2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10020794

Causal Discovery for Feature Selection in Physical Process-Based Hydrological Systems

Paras Sheth*, Ting Liu†, Durmus Doner*, Qi Deng§, Yuhang Wei‡, Rebecca Muenich†, John Sabo§,
K. Selçuk Candan*, Huan Liu*
*School of Computing and Augmented Intelligence, Arizona State University
†School of Sustainable Engineering and the Built Environment, Arizona State University
‡Future H2O, Julie Ann Wrigley Global Futures Laboratory, Arizona State University
§Department of River-Coastal Science and Engineering, ByWater Institute, Tulane University
Email: *{psheth5, ddoner, candan, huanliu}@asu.edu, †{tliu154, rmuenich}@asu.edu,
‡{qdeng6, ywei66}@asu.edu, §jsabo1@tulane.edu

***Abstract*—Physical process-based hydrological models are widely adopted to simulate the water quantity or quality. One of the most commonly used hydrological models is Soil and Water Assessment Tool (SWAT). SWAT models for a large watershed can have over tens of thousands of Hydrological Resource Units (HRUs) which necessitates considerable computational resources. One way to speed up applications of the SWAT model could be to leverage machine learning techniques to identify the crucial features for the prediction task – feature selection. However, majority of the feature selection techniques rely on correlations or some form of a score metric (e.g. mutual information). Furthermore, since correlation does not imply causation, it is important to identify the causal features to improve the prediction accuracy while enhancing the interpretability of machine learning models. However, the SWAT model uses multiple data inputs and features that typically vary by space/HRUs, but may or may not vary over time. This makes it difficult to directly utilize causal discovery models to infer the causal relations. Furthermore, due to the lack of the ground truth causal graph for the SWAT model it is difficult to comment on the validity of the learned causal relations. To overcome these problems, we propose a novel framework that first infers the causal relations for the daily scale of the SWAT data using causal discovery algorithms. Then, it utilizes a community detection module to group similar features together for better interpretability. Finally, it identifies the stable causal relations that appear most often across all the timesteps and leverage them for the prediction of the water quantity. By utilizing only the causal features for the prediction of the target variable can lead to high accuracy as it removes the reliance on spurious correlations. Furthermore, we conduct extensive experiments to validate the effectiveness of the proposed framework along with a real-world case study to evaluate whether the selected features are interpretable or not.**

***Index Terms*—causal discovery, feature selection, hydrological systems, neural networks, SWAT models**

I. Introduction

The physical process-based hydrology models use pre-set equations with transient and spatial features to simulate the river stream water quantity and quality [1]. These models have unique advantages, such as the ability to comprehend the mechanics driving the hydrologic model, the ability to explain observed and simulated phenomena, and the ease of transferring the calibrated model features to other regional settings [2]. However, the improvement of the hydrologic models often require more descriptive features and a higher spatial and temporal resolution [3] which increases the complexity and computational needs of the model [4], [5].

Soil and Water Assessment Tool (SWAT) is a common physical process-based hydrology model that simulates water quantity and quality [6]. SWAT considers multimodal features including fixed-spatial and transient data. The fixed-spatial inputs include slope (typically derived from digital elevation models or DEM), land use, and soil type. On the other hand, the transient data includes precipitation, temperature, land management schedules, and optional wind speed, relative humidity and solar radiation. The SWAT model first calculates the location-based water quality and quantity yield at the Hydrological Resource Units (HRUs). Their calculations are then aggregated into sub-basins and routed through river networks based on the structure of the watershed, where subsequent transformations can occur. Figure 1(d) shows the brief flowchart of SWAT calculations on water quantity. The HRUs in each sub-basin are defined by the user inputs mentioned above in the format of features shown in Table I. The total number of HRUs within a sub-basin depends on the unique combinations of the three fixed-spatial inputs within each sub-basin. As the SWAT model database has more than 100 soil and land uses available to simulate (besides, users can add new ones to the database), and the model usually consists of more than one sub-basin, the SWAT models typically developed for a large watershed can have over 50,000 HRUs. This spatial discretization when coupled with daily time steps over a decade's horizon, may require considerable time and computational resources.

The most common method to simplify the SWAT model is to merge minor HRU classes within major HRU classes by setting a land area threshold. Only those classes larger than the area threshold are included in the model [7]. However, there are only a limited number of studies to justify which threshold value is proper without losing accuracy of the model [8]. Additionally, the area alone does not necessarily determine the importance of different slope, soil type, and land use to

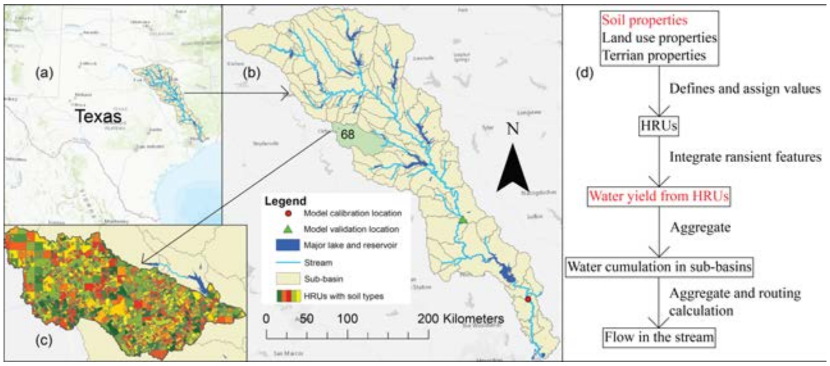

Fig. 1: (a) The location of the Trinity River basin in Texas. (b) The spatial elements of the Trinity basin SWAT model such as the sub-basins, river stream, and reservoirs (c) The HRUs in this study within the sub-basin 68. Different colors represent various soil types. (d) The brief steps of SWAT river flow simulation. The target features in this study are in red.

the hydrological process. A *better approach* to simplify the SWAT model would be to identify those features that are vital for SWAT outputs, especially to the water quantity output (i.e., streamflow). However, the complexity of the SWAT model makes it nearly impossible to analytically determine the features which cause the alteration of the simulated river flow. Classical feature selection techniques select features based on the correlations between predictive features and the target variable and fail to capture causal relationships between them. Furthermore, since causal relationships help in understanding the underlying mechanism of a system [9], [10], identification of causal features can lead to building interpretable and robust prediction models. Thus, by integrating causality in the feature selection approach one can aim to learn and identify interpretable features vital for the prediction task. In order to extract causal features, one needs to first infer the causal relations from the observed data, which is a challenging task.

There has been a wide range of works developed to identify and learn the causal features present in temporal data [11]. In SWAT models, however, there exist features that vary across HRUs but are constant across time (eg. soil parameterss) and there also exist features that vary across time but are constant across HRUs within a sub-basin (eg. temperature). Due to this conflicting nature of the features it is infeasible to simply apply traditional [12], [13] or temporal causal discovery methods to infer the causal relationships. Furthermore, since the SWAT model contains multitude of features it would be beneficial to identify the group of causal features that affect the river flow for better interpretation.

In this paper we propose a novel framework, named *Causality Aware Physical Systems (CAPS)*[1], to overcome the aforementioned problems. To deal with the conflicting nature of features, CAPS first slices the observational data into daily slices making it feasible to utilize traditional causal discovery models to infer the causal relations. CAPS relies on the assumption that each day can be considered as an independent unit and no inter-day causal edges exist in the daily causal graphs. Next, CAPS infers causal relationships from the daily slices and infers a daily causal graph. To enhance the interpretability of the features, CAPS utilizes community detection strategy to group similar causal features together. Finally, after inferring the causal communities for each day, CAPS identifies the stable relationships across time. These stable relations are then considered as causal features. This is supported by various studies that suggest, causal relationships are usually stable [14], [15]. Our main contributions are as follows:

- We study a novel problem setting of learning causal relations when a major subset of features are constant across time, while another subset of features is constant across the HRUs within a sub-basin;
- We propose a novel framework for causality aware feature selection for the SWAT model for better prediction and interpretation; and
- We demonstrate the effectiveness of the proposed framework with comparative analysis and real world case study for SWAT models. More specifically, based on the arguments that (a) causal features are generally more interpretable than other features [16] and that (b) they maybe lead to better prediction performance [17], [18], we hypothesize that if the features learned by CAPS can be demonstrated to be more interpretable compared to features learned by other methods and if the resulting prediction accuracy is better than the traditional feature selection techniques, then we conclude that CAPS is able to identify causally important variables w.r.t a given target variable; in this case the river flow.

II. Methods

For physical process-based hydrological models, the identification of streamflow alteration drivers is done via numeric experiment where model fitness to data is evaluated with different ensembles of input data, model structure, and hyperparameter settings [19]. The authors in [20] revealed how river water content in lake is controlled by catchment size and climate forcing variation via perturbation of basin sizes and layouts. The driving force of drainage density on land-river water and carbon dynamics was addressed in [21].

[1] https://github.com/paras2612/CAPS

From data-driven approaches, such drivers can be selected via various Feature Selection (FS) methods. The common three categories of FS are-*filter*, *wrapper*, and *embedding* methods [22] have been widely applied in machine learning hydrological models [23], [24]. Filter methods such as Information Gain, Correlation, serve as the pre-process for selecting features and are independent of predictor choice. Recent example studies are testing effectiveness of eight filter-based methods in boosting forecasting performance on monthly streamflow [25], satisfying relative humidity predictability via combination of extreme gradient boost (XGBoost) and machine learning predictors. Wrapper methods evaluate variables based on their contribution to predictor power and treat the learning model as pure black box, such as Recursive Feature Elimination (RFE) used to improve daily streamflow prediction [26]. Embedded methods integrate features selection in the training process and is optimized together with the learning process, such as Lasso Regularization, Random Forest Importance (RFI) are used for groundwater quality prediction [24].

Since for physical processes it is crucial to understand why the river flow behaves as it does and what are the critical features on which experts should focus on, when analyzing the river flow, its been discovered that the non-causal feature selection methods may not yield the best results [27], [28]. One possible solution to this problem is to leverage causal discovery algorithms that aim at discovering the causal relationships from observational data under certain assumptions.

Causal discovery methods are categorized as constraint based [29], [30] and score based [31]. A series of work have focused on leveraging causal discovery algorithms for feature selection in machine learning models [32], [33]. For instance, the authors in [34] aimed to learn a candidate Markov blanket (CMB [9]) with causal information. Then, utilized neighborhood conditional mutual information to delete the false positives in CMB. However, this method was designed for online streaming data rendering it inapplicable for our dataset. Because, the online streaming data possesses features that vary across the stream, whereas in our case some features vary across time (soil features) and some features vary across HRUs but are constant across time (temperature). Another interesting work is a survey of causality-based feature selection methods [9]. However, we hypotheisze that, these methods may not yield the CMB in our setting due to the conflicting nature of features in our dataset. For instance, the SLL algorithm [32] is a score-based variant of the divide-and-conquer Markov Boundary learning algorithms. In the causal structure learning phase, SLL employs a Bayesian Network structure learning algorithm. Then SLL implements the symmetric check using the AND and OR rules.

III. Preliminaries

A. *Causal Discovery*

Discovering and utilizing causal relations is a vital effort in numerous branches of science [31], [35], [36]. Generally, Directed Acyclic Graphs (DAGs) are used to represent the causal relations present in a system.

Let $X = X_1, \ldots, X_p$ refer to a set of p random variables (features). In a causal DAG, nodes represent random variables, and edges represent causal links. For instance, if $X_i \rightarrow X_j$ exists in a DAG, it implies that X_i is a cause of X_j. Since recovering the true causal structure along with edge directions is not always possible from observational data alone [37], causal discovery methods are accompanied by certain common assumptions: Let $G = (X, E)$ represent the DAG containing X nodes and E edges, and P be the joint probability distribution of X.

Assumption 1 (Causal Markov Condition [33]): The Causal Markov Condition holds true iff given the set of all its parents, a node in G is independent of all its non-descendants.

Assumption 2 (Faithfulness [33]): There is no conditional independence between G and P unless entailed by the Causal Markov Condition.

Assumption 3 (Causal Sufficiency [33]): All the information to learn the causal graph is present in the dataset. In other words, there are no unmeasured common causes (confounders) between the different features.

B. *Hydrological Systems*

The SWAT model (SWAT2012 rev. 681) in this work describes the Trinity River Basin in eastern Texas. The model consists of 107 sub-basins and 39,104 HRUs (including 14,168 agricultural HRUs) with a total area of 40380 km^2 (Figure 1(a,b,c)) and 15 reservoirs [38]. Contrary to the typical SWAT method that removes the minor HRU classes based on the overall percentage area threshold, agricultural area (not urban) HRUs are defined by parcel boundaries (except urban areas) to investigate cropland management schemes on the field level in this work. We generated the HRUs by integrating over land parcel data (TNRIS [39]) and land use data (CDL [40]). We determine the agricultural plant species summarized from the land use data from 2009 to 2013, with one or two-year's cycle for crop rotations. The type with the most significant area then defines the soil [41] in each HRU. The land slope of each HRU is calculated from the DEM [42]. The urban areas were not the focus of the study and therefore the urban HRUs were generated primarily for computation efficiency so spatially disconnected grid cells could be aggregated into the same HRU as long as they have the same land use and soil type.

Both the model transient input and output are in daily steps. Model flow calibrations were performed near the outlet of the Trinity River from 2002 to 2003[2]. The simulated model achieved streamflow calibration statistics of R^2= 0.76, Nash Sutcliffe efficiency (NSE)= 0.76, and percent bias (PBAIS) = -0.44 compared to observations (sub-basin 99, Figure 1(b)). The model validation is performed in the middle part of the Trinity river basin (sub-basin 85, Figure1(b)) to demonstrate that the calibrated model feature values are valid throughout the different parts of the watershed. The validation point has R^2= 0.74, NSE= 0.67, and PBAIS= 11.44 compared to the daily

[2] https://waterdata.usgs.gov/usa/nwis/uv?08066500

Source Data	Feature Type	Notation	SWAT Feature Name	Description	Symbol
Soil data	Soil-related features	X_1 - X_4	Depth.................mm.1,2,3, DEP_IMP	Soil depth parameters 1-4	SDEP1-4
		X_5 - X_7	Bulk.Density.Moist..g.cc.1,2,3	Soil density parameters 1-3	SDEN1-3
		X_8 - X_{11}	Ave..AW.Incl..Rock.Frag1,2,3,Crack.volume.potential.of.soil	Soil particle size parameters 1-4	SPS1-4
		X_{12} - X_{14}	Ksat...est.........mm.hr.1,2,3	Soil hydraulic parameters 1-3	SH1-3
		X_{15} - X_{17}	CF, CFH, CFDEC	Soil-water interaction parameters 1-3	SWI1-3
Calibrated	Groundwater-related features	X_{18} - X_{22}	GW_DELAY, GWQMN, GW_REVAP, REVAPMN, RCHRG_DP	Groundwater vertical flow parameters 1-5	GWV1-5
		X_{23} - X_{25}	ALPHA_BF, GW_SPYLD, ALPHA_BF_D	Groundwater horizontal flow parameters 1-3	GWH1-3
DEM / slope	Location-related features	X_{26} - X_{28}	HRU_FR, DIS_STREAM, AREAkm2	HRU's size and distance to the stream	LRF1-3
	Terrain-related features	X_{29} - X_{31}	SLSUBBSN, SLSOIL, HRU_SLP	Slope parameters 1-3	TSLP1-3
Land use (partially)		X_{32} - X_{36}	OV_N, LAT_TTIME, CANMX, SURLAG, R2DJ	Land surface shape parameters 1-5	TLSS1-5
		X_{37} - X_{39}	ESCO, EPCO, CN2	Land surface water loss parameters 1-3	TLSW1-3
Climate data	Transient features	X_{40}	Precipitation	Daily precipitation (mm)	PCP
		X_{41}	Temperature	Average daily temperature (\textcelsius)	TMP
	Model output	X_{42}	WYLD(mm)	Water yield (generated) from HRU	WYLD

TABLE I: The notations of the SWAT features.

streamflow rate observations. All the HRUs in sub-basin 68 are selected as the experiment object in this work. The reason to select this sub-basin is that it is an agricultural-dominated region with a good variation of soil types. Besides, the sub-basin is one of the headwater sub-basins of the Trinity River, and the streamflow in this basin is sourced internally from the HRUs without interactions with other sub-basins. In this sub-basin, there are 1,002 types of soil, 1 class of slope, and 10 types of land use resulted in 2,715 HRUs. In comparison, the entire basin contains 15,394 types of soil, 1 class of slope, and 20 types of land uses yielding 39,104 number of HRUs. No HRUs from other sub-basins are selected in this study because a single sub-basin is sufficient to represent both spatial and transient variations: as the transient data is differentiated among the different days. However, future work may include the HRUs multiple sub-basins.

Most of the water quantity input and output calculations in the SWAT model are within the HRUs. Except for with reservoirs which have big impact on flow. The alteration of water output from the HRUs will affect the streamflow simulation results. In a reverse way, any feature that is not impacting the water quantity output in the HRUs will also not affect the simulation results (Figure 1(d)). The previously-mentioned data inputs convert into HRU-level features in the SWAT model include soil, groundwater, location, terrain-related, and transient features show in I. Among the features, the groundwater-related features and parts of the terrain-related features are mostly determined by assumptions and model calibrations. In comparison, the value of soil-related features are derived from high precision data, with over 10,000 of feature combinations, and is static over time. We target to identify the soil-related features which could potentially be simplified without altering the river streamflow simulation result as the demonstration of the method in this work. The result will help identify the soil classifications that differed only from those of non-water-yield-causal soil-related features. These soil classifications will be merged in future works to reduce the computation load on the SWAT model for water quantity simulations without losing accuracy.

There are in-total 42 features on the HRU-level ($X_{i,h,t}$). Here, i represents the features ($i = 1, 2, \dots, 42$), h represents the HRU ($h = 1, 2, \dots, 2715$), and t represents the time step (day, $t = 1, 2, ..., 730$). Among the 42 features, there are 17 soil-water-related features ($X_{1-17,h,t}$), 2 transient features ($X_{40-41,h,t}$), the water yield representing the model output on the HRU-level ($X_{42,h,t}$), and 20 other features related to groundwater, location, and terrain shape ($X_{18-39,h,t}$) (Table I). Except for the two transient features ($X_{40-41,h,t}$) and the water yield ($X_{42,h,t}$) which may change over different time steps and across different HRUs, all the other features always have the same value regardless of time (i.e., $X_{1-39,h,t} \equiv X_{1-39,h,t+1}$). We focus on the typical causal relationship between the individual ($X_{i,h,t} \rightarrow X_{42,h,t}, i = 1, 2, ..., 17$) or grouped ($X_{[i1,i2,i3,...],h,t} \rightarrow X_{42,h,t}, [i1, i2, i3, ...] \in 1, 2, ..., 17$) soil-related features and the water yield throughout all the HRUs (h) in 730 time-steps (t) (all the days in 2002-03).

IV. Proposed Method

In this work we aim to leverage causal discovery methods to improve feature selection in the SWAT model. As mentioned earlier, the SWAT model consists of multiple HRUs, which in turn contain multitude of features. The observed features vary across two domains, namely, the HRUs and time. Features such as temperature and precipitation rate vary across time but remain constant across the HRUs within a sub-basin, whereas features such as soil type vary across HRUs but are constant across time. Due to this conflicting nature of the features it is difficult to apply any form of temporal or traditional causal discovery models on the whole dataset. To overcome this problem we propose a novel framework called CAPS that works in three steps. First, it utilizes traditional causal discovery on daily data to capture the daily causal relationships. Second, it performs community detection to identify and capture stable relations. Finally, it identifies the important features and uses them to perform the prediction task. An overview of our proposed framework can be seen in Figure 2.

A. Causal Discovery for Feature Selection

Our fundamental step is to discover the causal relations underneath the observed data such that the features either vary across the HRUs or across time or a combination of both. This section shows how to construct the causal graph from the data.

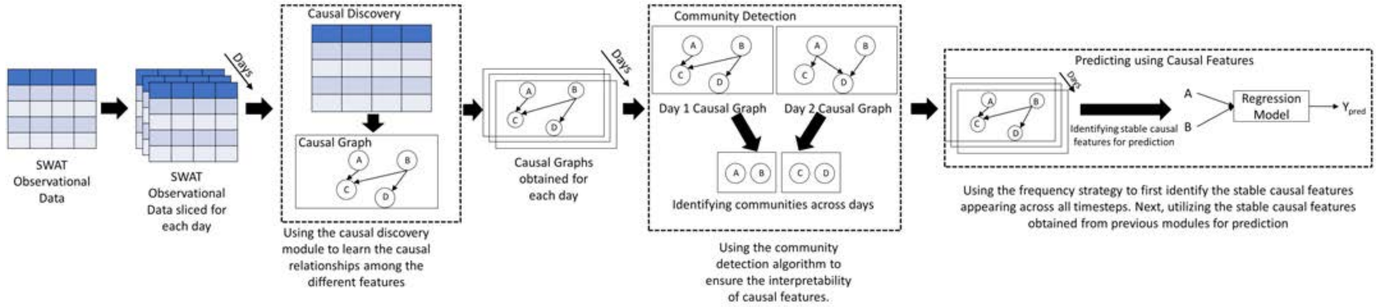
SWAT Observational Data
SWAT Observational Data sliced for each day
Days
Causal Discovery
Causal Graph
A
B
C
D
Using the causal discovery module to learn the causal relationships among the different features
Causal Graphs obtained for each day
Days
Community Detection
Day 1 Causal Graph
Day 2 Causal Graph
Identifying communities across days
Using the community detection algorithm to ensure the interpretability of causal features.
Predicting using Causal Features
Days
Identifying stable causal features for prediction
Regression Model
Y_{pred}
Using the frequency strategy to first identify the stable causal features appearing across all timesteps. Next, utilizing the stable causal features obtained from previous modules for prediction

Fig. 2: An overview of the architecture of the proposed CAPS. First the data is sliced to represent daily data. Then traditional causal discovery approaches are utilized to learn the causal relationships daily. Next, the causal community detection module aims to utilize the daily causal graphs and group similar features together. Finally, by utilizing the frequency strategy CAPS identifies the stable causal relations with respect to the target variable and use them to perform the prediction task.

We first proceed with splitting the observed data $\mathbf{X}$ into daily segments, such that given a set of features $\mathbf{X} = \{X_{i,h,t}\}$ where t represents the time-steps, h represents the HRUs and i represents features. Each slice contains the observed features across all the HRUs for each day – since all the features in a given slice share the same timestamp, we drop the index t and use the short hand $X_{i,h}$ to refer to features when it is clear from the context that they are in the same time slice. After obtaining this daily slice we build on the state-of-the-art PC algorithm [12] to identify the causal relationships present in the observational data and learn causal graphs for each day. The PC algorithm functions in two steps. First, it learns a skeleton graph from the observed data that consists of undirected edges only. Next, it orients the undirected edges to form an equivalence class of Directed Acyclic Graphs. The PC algorithm utilizes the principle of d-seperation [43] to identify the causal variables. d-separation is a criterion for deciding, from a given causal graph, whether a set $X_{i,h}$ of variables is independent of another set $X_{j,h}$, given a third set $X_{k,h}$

Definition 1: In a Directed Acyclic Graph (DAG) G, vertex sets $X_{i,h}$ and $X_{j,h}$ are said to be d-connected relative to a set of vertices $X_{k,h}$ such that X_i, $X_j \notin X_k$ if (i) there exists a collider free path between $X_{i,h}$ and $X_{j,h}$ that traverses no member of $X_{k,h}$; and (ii) if a collider $X_{c,h}$ is a member of $X_{k,h}$ or has a descendant in $X_{k,h}$ then $X_{c,h}$ no longer blocks any path between $X_{i,h}$ and $X_{j,h}$.

$X_{i,h}$ and $X_{j,h}$ are d-separated by $X_{k,h}$ in G if and only if they are not d-connected by $X_{k,h}$ in G.

A variable is referred as collider if there exists a *v-structure* in the graph, such as $X_{i,h} \rightarrow X_{k,h} \leftarrow X_{j,h}$. In the skeleton learning step, the PC algorithm starts with a complete undirected graph of the observed data, where the features of the data are represented via nodes. The algorithm conducts conditional independence tests to decide whether an edge should be removed or not. The conditional independence tests are conducted as follows:

- For each $X_{i,h}$ and $X_{j,h}$, the PC algorithm checks if $X_{i,h} \perp\!\!\!\perp X_{2,h}$;
- For each $X_{i,h}$ and $X_{j,h}$, and each third variable $X_{k,h} = \{X_{m,h}\}$, the algorithm checks if $X_{i,h} \perp\!\!\!\perp X_{j,h} | X_{k,h}$; if so, edge is removed between $X_{i,h}$ and $X_{j,h}$.
- For each $X_{i,h}$ and $X_{j,h}$, and each third and fourth variable $X_{k,h} = \{X_{m,h}, X_{n,h}\}$, the algorithm checks if $X_{i,h} \perp \perp X_{j,h} | X_{k,h}$; if so, removes the edge.
- . . .
- For each $X_{i,h}$ and $X_{j,h}$, if they are still connected, the algorithm checks if $X_{i,h} \perp\!\!\!\perp X_{j,h} | X_{k,h}$ where $X_{k,h} \in X \setminus \{X_{i,h}, X_{j,h}\}$ if so, edge is removed between $X_{i,h}$ and $X_{j,h}$.

This is how PC algorithm leverages d-seperation. The conditioning set $X_{k,h}$ starts with the value of $d = 0$ and is progressively increased (by one) at each new level until d is greater than the size of the adjacent sets of the testing vertices. In each iteration, the neighborhood of all the nodes are updated dynamically with the removal of each edge.

After obtaining the causal graph G_t for every day in consideration, we proceed with the collection $\mathbf{G}$ of per-day causal graphs to the community detection module to identify features shared across different days.

B. Community Detection on Causal Features

Since the causal graphs learned from the daily slice of the SWAT data contains multiple set of features we aim to identify similar and stable sets of features that are potential causes of the variable of interest across the whole time span. Causal features tend to be more interpretable in nature [16]. To enhance the interpretability of the causal features, i.e. to ensure whether causally similar features have similar cause-effect relations, we apply community detection on each causal graph. After identifying the communities, we focus on those communities that contain the nodes that affect the variable of interest. After identifying these communities we find communities that occur frequently across the whole time span and the features from the maximally appearing communities are utilized as the optimal subset of features causing the output. In this section we explain how the community detection algorithm is used to identify the feature communities.

The community detection algorithm we use for this purpose is [44]. This algorithm is designed to identify the communities in a DAG. Since causal graphs are represented as DAGs, this community detection algorithm serves as the perfect candidate to identify the causal communities. This algorithm works in two steps. First, it aims to define antichains. In a DAG, a natural property of nodes at the same level is that they are not connected, making them suitable candidates to be clustered in the same community. These nodes are known as antichains. Although the nodes act as candidates for being clustered in a community, they need to be similar in some sense. In the second step, the algorithm conditions the similarities between two nodes in an antichain to deduce if they should be clustered together or not. The algorithm begins by partitioning the DAG into antichains that have large neighborhood overlaps. The algorithm encapsulates the antichain and neighborhood similarities in a function called Siblinarity $S(\mathfrak{F})$, which measures the quality of a given partition $\mathfrak{F}$ into antichains, $\mathcal{F}$. It is defined as,

$$S(\mathfrak{F}) = \sum_{\mathcal{F}\in\mathfrak{F}} \sum_{n\in\mathcal{F}} \sum_{m\in\mathcal{F}\setminus n} \left(\mathrm{sim}(n,m) - \mathrm{sim}_{\text{null}}\,(n,m)\right) \quad (1)$$

where $sim(n,m)$ is some measure of the similarity of two nodes n and m. The second term, $sim_{null}(n,m)$, is the expected value of similarity of these two nodes in some suitable null model which is generally some randomised version of the DAG. As m and n are in the same antichain $\mathcal{A}$ there is no path between nodes contributing to siblinarity. The logical choice for the similarity measure is the neighborhood overlap between nodes n and m, $\mathrm{sim}(n,m) = |\mathcal{N}(n) \cap \mathcal{N}(m)|$ where $\mathcal{N}(n)$ is the neighbourhood of node n. The Eq.(1) can be written as,

$$\begin{gathered} S(\mathfrak{F}) = \textstyle\sum_{\mathcal{F}\in\mathfrak{F}} \sum_{n\in\mathcal{F}} \sum_{m\in\mathcal{F}\setminus n} \left(\tilde{A}_{nm} - \frac{\kappa_n \kappa_m}{W}\right), \\ \kappa_n := \textstyle\sum_m \tilde{A}_{nm}, W = \sum_{n,m} \tilde{A}_{nm} \end{gathered} \quad (2)$$

Here the adjacency matrix $\mathbf{A}$ for the DAG is defined so that A_{nm} is the weight of the edge from n to m. The neighbourhood overlap is captured by the matrix $\tilde{\mathbf{A}}$ which is the product of the adjacency matrix $\mathbf{A}$ of the DAG and its transpose. W is the total strength of edges in that graph. If the value of W is small, and the siblinarity of a given antichain is large, we can expect that the overlap of neighbourhoods of nodes in the antichain is sparse.

C. Prediction using Causal Features

We obtain the set of daily causal features from the causal discovery module, namely $\mathcal{X}_t^{causal}$, represents the day. We then utilize a frequency strategy to identify the stable causal relations. We first sort the causal features based on the frequency with which they appear across all days. Next, we identify the top k most frequent causal relationships affecting the target variable directly. The value of k is calculated based on a thresholding factor, α . We set $\alpha = 0.8$ implying the causal relationships that appear at least 80% of the time are selected for prediction. After identifying the set of causal features we assess the predicting capabilities of the target variable, i.e. the water quantity. Given the data for each day $\mathcal{D} = \{(y_h, \mathbf{x}_{i,h})$ where y_h is the water quantity response for the hth HRU, measured on a continuous scale; and $\mathbf{x}_{i,h} = \mathcal{X}_{i,h}^{causal}$ is the associated predictor vector. After obtaining the predicted value $\hat{y}_h$ we compute the error on prediction by utilizing mean squared error. The overall loss function is,

$$\mathcal{L} = \frac{1}{n} \sum_h (y_h - \hat{y}_h)^2. \quad (3)$$

where y_h is the actual value and $\hat{y}_h$ is the predicted value.

V. Experiments

We conducted a series of experiments to understand whether causal discovery can aid in selecting optimal features for understanding physical processes – more specifically, predicting water quantity. Ideally, a causal discovery method is evaluated against a ground truth causal graph [45]. Since the true causal graph for physical process based hydrological models is not available we follow the recent advances [10], [17], [18] and hypothesize that, utilizing only the causal features in predicting the water quantity will yield a lower error rate compared to utilizing features identified via different feature selection techniques. Furthermore, since causal features are more interpretable than other features [16] we also conduct a real-world case study to evaluate the interpretability of the causal features. We answer the following research questions:

- RQ.1 Can causal discovery methods be utilized for feature selection in physical process-based hydrological models?
- RQ.2 Is CAPS able to cluster related variables together?
- RQ.3 Are the features learned by causal discovery based feature selection method more intuitive and interpretable compared to other feature selection techniques?
- RQ.4 Which soil-related features do not cause variation in water yield (WYLD) values?

A. Data Preperation

The SWAT dataset consists of HRU level data measured 730 days. The dataset contains 2,715 HRUs where each HRU is measured across 42 features. We consider the data for each day separately to learn daily causal graphs.

B. Experimental Setup

We build on the PC algorithm [12] for performing causal discovery and antichain algorithm for DAGs [44] for performing community detection. We implement the causal discovery module with the help of pcalg package in R programming language, where we use the Gaussian conditional independence testing [46] and we implement the community detection algorithm using python. We used grid search to find the optimal number of features to be selected by the wrapper methods. For the correlation, mutual information and random forest importance methods, the features were selected based on a threshold value, i.e. if the score metric is greater than 0 we selected those features for prediction. For the KNN regressor we selected the neighbors as 20 and for the SVR we selected the margin of tolerance (ϵ) as 0.2 and regularization parameter C as 1. We chose the Chi-Square test for computing the statistically significant features as it measures if two variables are statistically dependent on each other or not. The data is split w.r.t the days, i.e. since the total dataset consists of 730 days' worth of measurements, the training is done on 610 days' data and evaluated on the remaining 120 days.

C. Baselines

Feature selection approaches are usually divided into four categories: filter methods [47], wrapper methods [48], embedded methods [49], and causal methods [9].

Filter Methods: These methods are generally faster and computationally less expensive than other methods. Examples of these include *Mutual Information* [50], and *Correlation* [51]. The logic behind using correlation is that usually predictive variables are highly correlated with the target variable.

Wrapper Methods: Wrapper methods leverage the space of all possible subsets of features for identifying the optimal feature subset for predicting a target variable. *Forward Feature Selection* [52] method starts with the best performing variable against the target. It iteratively adds other variables that give best performance in combination with the previous variables. *Backward Feature Selection* [53] is the opposite of Forward Feature Selection. *Recursive Feature Elimination* [54] method selects features by recursively considering smaller and smaller sets of features. First, an estimator is trained on all the features to receive importance scores and the least important features are pruned.

Embedded Methods: These methods encompass the benefits of both the wrapper and filter methods, by including interactions of features. *Lasso Regularization* [55] relies on L1 constraint to shrink coefficients of some of features to zero; features with zero coefficients are removed from the model. *Random Forest Importance* [56] is a decision tree-based stretegy. Nodes with the greatest decrease in impurity are closer to the root node, while nodes with the least decrease in impurity occur closer to the leaf nodes. Thus, by pruning trees below a particular node, it identifies important features.

Causal Methods: These methods leverage causal discovery algorithms to identify and select features that cause the target variables. For instance, the the SLL-MB algorithm [32] is a score-based variant of the divide-and-conquer Markov Boundary learning algorithms. In the causal structure learning phase, SLL employs a Bayesian Network structure learning algorithm. Then SLL implements the symmetric check using the AND and OR rules.

We implemented all traditional baselines using the sklearn library, and the causal baseline with the help of pyCausalFS.

D. Evaluation Measures

To evaluate the quality of features selected by different approaches, we employ several common predictive models including Linear regression, K-Nearest Neighbor Regressor (KNN) [57], and Support Vector Regressor (SVR) [58]. The use of multiple models in the evaluation avoids the possibility of model biases against specific data sets. To ensure fair comparison against CAPS, for each baseline, we first identify the important features for the daily slice of the data. Then, we utilize the frequency strategy (also used in CAPS) to identify the most frequently appearing features across all time steps and perform prediction on this stable feature subset. The performances are assessed by the following criteria:

- prediction accuracy measures; and
- proportion of features selected relative to the original number of features.

For assessing the predictive performances of different models, we utilize the following evaluation metrics (the lower the values, the better the performance):

- MAE: mean absolute error; $MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \widehat{y_t}|$
- MSE: mean square error; $MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \widehat{y_t})^2$
- RAE: relative absolute error; $RAE = \frac{\sum_{t=1}^{n}|y_t-\widehat{y_t}|}{\sum_{t=1}^{n}|y_t-\bar{y}|}$.

E. Performance Comparisons (RQ.1)

We compare the different baseline approaches with CAPS on the SWAT dataset for three machine learning techniques – Linear regression, K-Nearest Neighbor (KNN) [57], and Support vector Regressor (SVR) [58]. Table II demonstrates prediction performances across three different metrics (MSE, MAE, and RAE). The table also lists the number of statistically significant features based on the Chi-Square Test, the features selected by the different feature selection methods and their proportion with respect to all the features. From this table, we can make the following observations with respect to **RQ.1**:

- CAPS consistently yields the best performance across all models and all accuracy metrics.
- Since CAPS utilizes causal discovery algorithms (PC in this case), CAPS selects the maximum number of statistically significant features compared to the other methods as shown in Table II. Moreover, the second highest statistical features are selected by SLL-MB which also utilizes a causal discovery method to select the optimal features.

The reason why CAPS is able to achieve better performance is that traditional methods such as Correlation Method, selects features solely based on the correlation score. However, high correlation might not necessarily reflect causal relation [59], i.e. the correlation might be due to other unobserved factors. CAPS on the other hand, utilizes causal discovery methods that are more advanced than correlation measures because unlike correlation that implies there is a statistical association between variables causation suggests that a change in one variable causes a change in another variable. Thus, the features selected by CAPS represent those features that have a significant impact on the target variable.

In addition to the above key observation, we also make several lesser observations from the results:

- Among the four categories of baselines, we observe that SLL-MB from causal methods serve as the strongest baseline. This is because, similar to CAPS, SLL-MB also aims at utilizing score based causal discovery methods for feature selection. However, We believe SLL-MB does not outperform CAPS because of the conflicting nature of the features present in the SWAT dataset. Since SWAT contains features that vary across time but are constant

Metric	Model	Original	Filter Methods		Wrapper Methods			Embedded Methods		Causal Methods	
			Mutual Info.	Corr.	Forward Feature Selection	Backward Feature Selection	Recursive Feature Elimination	Lasso Reg.	Random Forest Importance	SLL-MB	CAPS (ours)
MSE	**Linear Regression**	0.89±0.07	0.91±0.05	0.88±0.03	0.89±0.04	0.93±0.04	0.87±0.03	0.92±0.06	0.88±0.04	0.87±0.02	**0.85±0.01**
	K-Nearest Neighbor Regressor	0.92±0.05	0.94±0.04	0.91±0.02	0.92±0.03	0.94±0.05	0.89±0.02	0.90±0.04	0.90±0.03	0.89±0.01	**0.87±0.02**
	Support Vector Regressor	0.86±0.05	0.88±0.05	0.86±0.01	0.88±0.03	0.90±0.06	0.85±0.03	0.92±0.04	0.87±0.03	0.86±0.03	**0.84±0.01**
MAE	**Linear Regression**	0.41±0.04	0.43±0.03	0.41±0.02	0.41±0.03	0.43±0.05	0.40±0.03	0.43±0.04	0.40±0.02	0.39±0.02	**0.38±0.01**
	K-Nearest Neighbor Regressor	0.28±0.04	0.29±0.03	0.26±0.03	0.28±0.04	0.29±0.04	0.27±0.03	0.28±0.03	0.26±0.02	0.25±0.03	**0.24±0.02**
	Support Vector Regressor	0.23±0.03	0.26±0.02	0.22±0.02	0.24±0.04	0.25±0.02	0.21±0.02	0.23±0.04	0.21±0.03	**0.19±0.02**	**0.19±0.02**
RAE	**Linear Regression**	0.86±0.06	0.94±0.04	0.88±0.03	0.89±0.05	0.95±0.03	0.89±0.02	0.94±0.03	0.88±0.02	0.86±0.01	**0.83±0.01**
	K-Nearest Neighbor Regressor	0.59±0.05	0.61±0.03	0.58±0.01	0.59±0.03	0.61±0.04	0.57±0.03	0.60±0.03	0.56±0.03	0.56±0.03	**0.54±0.02**
	Support Vector Regressor	0.54±0.04	0.57±0.03	0.55±0.02	0.54±0.03	0.56±0.03	0.53±0.02	0.57±0.04	0.52±0.02	0.51±0.01	**0.49±0.01**
# of Statistically Significant features		22	9	13	10	7	10	9	10	14	20
# of Features Selected		42	20	13	18	20	15	15	15	18	20
Proportion of Features Selected (%)		100	48	31	43	48	36	36	36	43	48

TABLE II: Prediction performances using linear regression, KNR, and SVR with features selected by different methods averaged across all days (bold font highlights best results). The statistical significant features are computed via the Chi-Square test, which tests for independence between the features and the target variable.

Latent Communities		
Community #1	**Community #2**	**Community #3**
'SDEP1,2,3', 'SDENI,1,2,3', 'SH1,2', 'SPS2,3'	'SH3', 'SPS1', 'TSLP1,3', 'TLSW3', 'LRF1,3', 'TLSS1'	'WYLD'

TABLE III: Latent communities identified by the community detection module. The feature symbols are same as Table I.

Method		Selected Features			Percentage of features shared with CAPS
		Transient Features	Soil-related Features	Other Features	
Filter Methods	**Mutual Info**	**'PCP'**	**'SH1,3', 'SWI1,2,3', 'SPS2,3'**, 'SPS4'	'TLSW1,2', 'GWH1', 'TSLP2,' **'TSLP3'**, 'GWV1,2', 'TLSS4,5', 'DEP_IMP', 'LRF2'	45
	Correlation	**'PCP', 'TMP'**	**'SDEP1,2', 'SH1', 'SPS1,2,3', 'SDEN1,2,3'**	**'TSLP1', 'TLSW3'**	65
Wrapper Methods	**Forward Feature Selection**	-	**'SDEP1,2,3', 'SDEN1,2,3','SPS2,3, 'SH3'**, 'SPS4'	**'TLSW3'**, 'TLSW1', 'GWV1,2,3,4,5', 'GWH1'	50
	Backward Feature Selection	**'TMP'**	**'SWI1,2,3'**	'GWV4,5', 'GWH2,3', 'LRF1,2' **'TSLP1,3'**, 'TSLP2', **'TLSS1'**,'TLSS2,3,4', 'TLSW1,2', 'DEP_IMP'	35
	Recursive Feature Elimination	**'PCP'**	**'SDEP1', 'SDEN1,3', 'SPS3', 'SH1,2,3'**,'SPS4'	**'TLSW3'**, 'LRF1,2', 'TSLP2', **'TLSS1'**,'TLSS3'	50
Embedded Methods	**Lasso Regularization**	**'PCP'**	**'SPS1,2,3', 'SWI1,2', 'SDEP3'**	'TSLP2', **'TSLP3'**, 'LRF1', **'TLSS1'**, 'TLSS2', 'GWV1,2', 'GWH1'	45
	Random Forest Importance	**'TMP'**	**'SPS2', 'SH1,3', 'SDEN3', 'SDEP1,2,3'**, 'SPS4'}	'GWV4,5', 'GWH2,3', **'TSLP1,3'**	50
Causal Methods	**SLL-MB**	**'TMP', 'PCP'**	**'SDEP1,2', 'SDEN3','SPS1,2,3', 'SH1,2,3'**,'SPS4'	'GWH1,2', **'TSLP1,3', 'TLSW3'**, 'TLSW1'	70
	CAPS (Ours)	**'TMP', 'PCP'**	**'SPS1,2,3', 'SH1,2,3', 'SDEN1,2,3', 'SDEP1,2,3', 'SWI1,3'**	**'TSLP1,3', 'TLSS1', 'TLSW3'**	100

TABLE IV: Features selected by different methods. The feature symbols are consistent with the symbols shown in Table I. Bold symbols represent statistically significant features.

across HRUs and features that vary across HRUs but are constant w.r.t time, SLL-MB ignores certain features while learning the causal structure.

Among the traditional baselines, random forest importance outperforms the other methods, likely due to its capabilities of capturing non-linear relations among variables [60]. While causal and conventional baselines have a close margin in terms of their prediction performances, when we compare the statistically significant features picked by the causal techniques with the traditional methods, we are able to observe a major difference: Because causal approaches rely on conditional independence testing during the feature selection phase, they are select features on which the target variable depends.

- Among the three different models we observe that Support Vector Regressor outperforms K-Nearest Neighbor Regressor and Linear Regression Models.

SVR's superior performance can be attributed to the following qualities: (1) SVR uses kernel trick to solve complex solutions and capture both linear and non-linear solutions. (2) SVR is better equipped to deal with outliers with the aid of the soft margin constant, and (3) SVR uses a convex optimization function, allowing global minima to be achievable.

F. Quality of Variable Clusters – a Case Study (RQ.2)

The causal discovery module identifies daily causal relations and encapsulates these relations in a graphical representation, rendering a collection of causal graphs $\mathbf{G} = \{G_t : t = \{1, \ldots, 730\}$. The community detection module, then, helps group related features together. As discussed in Section IV-B, variables having similar causal neighborhoods are assigned into the same community. Table III presents the communities.

To assess whether the learned communities are able to group causally similar features together, we relied on the expertise of domain experts. The feedback we received was as follows:

- Latent Community $\#1$ is intuitive in that it groups soil-related features together.
- Community $\#3$ is also intuitive in that it consists only of the SWAT output variable "*water yield*" ('WYLD'), which is caused by other variables, but does not cause any others.
- Almost all features except 'SH3' and 'SPS1' (which are soil related) in Community $\#2$ are terrain or location related.

Overall, the variable groups identified by CAPS are physically meaningful. The existence of two soil related features 'SH3' and 'SPS1' in Community $\#2$, among land-use related variables, indicate either a non-trivial causal relationship between the land variables and these two soil variables or SWAT data may contain hidden confounders between soil-related and the land-related features. If the latter is true, this can be alleviated by incorporating causal discovery models such as [30], that account for the presence of the unobserved confounders. We leave this for future investigation.

G. Causal Interpretability – a Case Study (RQ.3 and RQ.4)

Causal features should be more interpretable and intuitive when compared to non-causal features, as causal processes reflect how humans perceive the data [61]. Thus, to verify whether CAPS is able to select intuitive and interpretable features, within the context of predicting water quantity we conduct a case study utilizing the domain knowledge of human experts, i.e. a team of three hydrologists.

As mentioned earlier, the SWAT model consists of three types of features: transient features (precipitation, temperature, and the model output WYLD), soil-related features (such as soil depth, soil moist levels, and rock fragment levels), and other features (including groundwater-related, location-related and terrain-related features in TableI). For this study, we first divided the set of features identified by each feature selection algorithm to these three categories. The set of features selected by each method and the shared features across each method and CAPS are shown in Table IV. We then prepared a data frame consisting of two columns, $Method$ and $FeaturesSelected$. The $Method$ column contained an integer identifiers that map to the method names. The underlying mapping was not exposed to the hydrologists. The $FeaturesSelected$ column contained the set of features selected by each method under each of the three categories. This dataframe was then shared with the hydrologists, who were tasked to identifying which sets of features were more crucial for predicting the water quantity.

In this blind test, all three members of the hydrologist team identified the features included in CAPS method as the best set of features for predicting water quantity output on the HRU level. Here is the summary explaining their choice:

- First, both the 'PCP' and 'TMP' are essential features for hydrology processes: as the precipitation is the only water source to the HRUs and the temperature controls the water evaporation process. Any methods without the inclusion of these transient data misses important water balance element.
- Second, methods consisting of features describing land surface conditions including the curve number ($'TLSW3'$), the Manning's roughness coefficient ($'TLSS1'$), and the slope length and steepness conditions ($'TSLP1'$ and $'TSLP3'$) are more meaningful and interpretable. Because these features have a more direct and significant influence on water quantities compared to the groundwater transport ($'GWV1-5'$ and $'GWH1-3'$) and evaporation and plant-uptake processes ($'TLSW1-2'$, and $'TLSS3'$).
- Third, the features with zero variations among different HRUs and days are out of the causal model's ability and thus should not exist in the result. These features include: $'GWV1,2,4'$, $'GWH3'$, $'LRF2'$, $'TLSS2,4,5'$, $'TSLP2'$, $'TLSW2'$, $'SDEP4'$, and$'SWI1-2'$.

This confirms CAPS ability to select meaningful sets of features and further explains the higher prediction accuracies reported in Table II. Furthermore, we observed that among the traditional baselines Correlation method performed the best in terms of interpretable features. This can be attributed to the fact that the correlation method captures the linear association between the features and the target. However, since this method fails to capture the non-linear associations, it fails to capture all the interpretable features which are essential for predicting the river flow as per the domain experts. Also, the high overlap between SLL-MB and CAPS suggests that causal methods are able to select more interpretable features when compared to traditional methods.

VI. Conclusion and Future Work

In this paper, we have presented a causal feature selection method, named Causality-Aware Physical Systems (CAPS) for physical process-based hydrological systems such as SWAT. The SWAT model consists of 17 soil-related features. To reduce the complexity of the SWAT model, CAPS eliminates features by identifying causally unimportant soil-related features that do not influence the water yield (WYLD). To highlight the efficacy of CAPS we conducted multiple experiments evaluating the predictive capabilities and the interpretability of features. We also conducted a case study to highlight that the features selected by CAPS are mostly aligned with the features selected by the human experts. Such results suggest that any soil types which do not differ across the 12 causal features selected by CAPS, may be merged into a single soil type.

A possible direction to extend this work would be to develop a component that could account for confounders. Another direction can be to capture the inter-day causal relationships among the features and identify whether features on the current day have any causal impact on the features the next day.

VII. ACKNOWLEDGEMENTS

This material is based upon work supported by, or in part by Army Corps of Engineers (USACE) grant GR40695 as well as by the National Science Foundation (NSF) grant 1909555.

References

[1] S. Fatichi *et al.*, "An overview of current applications, challenges, and future trends in distributed process-based models in hydrology," *Journal of Hydrology*, 2016.

[2] T. Xu and F. Liang, "Machine learning for hydrologic sciences: An introductory overview," *Wiley Interdisciplinary Reviews: Water*, 2021.

[3] R. Shrestha, Y. Tachikawa, and K. Takara, "Input data resolution analysis for distributed hydrological modeling," *Journal of hydrology*, 2006.

[4] L. Hoang, R. Mukundan, K. E. Moore, E. M. Owens, and T. S. Steenhuis, "The effect of input data resolution and complexity on the uncertainty of hydrological predictions in a humid vegetated watershed," *Hydrology and Earth System Sciences*, 2018.

[5] R. Orth, M. Staudinger, S. I. Seneviratne, J. Seibert, and M. Zappa, "Does model performance improve with complexity? a case study with three hydrological models," *Journal of Hydrology*, 2015.

[6] S. L. Neitsch, J. G. Arnold, J. R. Kiniry, and J. R. Williams, "Soil and water assessment tool theoretical documentation version 2009," TWRI, Tech. Rep., 2011.

[7] M. Winchell *et al.*, "Arcswat interface for swat 2005," *User'sGuide, Blackland Research Center*, 2007.

[8] P. V. Femeena, R. Karki, R. Cibin, and K. Sudheer, "Reconceptualizing hru threshold definition in the soil and water assessment tool," *JAWRA*, 2022.

[9] K. Yu *et al.*, "Causality-based feature selection: Methods and evaluations," *CSUR*, 2020.

[10] Y. Sun *et al.*, "Using causal discovery for feature selection in multivariate numerical time series," *Machine Learning*, 2015.

[11] R. Moraffah *et al.*, "Causal inference for time series analysis: Problems, methods and evaluation," *KAIS*, 2021.

[12] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*, 2000.

[13] M. J. Vowels, N. C. Camgoz, and R. Bowden, "D'ya like dags? a survey on structure learning and causal discovery," *arXiv:2103.02582*, 2021.

[14] K. W. Soo and B. M. Rottman, "Causal strength induction from time series data." *Journal of Experimental Psychology: General*, 2018.

[15] Z. Li, R. Cai, T. Z. Fu, and K. Zhang, "Transferable time-series forecasting under causal conditional shift," *arXiv:2111.03422*, 2021.

[16] G. Xu, T. D. Duong, Q. Li, S. Liu, and X. Wang, "Causality learning: a new perspective for interpretable machine learning," *arXiv:2006.16789*, 2020.

[17] I. Guyon *et al.*, "Causal feature selection," in *Computational methods of feature selection*, 2007.

[18] X. Zhang *et al.*, "A causal feature selection algorithm for stock prediction modeling," *Neurocomputing*, 2014.

[19] W.-P. Tsai, K. Fang, X. Ji, K. Lawson, and C. Shen, "Revealing causal controls of storage-streamflow relationships with a data-centric bayesian framework combining machine learning and process-based modeling," *Frontiers in Water*, 2020.

[20] X. Ji *et al.*, "Seasonal and interannual patterns and controls of hydrological fluxes in an amazon floodplain lake with a surface-subsurface process model," *Water Resources Research*, 2019.

[21] C. Shen, W. J. Riley, K. R. Smithgall, J. M. Melack, and K. Fang, "The fan of influence of streams and channel feedbacks to simulated land surface water and carbon dynamics," *Water Resources Research*, 2016.

[22] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *MIPRO*, 2015.

[23] V. Moreido, B. Gartsman, D. P. Solomatine, and Z. Suchilina, "How well can machine learning models perform without hydrologists? application of rational feature selection to improve hydrological forecasting," *Water*, 2021.

[24] V. F. Rodriguez-Galiano, J. A. Luque-Espinar, M. Chica-Olmo, and M. P. Mendes, "Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods," *Science of the total environment*, 2018.

[25] K. Ren, W. Fang, J. Qu, X. Zhang, and X. Shi, "Comparison of eight filter-based feature selection methods for monthly streamflow forecasting–three case studies on camels data sets," *Journal of Hydrology*, 2020.

[26] G. B. Reis *et al.*, "Effect of environmental covariable selection in the hydrological modeling using machine learning models to predict daily streamflow," *Journal of Environmental Management*, 2021.

[27] M. Ombadi, P. Nguyen, S. Sorooshian, and K.-l. Hsu, "Evaluation of methods for causal discovery in hydrometeorological systems," *Water Resources Research*, 2020.

[28] N. K. Singh and D. Borrok, "A granger causality analysis of groundwater patterns over a half-century," *Scientific reports*, 2019.

[29] I. Tsamardinos, C. F. Aliferis, A. R. Statnikov, and E. Statnikov, "Algorithms for large scale markov blanket discovery." in *FLAIRS*, 2003.

[30] P. Forré and J. M. Mooij, "Constraint-based causal discovery for nonlinear structural causal models with cycles and latent confounders," *arXiv:1807.03024*, 2018.

[31] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in genetics*, 2019.

[32] T. Niinimaki and P. Parviainen, "Local structure discovery in bayesian networks," *arXiv:1210.4888*, 2012.

[33] T. D. Le, T. Hoang, J. Li, L. Liu, H. Liu, and S. Hu, "A fast pc algorithm for high dimensional causal discovery with multi-core pcs," *IEEE/ACM transactions on computational biology and bioinformatics*, 2016.

[34] L. Li, Y. Lin, H. Zhao, J. Chen, and S. Li, "Causality-based online streaming feature selection," *Concurrency and Computation: Practice and Experience*, 2021.

[35] J. Pearl, *Causality*. Cambridge university press, 2009.

[36] M. J. Van der Laan, S. Rose *et al.*, *Targeted learning: causal inference for observational and experimental data*, 2011.

[37] G. F. Cooper and C. Yoo, "Causal discovery from a mixture of experimental and observational data," *arXiv:1301.6686*, 2013.

[38] "Water data for texas: Texas reservoirs," Texas Water Development Board, 2019. [Online]. Available: https://waterdatafortexas.org/reservoirs/statewide

[39] "Texas land parcels," Texas Natural Resources Information System, 2019. [Online]. Available: https://tnris.org/stratmap/land-parcels/

[40] "Cropland data layer," United States Department of Agriculture, 2013. [Online]. Available: https://nassgeodata.gmu.edu/CropScape/

[41] "Soil survey geographic (ssurgo) database," United States Department of Agriculture, 2021. [Online]. Available: https://sdmdataaccess.sc.egov.usda.gov

[42] "Shuttle radar topography mission (srtm) global," NASA Shuttle Radar Topography Mission (SRTM), 2013. [Online]. Available: https://sdmdataaccess.sc.egov.usda.gov

[43] J. Ramsey, J. Zhang, and P. L. Spirtes, "Adjacency-faithfulness and conservative causal inference," *arXiv:1206.6843*, 2012.

[44] V. Vasiliauskaite *et al.*, "Making communities show respect for order," *Applied Network Science*, 2020.

[45] L. Cheng *et al.*, "Evaluation methods and measures for causal learning algorithms," *IEEE TAI*, 2022.

[46] M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann, "Causal inference using graphical models with the r package pcalg," *Journal of statistical software*, 2012.

[47] N. Sánchez-Marono, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection–a comparative study," in *ICIDEAL*, 2007.

[48] N. Nnamoko, F. Arshad, D. England, J. Vora, and J. Norman, "Evaluation of filter and wrapper methods for feature selection in supervised machine learning," *Age*, 2014.

[49] T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff, "Embedded methods," in *Feature extraction*, 2006.

[50] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural computing and applications*, 2014.

[51] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, 1999.

[52] D. Ververidis and C. Kotropoulos, "Sequential forward feature selection with low computational cost," in *ESPC*, 2005.

[53] S. Abe, "Modified backward feature selection by cross validation." in *ESANN*, 2005.

[54] X.-w. Chen and J. C. Jeong, "Enhanced recursive feature elimination," in *ICMLA*, 2007.

[55] V. Fonti and E. Belitser, "Feature selection using lasso," *business analytics*, 2017.

[56] B. H. Menze *et al.*, "A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC bioinformatics*, 2009.

[57] J. G. Cleary and L. E. Trigg, "K*: An instance-based learner using an entropic distance measure," in *Machine Learning Proceedings*, 1995.

[58] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, 2004.

[59] S. M. Stigler, "Correlation and causation: A comment," *Perspectives in Biology and Medicine*, 2005.

[60] N. Ludwig, S. Feuerriegel, and D. Neumann, "Putting big data analytics to work: Feature selection for forecasting electricity prices using the lasso and random forests," *Journal of Decision Systems*, 2015.

[61] S. A. Sloman and D. Lagnado, "Causality in thought," *Annual review of psychology*, 2015.