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. We compare predictive models that incorporate causal information derived from the relationships between catchment, climate, and runoff characteristics. 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, particularly for low flow duration, high flow duration, low flows, and high flows. 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-

25 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

30 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 (Ficchì et al., 2019).

Although all processes in a catchment contribute to its runoff response, each runoff property (or signature) is directly in-

35 fluenced 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.

40 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. Classifica-

45 tion 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

50 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

55 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 the 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-investigated cause-and-effect relationship, a confounding variable is an unknown or unmeasured variable that influences both the supposed effect and the 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 the conditional dependencies among ~~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 constraint-based ~~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).

One of the most widely-used causal discovery algorithms is PC, named after its authors Peter Spirtes and Clark Glymour (Spirtes et al., 2001). The PC algorithm is a constraint-based method for causal discovery, which means it infers causal relationships by testing for conditional independencies in the data. It operates under the assumption of causal sufficiency, that is, all relevant variables are measured and there are no unobserved confounders. Given this assumption and a sufficiently large

sample size, the algorithm guarantees asymptotically correct results Glymour et al. (2019). Although this method is used for discovering Directed Acyclic Graphs (DAGs), its results do not fully identify the true causal structure; instead, it outputs a Markov equivalence class, a set of causal graphs that encode the same conditional independencies. These equivalence classes are typically represented using Completed Partially Directed Acyclic Graphs (CPDAGs) (Peters et al., 2017). Due to its simplicity, computational efficiency, and strong performance, the PC algorithm has been widely adopted across various fields, such as climate science (Ebert-Uphoff and Deng, 2012; Deng and Ebert-Uphoff, 2014), medicine (Sanchez-Romero et al., 2023), and epidemiology Petersen et al. (2021).

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 (Bayesian Network) 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 PCPeter 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,

4

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)~~1) identify causal relationships between catchment attributes, climate characteristics, and runoff signatures (network structure) using PC~~Peter and Clark (PC)~~ causal discovery algorithm~~method~~ (Spirtes et al., 2001), 2)~~3)~~ execute models using both the causal parents (causal models) and all selected variables (non-causal models) for entire catchments and subset of catchments, 3)~~4)~~ evaluate the robustness of the causal and non-causal models.

~~**Data and methods**~~

2   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 includes 41 catchment and climate attributes, including both categorical and continuous variables. In this study, we perform clustering, causal discovery, and prediction, each of which relies on specific assumptions that must be satisfied when applied to our data. Ensuring these assumptions hold is essential for the validity and reliability of the results. All 41 attributes were used in the clustering analysis. However, for the causal discovery and prediction tasks, we used a subset of 22 continuous variables that are most relevant to all runoff signatures. We performed three types of correlation analyses, Pearson, Kendall, and Spearman, to examine the relationships between catchment and climate attributes and runoff signatures. To further assess the predictive power of these attributes, we used the Random Forest algorithm to evaluate feature importance. Variable importance was ranked using the out-of-bag (OOB) approach, based on the Mean Decrease in Accuracy (IncMSE) score. This metric quantifies the increase in prediction error when a variable is excluded from the model, thereby indicating its relative contribution to predicting runoff signatures. The random forest algorithm was implemented using the randomForest package in R (Liaw et al., 2015). The results of this analysis are provided in the Supplementary Materials.

It is important to note that when constructing a causal graph (Directed Acyclic Graph, or DAG) and performing predictions using Bayesian Network methods, which are part of the probabilistic graphical model framework, the selected variables (nodes)

must not be deterministic functions of one another, as this would violate the conditional independence assumptions underlying the DAG structure (Koller and Friedman, 2009). For this reason, the aridity index, which is the ratio of precipitation to potential evapotranspiration, is excluded from the analysis. The issue of determinism also arises when variables are complementary, such as soil texture components, which are represented as percentages of sand, silt, clay, water, organic matter, and other contents that together sum to 100%. To ensure valid causal discovery and prediction, we use only a subset of these variables, specifically sand, silt, and clay, in our analysis.

Table 1. Catchment and climate attributes used for clustering, as well as for causal discovery and prediction of all runoff signatures. Variables shown in bold are those selected for causal discovery and prediction. They were 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	km²
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	m²

Table 1. (continued) Catchment and climate attributes used for clustering, as well as for causal discovery and prediction of all runoff signatures. Variables shown in bold are those selected for causal discovery and prediction. They were, 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 material (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 (target variables) 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)

3 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 22~~selected~~ variables as predictors (Table 1). 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 is also~~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 performe~~performed~~ comparably to or better than non-causal models across different environments, it indicates that causal parents suffice to explain the target variable. In this situation, we can also conclude that the causal discovery is able to recover the direct causes of the runoff signature. In cases where causal models outperform~~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.

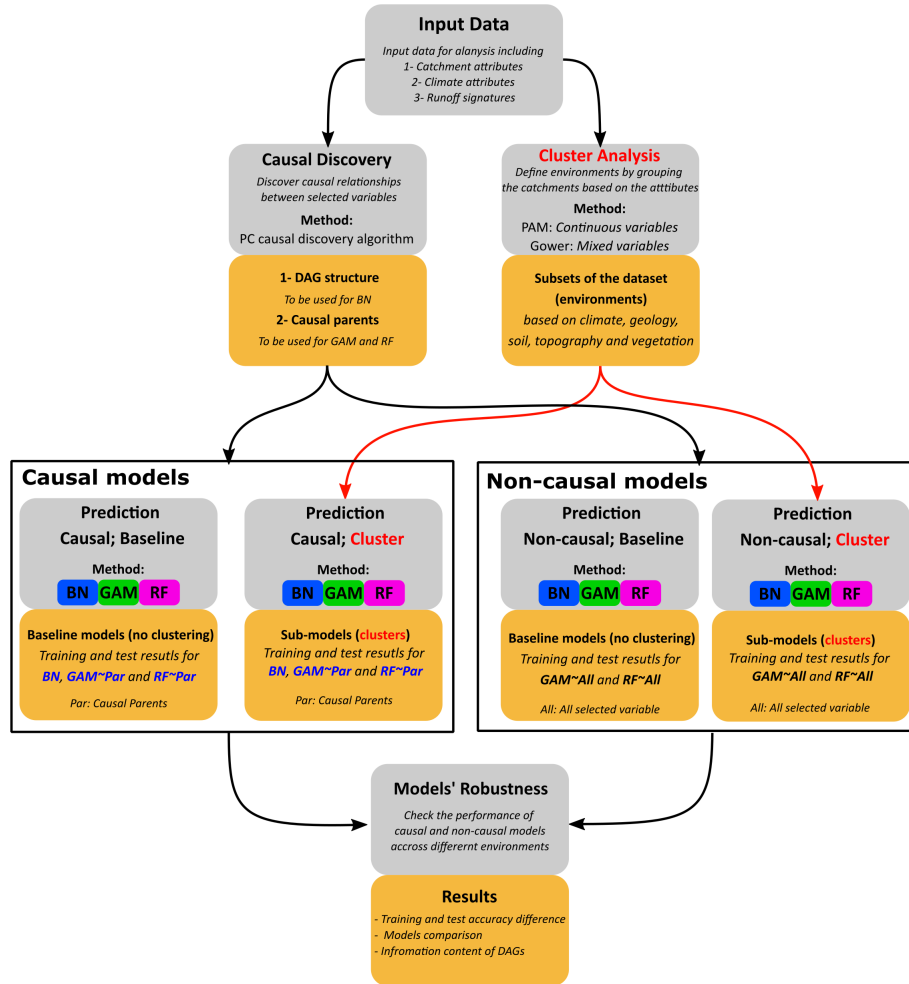The steps are explained in the following sections.

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 2) reduce the dimensionality of the causal discovery problem (Runge et al., 2023). Including all 41 climate and catchment

200

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 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:

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.

3.1 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 shares 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 the same across all environments if the distribution of X changes. This information will help investigate the performance of the causal compared to the 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 noise than the K-means 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.

3.2 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., 2018a). Here, we try to find causal structures from the observational data without specifying the underlying physical equations using a causal discovery method. Causal discovery is applied to each runoff signature along with 22 variables from the CAMELS dataset to identify the causal graph associated with each runoff signature. ~~The causal discovery method is applied to the selected variables for the whole dataset and each runoff signature.~~

3.2.1 PC causal discovery algorithm

In this~~This~~ study,~~uses~~ the constraint-based PC causal discovery algorithm (Spirtes et al., 2001), named after its authors Peter Spirtes and Clark Glymour is used. The PC algorithm recovers the causal graph from observed data by testing which variables are conditionally independent of each other. These independencies are seen as constraints to satisfy the true data-generating process. ~~This method identifies the DAG under faithfulness and Markov assumptions.~~ The PC algorithm outputs a Completed

Partially Directed Acyclic Graph (CPDAG), which represents a Markov equivalence class of causal structures. This means that instead of identifying a single, unique Directed Acyclic Graph (DAG) that fully describes the causal relationships among variables, the PC algorithm returns a set of DAGs that are indistinguishable based on observed statistical dependencies alone. These DAGs all share the same conditional independence, that is, the probability distribution over the variables satisfies the Markov property with respect to each of them. The CPDAG captures this equivalence class by showing which edges are directed and which are only known to exist but not in which direction (undirected edges).

To recover a valid CPDAG using the PC algorithm, certain assumptions must be satisfied to ensure the causal structure is identifiable from the data. These include the causal Markov condition, which links the graph structure to statistical independencies, the faithfulness condition, which ensures that all observed independencies are due to the graph structure or the distribution is faithful to DAG (Peters et al., 2017), and causal sufficiency, which assumes there are no unmeasured confounders.

To construct the causal graph, PC begins with the assumption that every variable is potentially connected to every other variable, forming a fully connected undirected graph. Then, the algorithm starts removing edges between variables based on statistical tests of conditional independence. The key idea is that if two variables are conditionally independent given some set of other variables, then there is no direct causal link between them, so the edge connecting them can be removed. By checking different sets of conditioning variables and removing edges accordingly, the PC algorithm recovers the underlying skeleton of the graph, that is, the undirected structure that shows which variables might have a direct relationship.

Once the skeleton is established, the PC algorithm proceeds to orient edges using specific rules based on the results of the conditional independence tests. A key orientation rule involves identifying v-structures or unshielded colliders. If variables A and B are both connected to a third variable C (i.e., A–C–B), and A and B are not connected to each other, and C was not in the conditioning set that rendered A and B independent, then the structure is oriented as a collider (A $\longrightarrow$ C $\longleftarrow$ B). After identifying all the v-structures, the PC algorithm applies a set of logical rules to orient the edges throughout the graph while avoiding cycles and introducing new v-structures. However, some edges remain undirected when the available conditional independence information is insufficient to determine a unique orientation without introducing ambiguity or inconsistency with the test results. These undirected edges represent causal relationships whose direction cannot be resolved from the data alone without further assumptions (Meek, 1995; Spirtes et al., 2001).

The results of the PC causal discovery algorithm depend on the alpha value (also called the significance level or threshold). The alpha value is used during conditional independence tests to decide if an edge should be removed. If the p-value of a test exceeds the alpha threshold, the algorithm considers the variables to be conditionally independent and removes the edge. Therefore, the value of alpha directly affects the sparsity and structure of the resulting causal graph. The variation in alpha value affects the skeleton and orientation of the graph produced by the PC algorithm (Kalisch et al., 2012). Furthermore, the sample size affects the results of the PC, as the algorithm relies on statistical tests of conditional independence to determine the graph structure. As the sample size increases, the accuracy of these tests improves, making it more likely for the algorithm to recover the correct causal structure, or more precisely, the correct Markov equivalence class. Identifying the correct structure of a Bayesian network, which relies on independence testing similar to that derived from the PC algorithm, requires substantially more data than approximating the underlying distribution. Underfitting, where true edges are missed, is especially likely when

14

sample sizes are small, as shown in Zuk et al. (2012). Thus, Considering these issues, we aim to address the choice of alpha values and sample sizes in the following sections..

3.2.2 Background knowledge and edge assumptions

305 Background knowledge can significantly enhance causal discovery by reducing ambiguity in the orientation of edges and narrowing the space of plausible causal graphs. According to Perković et al. (2017), incorporating background knowledge, such as known edge directions or variable orderings, into a CPDAG results in a CPDAG which they refer to as maximally oriented partially directed acyclic graphs (maximal PDAGs). Maximal PDAG often contains fewer undirected edges and thus represents a smaller Markov equivalence class. This improves the identifiability of causal relationships and allows for more accurate

310 estimation of causal effects. Similarly, Bang and Didelez (2025) demonstrate that tiered background knowledge, derived from temporal or hierarchical structures, can be encoded as forbidden edge directions to guide constraint-based algorithms. Their algorithms show that such knowledge not only improves the clarity but also reduces the number of independence tests and possible equivalence classes, enhancing both computational efficiency and discovery accuracy. To obtain maximal PDAG and avoid implausible edge orientations, we define some edge assumptions by blacklisting the implausible edges before running

315 PC algorithm. The edge assumptions are as follows:

- Climate variables cannot cause topography variables.

- Soil variables cannot cause climate variables.

- Geological variables cannot cause climate variables.

- Climate variables cannot cause geological variables.

320 - Vegetation variables cannot cause climate variables, except for potential evapotranspiration.

- Vegetation variables cannot cause topographic variables.

- Vegetation variables cannot cause geological variables.

- Soil variables cannot cause topographic variables.

- Runoff signatures cannot have child nodes.

325 It is important to note that many of these causally implausible links, such as climate influencing topography, may become plausible when considered over geological timescales. ~~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.~~ The final assumption, which treats the runoff signature (the target variable) as a sink node by preventing it from causing other variables, implies that its causal parents correspond to its Markov and stable blankets.~~This setting makes the causal parents of the signatures their Markov and~~

330 ~~stable blankets.~~ The Markov blanket of a node consists of its parents, its children, and the parents of its children. Conditioning

15

on the Markov blanket of a node makes the node independent of the rest of the DAG (Pearl, 1988). This setting also makes the target variable's causal parents equivalent to its stable blanket for regression. ~~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).

3.2.3 Implementation and evaluation of the PC algorithm's results

PC algorithm assumes that the variables are normally distributed. Therefore, the Box-Cox transformation is applied to the data (Dutta and Maity, 2020). Since the results of the PC algorithm can depend on the order of the input variables, we use the PC-stable variant (Colombo and Maathuis, 2014), which addresses this issue by ensuring order-invariant outputs. The bnlearn R package (Scutari, 2009) is used to apply the PC algorithm. Mutual information with the James-Stein estimator (Hausser and Strimmer, 2009) ~~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.~~

In this study, we run the PC algorithm for 22 variables along with a runoff signature. We do this for 11 runoff signatures. The sample size for each variable is roughly 670 data points. Since the result of the PC algorithm is sensitive to the sample size and significance threshold (alpha value), we set the significance threshold to 0.2 to allow for a more inclusive initial edge selection. The high significance level helps reduce the risk of Type I error, that is, missing true causal edges; however, it increases the risk of Type II errors, which is the appearance of false positive edges. The edge assumptions, defined as a blacklist, help prevent the occurrence of a large number of false positive edges. We evaluate the resulting CPDAG following the approach proposed by Petersen et al. (2021). First, we assess the stability of the discovered edges by performing 1,000 bootstrap resamples (Scutari

365 and Nagarajan, 2013) of the data and applying the PC algorithm to each resample, using a stricter significance threshold of 0.05 for conditional independence tests. We then measure the strength of each edge based on its frequency across the bootstrap iterations. The resulting edge strength estimates, which represent the proportion of bootstrap samples in which each edge appears, are then mapped onto the initial CPDAG obtained from the PC algorithm. This approach enabled us to evaluate the stability and sparsity level of the causal links if we had a dataset with a larger sample size. Then, we use regression modelling

370 with cubic splines as a heuristic test for conditional independence, assessing the statistical significance of each edge present in the CPDAG produced by the PC algorithm. We fit cubic spline regression models for each pair of variables (X_i, X_j), using a conditioning set $\mathbf{Z} = \{X_1, \ldots, X_m\} \in \mathbf{X} \setminus \{X_i, X_j\}$. We regress X_i once on $\mathbf{Z}$, and once on $\{X_j\} \cup \mathbf{Z}$ (forward direction). To account for potential asymmetry in the relationships, we repeat the same procedure by regressing X_j on $\mathbf{Z}$ and on $\{X_i\} \cup \mathbf{Z}$ (backwards direction). In total, we fit four models for each pair of adjacent variables in the CPDAG, which are as follows:

375
$$M_{0a} : g(X_i) = \sum_{k=1}^{m} s_k(Z_k)$$

$$M_{1a} : g(X_i) = s_0(X_j) + \sum_{k=1}^{m} s_k(Z_k)$$

$$M_{0b} : \tilde{g}(X_j) = \sum_{k=1}^{m} \tilde{s}_k(Z_k)$$

380

$$M_{1b} : \tilde{g}(X_j) = \tilde{s}_0(X_i) + \sum_{k=1}^{m} \tilde{s}_k(Z_k)$$

where g and $\tilde{g}$ are identity functions, s and $\tilde{s}$ are cubic splines. The likelihood ratio tests are applied to compare M_{0a} and M_{1a} as well as M_{0b} and M_{1b} to estimate the conditional independence between pairs of adjacent nodes. We consider variables

385 (X_i, X_j) adjacent if at least one of the following null hypotheses is rejected.

$$H_{0a} : M_{0a} = M_{1a} \quad and \quad H_{0b} : M_{0b} = M_{1b}$$

Since no distributional assumptions are made, this test establishes a necessary, but not sufficient, condition for conditional independence between variables (Petersen et al., 2021). Here, we report the p-values obtained from the likelihood ratio test for each edge in the DAGs, without assuming a specific alpha threshold. The alpha value can be set to 0.05 to be compared with the significance threshold used in bootstrapping DAGs, or to a value between 0.05 and 0.2 to be compared with the edges

390 identified in the initial PC algorithm run. The p-value provides an insight into the trade-off between statistical significance and edge strength.

17

The DAG for each runoff signature is derived from the corresponding CPDAG by orienting the undirected edges. It is done in a way that avoids introducing new unshielded colliders—structures in which two parent nodes point to a common child without being connected to each other. Introducing such colliders would change the set of conditional independencies encoded by the original CPDAG.

3.3 Prediction models

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).

3.3.1 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.

3.3.2 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). we used cubic regression splines for the smooth terms. The outcome variable is continuous, and we used the default identity link function with a Gaussian error distribution. The GAMs were fitted using Restricted Maximum Likelihood to estimate the smoothing parameters. 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).

3.3.3 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 com-
425 putational 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
430 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
435 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.

4 Results

440 ### 4.1 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
445 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
450 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, is 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.

485 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.

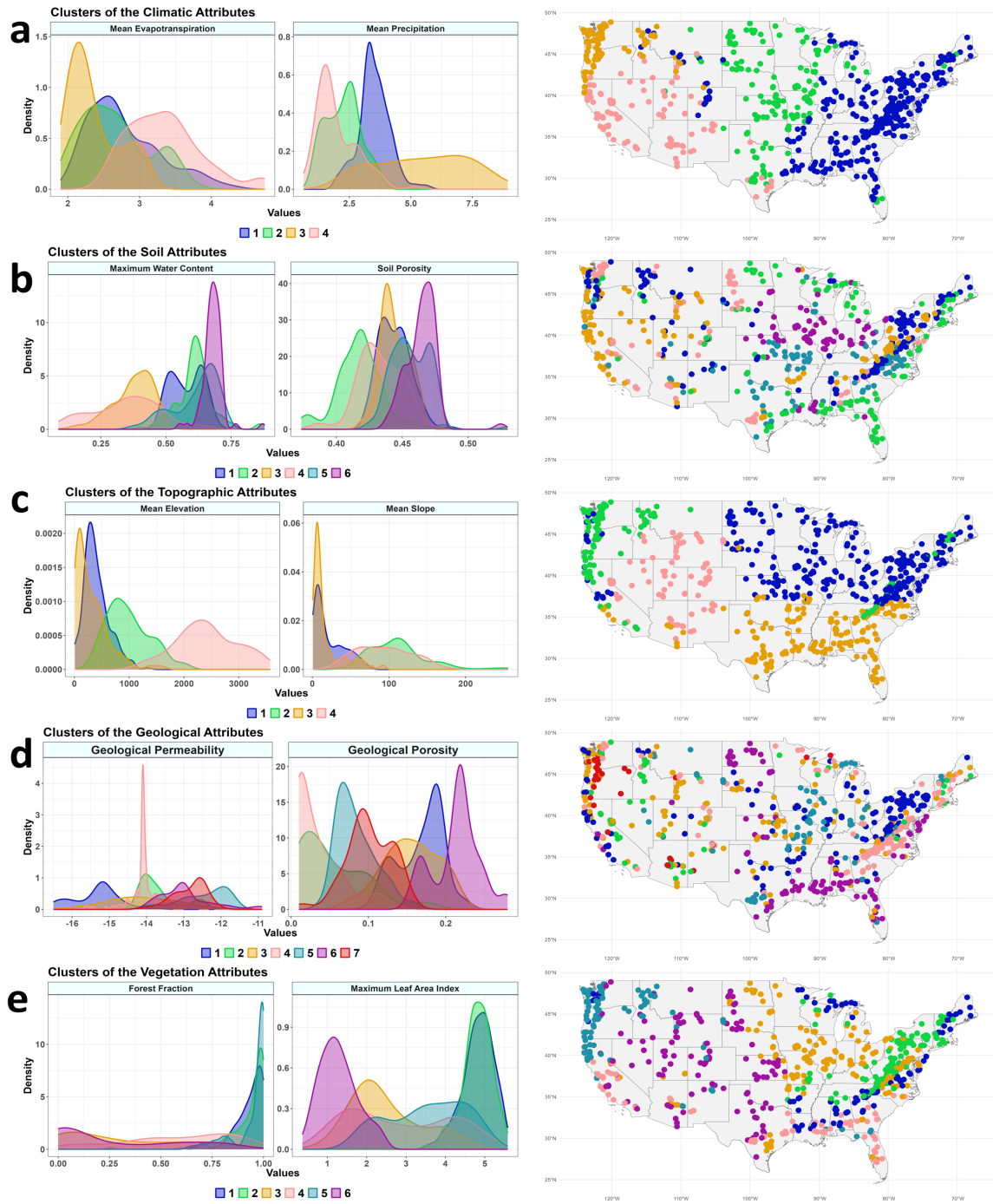| No	Category	Method	No. of cluster	No. of Catchments
1	Climate	Gower	4	334, 144, 87, 103
2	Soil	PAM	6	154, 123, 138, 88, 95, 70
3	Topography	PAM	4	282, 119, 117, 90
4	Geology	Gower	7	149, 53, 123, 116, 64, 104, 42
5	Vegetation	Gower	6	89, 131, 149, 69, 105, 128

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.

4.2 Identification of Causal Links

PC algorithm results identify the causal links between all variables. The output of the PC algorithm is a Completed Partially Directed Acyclic Graph (CPDAG), which may contain undirected edges. In all CPDAGs obtained in this study, the connection between mean elevation and mean slope remains undirected. To derive a fully directed acyclic graph (DAG) from the CPDAG, we orient this edge from mean elevation to mean slope. As shown in Fig. 3b, this orientation does not introduce any unshielded colliders or cycles in the graph. Therefore, we ensure that the resulting DAG belongs to the same Markov equivalence class as the CPDAG produced by the PC algorithm.~~, 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 → frac_forest). However, the causal direction between climatic or vegetation variables, such as the direction between high precipitation frequency and low precipitation frequency, is not clearly definable and differs 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. 3a shows the obtained DAG for the baseflow index. The signature (red node) has four~~five~~ direct causes or parents (yellow nodes). Fig. 3c shows the ~~The~~nodes that form the independent causal mechanism for the baseflow index, which are shown by the green line. The identified causal parents of the base flow index include variables related to catchment storage, such as groundwater and snow storage, which are physically meaningful. For each recovered edge, we report both the edge strength and the p-value from the likelihood ratio test, as shown in Fig. 3a. The causal models, $\sim$Par, are trained within the causal mechanism to predict the baseflow index (Fig. 3b). 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. 3a are controlled by topographic attributes, which are~~gauge latitude,~~ mean elevation~~,~~ and mean slope. These variables are independent of other categories in the~~this~~ DAG since they do not have any parents delong to the other categories. Furthermore,~~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 (see Supplementary Materials Section 1). ~~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.~~
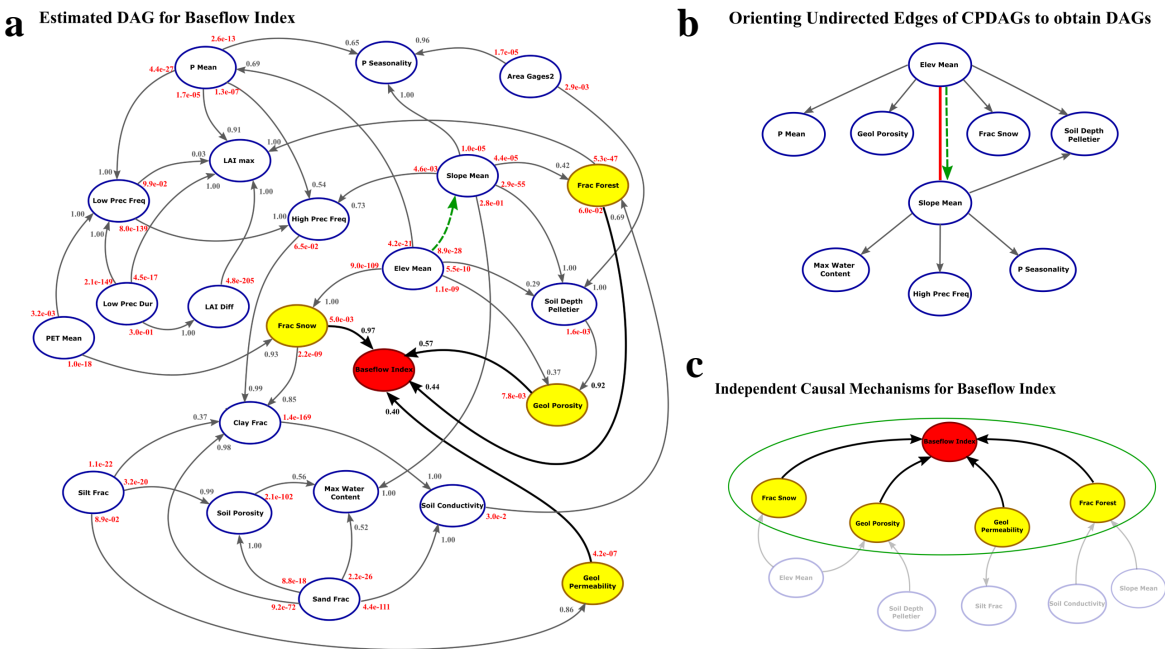
Figure 3. a) Directed Acyclic Graph (DAG) for the Baseflow Index. Arrows indicate the causal links between variables. The green dashed arrow represents an oriented edge that was originally undirected in the CPDAG derived from the PC algorithm. The red node denotes the target variable (runoff signature), the yellow nodes represent its causal parents. Red numbers at the beginning of each arrow correspond to the p-values from the likelihood ratio tests, and the grey (or black for the target variable) numbers indicate the edge strengths derived from 1000 bootstrap resamples.~~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. b) The part of the CPDAG that contains an undirected edge. Orienting this edge does not introduce any new unshielded colliders. c) The independent causal mechanism for the baseflow index, which is represented by the~~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 p-value of the likelihood ratio test, and the edge strength for each runoff signature.~~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 56 variables~~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.

24

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

Table 4. Causal parents of the runoff signatures and their numbers derived from the PC algorithm. P-values are obtained from the likelihood ratio test, and the edge strengths are derived from the frequency of edges that appear in 1000 bootstrap resamples.

Signature	Detected causal parent by PC (p-value; Edge strength)
baseflow_index	geol_permeability (4.2e−07; 0.40), frac_snow (5.0e−03, 0.97), frac_forest (6.0e−02, 0.44), geol_porosity (7.8e − 03, 0.57)
high_q_dur	p_mean (3.7e − 08, 0.56), lai_diff (5.6e − 04, 0.95)
high_q_freq	low_prec_freq (2.4e − 01, 0.79), geol_porosity (6.8e − 02, 0.52)
low_q_dur	low_prec_dur (3.0e − 08>, 0.74), max_water_content (2.1e − 03, 0.76), frac_snow (2.9e − 04, 0.88)
low_q_freq	frac_snow (2.7e−06, 0.97), low_prec_freq (1.6e−02, 0.45), frac_forest (1.0e0, 0.71), geol_porosity (8.0e − 01, 0.66), geol_permeability (2.5e − 06, 0.27)
q_mean	p_mean (5.3e−160, 1.0), p_seasonality (3.2e−03, 0.85), area_gages2 (4.8e−02, 1.0), frac_forest (6.2e − 02, 0.63), geol_porosity (1.0e − 06, 0.79)
Q5	p_mean (4.3e − 18, 0.77), low_prec_freq (2.7e − 10, 0.67), slope_mean (6.8e − 02, 0.82), geol_porosity (1.6e − 01, 0.57), geol_permeability (3.9e − 05, 0.26)
Q95	p_mean (7.6e − 96, 1.00), p_seasonality (1.5e − 01, 0.98), low_prec_freq (1.4e − 03, 0.81), area_gages2 (6.3e − 02, 0.35), frac_forest (1.3e − 01, 0.57)
runoff_ratio	p_mean (1.0e − 34, 0.83), p_seasonality (6.8e − 05, 0.97), area_gages2 (2.6e − 01, 1.00), geol_porosity (4.8e − 06, 0.97)
slope_FDC	p_mean (3.7e − 07, 0.92), pet_mean (4.1e − 09, 0.40), low_prec_freq (3.7e − 06, 0.39)
stream_elast	pet_mean (7.1e − 02, 0.32), clay_frac (2.5e − 01, 0.41), frac_snow (1.6e − 02, 0.79), frac_forest (4.7e − 05, 0.81), area_gages2 (1.4e − 01, 0.41)

The obtained DAGs, presented in the Supplementary Materials, reveal consistent causal relationships between catchment and climate attributes across all runoff signatures, although the strength and significance of the edges vary. In all cases, topographic variables directly influence climate, vegetation, soil, and geological attributes. Climate variables influence vegetation and all runoff signatures (Table 4). Only two edges connect climate and soil variables, specifically, from high precipitation frequency and the fraction of snow to the clay fraction. Across all DAGs, these edges consistently exhibit high strength and statistical significance. Soil variables influence vegetation, specifically the fraction of forest, as well as geological variables. However, these variables act as causal parents for only two runoff signatures, namely low flow duration and streamflow elasticity. Vegetation attributes do not drive other catchment attributes. They are influenced by climate, topography, and soil variables. However, they directly affect six runoff signatures, including baseflow index, high flow duration, low flow frequency, mean flow, high flow, and streamflow elasticity (Table 4). The geological variables are influenced by topography and soil. They are among the causal parents of baseflow index, high flow frequency, low flow frequency, mean flow, low flow, and runoff ratio (Table 4). The edges

in the obtained DAGs are generally characterized by high strength and statistical significance. However, the link between low precipitation frequency and maximum leaf area index stands out as the weakest, with a strength of less than 1% and a marginal significance level (p-value slightly below 0.1). Notably, this link is absent in the DAGs corresponding to high flow duration, high flow frequency, low flow frequency, and the slope of the flow duration curve.

4.3 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 the 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 500~~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, and the slope of flow duration curve ~~and streamflow elasticity~~ with 38%~~21%,~~ ~~15%,~~ 53%~~39%,~~ and 25%~~13% and 15%~~ increases in distance, respectively, caused by using causal model (Fig. 5). The causal model slightly reduces the gap between training and test results for low-flow duration, low-flow frequency, and high-flow magnitude, with improvements of 7%, 6%, and 1%, respectively. ~~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 41%~~29%~~ for the slope flow duration curve~~runoff ratio~~ to 87%~~90%~~ for the high flow frequency~~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.
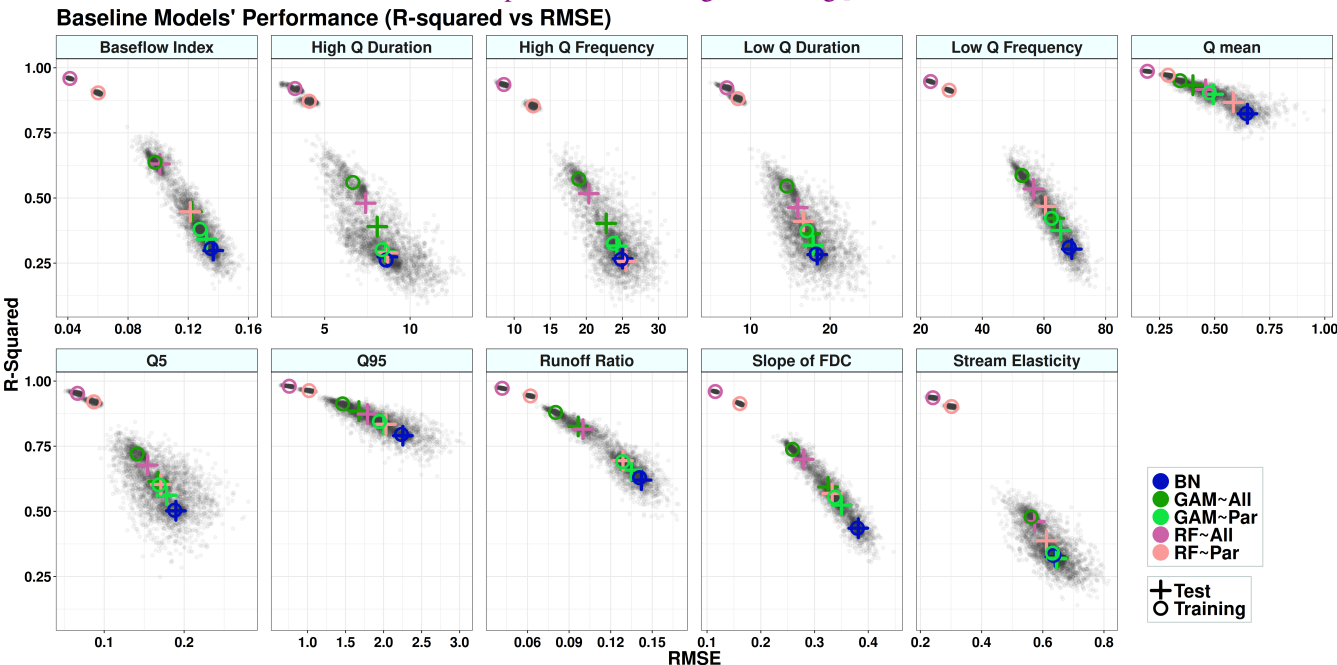
Figure 4. Performance of the models: R-squared vs. RMSE. Each coloured circle and cross represent the centroid of a set of 500100 data points (grey dots) generated from the models' execution. Circles indicate the training results, and crosses indicate the test results. In the legend, "All" refers to 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.

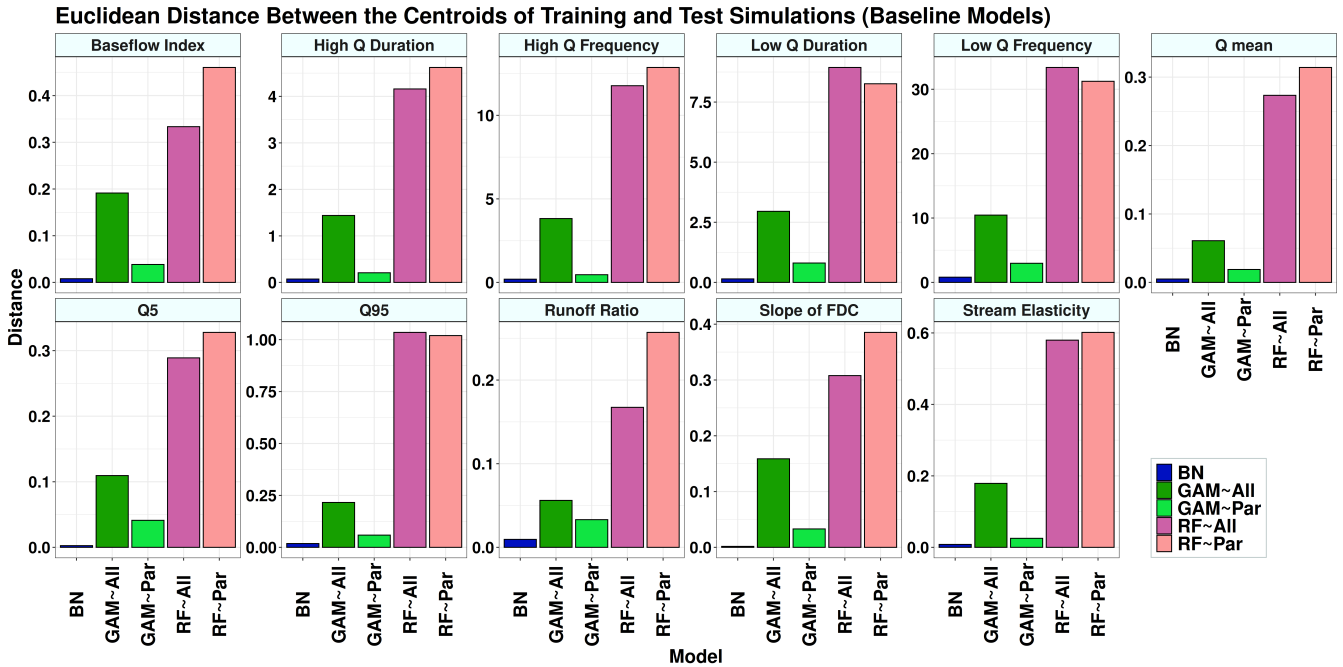[Comment: The figure has been updated for the new set of predictor variables.]

Figure 5. The Euclidean 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.

565   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

570   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.

4.4   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, tables
575   in Appendix A, and the sections 2 and 3 of the Supplementary Materials~~and figures in Appendix B~~. According to the results, GAM∼All shows high accuracy during training across most clusters, but performs poorly during testing. The distance between training and test for GAM∼Par is lower than GAM∼All in all clusters, and in most cases. This may be due to overfitting in GAM∼All when the sample size is small, resulting in its performance being statistically insignificant from, or in most cases

lower than, GAM∼Par across all environments in the test mode. On the other hand, RF∼All shows the highest performance

580 in most cases in both training and testing modes. However, in many cases, RF∼Par performs comparably to RF∼All, despite using significantly fewer predictors. In the case of the BN model, which is linear, it generally exhibits the lowest accuracy compared to GAM and RF. However, it also shows the smallest drop in accuracy between training and testing simulations (see Supplementary Materials Section 2, Table S1, and Table S2). The simulation results for each runoff signature are discussed in the following sections.

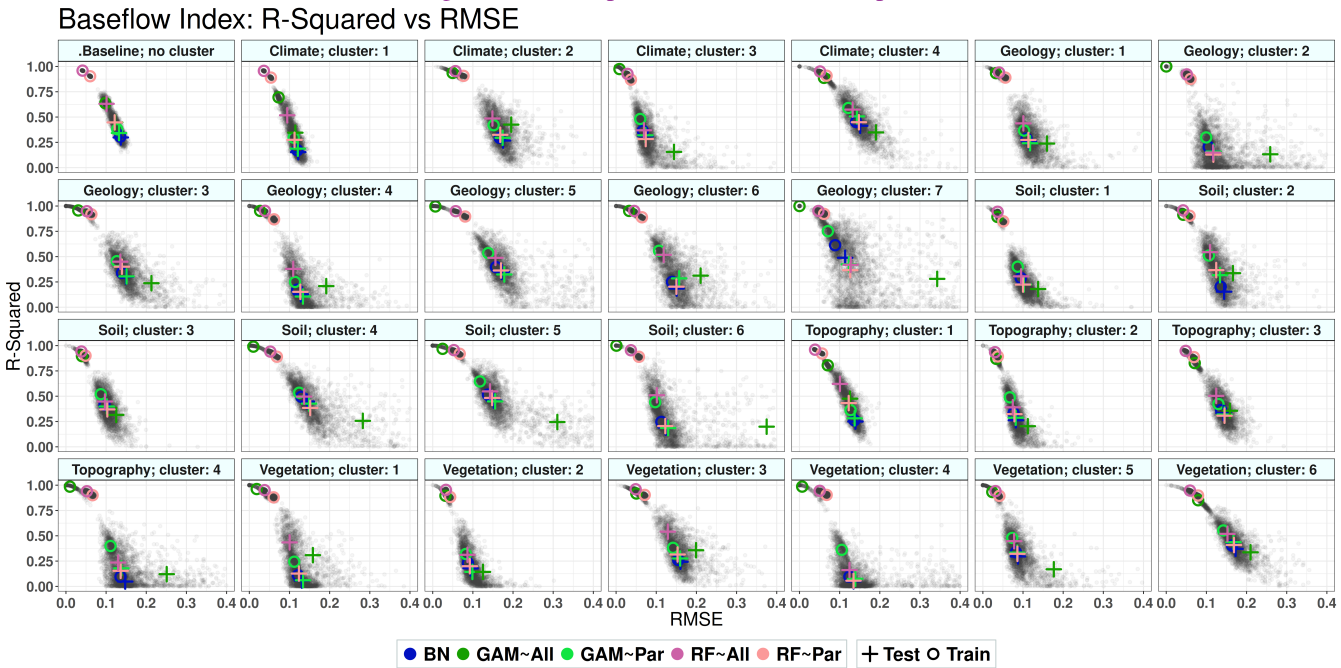[Comment: The figure has been updated for the new set of predictor variables.]

Figure 6. Performance of the models for baseflow index: R-squared vs. RMSE. Each coloured circle and cross represents the centroid of 500100 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.
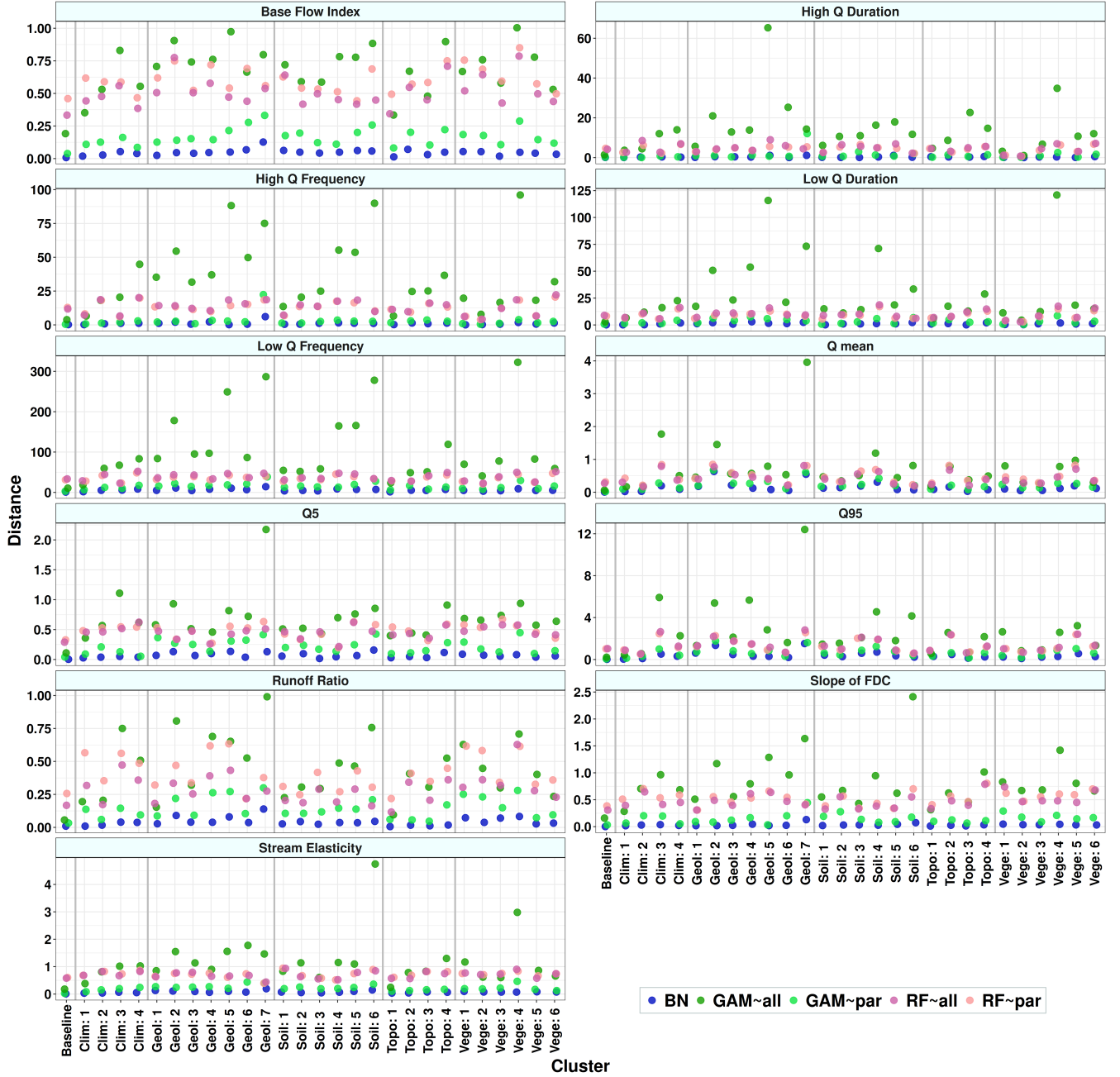
Figure 7. The Euclidean distance between the training and test simulation for runoff signatures across different environments for each sub-model. In the legend, "All" refers to using all variables as predictors (non-causal model), and "Par" refers to using only parent variables as predictors (causal 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.

585   1. **Baseflow Index:** The four parents of this signature belong to climate, ~~soil,~~ vegetation, and geology categories (Table 4). The identified causal parents exhibit both high statistical significance and strong edge strength (Table 4 and Fig. S23). ~~The obtained DAG, Fig. 3, indicates the topographic attributes directly control the snow fraction and indirectly control the forest fraction and geological porosity.~~ The~~It can be seen in Fig. 6 that the~~ models in the climate, topography and some clusters of soil~~topographic and climatic~~ groups perform well compared to the baseline (Fig. 6). ~~According to Fig.~~
590   ~~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
595   cluster numbers ~~1,~~ 3, 4, and 5, or Geology 2 and 7~~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 GAM~Par~~causal models~~ is improved from a -24%~~20%~~ drop for the baseline model to +9% for geology, +27% for soil, and +4% for vegetation categories in the~~less than a 5% drop for~~ sub-models (Table 5). For the climate and topography categories, the accuracy drop caused by using GAM~Par is 6% and 4% smaller in the sub-models compared to the
600   baseline model. The use of causal RF (RF~Pa) results in a 25% drop in accuracy in the baseline model. This reduction becomes more pronounced in the sub-models, except for the soil category, where the accuracy drop is 2% smaller than in the baseline. ~~The R-squared is increased using parents for GAM in geology, soil and topography categories (Table 5).~~

       2. **High Flow Duration:** This signature has two causal~~3~~ parents belonging to climate, ~~soil,~~ and vegetation categories (Table 4). The two causal parents of the high flow duration exhibit both high statistical significance and strong edge strength
605   (Table 4 and Fig. S25). ~~The obtained DAG shows that the parent from the soil category is an independent variable (Fig. S29). The effect of this parent can be seen in Fig. S26, where the highest accuracy of models is 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. S26) compared to the baseline.~~ The models perform well across some clusters of
610   the soil and geology categories compared to the baseline (Fig. S26). 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. S26). 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 (Fig. 7).~~, and in~~ In many~~most~~ cases, the difference between causal and non-causal RF
615   models is negligible (Table S2, Fig. S26). Although BN shows less accuracy compared to GAM and RF, it outperforms these models in some soil and geology clusters. ~~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 that of baseline models (Table 5).~~ The use of causal parents as predictors leads to a 26% and 40% drop in R squared in the GAM

and RF simulations, respectively. Although the causal models use only two predictors compared to 22 in the non-causal models, the inclusion of causal parents increases the accuracy of GAM by up to 50% in the geology, soil, topography, and vegetation categories. Additionally, they help reduce the accuracy drop of RF in the sub-models relative to the baseline model.

3. **High Flow Frequency:** This signature has ~~only~~ two parents belonging to the climate and geology~~vegetation~~ categories (Table 4). Unlike high flow duration, the causal parents of the high flow frequency does not show high statistical significance with p-value of 0.24 and 0.06 for low precipitation frequency and geological porosity, respectively; however, they show acceptable edge strength with 79% for the former and 52% for latter (Table 4 and Fig. S27). ~~The obtained DAG (Fig. 27) indicates that topographic attributes influence the causal parents.~~ Models perform well across some clusters of climate, soil, and geology categories.~~most clusters based on climate and topography.~~ However, there is no single category within which all models outperform the others (Fig. S28). For instance, the models perform well in Vegetation Cluster 5 (Fig. S28), 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. S28). However, GAM~Par demonstrates 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. S28; 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 three~~five~~ environments, namely, Geology 5, 7 and soil 5~~4, 5, and 7, Soil 5, and Vegetation 4~~. For the rest of the environments, RF~All is more accurate (Table S2). In most cases, the accuracy of GAM~Par is higher than RF~Par.~~The accuracy of RF~Par and GAM~Par models is comparable to RF~All.~~ Using causal parents leads to a decrease in the R-squared in the baseline model for both GAM and RF by -20% and -50%, respectively. While this behaviour remains for RF across sub-models, the causal parents increase the accuracy of the GAM by up to 65% (Table 5). ~~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 three~~3~~ parents belonging to climate and soil~~, soil, and vegetation~~ categories (Table 4). All causal parents of the low flow duration exhibit both high statistical significance and strong edge strength (Table 4 and Fig. S29). ~~The DAG shows that parents are controlled by the topographic variables (Fig. S29).~~ Training and test simulation performed well across all topographic clusters except for cluster number 4, where catchments have high elevations (Fig. 2 and Fig. S30). The signature also shows high predictability in clusters with high precipitation intensity (Climate Cluster 3) and clusters with low soil porosity (Soil Cluster 2) or clusters with low maximum water content (Soil Cluster 3). 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). Using causal parents results in a decrease in R-squared values of approximately

10% for both GAM and RF. However, causal parents improve the accuracy of GAM across all categories by up to 85%, whereas for RF, the pattern is reversed, with accuracies dropping by up to 15%. ~~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 five~~5~~ parents, two~~3~~ belonging to climate, one to vegetation, and two~~one~~ to geological categories (Table 4). Three out of the five causal parents of low flow frequency exhibit both high statistical significance and strong edge strength. However, despite the relatively high edge strengths, 71% for forest fraction and 66% for geological porosity, the p-values from the likelihood ratio test are large (Table 4 and Fig. S31). ~~The topographic variables control the causal parents, according to the obtained DAG (Fig. S31).~~ Models perform well across most clusters of climate and topography categories (Fig. S30). 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, the performance of the BN, GAM~Par, RF~Par, and RF~All are comparable across most clusters (Fig. S32).~~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 clusters 1 and 2, GAM~All outperform RF~Par and GAM~Par.~~ Causal parents lead to a drop in the R-squared of both GAM and RF models. This pattern is the same for RF across the sub-models; however, it leads to improving the accuracy of the GAM by up to 33% (Table 5). ~~BN has the smallest difference between training and test simulations. 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 five parents of the mean daily runoff belong to climate, topography, vegetation, and geology categories (Table 4 and Fig. S33). All five causal parents of mean daily flow exhibit both high statistical significance and strong edge strength, except for forest fraction, which has a p-value of 0.06 in the likelihood ratio test (Table 4 and Fig. S33). 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. S34). 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, especially, for the GAM (Table S1 ~~and Table S2~~). ~~The accuracy is also lower for categories compared to the baseline (Table 5).~~ For mean daily flow, the reduction in R-squared resulting from using causal parents as predictors is minimal, even in the baseline models, approximately -8% for GAM and -5% for RF. While the accuracy drop in RF increases to as much as -15%, using causal parents improves GAM accuracy by up to 12%.

7. **Low Flow (Q5):** The five parents of low flow belong to climate, geology, and topography categories (Table 4 and Fig. S35). Among the causal parents of low flow, which generally exhibit both high statistical significance and strong edge strength, mean slope and geological porosity have higher p-values in the likelihood ratio test compared to the other parents, although their edge strengths remain high (Table 4 and Fig. S35). The models' test results are comparable

to the baseline models in Geology Cluster 2 and 4 and Soil Cluster 2 and 4 (Fig. S36). GAM∼All is outperformed by GAM∼Par and other models in test simulation (Fig. S36, 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
690   Fig. S36, 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 is negligible for most environments (Table S2). BN has the smallest difference between training and testing, and its performance is comparable to GAM and RF in most cases.~~it outperforms GAM models in Climate Cluster 3, Geology Cluster 6 and 7, Soil Cluster 1 and~~
695   ~~4, Topographic Cluster 4 and Vegetation Cluster 6.~~ Using parents as predictors increases the accuracy of GAM in the climate, geology, soil, and vegetation categories by 13%, 34%, 41%, and 31%~~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).~~ For the RF, the drop in accuracy remains consistent across baseline and sub-models by around -10% to -15% Table 5).

700 8. **High Flow (Q95):** High flows are among the most identifiable signatures. According to the obtained DAG, high flow is~~flows are~~ controlled by five causal parents belonging to vegetation (land cover), climate, and topography category~~topographic variables~~ (Table 4 and Fig. S37). Among all the causal parents of high flow, precipitation seasonality has the highest p-value (0.15), despite its strong edge strength (98%). On the other hand, the catchment area shows the lowest edge strength (35%) and a relatively high p-value (0.06). The rest of the causal parents of this runoff signature
705   exhibit high statistical significance and high edge strength (Table 4 and Fig. S37). The models showed high accuracy across most clusters~~all clusters of the topographic categories~~. Unlike other signatures, the the accuracy of RF∼All and RF∼Par models, which are the most accurate overall, are comparable to~~outperformed by~~ GAM and BN in certain cases (Fig. S37). The difference between training and test simulations is improved in all clusters when using parents for GAM, except for climate cluster 1 and 2 as well as topography cluster 1~~, and most of the clusters for RF~~ (Table S1). ~~Table 5~~
710   ~~indicates that~~ The drop~~the difference~~ in the models' accuracy, caused by using the causal parents, is smaller compared to the other signatures~~negligible when using causal parents~~. The R-squared is improved among geology,~~and~~ soil, and vegetation 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 topography ~~and vegetation~~ categories (Table 4 and Fig. S39). Although catchment area has a high p-value as a causal parent of stream elasticity, its edge
715   strength is 100%. All other causal parents of the runoff ratio exhibit both high statistical significance and strong edge strength (Table 4 and Fig. S39). ~~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. S39). 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:** All three~~The~~ parents of the slope of the flow duration curve belong to climate cate-gory~~and vegetation categories, which, according to the DAG, are controlled by topographic variables~~ (Table 4, Fig. S41). All causal parents of the slope of the flow duration curve exhibit both high statistical significance and relatively strong edge strength (Table 4 and Fig. S41). Models in topographic clusters performed well except for the cluster~~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). The~~However, this~~ difference between~~for~~ RF~Par and RF~All is statistically signif-icant except for two~~insignificant for only 8~~ clusters, namely Geology Cluster 4 and Vegetation Cluster 1 (Fig. S42, Table S2). ~~However, the accuracy of RF~Par and RF~All is comparable in most cases.~~ Using causal parents as predictors for the slope of the flow duration curve, increase the accuracy of the sub-models GAMs by 42%, 25%, and 33% for geology, soil, and vegetation categories; however, for the causal RF models lead to decrease in R-sqaured compared to the non-causal RFs (Table 5).~~GAM~Par performs better than GAM~All in the geology category by increasing the R-squared by (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 five~~three~~ parents of this signature belong to climate, soil, topography, and veg-etation categories (Table 4 and Fig. S43). ~~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 their clusters.~~ Among the causal parents, snow fraction and forest fraction exhibit both high statistical significance and relatively strong edge strength. In contrast, the remaining causal parents have either relatively low edge strength or high p-values in the likeli-hood ratio test (Table 4 and Fig. S43). In most cases, the performance of RF~All and RF~Par models in the test mode is comparable (Fig. S44), especially in the geology cluster numbers 2, 3, 5, and 7 (Table S2). The same as other signatures, GAM~All performs well only in the training simulation. The~~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 statistically significant for the GAMs (Table S1). 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). According to the Table 5, the causal parents lead to increase in accuracy of the GAM model by 17%, 75%, 14%, and 15% for climate, geology, soil, and vegetation categories~~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).~~

36

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	-24.44	-28.57	-17.65	-38.00	8.7	-35.71	26.92	-26.90	-20	-32.00	4.00	-38.1
high_q_dur	-25.64	-39.58	-15.15	-30.23	50.00	-19.57	26.92	-23.81	15.38	-32.50	17.39	-28.95
high_q_freq	-20.00	-50.00	-25.93	-62.79	65.00	-30.77	21.74	-41.03	5.88	-31.11	15.88	-50.00
low_q_dur	-11.11	-10.87	11.54	-17.65	85.00	-7.89	7.41	-14.71	0.00	-14.29	18.18	-23.53
low_q_freq	-9.52	-12.96	0.00	-27.27	42.86	-10.86	33.33	-14.63	-9.78	-12.00	25.00	-26.32
q_mean	-3.23	-5.43	-11.84	-15.79	12.33	-8.75	7.5	-4.71	4.88	-10.71	1.37	-12.00
Q5	-8.20	-11.76	12.90	-11.11	34.37	-11.54	41.18	-3.85	-5.26	-15.09	30.77	-13.64
Q95	-5.68	-4.60	-9.23	-12.50	28.33	-1.37	14.29	-2.47	-9.23	-12.50	4.76	-10.45
runoff_ratio	-20.48	-15.85	-38.33	-35.38	13.46	-23.19	1.61	-17.33	0.00	-17.11	-9.09	-23.81
slope_FDC	-11.86	-18.57	-12.50	-33.96	42.42	-12.96	25.00	-15.00	-4.44	-23.21	33.33	-22.73
stream_elast	0.00	-15.22	16.67	-24.14	75.00	-16.67	14.29	-15.62	-3.70	-27.03	15.00	-23.33

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. In contrast, BN in the baseline mode ranks 11th, which is the lowest among all models in this mode. This suggests that BN is either not sensitive or only weakly sensitive to sample size. For the other models, the baseline mode consistently ranks among the top 10 in terms of performance. Examining the top 10 rankings of the models across all environments reveals distinct patterns. For BN, the top-performing clusters include 2 from the climate category, 3 from geology, 2 from soil, and 1 from topography. For GAM~Par, the top 10 includes the baseline, 1 cluster each from climate and topography, 2 from geology and vegetation, and 3 from the soil category. For GAM~All, the top ranks comprise the baseline, 3 clusters from climate, and 2 clusters each from soil, topography, and vegetation, with none from geology. In the case of RF~Par, the top 10 includes 1 cluster each from the baseline, climate, topography, and vegetation, 2 from geology, and 4 from the soil category. Lastly, for RF~All, the top-performing clusters include 1 from the baseline, 1 cluster each from geology and topography, 2 from climate and vegetation, and 3 from the soil category. Overall, across the top 10 rankings of all models, clusters associated with the soil category appear most frequently, with 15 occurrences. This is followed by the climate category with 9 occurrences, geology and vegetation with 8 each, and topography with 6. Among all environments, Climate 4, followed by Vegetation 6, are the catchment groups where all models consistently achieve high rankings. Catchments in both climate 4 and Vegetation 6 are characterized by relatively

low precipitation, high evapotranspiration, and low maximum leaf area index. On the other hand, Vegetation 4, followed by Topography 4, are the environments where catchments consistently exhibit the lowest model performance. Catchments in Vegetation 4, which are mostly located in the southern and southeastern U.S., show high variability in forest fraction and maximum leaf area index. In contrast, catchments in Topography 4 are primarily situated in the Rocky Mountains and are characterized by high elevation and steep slopes.

~~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 the 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.~~

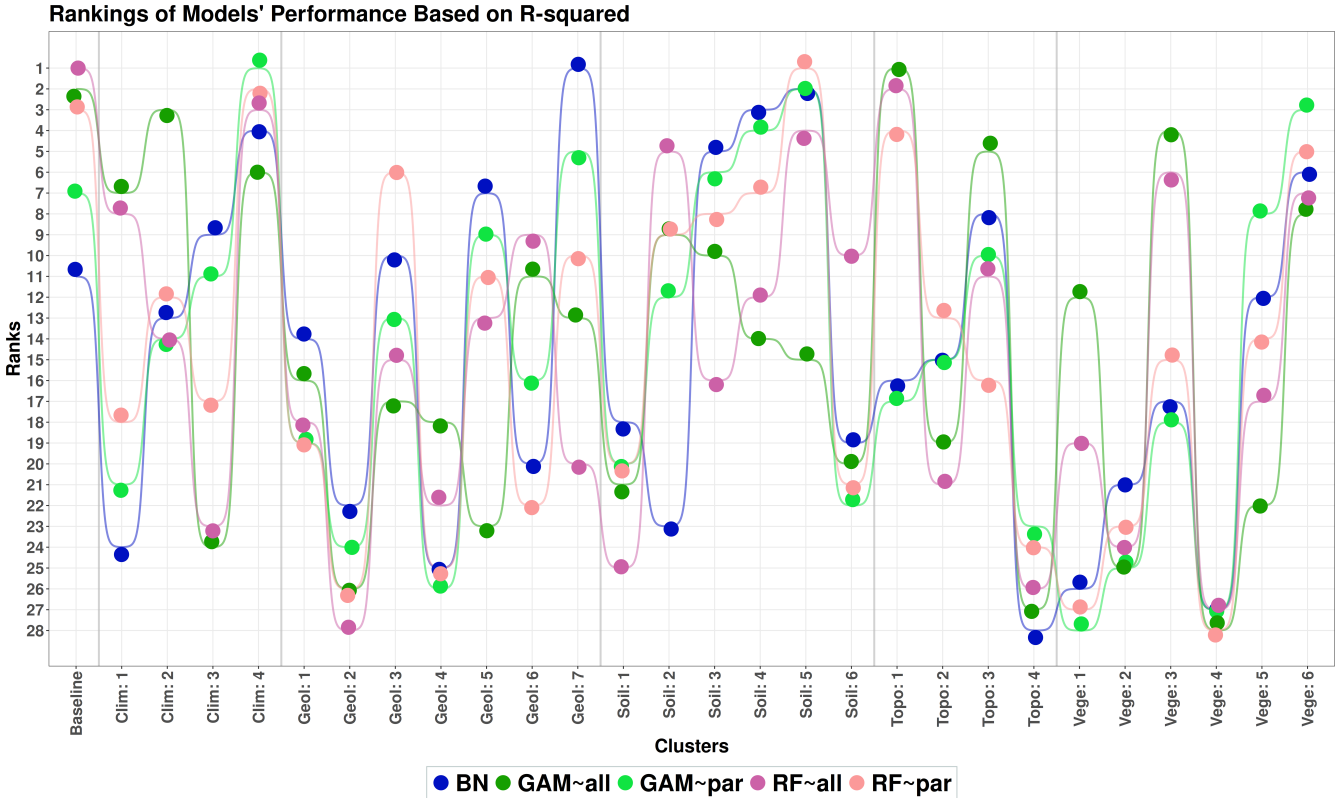[Comment: The figure has been updated for the new set of predictor variables.]

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.

5 Discussion

The aims of this study have tried to (1) recover the causal graph, represented as a Directed Acyclic Graph (DAG), from catchment attributes, climate characteristics, and runoff signatures; (2) predict runoff signatures using their causal parents as well as all variables in the DAGs; and (3) compare the predictive performance of models using only causal parents (which is an independent causal mechanism) versus those using all available variables.

785 PC-stable algorithm (Colombo and Maathuis, 2014) is used to uncover the underlying causal relationships between runoff signatures, catchment attributes, and climate characteristics. To improve the plausibility of the resulting graph, background knowledge is applied before running PC. Applying background knowledge reduces the risk of spurious or false positive edges, decreases the number of equivalence classes, and improves the stability of the learned causal graph (Meek, 1995; Spirtes et al., 2001; Perković et al., 2017; Bang and Didelez, 2025). The background knowledge is applied by blocking the implausible 790 edges. In hydrological systems, many causal directions are well understood from process-based reasoning and can be used to restrict the search space of the algorithm. By enforcing these structural constraints (e.g., forbidding reverse causality from runoff to climate), the stability of the learned causal graph is improved, and it allows us to derive a complex causal graph for the catchment area.

To mitigate the rate of false positives and negatives (Type I and II errors) in the PC algorithm due to our limited sample size 795 (Li and Wang, 2009), we start the causal discovery analyses by applying a relatively lenient significance threshold (Kalisch and Bühlman, 2007). This threshold increases the risk of false-positive edges. Therefore, to evaluate the quality of the fit, we assess the significance levels of dependencies between adjacent nodes and their corresponding strengths (Petersen et al., 2021). We perform a heuristic evaluation using cubic spline regression and the likelihood ratio test, which allows for examining nonlinear dependencies between adjacent nodes, particularly given that real-world hydrological systems often exhibit nonlinear 800 relationships (Kirchner, 2024). To further assess the strength and stability of the inferred links, we perform bootstrap resampling and learn the network structure 1000 times by the PC algorithm using a more stringent significance level (Scutari and Nagarajan, 2013). This approach helps us investigate potential false-positive edges (Type I errors) in the DAG obtained using the lenient significance threshold. Among all the inferred edges in the resulting DAGs, the link between low precipitation frequency and maximum leaf area index exhibits significantly low strength.

805 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 characteristics (mean, seasonality, and frequecy) ~~seasonality~~, vegetation cover, ~~mean precipitation,~~ 810 and catchment area ~~slope~~. This suggests that precipitation intensity, often driven by seasonality, influences runoff-generating mechanisms like the infiltration excess process (Nanda et al., 2019). Area ~~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. Additionally, the PC ~~However, the PC causal~~

~~discovery~~ results for low flows (Q5) includes two variables belong to~~were expected to identify geological~~ the geology, two
climate, and one topography ~~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). However, the causal parents identified by the PC method for the slope of the flow duration curve include only climate variables. While this flow signature is strongly influenced by catchment storage, shaped by geology, topography, and land cover (Dey et al., 2024), the method failed to capture these non-climatic influences. The identified causal parents of high flow frequency include high precipitation frequency and geological porosity. However, land cover is also recognized as a key driver of this flow signature (Zabaleta et al., 2018), though it was not captured by the causal discovery method. Therefore, while the PC algorithm identifies physically meaningful causal parents for most signatures, it occasionally fails to capture expected parent variables from certain categories that are known to influence those signatures. ~~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.~~ Furthermore, the derived causal parents by the PC are not necessarily the highest correlated variables with the runoff signatures. This highlights that there can be strong causal relationships between variables even when their statistical associations Gao et al. (2023).

The obtained DAGs~~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 of hydrological systems at the catchment scale. 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 soil and climate~~topographic~~ clusters for most signatures with consistent high accuracy rankings (Fig. 8). However, in Vegetation Cluster 4 and Topography Cluster 4 ~~characterized by high elevation, steep slopes, and low precipitation~~all models struggle to predict signatures accurately. In the catchments of Vegetation Cluster 4, forest fraction varies almost uniformly between 0 and 1, and as a result, the maximum leaf area index exhibits a similar pattern (Fig. 2). This high variability, combined with the small sample size of 69 catchments, leads to poor model performance (Table 3). On the other hand, catchments in the Topography Cluster 4 are characterized by high elevation and low precipitation. The low prediction accuracy in this cluster~~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 and~~,~~ Soil 5~~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.~~

The causal GAM model, GAM~Par, outperforms the non-causal GAM~All in most environments during testing, despite the latter exhibiting higher accuracy in the training mode. GAM~All achieves its best performance in the baseline mode, where the entire dataset is used for simulation. However, even in these baseline cases, its performance remains comparable to that of GAM~Par.~~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.~~

850   In the case of RF models~~For the RF models~~, RF∼All achieves the highest accuracy across most environments compared to the other models; however, the performance of the RF∼Par is still~~this difference is also~~ comparable to RF∼All.~~insignificant for half of the signatures~~. The difference in performance between RF∼All and RF∼Par is statistically significant in most cases, as determined by a non-parametric permutation test (see the Supplementary Materials Section 3). However, this significance does not merely represent the magnitude of difference between the models. The results of the test also depend on the spread of results

855 across the R-squared and RMSE space. For example, two models might differ in average accuracy, and if their performance varies widely during bootstrapped training and testing, the difference may not be statistically significant. Conversely, even a small difference in accuracy can be statistically significant if the performances of the models are consistently stable. 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

860 relationships between variables, although further investigation is needed. The difference between causal and non-causal RF models is mostly statistically significant across clusters of the baseflow index, high flow frequency, low flow frequency, mean daily flow, runoff ratio, the slope of the flow duration curve, and the stream precipitation elasticity.~~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

865 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. This pattern holds for GAM∼Par across most clusters, and for RF∼Par across the majority of clusters related to low flow duration, high flow duration, low flows, and high flows.

~~   Despite BN having lower accuracy than GAM and RF, it shows the smallest difference between training and test results~~

870 ~~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~~

875 ~~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~~

880 ~~cases, models using causal parents even achieve higher accuracy, such as predicting high flow duration in the geology category for GAM (Table 5).~~

~~   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~~

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). However, there is still room for further investigation, as the causal graphs obtained using the proposed methods require deeper analysis. In this study, we used the PC algorithm to identify the causal parents of runoff signatures. Other methods are specifically designed to target causal parents of a given variable, such as Invariant Causal Prediction (ICP), and could offer additional insights (see Peters et al. (2016); Heinze-Deml et al. (2018b); Kook et al. (2024)). The insights gained from causal discovery can~~interconnections~~ not only improve the understanding of hydrological systems at the catchment scale, 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 the uncertainty of such a model, particularly when combined with machine learning and classification algorithms (Herman et al., 2015; Singh et al., 2015; AghaKouchak et al., 2022).

6 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 catchment 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 500~~100~~ random samplings of training and test sets, resulting in a total of 140,000~~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 500~~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

920   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.

925   - The high accuracy of non-causal models, GAM~All and RF~All, in the baseline scenarios may be attributed to overfitting or 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.

- Signatures are most predictable when causal and non-causal models are trained on catchments with homogeneous soil properties.

930   - 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 is also the case for the soil and the climate categories where their variables are mostly independent of the other categories. This information helps identify environments where training models achieve higher accuracy, reduced uncertainty, and greater robustness.

935   - 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

940   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 offers an insight into the application ofhighlights the importance of causal inference methods in understanding runoff-generating mechanisms in hydrological systems.

945   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.

950 *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 500100 executions of each model.

Signature	R squared (Test Set)					RMSE (Test Set)				
	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	0.39	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

Table A1. R-squared and RMSE values for test simulations of baseline models. The values are an average of 500~~100~~ executions of each model.

Signature	R squared (Test Set)					RMSE (Test Set)				
	BN	GAM~All	GAM~Par	RF~All	RF~Par	BN	GAM~All	GAM~Par	RF~All	RF~Par
baseflow_index	0.30	0.45	0.34	0.63	0.45	0.14	0.12	0.13	0.10	0.12
high_q_dur	0.27	0.39	0.29	0.48	0.29	8.67	8.08	8.57	7.40	8.69
high_q_freq	0.27	0.40	0.32	0.52	0.26	25.01	22.76	24.22	20.33	25.48
low_q_dur	0.28	0.36	0.32	0.46	0.41	18.39	17.50	17.88	15.94	16.66
low_q_freq	0.30	0.38	0.42	0.54	0.47	69.02	63.31	65.40	56.62	60.50
q_mean	0.82	0.93	0.90	0.92	0.87	0.65	0.40	0.49	0.46	0.59
Q5	0.50	0.61	0.56	0.68	0.60	0.19	0.17	0.18	0.15	0.17
Q95	0.79	0.88	0.83	0.87	0.83	2.25	1.68	2.00	1.79	2.03
runoff_ratio	0.62	0.83	0.66	0.82	0.69	0.14	0.10	0.13	0.10	0.13
slope_FDC	0.44	0.59	0.52	0.70	0.57	0.38	0.33	0.35	0.28	0.33
stream_elast	0.33	0.32	0.32	0.46	0.39	0.64	0.65	0.64	0.57	0.61

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.~~

Signature	R squared values for causal models														
	Climate			Geology			Soil			Topography			Vegetation		
	BN	GAM	RF	BN	GAM	RF	BN	GAM	RF	BN	GAM	RF	BN	GAM	RF
baseflow_index	0.24~~0.31~~	~~0.35~~	**0.44**	0.25~~0.28~~	~~0.31~~	~~0.36~~	0.30~~0.33~~	~~0.33~~	~~0.38~~	0.24~~0.32~~	~~0.36~~	~~0.41~~	0.22~~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 A2. The R-squared values of causal models for each category, which are~~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	0.24	0.28	0.31	0.25	0.25	0.27	0.30	0.33	0.34	0.25	0.28	0.34	0.22	0.24	0.26
high_q_dur	0.24	0.28	0.30	0.35	0.39	0.37	0.33	0.33	0.32	0.26	0.30	0.27	0.23	0.27	0.27
high_q_freq	0.17	0.20	0.16	0.30	0.33	0.27	0.26	0.28	0.23	0.30	0.36	0.31	0.19	0.22	0.18
low_q_dur	0.24	0.29	0.28	0.40	0.37	0.35	0.30	0.29	0.29	0.43	0.41	0.42	0.23	0.26	0.26
low_q_freq	0.25	0.28	0.32	0.32	0.30	0.35	0.30	0.32	0.35	0.33	0.37	0.44	0.23	0.25	0.28
q_mean	0.61	0.67	0.64	0.80	0.82	0.73	0.83	0.86	0.81	0.83	0.86	0.75	0.72	0.74	0.66
Q5	0.31	0.35	0.40	0.40	0.43	0.46	0.45	0.48	0.50	0.35	0.36	0.45	0.34	0.34	0.38
Q95	0.58	0.59	0.56	0.75	0.77	0.72	0.77	0.80	0.79	0.58	0.59	0.56	0.65	0.66	0.60
runoff_ratio	0.30	0.37	0.42	0.58	0.59	0.53	0.62	0.63	0.62	0.66	0.68	0.63	0.46	0.50	0.48
slope_FDC	0.26	0.35	0.35	0.42	0.47	0.47	0.44	0.50	0.51	0.37	0.43	0.43	0.25	0.32	0.34
stream_elast	0.22	0.21	0.22	0.28	0.28	0.30	0.25	0.24	0.27	0.25	0.26	0.27	0.23	0.23	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

Table A3. The Rsquared values of non-causal models for each category which are~~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	0.34	0.50	0.23	0.42	0.26	0.46	0.35	0.50	0.25	0.42
high_q_dur	0.33	0.43	0.26	0.46	0.26	0.42	0.26	0.40	0.23	0.38
high_q_freq	0.27	0.43	0.20	0.39	0.23	0.39	0.34	0.45	0.19	0.36
low_q_dur	0.26	0.34	0.20	0.38	0.27	0.34	0.41	0.49	0.22	0.34
low_q_freq	0.28	0.44	0.21	0.39	0.24	0.41	0.41	0.50	0.20	0.38
q_mean	0.76	0.76	0.73	0.80	0.80	0.85	0.82	0.84	0.73	0.75
Q5	0.31	0.45	0.32	0.52	0.34	0.52	0.38	0.53	0.26	0.44
Q95	0.65	0.64	0.60	0.73	0.70	0.81	0.65	0.64	0.63	0.67
runoff_ratio	0.60	0.65	0.52	0.69	0.62	0.75	0.68	0.76	0.55	0.63
slope_FDC	0.40	0.53	0.33	0.54	0.40	0.60	0.45	0.56	0.24	0.44
stream_elast	0.18	0.29	0.16	0.36	0.21	0.32	0.27	0.37	0.20	0.30

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, ENVIRON-MENTAL 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.

Bang, C. W. and Didelez, V.: Constraint-based causal discovery with tiered background knowledge and latent variables in single or overlapping datasets, arXiv preprint arXiv:2503.21526, 2025.

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. and Maathuis, M. H.: Order-Independent Constraint-Based Causal Structure Learning, JOURNAL OF MACHINE LEARNING RESEARCH, 15, 3741–3782, 2014.

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.

1000   Deng, J., Shan, K., Shi, K., Qian, S. S., Zhang, Y., Qin, B., and Zhu, G.: Nutrient reduction mitigated the expansion of cyanobac-
        terial 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.

        Deng, Y. and Ebert-Uphoff, I.: Weakening of atmospheric information flow in a warming climate in the Community Climate System Model,
        GEOPHYSICAL RESEARCH LETTERS, 41, 193–200, https://doi.org/10.1002/2013GL058646, 2014.

1005   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.

        Dey, P., Mathai, J., Sivapalan, M., and Mujumdar, P. P.: On the regional-scale variability in flow duration curves in Peninsular India, HY-
        DROLOGY AND EARTH SYSTEM SCIENCES, 28, 1493–1514, https://doi.org/10.5194/hess-28-1493-2024, 2024.

        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
1010   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.

        Ebert-Uphoff, I. and Deng, Y.: Causal Discovery for Climate Research Using Graphical Models, JOURNAL OF CLIMATE, 25, 5648–5665,
        https://doi.org/10.1175/JCLI-D-11-00387.1, 2012.

1015   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,
1020   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.

1025   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 HY-
        drogeology MaPS (GLHYMPS) of permeability and porosity, GEOPHYSICAL RESEARCH LETTERS, 41, 3891–3898,
        https://doi.org/10.1002/2014GL059856, 2014.

1030   Glymour, C., Zhang, K., and Spirtes, P.: Review of Causal Discovery Methods Based on Graphical Models, FRONTIERS IN GENETICS,
        10, https://doi.org/10.3389/fgene.2019.00524, 2019.

        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–&,
1035   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, 1040   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.

Hausser, J. and Strimmer, K.: Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks., Journal of Machine Learning Research, 10, 2009.

1045   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, 2018a.

Heinze-Deml, C., Peters, J., and Meinshausen, N.: Invariant Causal Prediction for Nonlinear Models, JOURNAL OF CAUSAL INFERENCE, 6, https://doi.org/10.1515/jci-2017-0016, 2018b.

1050   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.

1055   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, 1060   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 1065   Research, 8, 2007.

Kalisch, M., Maechler, M., Colombo, D., Maathuis, M. H., and Buehlmann, P.: Causal Inference Using Graphical Models with the R Package pcalg, JOURNAL OF STATISTICAL SOFTWARE, 47, 1–26, 2012.

Kirchner, J. W.: Characterizing nonlinear, nonstationary, and heterogeneous hydrologic behavior using ensemble rainfall-runoff analysis (ERRA): proof of concept, HYDROLOGY AND EARTH SYSTEM SCIENCES, 28, 4427–4454, https://doi.org/10.5194/hess-28-4427-1070   2024, 2024.

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

Kook, L., Saengkyongam, S., Lundborg, A. R., Hothorn, T., and Peters, J.: Model-Based Causal Feature Selection for General Response Types, JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, https://doi.org/10.1080/01621459.2024.2395588, 2024.

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.

Li, J. and Wang, Z. J.: Controlling the False Discovery Rate of the Association/Causality Structure Learned with the PC Algorithm, JOURNAL OF MACHINE LEARNING RESEARCH, 10, 475–514, 2009.

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.

Meek, C.: Causal inference and causal explanation with background knowledge, arXiv preprint arXiv:1302.4972, 1995.

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 & SOFT-WARE, 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 TRANSAC-TIONS ON GEOSCIENCE AND REMOTE SENSING, 57, 1502–1513, https://doi.org/10.1109/TGRS.2018.2867002, 2019.

Perković, E., Kalisch, M., and Maathuis, M. H.: Interpreting and using CPDAGs with background knowledge, arXiv preprint arXiv:1707.02171, 2017.

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.

Petersen, A. H., Osler, M., and Ekstrom, C. T.: Data-Driven Model Building for Life-Course Epidemiology, AMERICAN JOURNAL OF EPIDEMIOLOGY, 190, 1898–1907, https://doi.org/10.1093/aje/kwab087, 2021.

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.

1150 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.

1155 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
1160   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 meth-
1165   ods: 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
1170   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.

1175 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.

Sanchez-Romero, R., Ito, T., Mill, R. D., Hanson, S. J., and Cole, M. W.: Causally informed activity flow models provide mechanistic insight
1180   into network-generated cognitive activations, NEUROIMAGE, 278, https://doi.org/10.1016/j.neuroimage.2023.120300, 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,
1185   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.

1190   Scutari, M.: Learning Bayesian networks with the bnlearn R package, arXiv preprint arXiv:0908.3817, 2009.

Scutari, M. and Nagarajan, R.: Identifying significant edges in graphical models of molecular networks, ARTIFICIAL INTELLIGENCE IN MEDICINE, 57, 207–217, https://doi.org/10.1016/j.artmed.2012.12.006, 2013.

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.

1195   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 func-
1200   tion, 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
1205   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.

1210   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.

1215   Todorovic, A., Grabs, T., and Teutschbein, C.: Improving performance of bucket-type hydrological models in high lati-
tudes 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,
1220   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.

Zabaleta, A., Garmendia, E., Mariel, P., Tamayo, I., and Antiguedad, I.: Land cover effects on hydrologic services under a precipitation gradient, HYDROLOGY AND EARTH SYSTEM SCIENCES, 22, 5227–5241, https://doi.org/10.5194/hess-22-5227-2018, 2018.

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.

Zuk, O., Margel, S., and Domany, E.: On the number of samples needed to learn the correct structure of a Bayesian network, arXiv preprint arXiv:1206.6862, 2012.