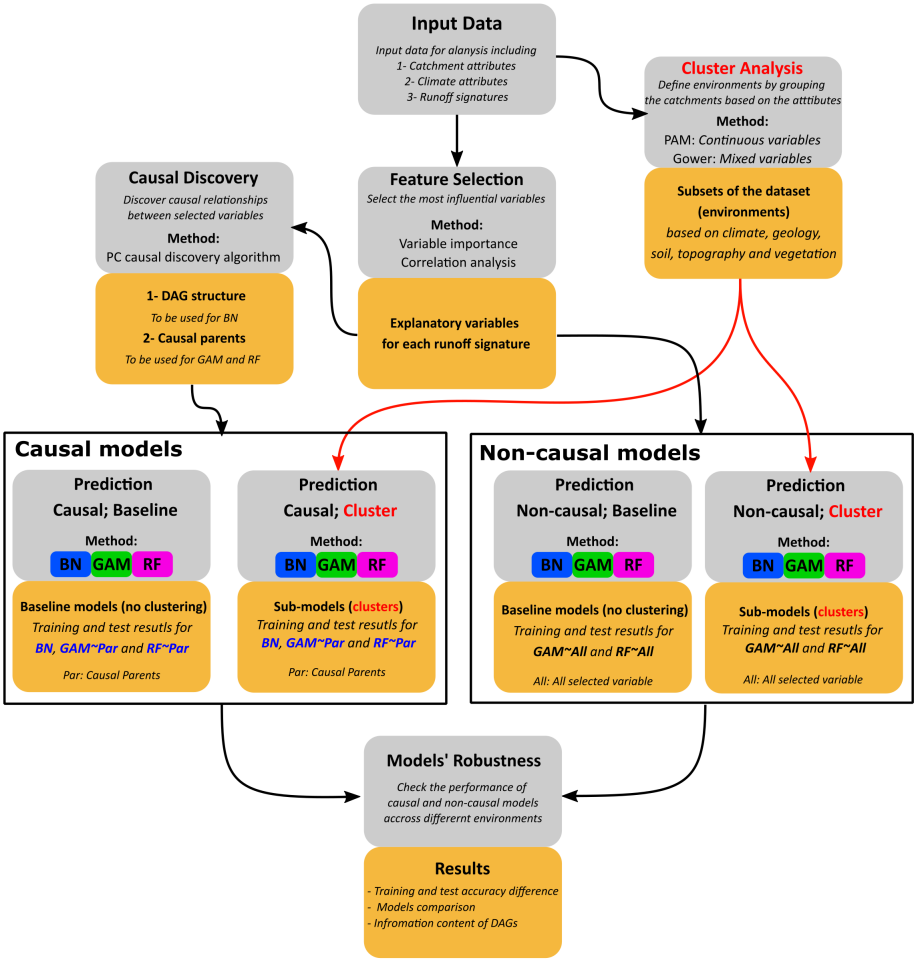

Figure 1. Flowchart depicting the steps followed in this study. Grey boxes indicate the procedures, orange boxes present the results of these procedures, blue text highlights where information about causality is utilized, and the red text and arrows highlight the cluster analysis and indicate where the 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. GAM∼Par and RF∼Par are causal models (GAM and RF) using only causal parent variables for prediction, while GAM∼All and RF∼All 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.

2.2.1 Feature selection

In this section, we conduct the variable selection to 1) identify the most influential factors explaining the target signature and
165 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 the dimensionality of the PC algorithm. Increasing the number of covariates in the PC algorithm can reduce the algorithm's detection power (Runge, 2018). 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, which requires that all common causes of the target variables are accounted for. 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 ranked correlation coefficients and variable importance. It should be noted that to develop the BN, which is a probabilistic graphical model, the selected variables (nodes) should not 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 discovered. 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 by the 3 methods are not always consistent, another feature selection procedure is conducted using the random forest method to investigate the feature importance. The variables are ranked using 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, for categories with a weak correlation to the runoff signature of interest, only the most correlated variable is chosen. For example, climatic variables often have a strong influence on runoff signatures, leading to the selection of multiple variables. In contrast, geological variables tend to have 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 CAMELS dataset provides five categories of catchment and climate attributes for each catchment (Table 1). 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. The causal parents and selected variables are, however, 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 remain the same across all environments for each runoff signature. Within each cluster or environment, covariate properties are assumed to be homogeneous with respect to specific attributes, allowing us to train and test models using variables with consistent properties. Defining environments as subsets of data is inspired by Peters et al. (2016). Here, we use clustering analysis to define these subsets, resulting in environments with specific properties. Therefore, clusters can be considered as subsets of data where the distribution of covariates shifts from one cluster to another. This variation across clusters provides a framework for exploring the underlying independent causal mechanisms of each runoff signature..

The causal independent 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 a set of variables X, and X is the cause of Y, the properties of conditional probability $P(Y|X)$ remains unchanged if the distribution of X changes. 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 CAMELS dataset. The K-medoids or Partitioning Around Mediods (PAM) clustering algorithm (Rdusseeun and Kaufman, 1987) is used for categories of attributes with continuous variables, namely, soil and topography. 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. 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 and it is a sink node, meaning that it does not have any child nodes. This setting makes the causal parents of the signatures their Markov and stable blankets. The Markov blanket of a node consists of its parents, its children, and the parents of its children. Conditioning on the Markov blanket of a node makes the node independent of the rest of the DAG (Pearl, 1988). Since the target variable (runoff signature) has no child nodes, its causal parents are also a stable 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).

PC algorithm assumes that the variables are normally distributed. 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 Monte 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 iterative 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 if they were not initially excluded. 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 that the DAGs resulting from this procedure represent the ground-truth causal links. In the absence of a known ground-truth DAG, the primary means of evaluating these graphs 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. While metrics like the Bayesian Information Criterion (BIC) or Structural Hamming Distance (SHD) could be used if reference DAGs were available, they are not applicable here. Instead, the legitimacy of the inferred graph was assessed based on our expert knowledge, judging the plausibility of the identified causal relationships. Despite these limitations, this iterative procedure produced stable and reproducible results.

The obtained DAG structures are used to predict runoff signatures using Bayesian Network (BN) methods. Additionally, Generalized Additive Models (GAM) and Random Forests (RF) 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).

2.2.4 Bayesian Network (BN)

Having the graph structure from the causal discovery algorithm, the data is fitted to the graph, and the parameters are estimated. Gaussian BN is used for inference purposes. Gaussian BN belongs to the family of continuous BNs, meaning the nodes are continuous variables. The conditional dependencies are linear and follow the joint Gaussian distribution. The prediction is made using averaging likelihood simulation with 500 random sampling numbers. Averaging likelihood simulation is a particle-based approximate method for inference in probabilistic graphical models. This method calculates the weight of samples according to the likelihood of evidence, which is a specific value of the signature of interest. It adds up these weights for each sample (Koller and Friedman, 2009). Since Gaussian BN is limited to capturing only linear relationships, other non-linear prediction methods are also employed in this study, which are explained in the following sections.

2.2.5 Generalized Additive Model (GAM)

The Generalized Additive Model (GAM) model (Hastie et al., 2009) is also chosen to handle non-linear relationships between predictors and runoff signatures. GAMs are extensions of Generalized Linear Models (GLMs), which can identify the linear and nonlinear relationship between response and explanatory variables. This method uses scatterplot smoothers (e.g., smoothing spline or kernel smoother) to fit the additive functions. In this study, the penalized regression spline is used as the smoother. This smoother prevents the model from overfitting where the coefficients of penalized spline decrease (Dubos et al., 2022). The calculation is done using mgcv R package (Wood, 2018). The model predicts the signatures once with all variables derived from feature selection (non-causal model) and once with only the causal parents of the signatures derived from the causal discovery section (causal model).

2.2.6 Random Forest (RF)

The last prediction model used in this study is Random Forest (RF). This method estimates response variables using multiple regression trees. Besides its ability to identify nonlinear patterns in the data, the likelihood of overfitting in RF is low because the model's prediction is an ensemble of multiple predictions. Therefore, it can deliver an accurate prediction with little computational effort. These features in the RF model help identify the issues of linearity and overfitting in BN and GAM models, respectively. The randomForest R package (Breiman, 2018) is used with the number of trees set to 500 to stabilize the prediction (Addor et al., 2018). Similar to GAM, RF is run twice: once using all selected variables as the predictors of the runoff signature (non-causal model) and once using only the causal parents as predictors (causal model).

For all models, BN, GAM, and RF, the environments are divided 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, the training and testing of models are conducted within the same environment, meaning that models trained for a specific environment are tested within that 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.

3 Results

3.1 Clustering results for each category

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. Fig. 2 illustrates each cluster's spatial extent of catchments along with two chosen variables. The obtained results from the cluster analysis for each category of attributes are as follows:

1. **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 members 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.

2. **Soil attributes:** The soil properties data, derived from the State Soil Geographic Database (STATSGO), provides information about the top 2.5 meters of soil. However, the CAMELS dataset only includes soil data for the top 1.5 meters. 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. 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.

3. **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. 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).

4. **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).

5. **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.

These clusters are subsets of the CAMELS dataset with specific properties and different numbers of catchments to be used for runoff signature prediction. They help evaluate the models' performance in different environments, analyse the effect of causal parents as predictors, and assess how the number of data points impacts the training and test simulations.

Table 3. Attribute categories, clustering methods, number of clusters, and catchments per cluster.

No	Category	Method	No. of cluster	No. of Catchments
1	Climate	Gower	4	334, 144, 87, 103
2	Soil	PAM	6	154, 123, 138, 88, 95, 70
3	Topography	PAM	4	282, 119, 117, 90
4	Geology	Gower	7	149, 53, 123, 116, 64, 104, 42
5	Vegetation	Gower	6	89, 131, 149, 69, 105, 128

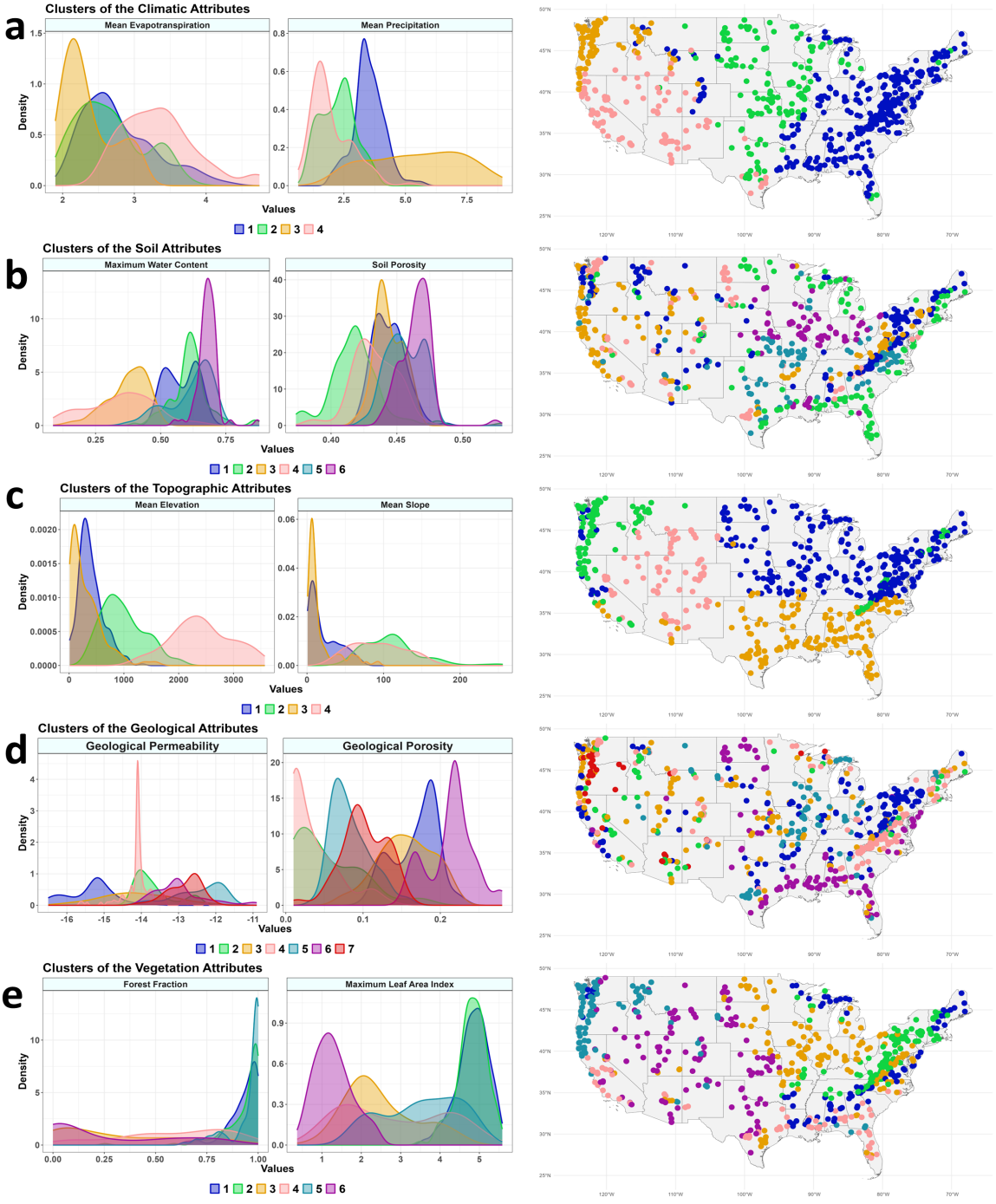

Figure 2. The spatial pattern of clusters (right column) and the density of two variables of its corresponding category (left column). The plots show spatial pattern of a) climate attributes, b) soil attributes, c) topographic attributes, d) geological attributes, and 5) vegetation or landcover attributes.

3.2 Identification of Causal Links

PC algorithm results identify the causal links between all variables, which are selected as the explanatory variables in the feature selection procedure. However, undirected links (edges) can be found in the PC results; therefore, the resulting graphs are usually partially directed. In this case, expert knowledge is used 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 if they were not initially excluded. For example, if PC finds a link between p_mean and frac_forest, the causal link should be an edge from precipitation to forest fraction (p_mean $\rightarrow$ frac_forest). However, the causal direction between climatic or vegetation variables, such as the direction between high precipitation frequency and low precipitation frequency, are not clearly definable and differ from one signature to another. This can be caused by the existence of unobserved variables between climatic or vegetation variables. Therefore, in such cases, the directions are left as determined by the algorithm.

Fig. 3 shows the obtained DAG for the baseflow index. The signature (red node) has five direct causes or parents (yellow nodes). The nodes that form the independent causal mechanism for the baseflow index are shown by the green line. The causal models, $\sim$Par, are trained within the causal mechanism to predict the baseflow index. The causal parents in the independent mechanisms also form the Markov and stable blankets for the baseflow index. The structure and variables of the DAG remain unchanged across all environments; only the values of the variables change across environments. DAGs can show the order in which the variables are connected. For instance, the climate and vegetation variables in Fig. 3 are controlled by topographic attributes, which are gauge latitude, mean elevation, and mean slope. These variables are independent in this DAG since they do not have any parents. It should be noted that the causal parents of the signatures, which are identified by the PC algorithm, are not necessarily the most influential variables derived from correlation and variable importance analysis. Gao et al. (2023) showed that there could be a strong causal relationship between variables with weak statistical associations. The highest correlated variable with a signature can differ across different catchments; however, the causal parents are a set of variables that are always the same and are independent of regions. The selected variables and DAGs for other signatures can be found in the supplementary materials.

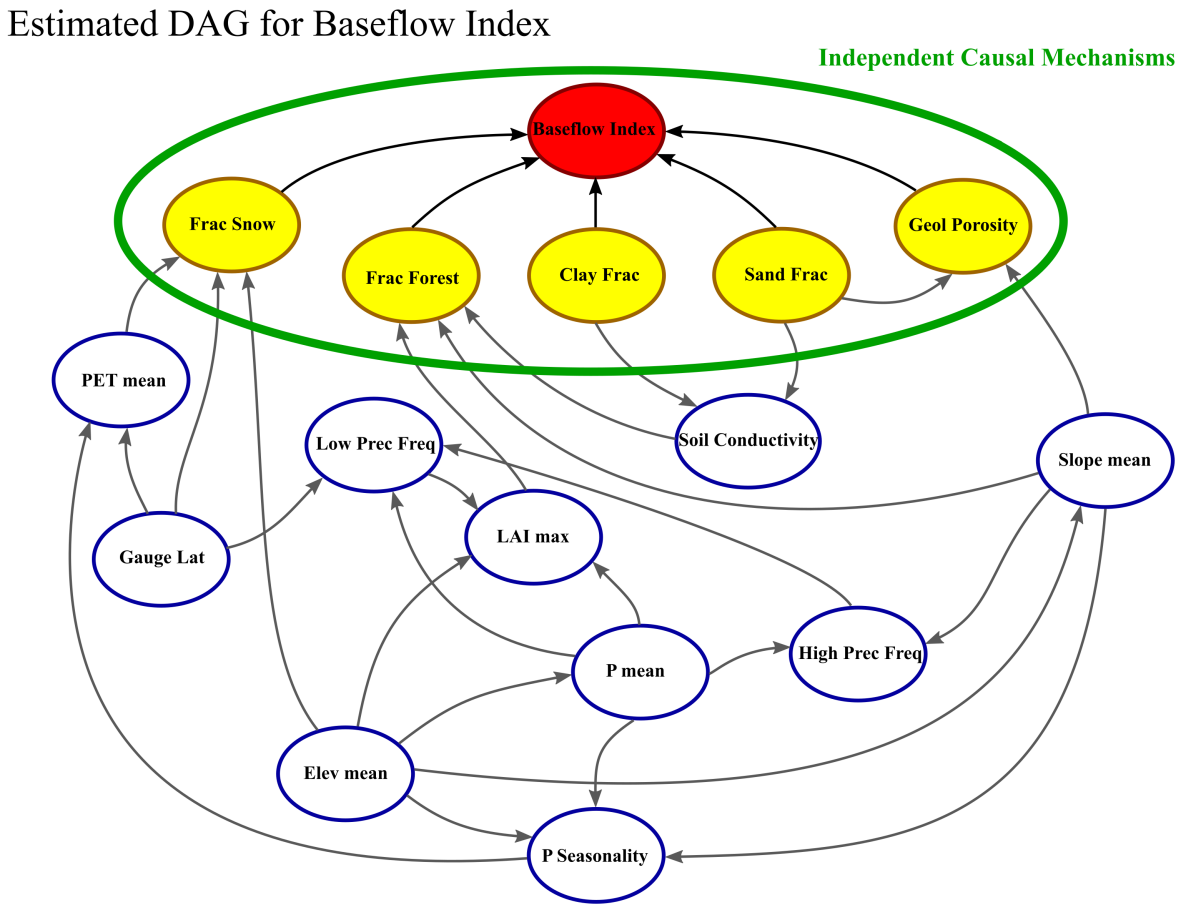

Figure 3. Estimated DAG for the baseflow index obtained from the PC method. Arrows (edges) show the causal links. The red node represents the target runoff signature, and the yellow nodes are the causal parents or direct cause of the target variable. The node variables are explained in Table 1. The red and yellow nodes are the causal mechanism for the baseflow index. The independent causal mechanism for the baseflow index is determined by the green line.

Table 4 shows the causal parents, the number of parents, and the number of all predictors chosen in the feature selection procedure for each runoff signature. The number of parents varies from 2 variables for high flow frequency to 6 for mean flow. We compared the performance of the models using only parents (causal models) to the models using all the selected variables as explanatory variables (non-causal models). The models are executed for the 671 catchments as baseline models and for each cluster as sub-models. The results reveal the models' behaviours in different environments (clusters) compared to the baseline models.

Table 4. Causal parents of the runoff signatures and their numbers derived from the PC algorithm.

Signature	Causal parents	No. of causal parents	No. of selected variables
baseflow_index	frac_snow, sand_frac, clay_frac, frac_forest, geol_porosity	5	15
high_q_dur	p_mean, silt_frac, lai_max	3	15
high_q_freq	low_prec_freq, frac_forest	2	16
low_q_dur	low_prec_dur, max_water_content, lai_diff	3	14
low_q_freq	frac_snow, low_prec_freq, low_prec_dur, frac_forest, geol_porosity	5	15
q_mean	p_mean, p_seasonality, low_prec_freq, area_gages2, frac_forest, geol_porosity	6	13
Q5	p_mean, low_prec_freq, slope_mean	3	15
Q95	p_mean, p_seasonality, low_prec_freq, slope_mean, frac_forest	5	14
runoff_ratio	p_mean, p_seasonality, frac_forest, geol_porosity	4	15
slope_FDC	p_mean, pet_mean, low_prec_freq, lai_max	4	12
stream_elast	high_prec_freq, clay_frac, frac_forest	3	14

3.3 Performance of the baseline models (prediction using the whole dataset)

The models' performance is evaluated according to the value of RMSE, R squared between observation and prediction, and the differences between the training and test results. The obtained results for each signature are shown in Fig. 4, Table A1, and Fig. 5. The results are derived from the simulation using the whole dataset (671 catchments), which we call baseline. Baseline models are considered the most accurate models, in which 75% of the whole dataset is used for training and 25% for test simulation. The training and test sets are randomly sampled 100 times, and models are executed after each sampling. The grey dots in Fig. 4 indicate the simulation results for each model's execution. The simulation for GAM and RF models is done

twice, once using all the predictors, which are shown by GAM∼All and RF∼All (non-causal models), and once using only causal parents as predictors, GAM∼Par and RF∼Par (causal models).

Fig. 4 and Table A1 show that reducing the number of predictors decreases the models' accuracy. Among all models, RF models are the most accurate despite showing the most significant drop in accuracy between training and testing simulations (Fig. 5). The R-squared values from the non-causal RF model (RF∼All), in which all selected variables are used as predictors, are compatible with the results obtained from the study of Addor et al. (2018). Using causal parents for RF simulations (RF∼Par) leads to a greater distance between training and test results compared to using RF∼All for some signatures. These signatures are baseflow index, low flows, runoff ratio, the slope of flow duration curve and streamflow elasticity with 21%, 15%, 39%, 13% and 15% increases in distance, respectively, caused by using causal model (Fig. 5). These differences are less significant for other signatures (less than 7%). Similar to the RF model, the accuracy of GAM models is decreased by reducing the number of predictors from all selected variables to parent variables (Table 4 and Fig. 5). However, unlike RF, the distance between the training and test accuracy in R squared versus RMSE space significantly decreases by using the causal model for the GAM (Fig. 5). This distance decreases from 29% for runoff ratio to 90% for streamflow elasticity (Fig. 5). Finally, BN is the least accurate model in capturing the variance since it is a linear model; however, it shows almost the same R squared and RMSE values in training and testing simulations. As seen in Fig. 5, BN has the shortest distance between training and test compared to the other two models.

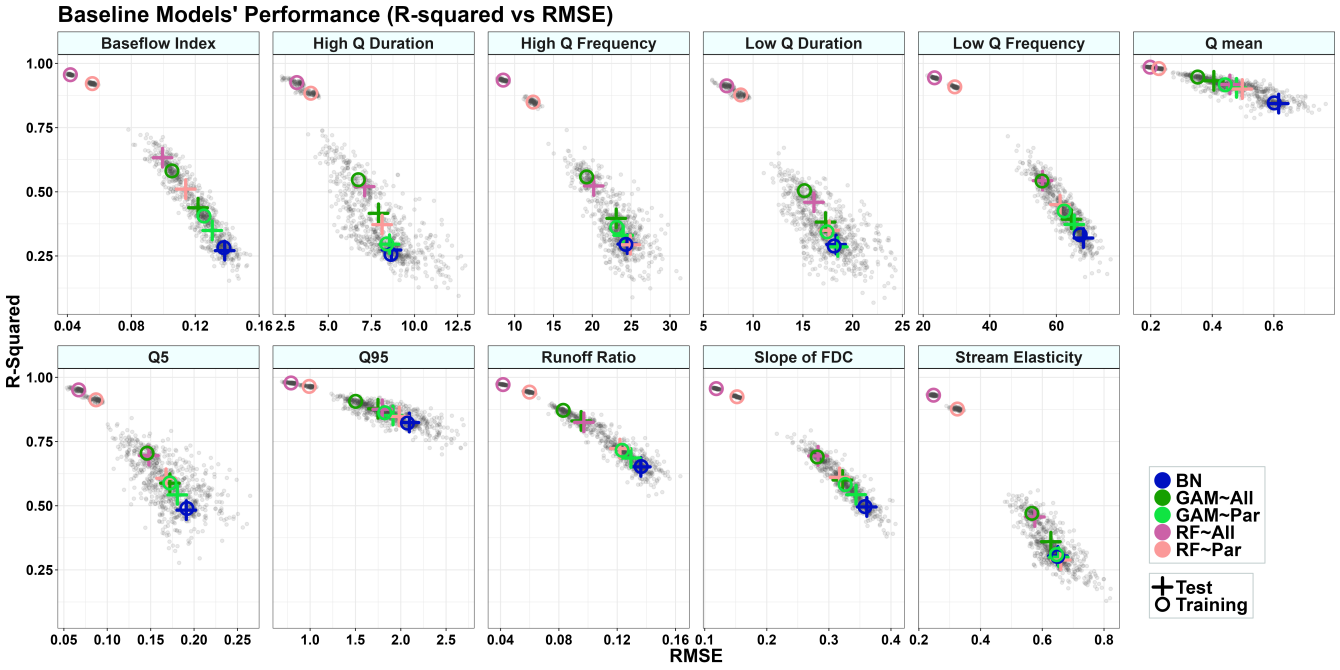

Figure 4. Performance of the models: R-squared vs. RMSE. Each coloured circle and cross represent the centroid of a set of 100 data points (grey dots) generated from the models' execution. Circles indicate the training results, and crosses indicate the test results. In the legend, "All" refers to using all variables as predictors (non-causal model), and "Par" refers to using only parent variables as predictors (causal model). BN refers to the Bayesian Network, GAM refers to the Generalized Additive Model, and RF refers to Random Forest.

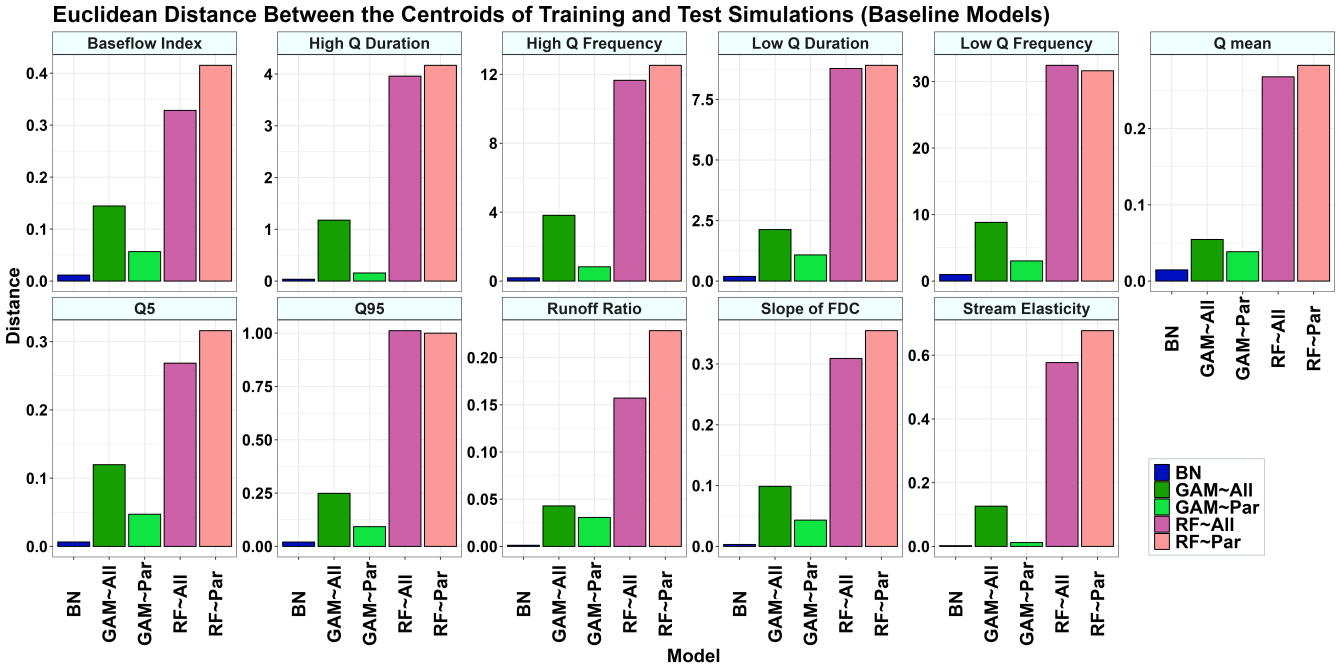

Figure 5. The Eucleadian distance between the centroid points of training and test simulations in Fig. 4. In the legend, "All" refers to using all variables as predictors, and "Par" refers to using only parent variables as predictors. BN refers to the Bayesian Network, GAM refers to the Generalized Additive Model, and RF refers to Random Forest.

We see that when the training set is large, the accuracy of the non-causal models is higher (GAM∼All and RF∼All). However, this pattern might not be the same if the size of the training set is reduced. Testing the models in different environments with different properties and sizes can help us understand how these models perform. In this study, environments are clusters of catchments, defined according to each category of attributes (Table 3) that result in homogeneous hydrological properties. The selected variables for the DAG structure and analysis are assumed to be the same, both with and without clusters. However, in the analysis based on clusters, the model's parameterization and predictions are derived from a smaller subset of data compared to the baseline models. The direct causes of signatures are assumed to be the same across all clusters. Therefore, causal models are assumed to result in robust prediction in different environments. This idea is investigated in the following sections.

3.4 The performance of models across different clusters (Sub-models)

The results of this simulation indicate different models' behaviours across clusters, which are shown in Fig.7, Table 5, and figures in Appendix B. The simulation results for each runoff signature are as follows:

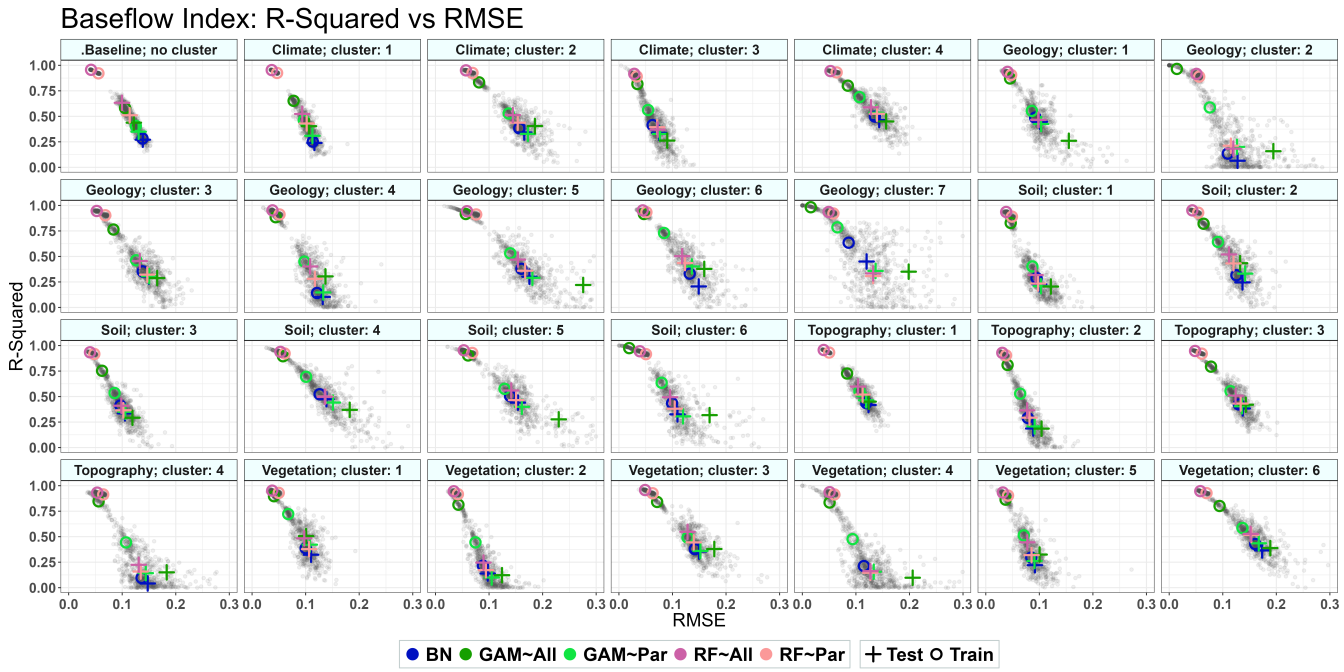

Figure 6. Performance of the models for baseflow index: R-squared vs. RMSE. Each coloured circle and cross represent the centroid of 100 data points (grey dots) generated from the models' execution. Circles indicate the training results, and crosses indicate the test results. In the legend, "All" refers to using all variables as predictors (non-causal model), and "Par" refers to using only parent variables as predictors (causal model). BN refers to the Bayesian Network, GAM refers to the Generalized Additive Model, and RF refers to the Random Forest. The results for other signatures are provided in supplementary materials.

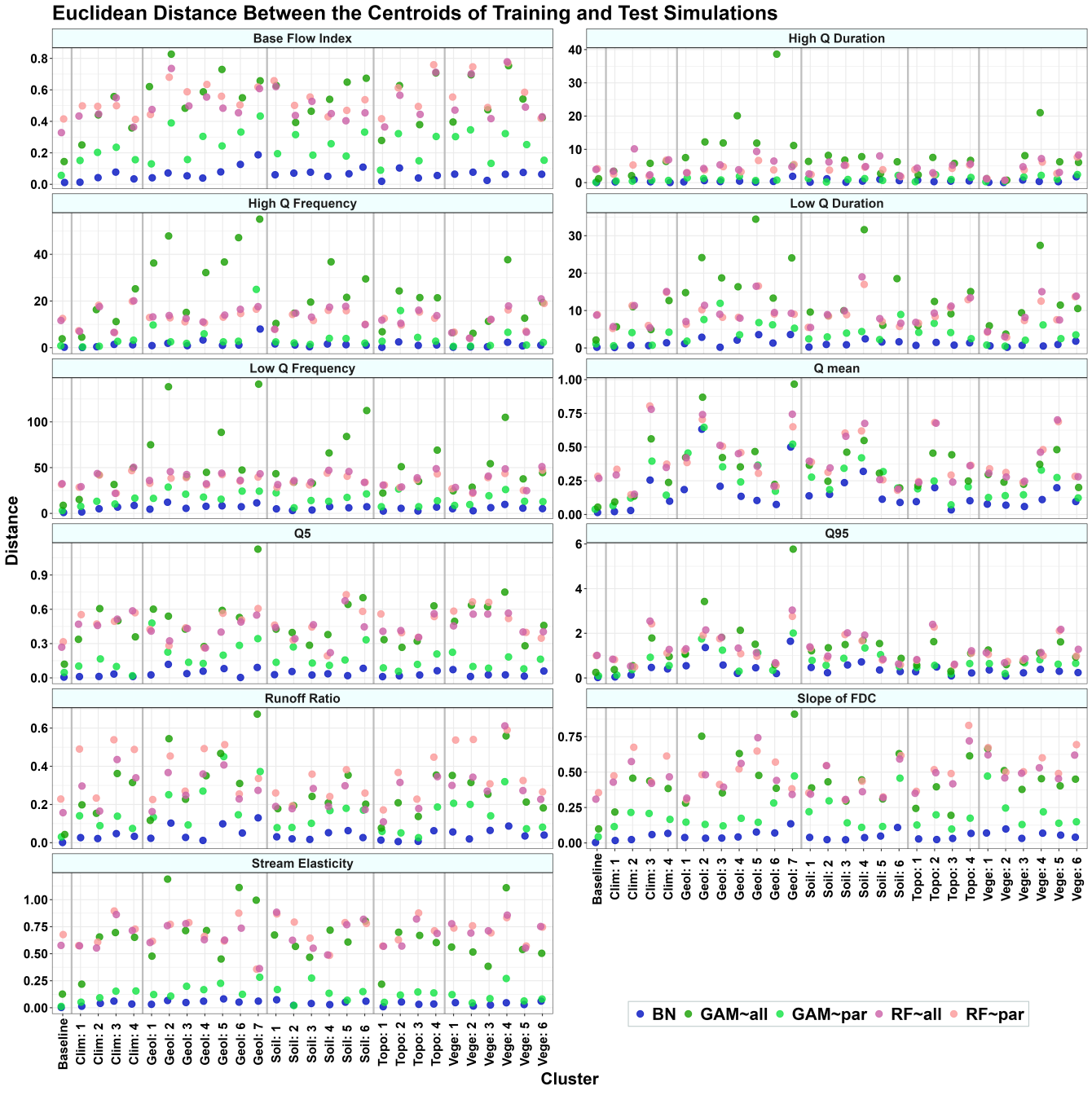

Figure 7. The Eucleadian distance between the training and test simulation for runoff signatures across different environments for each sub-model. In the legend, "All" refers to using all variables as predictors (non-causal model), and "Par" refers to using only parent variables as predictors (causal model). BN refers to the Bayesian Network, GAM refers to the Generalized Additive Model, and RF refers to the Random Forest. On the x-axis, Baseline means simulation without any clustering and is done for all 671 catchments. Clim stands for climate, Geol for geology, Topo for topography and Vege for vegetation. The numbers in front of these names on the x-axis represent the clusters' numbers.

1. **Baseflow Index:** The parents of this signature belong to climate, soil, vegetation and geology categories (Table 4). The obtained DAG, Fig 3, indicates the topographic attributes directly control the snow fraction and indirectly control the forest fraction and geological porosity. It can be seen in Fig. 6 that the models in the topographic and climatic groups perform well compared to the baseline. According to Fig. 6, GAM∼All demonstrates high accuracy in the training set. The distance between training and test for GAM∼Par is lower than GAM∼All in all clusters, and in most cases, the model's accuracy is higher than GAM∼All (Fig. 7). This can be due to the overfitting problem in GAM∼All, which makes the difference between GAM∼All and GAM∼Par insignificant for all environments (Tabls S1). Although RF∼All demonstrates the best performance, in most cases, the difference between the accuracy of the RF∼All and RF∼Par in the test set is negligible, for example, in soil category cluster numbers 1, 3, and 4 (Fig. 6; Table S2). Finally, BN has the lowest distance between training and test (Fig. 7) and in many cases, it outperforms GAM models (Fig. 6). The decrease in R-squared made by causal models is improved from a 20% drop for the baseline model to less than a 5% drop for sub-models (Table 5). The R-squared is increased using parents for GAM in geology, soil and topography categories (Table 5).

2. **High Flow Duration:** This signature has 3 parents belonging to climate, soil, and vegetation categories (Table 4). The obtained DAG shows the parent from the soil category is an independent variable (Fig. S7). The effect of this parent can be seen in Fig. S8, where the highest accuracy of models are among the clusters of soil category, namely Soil Cluster 1, 3, 4 and 6. Since the topographic attributes control the other two parents, namely mean daily precipitation and maximum leaf area index, the topography group of clusters also performed well with small uncertainty (spread of grey dots in Fig. S8) compared to the baseline. GAM∼All shows very high accuracy in the training sets, in some cases higher than random forest, and a significant drop in accuracy in the test sets (Fig. S8). In addition, the distance between training and the test is higher than GAM∼Par in all cases (Fig. 7). The causal GAM models show robust performance for all environments (Table S1). The distance between training and test simulation in RF∼Par is mainly smaller than RF∼All, and in most cases the difference between causal and non-causal RF models are negligible (Table S2). In addition, in Geology Cluster 5, the BN and GAM∼Par perform better than RF∼All. The accuracy difference between causal and non-causal sub-models is significantly smaller than those of baseline models (Table 5).

3. **High Flow Frequency** This signature has only two parents belonging to climate and vegetation categories (Table 4). The obtained DAG (Fig. S11) indicates that topographic attributes influence the causal parents. Models perform well across most clusters based on climate and topography. However, there is no single category within which all models outperform the others (Fig. S12). For instance, the models perform well in Vegetation Cluster 5 (Fig. S12), which are catchments with a high percentage of vegetation cover (Fig. 2). In general, GAM∼All does not show acceptable performance in the test set, and its accuracy in many cases is lower than linear BN (Fig. S12). However, GAM∼Par demonstrate a better performance by reducing the distance between training and test simulations (Fig 7) and increasing accuracy compared to GAM∼All across all clusters (Fig. S12; Table S1). Similarly, RF∼Par decreases the distance between the training and test across most of the clusters, although for the baseline models, this distance is smaller for RF∼All than RF∼Par (Fig.

7). However, the difference between RF∼All and RF∼Par is negligible in only five environments, namely, Geology 4, 5, and 7, Soil 5, and Vegetation 4. For the rest of the environments, RF∼All is more accurate (Table S2). The accuracy of RF∼Par and GAM∼Par models are comparable to RF∼All. Finally, GAM and RF show significantly smaller decreases in R-squared values across categories compared to the baseline models (Table 5). GAM∼Par improved the R-squared by 7.43% and 1.45% in geology and soil categories compared to GAM∼All.

4. **Low Flow Duration:** This signature has 3 parents belonging to climate, soil, and vegetation categories (Table 4). The DAG shows that parents are controlled by the topographic variables (Fig. S15). Training and test simulation performed well across all topographic clusters except for cluster number 4, where catchments have high elevations (Fig. 2 and Fig. S16). The signature also shows high predictability in clusters with high precipitation intensity (Climate Cluster 3) and clusters with low soil porosity (Soil Cluster 2). GAM∼Par performs better in different clusters than GAM∼All by reducing the distance between training and test simulation and increasing the model's accuracy (Table S1). This distance is almost the same across clusters for RF∼Par and RF∼All and, in some cases, smaller for RF∼Par and in most environments, the difference between RF∼Par and RF∼All is not significant (Table S2). The results show that the decrease in R-squared values due to using parents as predictors is significantly lower across categories for GAM and RF (Table 5).

5. **Low Flow Frequency:** This signature has 5 parents, 3 belonging to climate, one to vegetation, and one to geological categories (Table 4). The topographic variables control the causal parents, according to the obtained DAG (Fig. S19). Models perform well across most clusters of climate and topography categories (Fig. S16). In most cases, GAM∼All performs poorly compared to GAM∼Par (Table S1). The difference between training and testing is significantly reduced in GAM∼Par. This distance is also reduced in RF∼Par and, in many cases, performs as well as RF∼All. For example, in Soil Cluster 1, 3, 4, 5 or Geology Cluster 1, 2, 3, and 7, RF∼Par and GAM∼Par perform the same as RF∼All (Table S2). However, in Vegetation Cluster 1 and 2, GAM∼All outperform RF∼Par and GAM∼Par. BN has the smallest difference between training and test simulation. There are smaller drops in accuracy across categories when using parents for GAM and RF (∼Par). The accuracy of GAM∼Par is higher than GAM∼All in the geology, soil and vegetation categories (Table 5).

6. **Mean Daily Runoff:** The parents of the mean daily runoff belong to climate, topography, vegetation and geology categories (Table 4 and Fig. S23). This signature has the highest number of parents among other signatures and is the most predictable runoff signature. All models perform well across all clusters; however, unlike other signatures, BN and GAM models outperform RF in most cases, for example, Geology Cluster 2 (Fig. S24). In most cases, the difference between training and test simulations is smaller when using parents, which shows the benefits of using causal parents. In addition, the difference in model accuracy between simulations using only causal parent (∼Par) and those using all variables (∼All) is negligible across almost all clusters (Table S1 and Table S2). The accuracy is also lower for categories compared to the baseline (Table 5).

7. **Low Flow (Q5):** The parents of low flow belong to climate and topography categories (Table 4). The models' test results are comparable to the baseline models in Geology Cluster 2 and 4 and Soil Cluster 2 and 4 (Fig. S28). GAM∼All is outperformed by GAM∼Par and other models in test simulation (Fig. S28, Table S1). The obtained DAG shows that topographic variables are independent since they have no parents. These variables are also the drivers of the climatic variables. As shown in Fig. S28, models perform well across the topographic category. The difference between training and test simulation is improved in GAM∼Par compared to GAM∼All. This distance for RF∼Par is smaller than RF∼All across half of the clusters (Fig. 7) and the difference between causal and non-causal RF models are negligible for most environments (Table S2). BN has the smallest difference between training and testing, and it outperforms GAM models in Climate Cluster 3, Geology Cluster 6 and 7, Soil Cluster 1 and 4, Topographic Cluster 4 and Vegetation Cluster 6. Using parents as predictors increases the accuracy of GAM in the geology, soil and vegetation categories by 0.93%, 2.03%, and 5.0% (Table 5). The performance difference between RF∼Par and RF∼All is significantly smaller across categories than the baseline (Table 5).

8. **High Flow (Q95):** High flows are among the most identifiable signatures. According to the obtained DAG, high flows are controlled by vegetation (land cover), climate, and topographic variables (Table 4). The models showed high accuracy across all clusters of the topographic categories. Unlike other signatures, the RF∼All and RF∼Par models, which are the most accurate overall, are outperformed by GAM and BN in certain cases (Fig. S31). The difference between training and test simulations is improved in all clusters when using parents for GAM, except for climate cluster 1, and most of the clusters for RF (Table S1). Table 5 indicates that the difference in the models' accuracy is negligible when using causal parents. The R-squared is improved among geology and soil categories for GAM and RF models, where the signature is least predictable (Table 5).

9. **Runoff Ratio:** Runoff ratio has four parents belonging to climate, geology and vegetation categories (Table 4). The obtained DAG indicates that soil variables control the geological porosity (geological parent), and topographic variables control climate and vegetation variables. The models perform well across topographic and soil clusters, and models are more robust across those environments (Fig. S35). Causal models show negligible difference between training and test simulations for almost all clusters for GAM but not for RF (Fig. 5, Table S2). The difference between R-squared values is significantly lower across categories than the baseline models, especially in geology and soil categories (Table 5).

10. **Slope of Flow Duration Curve:** The parents of the slope of the flow duration curve belong to climate and vegetation categories, which, according to the DAG, are controlled by topographic variables (Table 4, Fig. S39). Models in topographic clusters performed well except for Topography Cluster 4, where there are catchments with a high elevation and steep slopes. RF∼Par and GAM∼Par perform almost the same across most of the clusters. In most cases, GAM∼Par reduced the difference between training and test simulations compared to GAM∼All (Table S1). However, this difference for RF∼Par and RF∼All is insignificant for only 8 clusters (Fig. S40, Table S2). However, the accuracy of RF∼Par and RF∼All are comparable in most cases. GAM∼Par performs better than GAM∼All in the geology category by increasing

the R-squared by 2.37% compared to the baseline model (Table 5). RF∼Par shows almost the same accuracy as RF∼All compared to the baseline in the soil category (Table 5).

11. **Stream Precipitation Elasticity:** The three parents of this signature belong to climate, soil and vegetation categories (Table 4). According to the obtained DAG, the topography controls the climate and vegetation parents. However, no dominant category exists where models perform well in all of its clusters. The same as other signatures, GAM∼All performs well in training simulation. However, GAM∼All shows the worst accuracy across the soil and geological clusters compared to the other models and the difference between causal and non-causal models is not significant (Table S1). It can be seen in Table 5, which indicates 5.38% and 8.08% increase in accuracy of GAM using causal parents in the geology and soil categories. The performance of RF∼All, RF∼Par and GAM∼Par are close and comparable in the test simulation (Fig. S43, Table 5, and Table S2). The distance between training and test simulation in GAM∼Par is smaller than GAM∼All. This pattern can be seen in only one-third of the clusters for RF models (Fig. 7).

Table 5. Comparison of R-squared values between causal and non-causal models presented as percentages. Negative values indicate a decrease in R-squared when using causal models compared to non-causal models. The R-squared values for each category are calculated using the weighted mean, with weights based on the proportion of catchments in each cluster relative to the total number of catchments. Red indicates a decrease in R-squared, while blue indicates an increase. The values of R-squared can be found in Table A2 and Table A3.

	Percentage of change in R squared made by using causal parents											
	Baseline		Climate		Geology		Soil		Topography		Vegetation	
Signature	GAM	RF	GAM	RF	GAM	RF	GAM	RF	GAM	RF	GAM	RF
baseflow_index	-20.22	-19.34	-2.39	-3.22	1.92	-2.11	1.43	-2.69	0.41	-4.27	-0.25	-3.12
high_q_dur	-28.89	-28.49	-4.11	-0.01	8.13	-2.03	3.46	-0.66	1.69	-6.52	0.62	-5.19
high_q_freq	-16.63	-43.84	-3.63	-13.34	7.43	-3.43	1.45	-6.59	-1.26	-7.52	-0.71	-8.37
low_q_dur	-25.29	-23.81	-2.28	-6.55	3.47	-2.58	-1.10	-3.97	-1.39	-5.05	-1.80	-4.55
low_q_freq	-5.18	-17.19	-3.25	-6.83	3.31	-1.18	2.5	-2.35	-0.71	-3.38	0.31	-5.13
q_mean	-2.95	-1.92	-2.39	-1.47	-0.29	0.19	-0.57	0.01	0.96	-0.58	-0.70	-0.58
Q5	-7.74	-12.71	-1.26	-4.09	1.99	-1.42	2.03	-1.99	-1.60	-4.75	5.0	-3.57
Q95	-2.97	-3.20	-3.28	-2.05	0.93	0.39	0.48	-0.09	-0.40	-0.57	-1.11	-0.66
runoff_ratio	-17.40	-12.40	-8.37	-7.32	-1.27	-2.07	-1.63	-2.06	-1.72	-3.03	-3.07	-3.31
slope_FDC	-9.50	-12.04	-2.88	-5.61	2.37	-0.91	-0.07	-1.48	-3.65	-4.80	-1.33	-3.30
stream_elast	-16.76	-36.78	-0.16	-6.24	5.38	-3.98	8.08	-5.16	-1.76	-7.67	3.45	-3.70

Fig. 8 displays the rankings of the overall performance of models across different environments for all signatures. RF∼All achieved the highest overall accuracy in the baseline mode where the whole dataset is used. The performance rankings of RF∼Par generally align with those of RF∼All across most clusters, with the exception of Climate 1 and Soil 5. GAM∼Par follows the same pattern as RF∼All except for clusters Climate 1 and 3, Geology 2 and Soil 5. The difference between the

rankings of the models for Climate 1 is the most significant for all models. The similar behaviour of causal models and RF∼All across clusters, particularly in the topography category, suggests that causal patterns as predictors perform comparably to using all variables as predictors. The model GAM∼All, despite experiencing overfitting in most clusters, shows strong performance across clusters Soil 2, 3, 4 and Topography 1, 2 and 3. Although BN model has a linear structure with the lowest accuracy in most cases, it follows the same behaviours as other models. In general, BN, GAM∼Par, RF∼Par, and RF∼All follow a similar ranking pattern; however, GAM∼All exhibits slightly different behaviour.

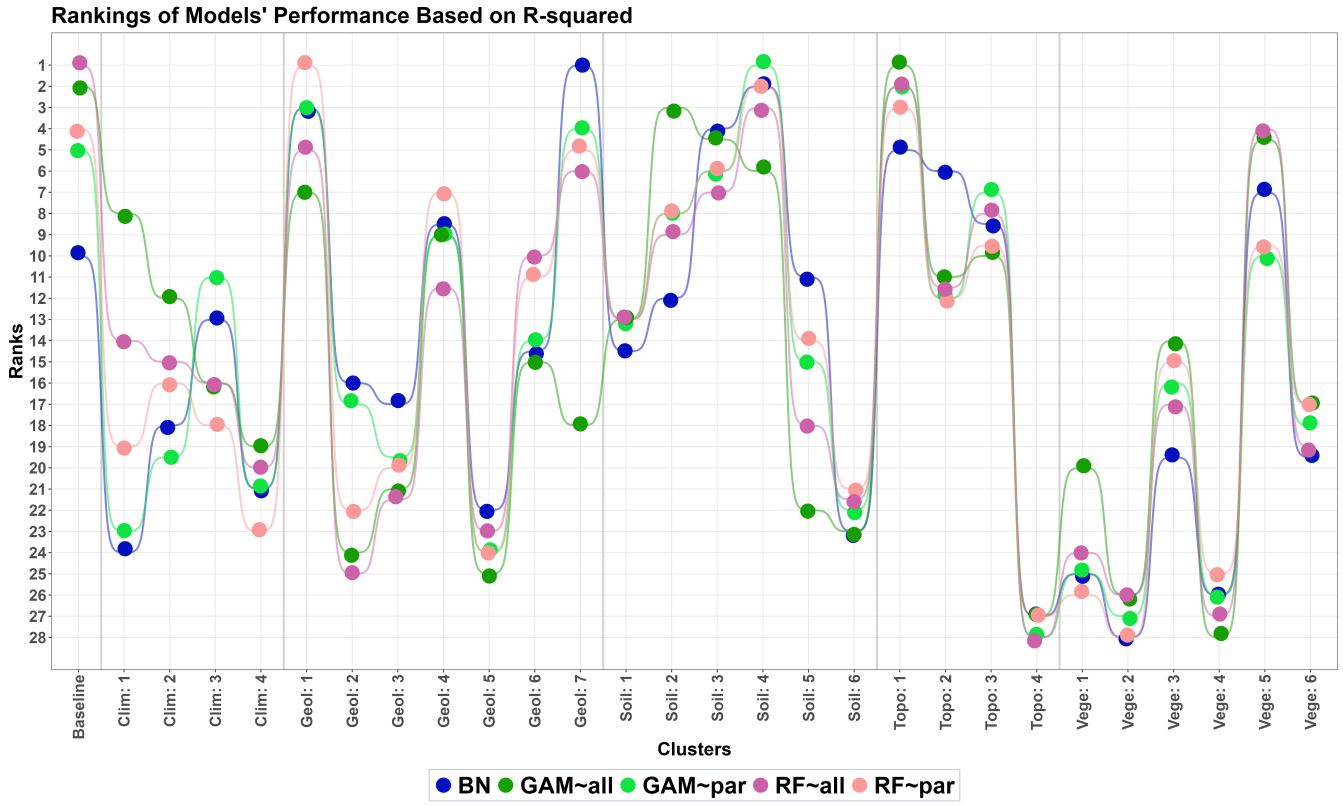

Figure 8. Rankings of model performance based on R-squared values obtained from evaluating their accuracy in predicting all signatures within each cluster. On the x-axis, Clim stands for climate, Geol for geology, Topo for topography and Vege for vegetation.

4 Discussion

For most runoff signatures, the Directed Acyclic Graphs (DAGs) indicate that topographic variables drive climate and vegetation and, in some cases, geological and soil variables. Also, they show that climate attributes influence all runoff signatures, a finding supported by various studies (E.g. Jehn et al. (2020); McMillan et al. (2022)). Models perform well across topographic clusters for most signatures with consistent accuracy rankings (Fig. 8). However, in Topography Cluster 4—characterized by high elevation, steep slopes, and low precipitation—all models struggle to predict signatures accurately. This issue aligns with

Viglione et al. (2013), who observed a decline in prediction model performance in arid catchments. Signatures prove to be more predictable in clusters characterized by high precipitation and low elevation, such as those in Climates 1 and 3. This indicates that even in catchments with low precipitation, the transfer of information from precipitation to runoff remains the predominant driver compared to other mechanisms (Neri et al., 2022). According to Fig. 8, models achieve high accuracy scores in regions with high precipitation, such as Geology 7, Topography 1, Soil 4, and Vegetation 5. The prediction results indicate that independent variables derived from causal discovery, such as topographic variables, can serve as effective criteria for catchment classification. Furthermore, the causal interconnections identified by DAGs improve model accuracy, reduce prediction uncertainty, and increase consistency between training and test simulations.

For the GAM model, the difference in accuracy between causal models, trained within the independent causal mechanism, and non-causal models, trained using all selected variables, is not significant across all runoff signatures. For the RF models, this difference is also insignificant for half of the signatures. The baseflow index, high flow frequency, runoff ratio, and the slope of the flow duration curve are the signatures that non-causal RF models outperform the causal models. For signatures where the difference is insignificant, using causal parents can enhance model parsimony by reducing the number of predictors, improve robustness by maintaining accuracy across environments comparable to non-causal models, and minimize accuracy reduction between the training and testing phases.

The causal parents identified by the PC algorithm align with the underlying physical processes for most of the signatures. For example, according to PC results, snow fraction drives the baseflow index and low flows, consistent with runoff-generating mechanisms during spring and summer (Gentile et al., 2023). In addition, vegetation, soil, and geological variables, which contribute to infiltration and groundwater flow, are causal parents of the baseflow index (Gnann et al., 2019). For high flows (Q95), drivers include precipitation seasonality, vegetation cover, mean precipitation, and slope. This suggests that precipitation intensity, often driven by seasonality, influences runoff-generating mechanisms like infiltration excess process (Nanda et al., 2019). Slope and vegetation cover also affect the time concentration and the magnitude of high flows in the catchment area (Sultan et al., 2022). In regions with high mean precipitation and low seasonality, saturation excess runoff mechanisms dominate high flows. However, the PC causal discovery results for low flows (Q5) were expected to identify geological variables as being important. Low flows are strongly governed by geological variables in addition to climate and topography (Laaha and Bloeschl, 2006; Giuntoli et al., 2013). The algorithm fails to identify any geological variables as a causal parent for this signature, likely due to the non-linearity of the low-flow process or limited sample size for the causal discovery.

The results of the baseline models indicate that the RF model is the most accurate, followed by GAM and BN. This finding is consistent with Pourghasemi and Rahmati (2018), who demonstrated the RF model's superiority over GAM when analyzing landslide causal factors. Reducing the number of predictors to causal parents decreases the accuracy of the baseline models. Although the models are expected to perform similarly across different environments, the results reveal significant uncertainty in the test simulations, primarily due to the smaller training set sizes across clusters. As emphasized by Riley et al. (2020), sample size plays a critical role in determining the accuracy and robustness of prediction models. In cases where the sample size is smaller than that of baseline models, non-causal models often fail to outperform their causal counterparts.

Despite BN having lower accuracy than GAM and RF, it shows the smallest difference between training and test results across all cases. This consistency may be due to the BN structure, which relies on conditional dependencies derived from the causal relationships between variables, although further investigation is needed. GAM models show completely different behaviour compared to the baseline when applied to clusters. Although GAM∼All is among the most accurate models in the training simulation, its test results have significantly lower accuracy than other models, likely due to overfitting when the training sample size is small. In contrast, GAM∼Par performs better, with higher accuracy and reduced uncertainty, suggesting that using causal parents makes GAM more robust across various environments. RF∼All display the highest accuracy among all models. However, for some signatures like high flow duration, it is outperformed by RF∼Par across most clusters. Additionally, for highly predictable signatures like mean daily flows and high flows, GAM and BN perform better than RF.

The high accuracy of GAM∼All and RF∼All in baseline models may result from the large number of data points in the training sets. However, with smaller sample sizes, the performance difference between causal and non-causal models becomes negligible. In some cases, models using causal parents even achieve higher accuracy, such as predicting high flow duration in the geology category for GAM (Table 5).

Finally, our results show that causal discovery enhances the representation of physical systems, making models more interpretable and parsimonious, as emphasized by Runge et al. (2019a) and Reichstein et al. (2019). The insights gained from causal interconnections not only improve the understanding of hydrological systems but also lead to more informed modelling practice (Slater et al., 2024). However, we still need theoretical developments to quantify the stability and robustness of uncertainty of such a model, particularly when combined with machine learning and classification algorithms (Herman et al., 2015; AghaKouchak et al., 2022; Singh et al., 2015).

5 Conclusions

This study investigates the application of causal discovery to represent the causal interconnections between variables in hydrological systems. The PC algorithm is used to identify the causal links between catchments attributes, climate indices, and 11 runoff signatures, producing a Directed Acyclic Graph (DAG) for each signature. DAGs reveal the connections between variables, including the direct causes (parents) of the target signatures. Three prediction models, BN, GAM, and RF, in five different settings, namely BN, GAM∼Par, GAM∼All, RF∼Par and RF∼All, are used to predict runoff signatures. These models are executed on the entire dataset as well as 27 clusters, with each configuration undergoing 100 random samplings of training and test sets, resulting in a total of 28,000 model executions. BN directly utilizes the DAG structure for prediction, while GAM and RF predict the target variable both by using all the variables in the DAG and by using only the causal parents (the variables that, together with the target variables, form the independent causal mechanism). Each model is run 100 times with random sampling of training and tests for each run. The dataset is then grouped into different clusters based on attribute categories. The clusters serve as new environments to train and test the models, allowing for an assessment of model performance when using causal parents as the explanatory variables. The major outcomes of this research are as follows:

- 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 $P(\text{Effect} \mid \text{Cause})$ 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.

- 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.

- 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.

- Causal inference methods contribute to improving prediction models' parsimony, interoperability and robustness in hydrological systems .

In conclusion, causal models maintain acceptable accuracy across environments with varying distributions of explanatory variables (covariates). The DAGs obtained from causal discovery enhance the interpretability of prediction models and offer more informed clustering criteria, which is valuable for regionalization purposes. This study focuses on investigating the direct causes of runoff signatures and their effects on prediction accuracy, but other criteria for selecting predictors from the DAG variables could be explored. For example, investigating the effect of variables with different topological ordering on the target variable, such as root nodes, ancestors of the target variables, etc. In addition, different causal discovery methods may yield alternative DAG structures, which merit further investigation. This work highlights the importance of causal inference methods in understanding runoff-generating mechanisms in hydrological systems.

While causal inference analysis has been extensively explored in fields such as computer science and medicine, its applications in hydrology are still in their infancy. There is a broad range of potential uses for causal models in hydrology, from identifying the drivers of hydrological anomalies (Tárraga et al., 2024) to linking extreme events with their cascading societal impacts (AghaKouchak et al., 2023). As research in this area progresses, the application of causal inference methods is likely to lead to more accurate and robust predictive models, offering valuable insights into complex hydrological variability.

Code and data availability. The codes are available on the GitHub repository at https://github.com/abbasizadeh/Catchment-Causal-Discovery. The CAMELS attributes are available at https://gdex.ucar.edu/dataset/camels.html.

Appendix A: The values of R squared and RMSE for the baseline models and R squared values for sub-models

A1 R-squared and RMSE values for test simulations of baseline models in Fig. 4

Table A1. R-squared and RMSE values for test simulations of baseline models. The values are an average of 100 executions of each model.

	R squared (Test Set)					RMSE (Test Set)				
Signature	BN	GAM∼All	GAM∼Par	RF∼All	RF∼Par	BN	GAM∼All	GAM∼Par	RF∼All	RF∼Par
baseflow_index	0.27	0.44	0.35	0.63	0.51	0.13	0.12	0.13	0.10	0.11
high_q_dur	0.27	0.42	0.30	0.52	0.37	8.66	7.91	8.55	7.11	8.12
high_q_freq	0.30	0.40	0.33	0.52	0.29	24.51	23.08	23.97	20.18	24.89
low_q_dur	0.29	0.38	0.28	0.46	0.35	18.33	17.27	18.50	16.11	17.64
low_q_freq	0.32	.039	0.37	0.54	0.45	68.13	64.52	65.36	55.84	61.11
q_mean	0.84	0.93	0.90	0.92	0.90	0.62	0.41	0.48	0.46	0.50
Q5	0.48	0.59	0.54	0.70	0.61	0.19	0.17	0.18	0.15	0.17
Q95	0.82	0.88	0.85	0.88	0.85	2.09	1.75	1.91	1.80	1.98
runoff_ratio	0.65	0.83	0.69	0.82	0.72	0.14	0.10	0.13	0.10	0.12
slope_FDC	0.50	0.60	0.54	0.69	0.61	0.36	0.32	0.34	0.28	0.32
stream_elast	0.30	0.36	0.30	0.46	0.29	0.65	0.63	0.65	0.58	0.66

A2 R squared values used to calculate values in Table 5

Table A2. The R-squared values of causal models for each category which is calculated using the weighted mean. The weights are the ratio of the catchments in each cluster to the total number of catchments.

	R squared values for causal models														
	Climate			Geology			Soil			Topography			Vegetation		
Signature	BN	GAM	RF	BN	GAM	RF	BN	GAM	RF	BN	GAM	RF	BN	GAM	RF
baseflow_index	0.31	0.35	**0.44**	0.28	0.31	0.36	0.33	0.33	0.38	0.32	0.36	0.41	0.27	0.30	0.35
high_q_dur	0.24	0.26	0.35	0.34	0.38	**0.39**	0.35	0.35	0.38	0.23	0.30	0.30	0.23	0.28	0.29
high_q_freq	0.21	0.24	0.20	0.32	**0.34**	0.32	0.25	0.29	0.24	0.30	0.31	0.32	0.19	0.22	0.19
low_q_dur	0.23	0.26	0.27	0.38	0.37	0.36	0.29	0.29	0.29	0.43	0.39	**0.40**	0.20	0.23	0.24
low_q_freq	0.23	0.28	0.32	0.33	0.31	0.35	0.34	0.32	0.35	0.37	0.37	**0.42**	0.21	0.24	0.25
q_mean	0.73	0.74	0.72	0.82	0.83	0.81	**0.86**	**0.86**	**0.86**	0.83	0.85	0.82	0.77	0.78	0.74
Q5	0.27	0.31	0.37	0.42	0.45	**0.49**	0.40	0.45	0.47	0.35	0.37	0.43	0.31	0.34	0.36
Q95	0.62	0.62	0.60	0.78	0.77	0.74	**0.81**	0.80	0.80	0.79	**0.81**	0.78	0.68	0.67	0.64
runoff_ratio	0.34	0.42	0.48	0.61	0.58	0.60	0.65	0.66	0.67	0.69	**0.70**	0.69	0.49	0.53	0.52
slope_FDC	0.31	0.37	0.41	0.46	0.46	0.49	0.48	**0.52**	0.55	0.41	0.43	0.47	0.27	0.33	0.37
stream_elast	0.30	**0.32**	0.31	0.28	0.27	0.27	0.25	0.23	0.22	0.26	0.28	0.27	0.25	0.26	0.23

Table A3. The Rsquared values of non-causal models for each category which is calculated using the weighted mean. The weights are the ratio of the catchments in each cluster to the total number of catchments.

	R squared values for non-causal models									
	Climate		Geology		Soil		Topography		Vegetation	
Signature	GAM	RF	GAM	RF	GAM	RF	GAM	RF	GAM	RF
baseflow_index	0.39	0.51	0.28	0.43	0.31	0.45	0.35	**0.48**	0.31	0.42
high_q_dur	0.33	0.37	0.26	**0.46**	0.30	0.43	0.27	0.39	0.26	0.38
high_q_freq	0.31	0.43	0.24	0.41	0.26	0.38	0.34	**0.45**	0.23	0.37
low_q_dur	0.29	0.37	0.28	0.41	0.30	0.37	0.41	**0.49**	0.24	0.31
low_q_freq	0.33	0.43	0.27	0.40	0.28	0.40	0.38	**0.50**	0.25	0.37
q_mean	0.81	0.76	0.85	0.80	**0.89**	0.86	0.82	0.84	0.81	0.77
Q5	0.33	0.45	0.41	**0.54**	0.41	0.51	0.42	**0.54**	0.31	0.45
Q95	0.70	0.65	0.74	0.73	0.78	0.81	**0.83**	0.79	0.71	0.67
runoff_ratio	0.63	0.66	0.65	0.69	0.74	0.76	0.76	**0.78**	0.64	0.65
slope_FDC	0.42	0.52	0.42	0.54	0.51	**0.60**	0.48	0.55	0.36	0.45
stream_elast	0.31	**0.39**	0.22	0.36	0.23	0.32	0.28	0.36	0.24	0.30

Author contributions. **Hossein Abbasizadeh**: Conceptualization, Methodology, Coding and computation, Data visualization, Writing – review & editing, Funding acquisition. **Petr Maca**: Supervision, Conceptualization, Methodology, Writing – review & editing, Funding acquisition; **Martin Hanel**: Methodology, Writing – review & editing, Funding acquisition; **Mads Troldborg**: Conceptualization, Methodology, Writing – review & editing; **Amir AghaKouchak**: Conceptualization, Methodology, Writing – review & editing.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This study was supported by the Internal Grant Agency of the Faculty of Environmental Sciences, Czech University of Life Sciences Prague (project No.2023B0026 and No.2024B0003) and the Ministry of Education, Youth and Sports of the Czech Republic (grant AdAgriF - Advanced methods of greenhouse gases emission reduction and sequestration in agriculture and forest landscape for climate change mitigation (CZ.02.01.01/00/22_008/0004635)).

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.

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., and Clark, M. P.: A Ranking of Hydrological Signatures Based on Their Predictability in Space, WATER RESOURCES RESEARCH, 54, 8792–8812, https://doi.org/10.1029/2018WR022606, 2018.

AghaKouchak, A., Pan, B., Mazdiyasni, O., Sadegh, M., Jiwa, S., Zhang, W., Love, C., Madadgar, S., Papalexiou, S., Davis, S., et al.: Status and prospects for drought forecasting: Opportunities in artificial intelligence and hybrid physical–statistical forecasting, Philosophical Transactions of the Royal Society A, 380, 20210 288, 2022.

AghaKouchak, A., Huning, L. S., Sadegh, M., Qin, Y., Markonis, Y., Vahedifard, F., Love, C. A., Mishra, A., Mehran, A., Obringer, R., et al.: Toward impact-based monitoring of drought and its cascading hazards, Nature Reviews Earth & Environment, 4, 582–595, 2023.

Aguilera, P. A., Fernandez, A., Fernandez, R., Rumi, R., and Salmeron, A.: Bayesian networks in environmental modelling, ENVIRONMENTAL MODELLING & SOFTWARE, 26, 1376–1388, https://doi.org/10.1016/j.envsoft.2011.06.004, 2011.

Arshad, A., Mirchi, A., et al.: Downscaled-GRACE data reveal anthropogenic and climate-induced water storage decline across the Indus Basin, Water Resources Research, 60, e2023WR035 882, 2024.

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.

Breiman, L.: Random forests, Machine learning, 45, 5–32, 2001.

Breiman, L.: randomForest: Breiman and Cutler's Random Forests for Classification and Regression, R package version, 4, 14, 2018.

Chagas, V. B. P., Chaffe, P. L. B., and Bloeschl, G.: Regional Low Flow Hydrology: Model Development and Evaluation, WATER RESOURCES RESEARCH, 60, https://doi.org/10.1029/2023WR035063, 2024.

Ciulla, F. and Varadharajan, C.: A network approach for multiscale catchment classification using traits, HYDROLOGY AND EARTH SYSTEM SCIENCES, 28, 1617–1651, https://doi.org/10.5194/hess-28-1617-2024, 2024.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, WATER RESOURCES RESEARCH., 47, https://doi.org/10.1029/2010WR009827, 2011.

Clausen, B. and Biggs, B.: Flow variables for ecological studies in temperate streams: groupings based on covariance, JOURNAL OF HYDROLOGY, 237, 184–197, https://doi.org/10.1016/S0022-1694(00)00306-1, 2000.

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

Delforge, D., de Viron, O., Vanclooster, M., Van Camp, M., and Watlet, A.: Detecting hydrological connectivity using causal inference from time series: synthetic and real karstic case studies, HYDROLOGY AND EARTH SYSTEM SCIENCES, 26, 2181–2199, https://doi.org/10.5194/hess-26-2181-2022, 2022.

Deng, J., Shan, K., Shi, K., Qian, S. S., Zhang, Y., Qin, B., and Zhu, G.: Nutrient reduction mitigated the expansion of cyanobacterial blooms caused by climate change in Lake Taihu according to Bayesian network models, WATER RESEARCH, 236, https://doi.org/10.1016/j.watres.2023.119946, 2023.

Desai, S. and Ouarda, T. B. M. J.: Regional hydrological frequency analysis at ungauged sites with random forest regression, JOURNAL OF HYDROLOGY, 594, https://doi.org/10.1016/j.jhydrol.2020.125861, 2021.

Dubos, V., Hani, I., Ouarda, T. B. M. J., and St-Hilaire, A.: Short-term forecasting of spring freshet peak flow with the Generalized Additive model, JOURNAL OF HYDROLOGY, 612, https://doi.org/10.1016/j.jhydrol.2022.128089, 2022.

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.

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

Ficchi, A., Perrin, C., and Andreassian, V.: Hydrological modelling at multiple sub-daily time steps: Model improvement via flux-matching, JOURNAL OF HYDROLOGY, 575, 1308–1327, https://doi.org/10.1016/j.jhydrol.2019.05.084, 2019.

Gao, B., Yang, J., Chen, Z., Sugihara, G., Li, M., Stein, A., Kwan, M.-P., and Wang, J.: Causal inference from cross-sectional earth system data with geographical convergent cross mapping, NATURE COMMUNICATIONS, 14, https://doi.org/10.1038/s41467-023-41619-6, 2023.

Geiger, D. and Heckerman, D.: Learning gaussian networks, in: Uncertainty in Artificial Intelligence, pp. 235–243, Elsevier, 1994.

Gentile, A., Canone, D., Ceperley, N., Gisolo, D., Previati, M., Zuecco, G., Schaefli, B., and Ferraris, S.: Towards a conceptualization of the hydrological processes behind changes of young water fraction with elevation: a focus on mountainous alpine catchments, HYDROLOGY AND EARTH SYSTEM SCIENCES, 27, 2301–2323, https://doi.org/10.5194/hess-27-2301-2023, 2023.

Giuntoli, I., Renard, B., Vidal, J. P., and Bard, A.: Low flows in France and their relationship to large-scale climate indices, JOURNAL OF HYDROLOGY, 482, 105–118, https://doi.org/10.1016/j.jhydrol.2012.12.038, 2013.

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.

Gnann, S. J., Woods, R. A., and Howden, N. J. K.: Is There a Baseflow Budyko Curve?, WATER RESOURCES RESEARCH, 55, 2838–2855, https://doi.org/10.1029/2018WR024464, 2019.

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

Guzha, A. C., Rufino, M. C., Okoth, S., Jacobs, S., and Nobrega, R. L. B.: Impacts of land use and land cover change on surface runoff, discharge and low flows: Evidence from East Africa, JOURNAL OF HYDROLOGY-REGIONAL STUDIES, 15, 49–67, https://doi.org/10.1016/j.ejrh.2017.11.005, 2018.

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.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H.: The elements of statistical learning: data mining, inference, and prediction, vol. 2, Springer, 2009.

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.

Herman, J. D., Reed, P. M., Zeff, H. B., and Characklis, G. W.: How should robustness be defined for water systems planning under change?, Journal of Water Resources Planning and Management, 141, 04015 012, 2015.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H. H. G., and Gascuel-Odoux, C.: Process consistency in models: The importance of system signatures, expert knowledge, and process complexity, WATER RESOURCES RESEARCH., 50, 7445–7469, https://doi.org/10.1002/2014WR015484, 2014.

Jackson-Blake, L. A., Clayer, F., Haande, S., Sample, J. E., and Moe, S. J.: Seasonal forecasting of lake water quality and algal bloom risk using a continuous Gaussian Bayesian network, HYDROLOGY AND EARTH SYSTEM SCIENCES, 26, 3103–3124, https://doi.org/10.5194/hess-26-3103-2022, 2022.

Jehn, F. U., Bestian, K., Breuer, L., Kraft, P., and Houska, T.: Using hydrological and climatic catchment clusters to explore drivers of catchment behavior, HYDROLOGY AND EARTH SYSTEM SCIENCES, 24, 1081–1100, https://doi.org/10.5194/hess-24-1081-2020, 2020.

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

Koller, D. and Friedman, N.: Probabilistic graphical models: principles and techniques, MIT press, 2009.

Kretschmer, M., Coumou, D., Donges, J. F., and Runge, J.: Using Causal Effect Networks to Analyze Different Arctic Drivers of Midlatitude Winter Circulation, JOURNAL OF CLIMATE, 29, 4069–4081, https://doi.org/10.1175/JCLI-D-15-0654.1, 2016.

Kuentz, A., Arheimer, B., Hundecha, Y., and Wagener, T.: Understanding hydrologic variability across Europe through catchment classification, HYDROLOGY AND EARTH SYSTEM SCIENCES, 21, 2863–2879, https://doi.org/10.5194/hess-21-2863-2017, 2017.

Laaha, G. and Bloeschl, G.: A comparison of low flow regionalisation methods -: catchment grouping, JOURNAL OF HYDROLOGY, 323, 193–214, https://doi.org/10.1016/j.jhydrol.2005.09.001, 2006.

Ladson, A. R., Brown, R., Neal, B., and Nathan, R.: A standard approach to baseflow separation using the Lyne and Hollick filter, AUSTRALASIAN JOURNAL OF WATER RESOURCES, 17, 25–34, https://doi.org/10.7158/W12-028.2013.17.1, 2013.

Ley, R., Casper, M. C., Hellebrand, H., and Merz, R.: Catchment classification by runoff behaviour with self-organizing maps (SOM), HYDROLOGY AND EARTH SYSTEM SCIENCES, 15, 2947–2962, https://doi.org/10.5194/hess-15-2947-2011, 2011.

Li, C. and Mahadevan, S.: Efficient approximate inference in Bayesian networks with continuous variables, RELIABILITY ENGINEERING & SYSTEM SAFETY, 169, 269–280, https://doi.org/10.1016/j.ress.2017.08.017, 2018.

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

Love, C. A., Skahill, B. E., England, J. F., and Karlovits, e. a.: Integrating Climatic and Physical Information in a Bayesian Hierarchical Model of Extreme Daily Precipitation, Water, 12, 2211, 2020.

Marcot, B. G. and Penman, T. D.: Advances in Bayesian network modelling: Integration of modelling technologies, ENVIRONMENTAL MODELLING & SOFTWARE, 111, 386–393, https://doi.org/10.1016/j.envsoft.2018.09.016, 2019.

Matos, A. C. d. S. and Oliveira e Silva, F. E.: Bayesian estimation of hydrological model parameters in the signature-domain: Aiming for a regional approach, JOURNAL OF HYDROLOGY, 639, https://doi.org/10.1016/j.jhydrol.2024.131554, 2024.

McMillan, H.: Linking hydrologic signatures to hydrologic processes: A review, HYDROLOGICAL PROCESSES, 34, 1393–1409, https://doi.org/10.1002/hyp.13632, 2020.

McMillan, H. K., Gnann, S. J., and Araki, R.: Large Scale Evaluation of Relationships Between Hydrologic Signatures and Processes, WATER RESOURCES RESEARCH, 58, https://doi.org/10.1029/2021WR031751, 2022.

Nanda, A., Sen, S., and McNamara, J. P.: How spatiotemporal variation of soil moisture can explain hydrological connectivity of infiltration-excess dominated hillslope: Observations from lesser Himalayan landscape, JOURNAL OF HYDROLOGY, 579, https://doi.org/10.1016/j.jhydrol.2019.124146, 2019.

Neri, M., Coulibaly, P., and Toth, E.: Similarity of catchment dynamics based on the interaction between streamflow and forcing time series: Use of a transfer entropy signature, JOURNAL OF HYDROLOGY, 614, https://doi.org/10.1016/j.jhydrol.2022.128555, 2022.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, HYDROLOGY AND EARTH SYSTEM SCIENCES, 19, 209–223, https://doi.org/10.5194/hess-19-209-2015, 2015.

Nguyen, T.-T., Huu, Q. N., and Li, M. J.: Forecasting Time Series Water Levels on Mekong River Using Machine Learning Models, in: 2015 Seventh International Conference on Knowledge and Systems Engineering (KSE), pp. 292–297, https://doi.org/10.1109/KSE.2015.53, 2015.

Nojavan, F. A., Qian, S. S., and Stow, C. A.: Comparative analysis of discretization methods in Bayesian networks, ENVIRONMENTAL MODELLING & SOFTWARE, 87, 64–71, https://doi.org/10.1016/j.envsoft.2016.10.007, 2017.

Olden, J. and Poff, N.: Redundancy and the choice of hydrologic indices for characterizing streamflow regimes, RIVER RESEARCH AND APPLICATIONS, 19, 101–121, https://doi.org/10.1002/rra.700, 2003.

Olden, J. D., Kennard, M. J., and Pusey, B. J.: A framework for hydrologic classification with a review of methodologies and applications in ecohydrology, ECOHYDROLOGY, 5, 503–518, https://doi.org/10.1002/eco.251, 2012.

Ombadi, M.: Causal Inference, Nonlinear Dynamics, and Information Theory Applications in Hydrometeorological Systems, University of California, Irvine, 2021.

Ombadi, M., Nguyen, P., Sorooshian, S., and Hsu, K.-l.: Evaluation of Methods for Causal Discovery in Hydrometeorological Systems, WATER RESOURCES RESEARCH, 56, https://doi.org/10.1029/2020WR027251, 2020.

Ouali, D., Chebana, F., and Ouarda, T. B. M. J.: Fully nonlinear statistical and machine-learning approaches for hydrological frequency estimation at ungauged sites, JOURNAL OF ADVANCES IN MODELING EARTH SYSTEMS, 9, 1292–1306, https://doi.org/10.1002/2016MS000830, 2017.

Ouarda, T. B. M. J., Charron, C., Hundecha, Y., St-Hilaire, A., and Chebana, F.: Introduction of the GAM model for regional low-flow frequency analysis at ungauged basins and comparison with commonly used approaches, ENVIRONMENTAL MODELLING & SOFTWARE, 109, 256–271, https://doi.org/10.1016/j.envsoft.2018.08.031, 2018.

Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., and Scholkopf, B.: Learning Independent Causal Mechanisms, in: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL 80, edited by Dy, J. and Krause, A., vol. 80 of *Proceedings of Machine Learning Research*, ISSN 2640-3498, 35th International Conference on Machine Learning (ICML), Stockholm, SWEDEN, JUL 10-15, 2018, 2018.

Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference, Elsevier, 1988.

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

Pearl, J., Glymour, M., and P.Jewell, N.: Causal inference in statistics: a primer, John Wiley & Sons Ltd, 2016.

Perez-Suay, A. and Camps-Valls, G.: Causal Inference in Geoscience and Remote Sensing From Observational Data, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, 57, 1502–1513, https://doi.org/10.1109/TGRS.2018.2867002, 2019.

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.

Pizarro, A. and Jorquera, J.: Advancing objective functions in hydrological modelling: Integrating knowable moments for improved simulation accuracy, JOURNAL OF HYDROLOGY, 634, https://doi.org/10.1016/j.jhydrol.2024.131071, 2024.

Pokhrel, P., Yilmaz, K. K., and Gupta, H. V.: Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures, JOURNAL OF HYDROLOGY, 418, 49–60, https://doi.org/10.1016/j.jhydrol.2008.12.004, 2012.

Pourghasemi, H. R. and Rahmati, O.: Prediction of the landslide susceptibility: Which algorithm, which precision?, CATENA, 162, 177–192, https://doi.org/10.1016/j.catena.2017.11.022, 2018.

Qian, S. S. and Miltner, R. J.: A continuous variable Bayesian networks model for water quality modeling: A case study of setting nitrogen criterion for small rivers and streams in Ohio, USA, ENVIRONMENTAL MODELLING & SOFTWARE, 69, 14–22, https://doi.org/10.1016/j.envsoft.2015.03.001, 2015.

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

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.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, NATURE, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, Jr., F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., and van Smeden, M.: Calculating the sample size required for developing a clinical prediction model, BMJ-BRITISH MEDICAL JOURNAL, 368, https://doi.org/10.1136/bmj.m441, 2020.

Rinderera, M., Ali, G., and Larsen, L. G.: Assessing structural, functional and effective hydrologic connectivity with brain neuroscience methods: State-of-the-art and research directions, EARTH-SCIENCE REVIEWS, 178, 29–47, https://doi.org/10.1016/j.earscirev.2018.01.009, 2018.

Rubin, D.: ESTIMATING CAUSAL EFFECTS OF TREATMENTS IN RANDOMIZED AND NONRANDOMIZED STUDIES, JOURNAL OF EDUCATIONAL PSYCHOLOGY, 66, 688–701, https://doi.org/10.1037/h0037350, 1974.

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., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Munoz-Mari, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schoelkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J.: Inferring causation from time series in Earth system sciences, NATURE COMMUNICATIONS, 10, https://doi.org/10.1038/s41467-019-10105-3, 2019a.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D.: Detecting and quantifying causal associations in large nonlinear time series datasets, SCIENCE ADVANCES, 5, https://doi.org/10.1126/sciadv.aau4996, 2019b.

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.

Sankarasubramanian, A., Vogel, R., and Limbrunner, J.: Climate elasticity of streamflow in the United States, WATER RESOURCES RESEARCH, 37, 1771–1781, https://doi.org/10.1029/2000WR900330, 2001.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., and Carrillo, G.: Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA, HYDROLOGY AND EARTH SYSTEM SCIENCES, 15, 2895–2911, https://doi.org/10.5194/hess-15-2895-2011, 2011.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J.: On causal and anticausal learning, arXiv preprint arXiv:1206.6471, 2012.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y.: Toward Causal Representation Learning, Proceedings of the IEEE, 109, 612–634, https://doi.org/10.1109/JPROC.2021.3058954, 2021.

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

Sendrowski, A. and Passalacqua, P.: Process connectivity in a naturally prograding river delta, WATER RESOURCES RESEARCH, 53, 1841–1863, https://doi.org/10.1002/2016WR019768, 2017.

Seydi, S. T., Abatzoglou, J. T., AghaKouchak, A., Pourmohamad, Y., Mishra, A., and Sadegh, M.: Predictive understanding of links between vegetation and soil burn severities using physics-informed machine learning, Earth's Future, 12, e2024EF004 873, 2024.

Singh, R., Reed, P. M., and Keller, K.: Many-objective robust decision making for managing an ecosystem with a deeply uncertain threshold response, Ecology and Society, 20, 2015.

Singh, S. K., McMillan, H., Bardossy, A., and Fateh, C.: Nonparametric catchment clustering using the data depth function, HYDROLOGICAL SCIENCES JOURNAL-JOURNAL DES SCIENCES HYDROLOGIQUES, 61, 2649–2667, https://doi.org/10.1080/02626667.2016.1168927, 2016.

Sivapalan, M.: Pattern, process and function: elements of a unified theory of hydrology at the catchment scale, Encyclopedia of hydrological sciences, 2006.

Slater, L., Blougouras, G., Deng, L., Deng, Q., Ford, E., Hoek van Dijke, A., Huang, F., Jiang, S., Liu, Y., Moulds, S., et al.: Challenges and opportunities of ML and explainable AI in large-sample hydrology, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2024.

Spieler, D. and Schuetze, N.: Investigating the Model Hypothesis Space: Benchmarking Automatic Model Structure Identification With a Large Model Ensemble, WATER RESOURCES RESEARCH, 60, https://doi.org/10.1029/2023WR036199, 2024.

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

Sultan, D., Tsunekawa, A., Tsubo, M., Haregeweyn, N., Adgo, E., Meshesha, D. T., Fenta, A. A., Ebabu, K., Berihun, M. L., and Setargie, T. A.: Evaluation of lag time and time of concentration estimation methods in small tropical watersheds in Ethiopia, JOURNAL OF HYDROLOGY-REGIONAL STUDIES, 40, https://doi.org/10.1016/j.ejrh.2022.101025, 2022.

Tárraga, J. M., Sevillano-Marco, E., Muñoz-Marí, J., Piles, M., Sitokonstantinou, V., Ronco, M., Miranda, M. T., Cerdà, J., and Camps-Valls, G.: Causal discovery reveals complex patterns of drought-induced displacement, iScience, 27, 2024.

Todorovic, A., Grabs, T., and Teutschbein, C.: Improving performance of bucket-type hydrological models in high latitudes with multi-model combination methods: Can we wring water from a stone?, JOURNAL OF HYDROLOGY, 632, https://doi.org/10.1016/j.jhydrol.2024.130829, 2024.

Vandenberg-Rodes, A., Moftakhari, H. R., AghaKouchak, A., Shahbaba, B., Sanders, B. F., and Matthew, R. A.: Projecting nuisance flooding in a warming climate using generalized linear models and Gaussian processes, Journal of Geophysical Research: Oceans, 121, 8008–8020, 2016.

Verma, T. and Pearl, J.: Causal networks: Semantics and expressiveness, in: Machine intelligence and pattern recognition, vol. 9, pp. 69–76, Elsevier, 1990.

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

Viglione, A., Parajka, J., Rogger, M., Salinas, J. L., Laaha, G., Sivapalan, M., and Bloeschl, G.: Comparative assessment of predictions in ungauged basins - Part 3: Runoff signatures in Austria, HYDROLOGY AND EARTH SYSTEM SCIENCES, 17, 2263–2279, https://doi.org/10.5194/hess-17-2263-2013, 2013.

Wang, Y., Yang, J., Chen, Y., De Maeyer, P., Li, Z., and Duan, W.: Detecting the Causal Effect of Soil Moisture on Precipitation Using Convergent Cross Mapping, SCIENTIFIC REPORTS, 8, https://doi.org/10.1038/s41598-018-30669-2, 2018.

Wood, S.: Mixed GAM computation vehicle with automatic smoothness estimation. R package version 1.8–12, 2018.

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

Yadav, M., Wagener, T., and Gupta, H.: Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, ADVANCES IN WATER RESOURCES, 30, 1756–1774, https://doi.org/10.1016/j.advwatres.2007.01.005, 2007.

Yang, M. and Olivera, F.: Classification of watersheds in the conterminous United States using shape-based time-series clustering and Random Forests, JOURNAL OF HYDROLOGY, 620, https://doi.org/10.1016/j.jhydrol.2023.129409, 2023.

Zachariah, M., Mondal, A., and AghaKouchak, A.: Probabilistic assessment of extreme heat stress on Indian wheat yields under climate change, Geophysical Research Letters, 48, e2021GL094 702, 2021.

Zazo, S., Molina, J.-L., Ruiz-Ortiz, V., Vélez-Nicolás, M., and García-López, S.: Modeling river runoff temporal behavior through a hybrid causal–hydrological (HCH) method, Water, 12, 3137, 2020.

Zhang, Y., Vaze, J., Chiew, F. H. S., Teng, J., and Li, M.: Predicting hydrological signatures in ungauged catchments using spatial interpolation, index model, and rainfall-runoff modelling, JOURNAL OF HYDROLOGY, 517, 936–948, https://doi.org/10.1016/j.jhydrol.2014.06.032, 2014.