Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Analyzing the effects of data splitting and covariate shift on machine learning based streamflow prediction in ungauged basins

Pin-Ching Li [a,*], Sayan Dey [b], Venkatesh Merwade [c]

[a] GSI Environmental Inc., Emeryville, CA, USA
[b] Taylor Geospatial Institute, Saint Louis University, St. Louis, MO, USA
[c] Lyles School of Civil Engineering, Purdue University, West Lafayette, IN, USA

ARTICLE INFO

This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Jie Chen, Associate Editor

Keywords:
Hydrology machine learning
Prediction in ungauged basins
Monte carlo analysis
Covariate shift

ABSTRACT

Machine learning (ML) models are alternatives to traditional hydrologic modeling for streamflow predictions in ungauged basins (PUB). The variability in watershed characteristics of ungauged basins; however, adds uncertainties to PUB frameworks based on ML models. These uncertainties arise from the inconsistency in the statistical distributions between the dataset used to train and test a ML model, known as covariate shifts, and the real-world (global) dataset on which the trained model is implemented. In real-world applications, covariate shift is a widespread issue for ML that has not been investigated in hydrological applications. This study evaluates the uncertainty in ML-based PUB method including Random Forest (RF) and Artificial Neural Network (ANN) under the influence of covariate shift. The Monte Carlo method is applied to aggregate simulations of RF and ANN according to various data splitting configurations as predictive distributions. The results indicate that ML performance is not robust under covariate shifts. ML performance is influenced by watershed characteristics displaying heterogeneity, such as drainage area, dam density, and urbanized area. 20–48% simulation results show a departure from the normal distribution under different covariate shift scenarios Furthermore, the efficiency and limitation of Random Forest models for PUB are highlighted by investigating their biased predictions in watersheds with varying dam density, drainage area, and meteorological variables, such as annual snowfall and annual precipitation.

1. Introduction

Many hydrologic studies including flow prediction rely on monitoring data from larger basins, and prediction in ungauged basins (PUB) provides an option to complement monitoring with information about smaller watersheds. Predicting streamflow at ungauged reaches is a challenging using conceptual or physically based hydrological models because of the heterogeneity of streamflow generation processes (Sivapalan et al. 2003; Hrachowitz et al 2013). In this context, Machine Learning (ML) is investigated as an alternative to conceptual or physically based modelling. ML has emerged as a promising method for PUB, demonstrating robustness and accuracy in predicting streamflow (Mosavi et al., 2018; Petty and Dhingra, 2018; Worland et al., 2018; Kratzert et al., 2019; Xiang et al., 2020; Adnan et al., 2021). PUB has been widely addressed using physics-based distributed hydrologic models involving basin/site-specific parameters (AghaKouchak and Habib, 2010; Razavi and Coulibaly, 2013; Hrachowitz et al., 2013; Pechlivanidis and Arheimer, 2015; Petty and Dhingra, 2018; Saksena et al., 2019) or using conceptual models utilizing lumped or semi-distributed approach with fewer model parameters (Das et al., 2008; Chang et al., 2017; Seibert et al., 2019; Darbandsari and Coulibaly, 2020). Physical-based approaches face challenges related to the unique physical interpretability of such models (Beven, 2006; Her and Chaubey, 2015; Her et al., 2019), while ML data-driven approaches face challenges associated with the preparation of data, ML algorithms, and their parameters (Schmidt et al., 2020; Underwood et al., 2023; Samadi et al., 2024).

ML models built for PUB generally follow a random data splitting process (Reitermanova, 2010), which divides the selected stations into training data for building ML models and testing data for evaluating ML models. Random sampling process can be insufficient for PUB due to the heterogeneity in hydrological datasets and unique watershed characteristics (Sivapalan et al., 2003; Hrachowitz et al., 2013). In traditional PUB studies, data splitting processes of conventional statistical models

[*] Corresponding author.
E-mail address: pcli@gsi-net.com (P.-C. Li).

https://doi.org/10.1016/j.jhydrol.2025.132731
Received 24 June 2024; Received in revised form 4 December 2024; Accepted 9 January 2025
Available online 26 January 2025

and conceptual models have been developed by dividing basins into subbasins (Kuchment and Gelfan, 2009; Janjić and Tadić, 2023; Wang et al., 2023). Using a random splitting process for heterogeneous datasets can lead to inconsistencies in the statistical distributions of input variables between training and testing data, known as covariate shift (Sugiyama et al., 2007; Reitermanova, 2010; Balogun and Attoh-Okine, 2021).

Although covariate shifts have been studied in applications related to image corruption, geometry detection, and credit card fraud, their impacts on hydrologic applications, particularly in the context of PUB, remain unexplored (Lucas et al., 2019; Schneider et al., 2020; Balogun and Attoh-Okine, 2021). For example, the distribution of drainage area and reservoir density in training data may be different from the testing data under covariate shift due to the random data splitting process. Under this scenario, ML models may underperform due to the mismatched generalization in the learning process of ML (Sugiyama et al., 2007; Balogun and Attoh-Okine, 2021). To investigate the influence of covariate shifts on ML's performance, Monte Carlo method is applied to address the uncertainty of ML models by creating multiple modeling scenarios by repeating the random data splitting process (Breuer et al., 2006; Moges et al., 2021). The results of the Monte Carlo method represent the distribution of ML model performance (Neal, 1992) and provide a better understanding of covariate shifts' implications on ML models for PUB.

Additionally, PUB models, ML or conceptual, cannot perfectly learn all the streamflow generation processes, surface runoff routing, and reservoir impact for basin-scale simulations (Hrachowitz et al., 2013; Khandelwal et al., 2020; Adombi et al., 2021; Fleming et al., 2021a; Nearing et al., 2021). Quantifying the preference features range and limitations of PUB models informs users and researchers about the appropriate physical settings for applying a specific PUB model. Under a data-driven framework, a ML model mines the inherent pattern within the data, which includes a combination of possible physical processes and the corresponding noise in streamflow generation. How a ML model learns the streamflow generation process depends on the collective understanding of the elements established by its algorithm (McGovern et al., 2019; Fleming et al., 2021b). The performance of ML models in streamflow prediction can be decoded by analyzing the minimum–maximum performance of ML models according to their input features.

Above discussion show that ML models have emerged as efficient approaches for large-scale applications in PUB, but their performance can be affected by the uncertainty in the random data splitting process (Addor et al., 2018; Nearing et al., 2021; Prieto et al., 2022), which can introduce covariate shift. The impact of covariate shifts on the performance of ML models in PUB needs to be investigated and quantified. Additionally, understanding the range for each variable that leads to better simulations, henceforth referred to as "preferred range," can help in identifying the appropriate physical settings of ML models for PUB (Prieto et al., 2022). This study addresses the mentioned issues by (i) evaluating the performance and incorporating uncertainty in ML models through Monte Carlo simulations, (ii) investigating the impacts of covariate shifts on the performance of ML models, and (iii) quantifying the preferred range of input variables for ML models and exploring their connection with possible streamflow generation regimes.

It is recognized that random splitting of a variable can generate two differently statistical distributions, which can introduce covariate shift and different results from ML application. The question that this study is attempting to address is how are the results affected with or without covariate shift. If the results are unaffected, then covariate shift can be ignored, but if they are affected, is it possible to identify which variables are problematic and quantify the their range of values, "preferred range", within which results are acceptable. Overall, the results from this study can lead to improved understanding of the impact of covariate shift in input data on streamflow prediction, which can have broader implications beyond ML.

2. Study area and dataset

2.1. Study areas

The Ohio River Basin (ORB, Fig. 1 and Table 1) is a two-digit Hydrological Unit Code (HUC2; 05) named by the United States Geological Survey (USGS) and is one of the 21 major water resource regions in the United States. The ORB encompasses watersheds from 11 U.S. states: Illinois, Indiana, Kentucky, Maryland, New York, North Carolina, Ohio, Pennsylvania, Tennessee, Virginia, and West Virginia. About 35 % of land use in the ORB is agricultural, and agricultural usage is primarily located in the states of Kentucky, Ohio, Indiana, and Illinois (Electric Power Research Institute (EPRI); 2010, 2014, 2016). The ORB has experienced extreme flood events since 1773 (National Water Service (NWS), 2017; Schlef et al., 2018; Gibson, 2020; Hauser, 2020), particularly in Indiana, Ohio, and Kentucky. These flood events highlight the need for streamflow prediction at ungauged reaches for developing flood mitigation and management strategies.

There are 598 USGS daily streamflow stations in the ORB (USGS, 2023), but only 431 USGS streamflow stations were selected here (explained in Section 2.2). The number and density of selected USGS stations from the ORB and five four-digit HUC (HUC4) level subregions within the ORB: Allegheny (0501), Monongahela (0502), Kentucky-Licking (0510), Wabash (0512), and Lower Ohio (0514) are shown in Fig. 1 and Table 1. The Other subregions within the ORB are not chosen due to insufficient number of streamflow stations (<30 available stations). The spatial scale and density of stations provide an overall understanding of the station network within study areas. The ORB has the largest study area (421,951 km^2), and the other watersheds have the same order of magnitude of study areas (10^4 km^2). The Wabash River is the largest HUC4 region (85,349 km^2) selected and has the greatest number of gauged reaches of the HUC4 regions included in this study (Table 1). Monongahela and Lower Ohio subregions are the smallest HUC4 regions selected. Monongahela has the highest stream gauging density (17 stations in 10^4 km^2), followed by Kentucky-Licking, Allegheny, Lower Ohio, Wabash, and the ORB (Table 1).

2.2. Datasets

This study builds ML models to predict daily streamflow series in ungauged reaches from 2010/10/01 to 2020/09/30. Accordingly, streamflow data from USGS stations are used to build and test the model. Ensuring the quality and coverage of streamflow data is crucial for constructing reliable data-driven PUB models (Razavi and Coulibaly, 2017; Yilmaz and Bihrat, 2019). The USGS stations, which are impacted by regulation of dams, and nested stations where the spatial correlation is high, were eliminated.

Dams in a streamflow network lead to the anthropogenic impact on streamflow generation processes (Hodgkins et al., 2023). The dam influence on streamflow cannot be fully mitigated, and it is difficult to define the 'natural' river because dams are ubiquitous in rivers and creeks. Stations with significant impact of dams need to be removed. Usually, a comparison of dam storage and annual hydrograph can delineate stations with "unnatural" flow. However, dam storage information is not always available which makes it difficult to implement such a threshold consistently across all stations. Therefore, in this study, streamflow data collected within 3.2 km of a dam or reservoir are eliminated to alleviate the impact of reservoir control on streamflow measurements following the investigation made in the ORB (Yilmaz and Bihrat, 2019). Additionally, the distributions of correlations between rainfall and streamflow for dam influenced stations and the final chosen stations (shown in Fig. S1) seems to suggest that there is some improvement in the correlation of rainfall and streamflow for the final chosen set.

Nested stations are eliminated from the streamflow stations due to the high spatial correlation between the stations shared the similar

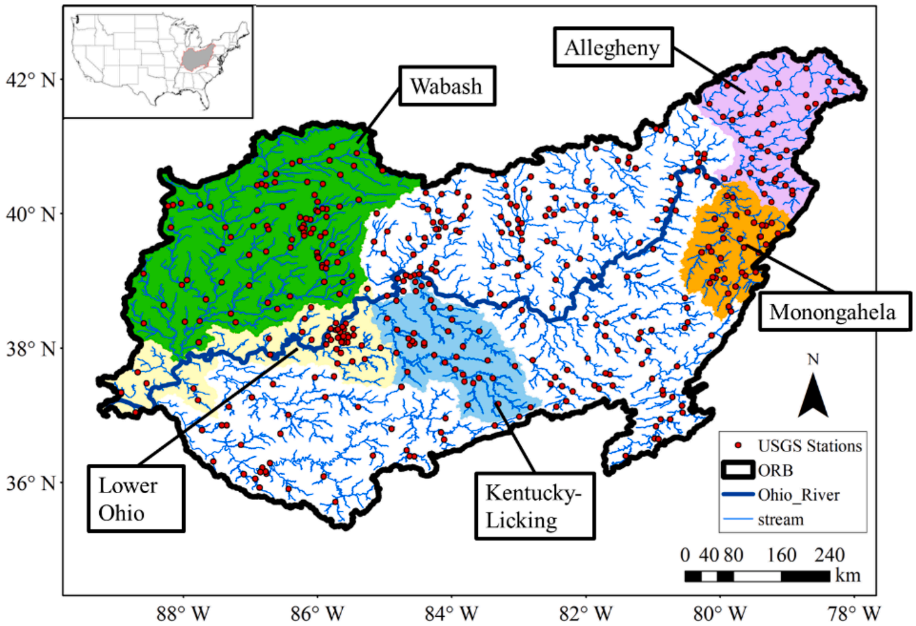

Fig. 1. Map of the ORB and the HUC4 level watersheds: Allegheny, Monongahela, Kentucky-Licking, Wabash, and Lower Ohio subregion. The distribution of USGS streamflow stations is also presented.

Table 1
Study area, number of available streamflow stations, and streamflow stations density of the study areas (following the selection process of USGS stations in Section 2.2).

Watershed	HUC level	Study area (km^2)	Number of available stations	Density of stations (stations/10^4 km^2)
Ohio Region	HUC2 (05)	421,951	431	10
Allegheny	HUC4 (0501)	30,376	37	12
Monongahela	HUC4 (0502)	19,104	33	17
Kentucky-Licking	HUC4 (0510)	27,638	35	13
Wabash	HUC4 (0512)	85,349	87	10
Lower Ohio	HUC4 (0514)	32,633	40	12

portion of watersheds. The spatial correlation of streamflow stations observed in connected streams (Krajewski et al., 2020; Krajewski and Sikorska-Senoner, 2021), known as spatial persistence, is minimized to ensure the quality of streamflow data. Krajewski et al. (2020) suggested that when the drainage area ratio of an upstream watershed to the corresponding downstream watershed is large (>0.6), the upstream and downstream station have similar streamflow patterns. As a result, when the drainage area ratio for a given station is greater than 0.6, its downstream station is removed from this study to avoid the influence of spatial correlation between stations.

Previous hydrological studies filtered the hydrological datasets with a missing rate ranging from 10 % to 20 % (Milly et al., 2005; Razavi and Coulibaly, 2017; Yilmaz and Bihrat, 2019). In addition, ML studies have highlighted the influence of missing data on model performance when the missing rate exceeds 5 % (Moorthy et al., 2014; Thomas and Rajabi, 2021). To minimize the uncertainty caused by missing data, a 5 % missing rate is chosen as the criteria for selecting USGS stations in this study. Thus, any USGS station having more than 5 % missing values or located within 3.2 km downstream of a reservoir was removed from this study.

PUB models are usually built using watershed characteristics and meteorological variables (Razavi and Coulibaly, 2013; Zhang et al., 2018; Schoppa et al., 2020). Both meteorological variables (Beck et al., 2017; Zhang et al., 2018) and watershed characteristics, including topography, soil, and land use conditions (Saadi et al., 2019; Schoppa et al., 2020) could act as major controls of PUB models. Table 2 shows the variables used in building ML models and evaluating the model performance in this study. Meteorological variables represent dynamic variation in climate, whereas watershed characteristics serve as static variables to capture streamflow generation processes. While anthropogenic modifications, such as dam density, are not commonly used in building PUB models, the impact of dams on hydrological response is investigated in the discussion section of this article (Nathan and Lowe, 2012; Yihdego and Webb, 2013; Fleming, B et al., 2021; Singh and Basu, 2022). Dam density is the density of georeferenced dams within a study area collected from the National Inventory of Dams (NID, 2022).

2.2.1. *Meteorological variables*

Rainfall and snowfall are the major meteorological forcing for streamflow generation processes. The meteorological variables include rainfall, snowfall, snow depth, and temperature. The ORB is in the middle latitude region, so snowfall and snow depth data are also included (Winkler et al., 2020; Schoppa et al., 2020). The temperature dataset indicates the possibility of snowmelt and frozen rivers (Prieto et al., 2019). Daily rainfall, snowfall, depth of snow, and temperature are downloaded from the National Oceanic and Atmospheric Administration (NOAA) Global Historical Climatology Network (GHCN) database (Menne et al., 2012). Streamflow is associated with current rainfall and the extension of the previous rainfall event, known as the antecedent rainfall (Istok and Boersma, 1986; Upreti and Ojha, 2021). Antecedent precipitation variables: precipitation in the previous one day, seven days, and thirty days are chosen to capture the influence of earlier rainfall (Table 2). In addition, the snow depth and snowfall affect the magnitude of streamflow when the snowmelt event happens (Winkler et al., 2020). The snowfall and snow depth at current time step are incorporated in the RF model (Table 2). The water equivalent from snow is considered by including the monthly snowfall and difference between snow depth at current time step and previous day to meteorological variables (Bergeron et al., 2016).

Table 2
Variables applied in the RF model. Dam density (with an asterisk) is not used to train the ML models but is used to disclose the impact of covariate shift and preferred range of ML models.

Categories	Variables	Code	Details	Range	Source
Meteorology	Rainfall	PRCP	Daily rainfall	–	GHCN
	Previous Rainfall (one day)	PRCP_lag1	Precipitation in the previous day (one day lag)	–	GHCN
	Previous Rainfall (seven days)	PRCP_7D	Seven days accumulation of precipitation	–	GHCN
	Previous Rainfall (one month)	PRCP_30D	Thirty days accumulation of precipitation	–	GHCN
	Snowfall	SNOW	Daily snowfall	–	GHCN
	Previous snowfall (one month)	SNOW_30D	Thirty days accumulation of snowfall	–	GHCN
	Snow depth	SNWD	Daily snow depth	–	GHCN
	Snow depth difference	SNWD_diff	The difference between the snow depth at current time step and previous day	–	GHCN
	Maximum temperature	TMAX	Maximum daily temperature	–	GHCN
	Minimum temperature	TMIN	Minimum daily temperature	–	GHCN
Topography	Drainage Area (km^2)	Area	Drainage area	7.76–251,229	USGS
	Average Elevation (m)	Elev	Averaged elevation	41.3–386.3	DEM from USGS
	Slope	Slope	Averaged Slope	7 × 10^{-3}-0.47	DEM from USGS
Land Use	Impervious Percentage	Imper	Averaged impervious percentage	0.11–56.0	NLCD
	Urbanized Area (%)	Urban	Averaged urbanized area	1.7–99.9	NLCD
Soil	Clay%	Clay%	Averaged clay percentage	16.0–45.1	STATSGO
	Sand%	Sand%	Averaged sand percentage	6.4–49.6	STATSGO
	Permeability (cm/hr)	Perm	Averaged permeability	1.0–15.1	STATSGO
	Hydraulic Conductivity (μm/s)	K	Averaged hydraulic conductivity	0.01–77.2	STATSGO
*Anthropogenic Modification	*Dam Density (dams/km^2)	–	Averaged dam density	0–0.122	NID

2.2.2. *Watershed characteristics*

Watershed characteristics are presumed to affect streamflow generation in response to meteorological forcing. ML models for PUB focus on generalizing the streamflow generation processes related to watershed characteristics, including soil properties, land use properties, and topography (Sivapalan et al., 2010; Razavi and Coulibaly, 2013). Here, the watershed characteristics for building the ML models include topography, soil, and land use. Topographic variables include slope, elevation, and drainage area of watersheds. The drainage area and elevation are the basic geometry and topography describing the watershed's latitude and size for understanding the overall capacity of surface water. The average slope of a watershed is related to the kinetic behavior and driving force of water flowing in the watershed due to the force of gravity. The average elevation and slope of watersheds are calculated from 30 m resolution Digital Elevation Model (DEM) from USGS 3D Elevation Program (USGS, 2022). The drainage area of watersheds is obtained from the metadata of USGS stations.

Soil properties and land use of watersheds provide information of rainfall abstraction and evapotranspiration processes. Watershed averaged land use and soil variables are prepared using the StreamCatchment (StreamCat) dataset (Hill et al., 2016; USEPA, 2022). Soil properties, including clay%, sand%, permeability, and hydraulic conductivity, are obtained from 0.5° resolution State Soil Geographic (STATSGO) dataset (Schwarz and Alexander, 1995; USDA, 2022). Finally, impervious percentages and urbanized area are obtained from 30 m resolution 2011 and 2016 National Land Cover Database (NLCD) from StreamCat (Table 2; MRLC, 2022).

The urbanized area in this study includes developed open area, developed low-intensity area, developed medium-intensity area, and developed high-intensity area. The difference between urbanized area and impervious percentage is that urbanized area represents only the percentage of developed area in a watershed, but the impervious percentage represents the percentage of developed area that is impervious and varies according to the intensity of development. The impervious percentage is 0–20 % of the total cover for developed open areas, 20–49 % for developed low-intensity areas, 50 %-79 % for developed medium-intensity areas, and 80–100 % for developed high-intensity areas (Wickham et al., 2013).

3. Methodology

3.1. *Data preparation*

In this study, the PUB models are built as lumped models requiring representative values of watershed characteristics and meteorological variables for each watershed. The boundary of each watershed is delineated using the DEM obtained from USGS. Watershed characteristics are derived by spatially averaging the raster datasets listed in Section 2.2. Meteorological variables for each watershed are spatially averaged using the Thiessen polygon method, which is typically used for estimating spatial averages of meteorological variables in flat topography, such as the ORB (Thiessen, 1911; McCuen, 2004). To evaluate the performance of a PUB model, the USGS stations, which are the outlets of the watersheds, are split into training/testing sets. An 80/20 split of streamflow stations for training/testing data is selected as presented in Table 3 because 80/20 has been reported as one of the best choices for dividing the dataset for the training/testing process in a machine learning model (Anifowose et al., 2017; Gholamy et al., 2018). The training set, which contains the remaining stations, is used to train the machine learning model (Fig. 2; Cibin et al., 2014; Athira et al., 2016;

Table 3
Number of streamflow stations used during the training and testing process and optimized hyperparameters of RF.

	No. of stations	Training stations	Testing stations	Number of trees	Maximum depth
ORB	431	344	87	1500	30
Allegheny	37	29	8	500	27
Monongahela	33	26	7	300	22
Kentucky-Licking	35	28	7	300	20
Wabash	87	69	18	1100	30
Lower Ohio	40	32	8	500	24

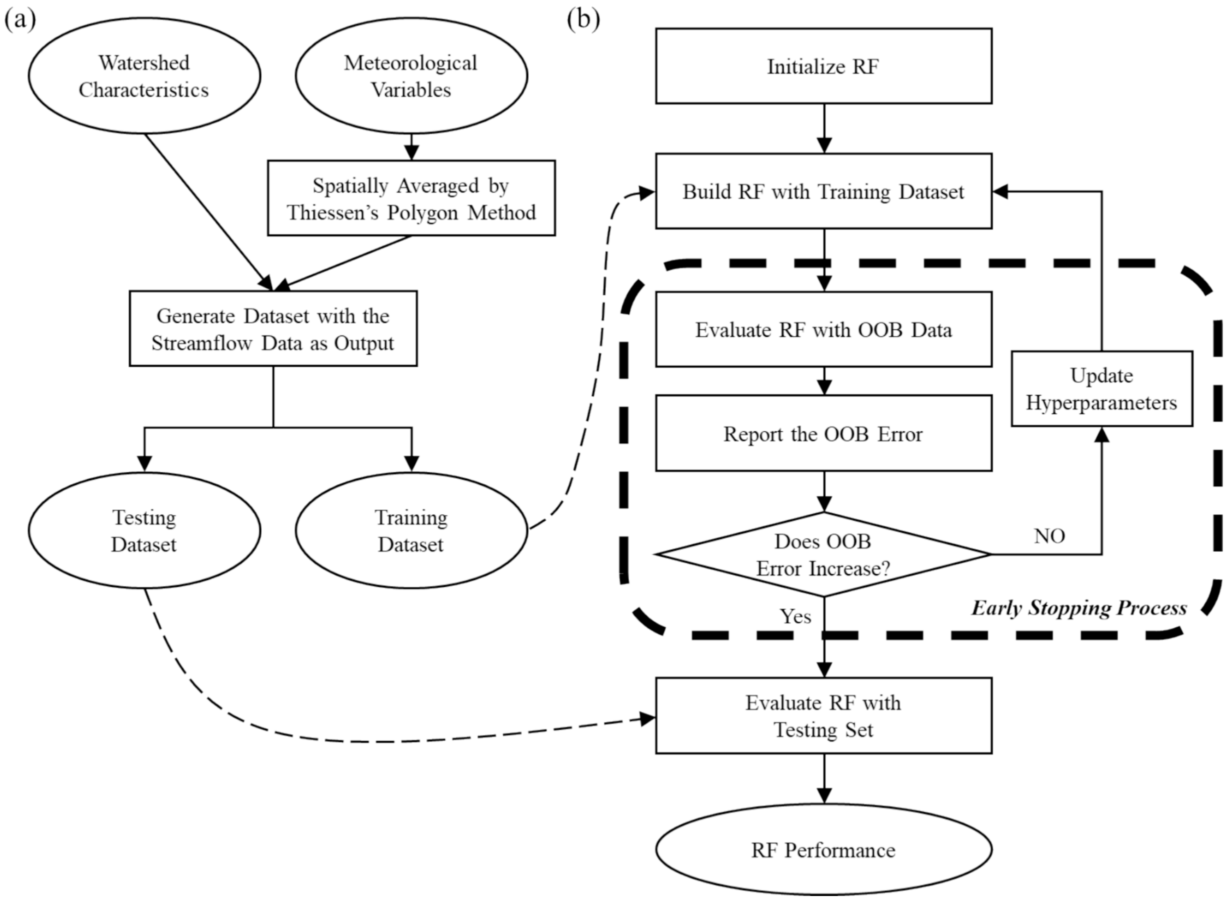
(a)
Watershed Characteristics
Meteorological Variables
Spatially Averaged by Thiessen's Polygon Method
Generate Dataset with the Streamflow Data as Output
Testing Dataset
Training Dataset
(b)
Initialize RF
Build RF with Training Dataset
Evaluate RF with OOB Data
Report the OOB Error
Update Hyperparameters
Does OOB Error Increase?
NO
Yes
Early Stopping Process
Evaluate RF with Testing Set
RF Performance

Fig. 2. The flow chart of applying RF models for PUB. (a) data preparation, (b) the training and testing process of RF with early stopping method. The training dataset is used in the training process for building RF models, and the testing dataset is used to evaluate the simulations (dashed lines). The solid arrows are the processed information passing to the next process. The heavy dashed line highlights the early stopping process as the iterative process to get the optimized hyperparameters of RF models.

Choubin et al., 2019). The testing set, hereafter referred to as pseudo ungauged reaches, is used for evaluating the model.

3.2. *Machine learning model – random Forest*

Random Forest (RF) is one of the most widely used algorithms for PUB due to its interpretability and non-parametric properties, facilitating its applications by engineers and researchers (Ziegler and König, 2014; Addor et al., 2018; Li et al., 2019; Prieto et al., 2019; Saadi et al., 2019; Araza et al., 2020; Desai and Ouarda, 2021; Esmaeili-Gisavandani et al., 2023). RF is capable of extracting similarities from vast datasets, which are typical of natural systems over large spatial, national, or global scales (Catal and Diri, 2009; Biau and Scornet, 2016; Carlisle et al., 2010; Miller et al., 2018; Mosavi et al., 2018; Saadi et al., 2019; Tyralis et al., 2019; Zhang et al., 2019). RF is an ensemble method that predicts physical quantity by averaging the predictions made among the decision trees (Breiman, 2001).

Decision trees in a RF model are built based on bootstrapped samples from training data. Each decision tree model is developed by splitting nodes into sub-nodes to make streamflow predictions. Each node in decision trees represents criteria based on specific input variables that determine the subsequent path toward streamflow predictions. The decision tree algorithm tends to have a high variance in simulation and often leads to overfitting problems. As a result, RF averages the predictions made among all the decision trees to prevent overfitting problems (Breiman, 2001). This study uses the Scikit-learn module in the Python development environment (Pedregosa et al., 2011) to develop the RF model.

The number of decision trees and the maximum depth of RF models are optimized using the out-of-bag (OOB) sample (Fig. 2). The OOB sample contains the left-out training data points during the training step. The early stopping method avoids over-fitting issues of RF models by finding optimized hyperparameters with the least out-of-bag error, which is in the form of mean squared error (Svetnik et al., 2003). During the optimization process, RF models are initialized by a set of hyperparameters, which is a pair of the number and maximum depth of decision trees. The optimized values of hyperparameters are found by grid search (Hinton, 2012), tries all pairs of the number of decision trees (from 100 to 1,800 at the step interval of 100) and maximum depth (from 10 to 50 at the step interval of one). Table 3 shows the optimized hyperparameters of RF models for all study areas.

Investigating the relative feature importance of variables in RF models provides insights into which variables contribute more to streamflow predictions made by RF (Fisher et al., 2019). RF's feature importance is used to determine the importance of input variables by measuring the reduction of impurity (variability of a variable within a decision tree node) by adding such variable during the regression process compared to the other variables (Breiman et al., 2017). The rank of feature importance reveals the relevance of variables in streamflow generation as learned by RF.

3.3. *Machine learning model – Artificial Neural Network (ANN)*

A fully connected shallow Artificial Neural Network (ANN) is the algorithm applied to predict streamflow in the ORB and is compared with the RF models for PUB in this study. ANN has been developed as one of the most popular ML models for PUB (Besaw et al., 2010; Chen et al., 2010; Valizadeh et al., 2017; Mosavi et al., 2018). This study develops an ANN model using the TensorFlow Keras module in the Python development environment.

The ANN model developed this study contains perceptron layers, activation functions between perceptron layers, and the regularization components (Fig. 3). Each layer of perceptron, known as a hidden layer, has multiple neurons. These neurons process the input information by multiplying the input from the previous layer by weights and then adding constant variables as the output. For the first perceptron layer, the input is the prepared datasets of watershed characteristics and meteorological variables (see Section 2.2). After processing, each perceptron layer passes its results to the subsequent layer, and this process continues until it reaches the final output layer. The weights and constant variables are determined during the training process.

Between each hidden layer, the activation functions decide whether the output of the neurons should be activated or not. Rectified linear activation units (ReLU) is chosen as the activation function due to its proven performance for regression and classification in ANN models (Goodfellow et al., 2016). ReLU is applied between each hidden layer to filter the negative values and produce the maximum values between the input values and zero (Eq. (1).

$$ReLU(x) = \max(0, x), x \in \mathbb{R} \tag{1}$$

which x is the output of each hidden layer in this study.

The hyperparameters indicates the overall structure of ANN including the number of neurons in each hidden layer and the parameters for regularization techniques (Table 4). The optimization process of ANN's hyperparameters searches a set of parameters that minimizes validation errors. The validation errors are the errors of ANN simulations by utilizing validation data, which is separated from the training data. The number of neurons for the two hidden layers is 1,000 and 100 respectively (Fig. 3). During the optimization process, the dropout and L2 regularization techniques are applied to prevent overfitting issues (Srivastava et al., 2014). The dropout randomly drops a percentage of neurons in a hidden layer during the training process to strengthen the learning ability of the other neurons and prevent the neurons from overly adapting the information of training data (Srivastava et al., 2014). The dropout rate is one of the hyperparameters in the ANN and ranges from 0 % to 10 % for shallow ANN models (Piotrowski et al., 2020). The other hyperparameter of regularization techniques is L2 regularization. L2 regularization adds an error term, which multiplies the L2 ratio with the square magnitude of the weight of neurons, in the loss function of ANN in the training process to avoid overfitting issues (Eq. (2). There are three values of the L2 regularization: 0.00001, 0.0001, and 0.001 for the optimization process. The optimized hyperparameters are found by the grid search method with the least error during the optimization process (Table 4).

Table 4
Optimized hyperparameters of ANN include the number of neurons, L2 ratio, and dropout rate of each hidden layer.

	Hidden Layer 1	Hidden Layer 2
Numbers of neurons	1000	100
L2 ratio	0.0001	0.0001
Dropout Rate	5 %	–

$$\text{L2 error} = \lambda \sum_{i=1}^{n} w_i^2 \tag{2}$$

where λ is the L2 ratio, w_i is the weight of neuron i in a hidden layer, and n is the number of neurons in the hidden layer.

3.4. Monte Carlo method for assessing uncertainty within data splitting process

The performance of ML models here may change with different training and testing data combinations due to the complexity of hydrological datasets. Therefore, any single set of training and testing data can be unreliable for evaluating the performance of ML models in hydrology (Araza et al., 2020; Schoppa et al., 2020; Prieto et al., 2022). To quantify the uncertainty in RF and ANN for PUB due to the data splitting process and covariate shift, Monte Carlo method is implemented as a bootstrap method for generating the empirical cumulative distribution function (CDF) of RF and ANN performance (Robert and Casella, 2013). One thousand training and testing data combinations, known as the Monte Carlo scenarios, are generated by repeatedly and randomly sampling the entire dataset (Fig. 2). In each scenario, 345 stations are randomly chosen as training set, and the rest of the stations (86) are chosen as testing set in the ORB. Repeating this process 1000 times leads to 1000 different combinations of training/testing set. Each station is expected to get selected in the training set around 800 times and in the

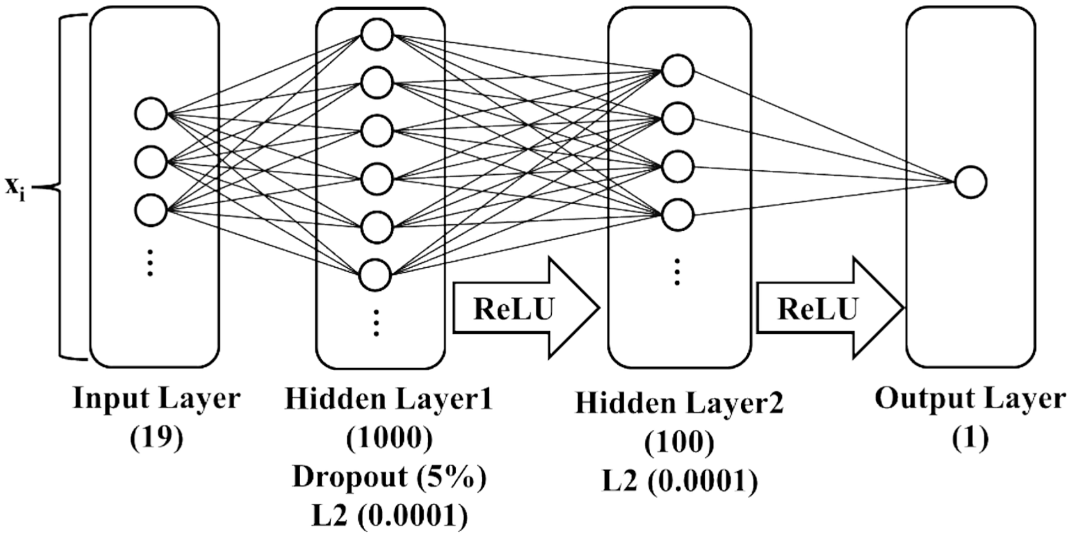

Fig. 3. The architecture of the shallow ANN applied to PUB. For each time step (from day 1 to day 3653), the ANN model generates a streamflow prediction for a station (training or testing). The accumulation of all 10 years of simulations becomes one complete prediction of the streamflow series at the station. x_i is the input variables, such as precipitation, snowfall, and drainage area, of a data point in a certain time step at the station.

testing set around 200 times under an 80/20 split chance for the training/testing dataset. The aggregate of simulation results from Monte Carlo scenarios represents predictive distributions of each reach: probability density functions (PDFs) of the performance of ML models using NSE. Due to the high computational demand of Monte Carlo analysis, high-performance computing (HPC) resources from the Extreme Science and Engineering Discovery Environment (XSEDE) (Towns et al., 2014) were used.

A covariate shift is present when input variables are distributed differently between a training and testing dataset. The presence of covariate shifts causes a trained ML model to fail to capture the inherent patterns in the testing set (Ramchandran and Mukherjee, 2021). The two-sample Kolmogorov–Smirnov (KS) test compares the distribution of variables in the training and testing dataset to determine whether the distributions are identical or independent (Raza et al., 2015). If the distributions of input variables are different between the training and testing datasets, the data splitting process is considered to include the covariate shift phenomenon. For example, suppose a Monte Carlo scenario has a training set entirely of stations with large drainage areas and a testing set containing only stations with small drainage areas. In that case, the inherent patterns learned by ML models are biased towards streamflow generation processes represented in large basins and may fail to estimate accurate streamflow for the testing set. This Monte Carlo scenario has suffered from inconsistency in the data splitting process due to covariate shift.

4. Result and discussion

4.1. Machine learning performance using Monte Carlo method

The overall performance of RF and ANN is evaluated using the Nash–Sutcliffe Efficiency (NSE) of streamflow simulations in a unit drainage area (Fig. 4). NSE reveals the ability of RF and ANN in detecting peak flows, and RMSE shows the overall error of ML performance. For most of the study areas, more than 80 % of RF simulations have satisfactory performance (NSE > 0.5; Moriasi et al., 2007). The RF simulations in Monongahela show the best overall performance with 97 % satisfactory simulations. In comparison, the ones in Kentucky-Licking only have 69 % satisfactory simulations, much lower than RF simulations in the other study areas. On the other hand, 52 % of ANN performance is unsatisfactory. Table 5 shows that ANN has the highest unsatisfactory rate, and more than 50 % of the pseudo ungauged reaches cannot get satisfactory simulations. The accuracy and overall performance of RF models are better than ANN models. There are four case studies of hydrographs of observations and predictions by RF and ANN models (Fig. 5 and Fig. 6). For the best scenarios, RF models are better in predicting high streamflow values than ANN models (Fig. 5). For the worst scenarios, ANN models would overestimate in the large streamflow case (Fig. 6). RF models have better simulation results for case studies of the best/worst performances in the ORB models. Table 5 shows NSE values at the 50 % cumulative probability (CP), which describes half of the simulation results exceeding the listed NSE values, of

Table 5
NSE values at 50 % CP, unsatisfactory rates, and the KS test results of ML performance. The unsatisfactory rate is defined as the number of pseudo ungauged reaches with unsatisfactory performance (NSE < 0.5; Moriasi et al., 2007) divided by the total number of available streamflow stations in the study areas. The KS tests (significant level is 0.05) are applied between the RF simulations in the ORB and the other simulations.

	NSE at 50% CP	Unsatisfactory Rate (%)	KS Test of NSE values (p value)	
			All Simulations	Station Means
ORB_RF	0.65	16	–	–
ORB_ANN	0.49	52	reject (p = 0.00)	reject (p = 0.00)
Allegheny	0.65	16	reject (p = 0.00)	fail to reject (p = 0.33)
Monongahela	0.71	3	reject (p = 0.00)	fail to reject (p = 0.18)
Kentucky-Licking	0.60	31	reject (p = 0.00)	fail to reject (p = 0.08)
Wabash	0.62	20	reject (p = 0.00)	fail to reject (p = 0.06)
Lower Ohio	0.70	13	reject (p = 0.00))	fail to reject (p = 0.15)

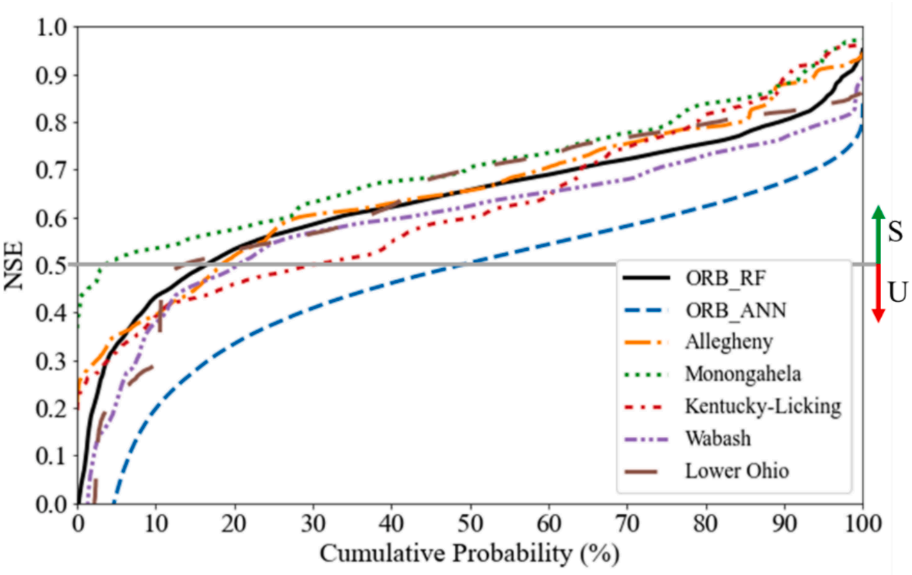

Fig. 4. CDF of the accumulated performance (NSE of depths of daily streamflow) using the ANN model for the ORB and RF models for the ORB, Allegheny, Monongahela, Kentucky-Licking, Wabash, and Lower Ohio subregion. S/U stands for satisfactory/unsatisfactory simulation results (NSE ≥ 0.5/ NSE < 0.5; Moriasi et al., 2007).

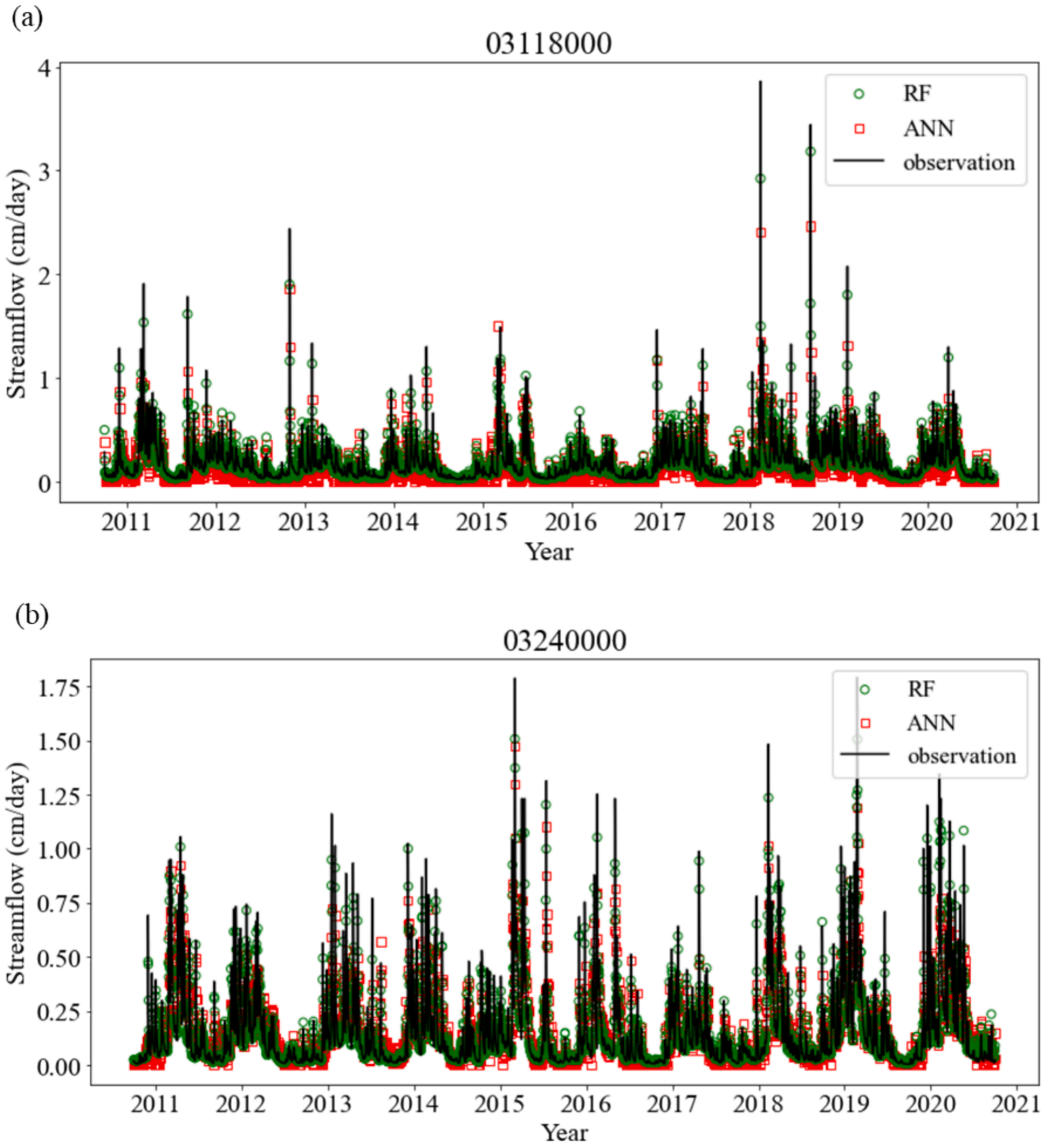

Fig. 5. The hydrographs of observation, RF simulations, and ANN simulations at (a) USGS 03118000 (NSE = 0.96 for RF and 0.74 for ANN) and (b) USGS 03240000 (NSE = 0.96 for RF and 0.78 for ANN), which are the stations with one of the best ML performances in the ORB.

all the CDFs of RF and ANN simulations. The KS test results show whether the distribution of RF and ANN simulation results in a study area is identical to the distribution of RF simulation results in ORB. The distribution of ANN simulation results is significantly different from the RF simulations (Table 5).

RF performance shows no specific correlation with the size of study areas with different spatial scales for PUB. Kentucky-Licking and Monongahela are the smallest study areas in this study. However, RF performs worst in Kentucky-Licking and has the lowest satisfactory rate (69 %), while it performs best in Monongahela with the highest satisfactory rate of simulations (97 %). In addition, the difference between the performance of RF models for ORB, Allegheny, Kentucky-Licking, and Wabash is small (Table 5). This comparison shows that the size of the study area may not be the only factor influencing the performance of RF. In order to have a more comprehensive investigation of RF performance, the distribution of RF simulations using Monte Carlo realizations is investigated (Fisher et al., 2019; Clark et al., 2021). This study provides two primary directions to investigate the performance of RF: (i) the statistics of RF's performance for pseudo ungauged reaches in study areas and (ii) the feature importance of RF and RF's preferred ranges of watershed characteristics (descriptions in Section 4.3).

Mean, variance, skewness, and kurtosis (actually excess kurtosis based on comparing the sample to a normal distribution) of predictive distributions, formed by aggregating RF simulation results among Monte Carlo scenarios, represent the pattern of RF's performance regarding pseudo ungauged reaches (Fig. 7). The mean values of predictive distributions show that most stations are expected to perform satisfactorily. In addition, more than 98 percent of stations show a low variance of NSE values, and only less than 2 percent of distributions have their variance reaching more than 0.01. Overall results from the mean and variance of the predictive distributions show that RF simulations are consistent at most stations.

Though the mean and variance of predictive distributions are consistent, the predictive distributions at some reaches are negatively

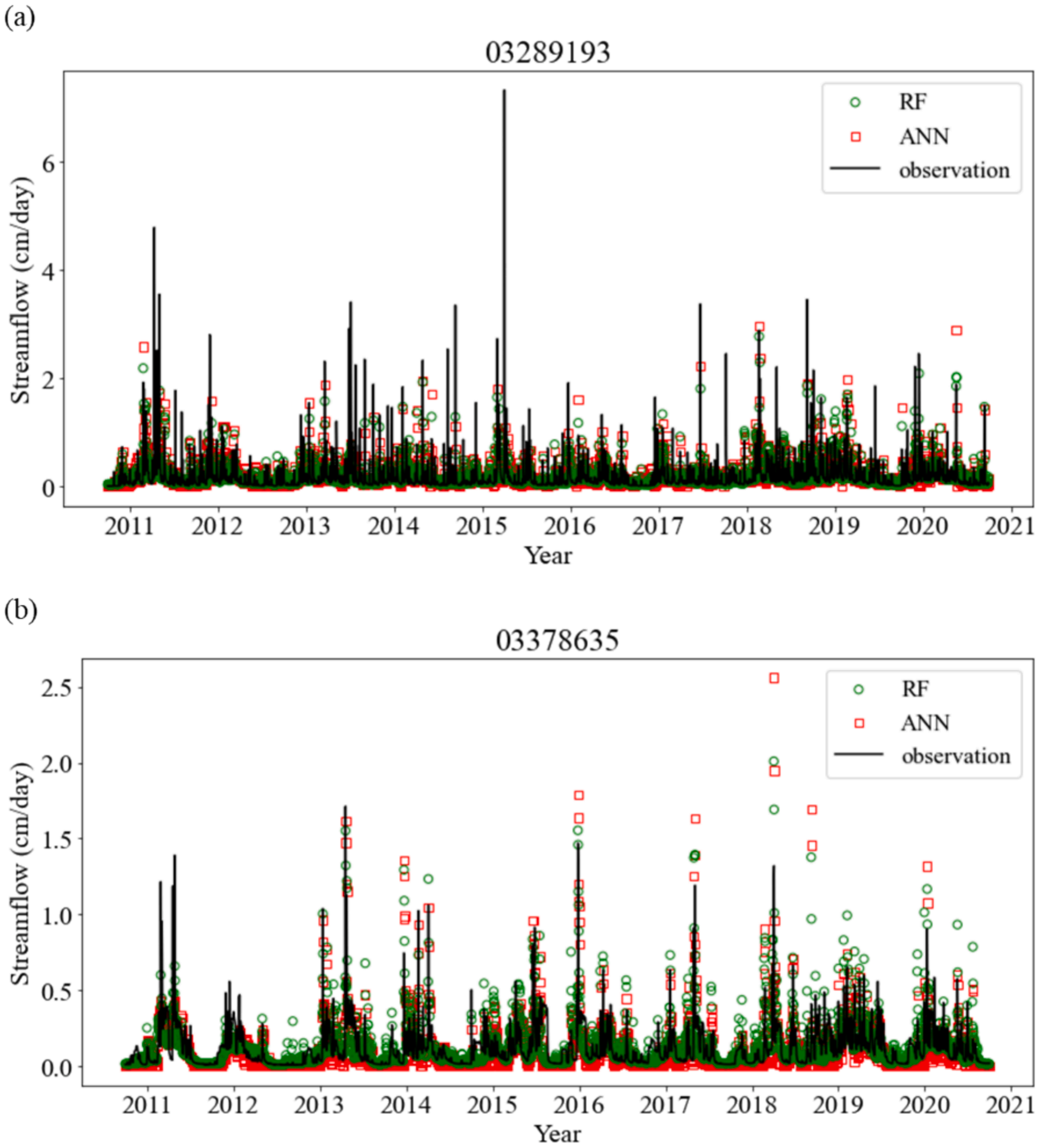

Fig. 6. The hydrographs of observation, RF simulations, and ANN simulations at (a) USGS 03289193 (NSE = 0.37 for RF and 0.34 for ANN) and (b) USGS 03378635 (NSE = 0.37 for RF and 0.24 for ANN), which are the stations with one of the worst ML performances in the ORB.

skewed and have high kurtosis (Fig. 8). The skewness of predictive distributions shows the lack of symmetry of their shape, and the excess kurtosis describes their peak and tail. Normal distribution tends to have zero skewness and kurtosis; however, the predictive distributions having negative skewness and high kurtosis tend to have a heavy tail and a higher chance of having outliers with poor performance (Westfall, 2014). Overall, the high skewness and kurtosis among Monte Carlo results show that relying on a single RF simulation for PUB may lead to potential outliers (extremes) and increased uncertainty in streamflow predictions.

4.2. *Influence of covariate shift on RF and ANN*

ML models suffer from covariate shifts of variables in the data splitting process in many real-world applications (McGaughey et al., 2016; Lucas et al., 2019; Schneider et al., 2020; Balogun and Attoh-Okine, 2021). PUB includes high variability of inputs (watershed characteristics) and might encounter covariate shifts when applying ML models. This section uses the Monte Carlo method to evaluate the effect of covariate shifts in ML models in the context of PUB in the ORB. Two criteria quantify the covariate shift phenomenon: (i) the total number of input variables suffering from covariate shift, henceforth referred to as global heterogeneity, and (ii) the influence of a specific covariate shift variable henceforth referred to as individual variable heterogeneity. The influence of covariate shift is quantified by comparing the mean, variance, skewness, and kurtosis of predictive distributions with and without covariate shift. 5,000 Monte Carlo scenarios are applied to investigate how predictive distributions vary under the covariate shift of a specific variable.

The numbers of global heterogeneity and individual variable heterogeneity from the Monte Carlo scenarios are shown in Fig. 9 and Table 6. The number of homogeneous/heterogeneous scenarios are similar in RF and ANN models due to the same data splitting method: random sampling. About 3,000 Monte Carlo scenarios are free from

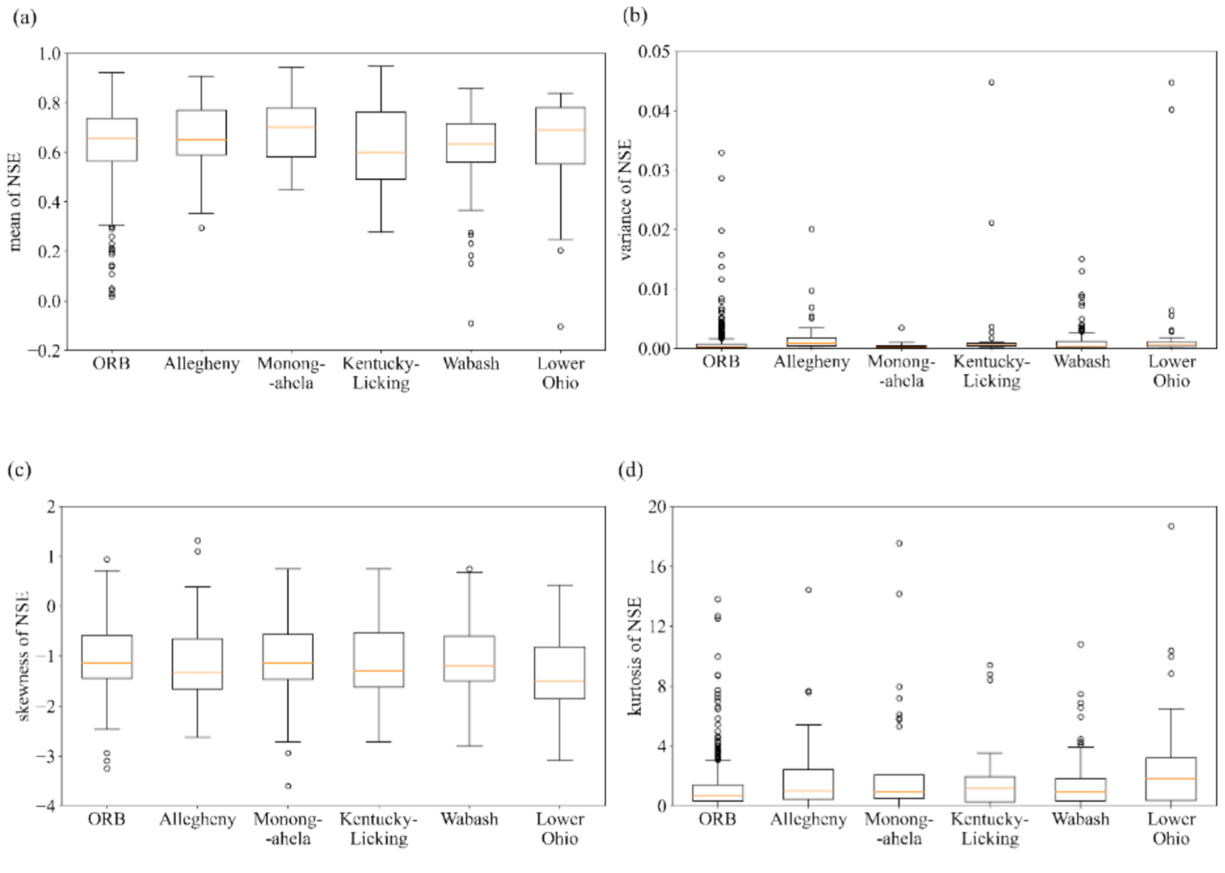

Fig. 7. Box plots showing (a) mean, (b) variance, (c) skewness, and (d) kurtosis of RF predictive distributions in different study areas.

covariate shifts, and the other Monte Carlo scenarios have at least one heterogeneous variable. Around 1150 scenarios are influenced by one covariate shift variable, which have the most frequent variables: drainage area, urbanized area, and dam density. The overall distributions of these variables are not normally distributed and show that most of the watersheds tend to have relatively low urbanized area, dam density, and drainage area (Fig. 10).

Covariate shifts result in predictive distributions' departures from normal distributions to non-normal distributions, known as the departure from normality (D'Agostino and Pearson, 1973; Das and Imon, 2016). In this study, the departure rate is defined as the ratio of the number of stations' predictive distributions suffering from the departures from normality to the total number of stations under covariate shift. Both global heterogeneity and individual variable heterogeneity result in departures from the normality of simulations (Fig. 11). Regarding the global heterogeneity, the departure rate of RF models increases from 18 % (one heterogeneous variable) to 22 % (three heterogeneous variables). ANN models maintain higher departure rates (28 % to 31 %) than RF models across different global heterogeneities. Overall, the results show that high global heterogeneity has considerable repercussions on kurtosis and skewness and can significantly add uncertainty and non-normality to the model performance.

The three most heterogenous variables– urbanized area, dam density, and drainage area – change the normality of more than 20 % of stations' predictive distributions of RF and ANN simulations (Fig. 11). RF and ANN have different departure rates under different individual variable heterogeneity. The covariate shifts caused by drainage area and dam density for RF and ANN simulations result in higher departure rates than the covariate shift with global heterogeneity of three random heterogeneous variables. The covariate shift caused by urbanized area for RF simulations has about 20 % departure rate, but the one for ANN simulations has 48 % departure rates, which shows the vulnerability of ANN models to the inconsistency of urbanized area between training and testing dataset.

The covariate shifts associated with drainage area, urbanized area, and dam density should be avoided because they have negative impacts on the normality of ML simulations and yield inconsistent patterns of biased streamflow generation regimes learned by ML models. Drainage area, urbanized area, and dam density are explicitly related to the streamflow generation regimes. Spatial scales of watersheds, described by their drainage area, are critical for building hydrological models (Betson, 1964; Jha et al., 2004). Furthermore, urbanized area and dam density represent the influence of anthropogenic activities on streamflow generation regimes of watersheds (Nathan and Lowe, 2012; Yihdego and Webb, 2013; Singh and Basu, 2022). As a result, the data prepared for RF (non-parametric ML) and ANN (parametric ML) models must avoid covariate shift of drainage area, dam density, and urbanized area due to the concern of the biased streamflow generation regime learned by ML models.

4.3. Role of variables in RF

How RF models learn streamflow generation processes from input variables and what the range of variables is favored for RF's best/worst performance reveal the applicability and limitations of RF. Feature importance represents the contribution of each variable in reducing the impurity of decision trees during the model-building process of RF. Most of the meteorological variables have greater feature importance (> 0.05) compared to the watershed characteristics (< 0.05) among all study areas (Fig. 12). The standard deviation of feature importance is relatively low compared to the mean of feature importance (<1%). Meteorological variables play a dominant role in building the RF model for PUB, while the watershed characteristics have much lower feature importance than the meteorological variables. The low feature

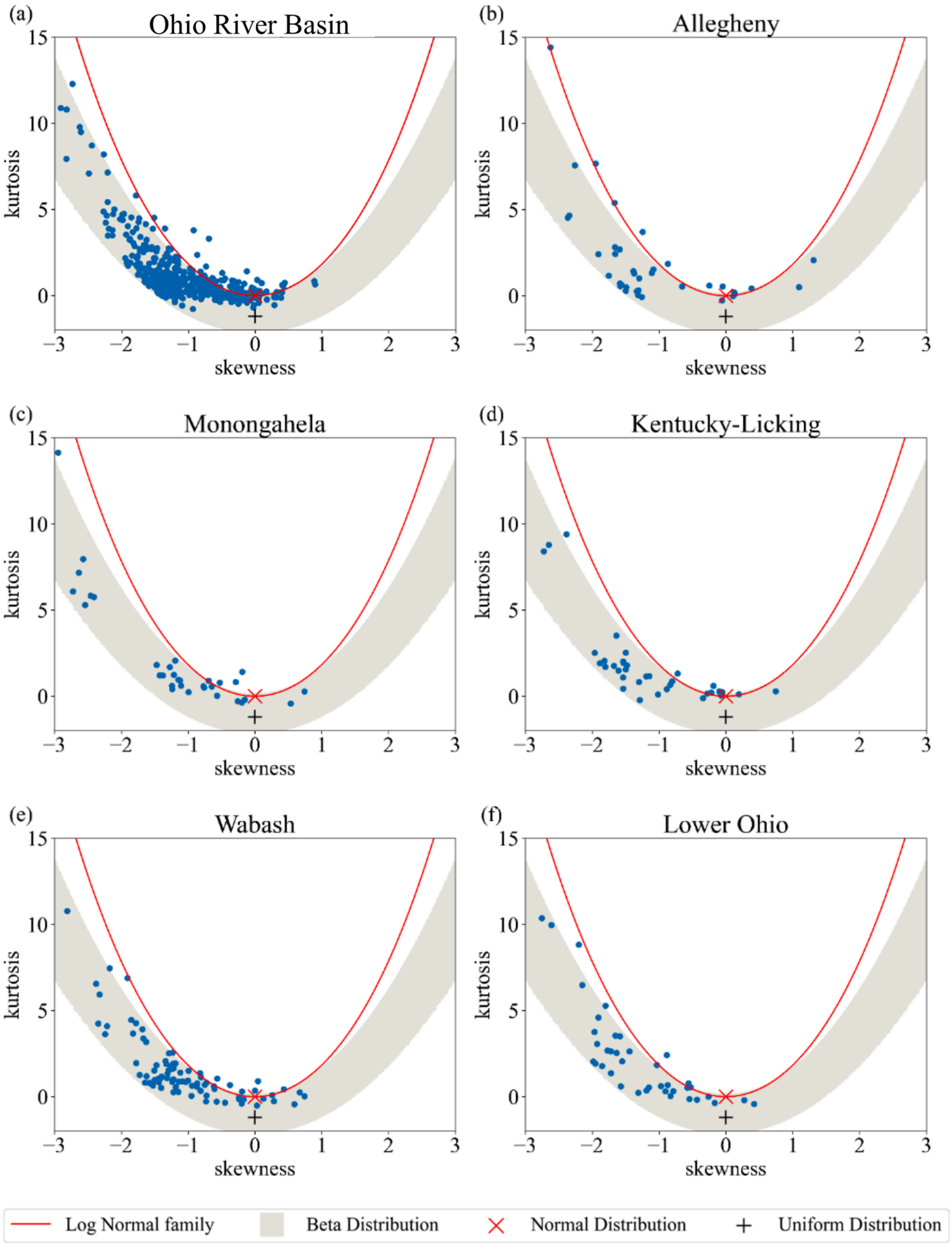

Fig. 8. Skewness-kurtosis plots of RF's performance in (a) Ohio River Basin, (b) Allegheny, (c) Monongahela, (d) Kentucky-Licking, (e) Wabash, and (f) Lower Ohio.

importance of watershed characteristics raises doubt about the ability of RF to apply watershed characteristics in predicting streamflow.

Among the meteorological variables, precipitation is noted as the most important variable, followed by short-term antecedent precipitations on the previous day or week in most of the study areas. These findings are consistent with Zhang et al. (2018), Beck et al. (2017), and Addor et al. (2018); climatic variables, such as precipitation and aridity, are among the most important variables for predicting the runoff signature by RF (Fig. 12). In areas with relatively higher annual snowfall, Allegheny and Monongahela, long-term snowfall (SNOW 30D) has higher feature importance than the other study area, and precipitation has lower feature importance than the other study area. The short-term snow variables, such as snowfall and snow depth, have much lower feature importance than the other meteorological variables and fail to contribute to building RF models even in the Allegheny and Monongahela (Fig. 12).

RF performs best when they have learned the streamflow generation processes and applied them to the PUB task in corresponding watersheds, while it performs worst when it fails to learn certain types of streamflow generation processes. Best and worst RF performance is defined as the top 10 percentile and bottom 10 percentile of expected NSE of predictive distribution among all stations. In the ORB, stations with expected NSE below the 10th percentile (NSE < 0.45) and above the 90th percentile of mean NSEs (NSE > 0.8) are assigned to the worst-performing group and best-performing group, respectively. For identifying significant differences in variables between the best/worst group

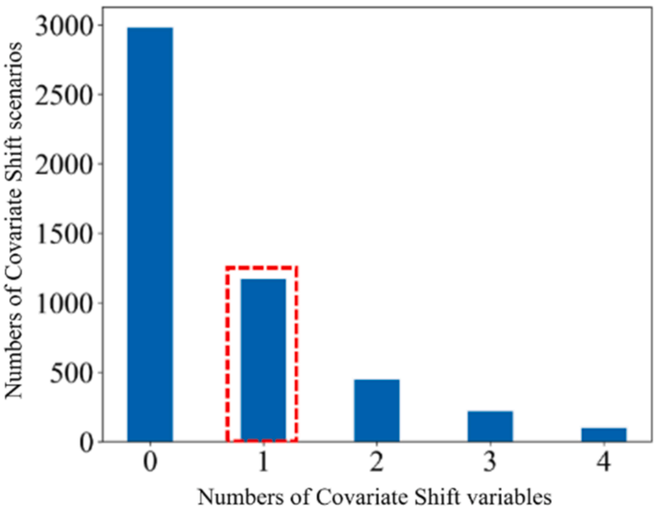

Fig. 9. The number or percentage of Monte Carlo scenarios with global heterogeneity of numbers of Covariate Shift variables.

Table 6
The percentage of covariate shift alone in Monte Carlo scenarios, which is visualized in Fig. 9 as the column selected by the dashed line.

Variables	Code	Details	Percentage of scenarios (%)
Annual Rainfall	–	Annual averaged precipitation	8.1
Annual Snowfall	–	Annual averaged snowfall	3.9
Annual Snow Depth	–	Annual averaged snow depth	3.5
Maximum temperature	TMAX	Maximum daily temperature	4.1
Minimum temperature	TMIN	Minimum daily temperature	4.1
Drainage area	Area	Drainage area	11
Averaged elevation	Elev	Averaged elevation	5.2
Slope	Slope	Averaged Slope	6.1
Impervious%	Imper	Averaged impervious percentage	5.6
Urbanized area (%)	Urban	Averaged urbanized area	9.9
Clay%	Clay%	Averaged clay percentage	5.2
Sand%	Sand %	Averaged sand percentage	5.4
Permeability	Perm	Averaged permeability	5.2
Hydraulic conductivity	K	Averaged hydraulic conductivity	12.3
Dam density	–	Averaged dam density	10.3

of stations and the overall stations, a two-sample KS test is applied with a 5 % significance level (Table 7).

The distribution of drainage area for overall stations is significantly different from the distribution of drainage area from stations with best simulations, while the drainage area distribution of the worst group does not show difference from the overall distribution of drainage area (Fig. 13a). About 60 % of reaches within the best group have upstream watersheds with drainage area from 1,000 to 10,000 km^2. Only about 20 % of reaches within the best group have their drainage area smaller than 1,000 km^2. On the contrary, 60 % of overall reaches are at the outlets of watersheds having drainage area smaller than 1,000 km^2, but only about 30 % are within 1,000 to 10,000 km^2, which is 30 % short compared to the best group. As a result, RF is more efficient in learning the streamflow generation processes for watersheds having larger drainage area (1,000 to 10,000 km^2) considering that larger basins mask the heterogeneity present in small ones (<1000 km^2). This preferred range of RF models in watershed scales may be related to the streamflow generation processes of watersheds with drainage area larger than 1,000 km^2, which have a higher signal-to-noise ratio of streamflow data compared to watersheds that are smaller than 1,000 km^2. Thus, RF models learn the signals of streamflow collected from watersheds with drainage area larger than 1,000 km^2 with less interference of noise than the ones from small watersheds (<1,000 km^2). In addition, the streamflow response, which is longer for meteorological forcings in watersheds with drainage area larger than 1,000 km^2, may dominate the streamflow generation processes learned by RF models. As a result, the streamflow generation processes learned by a ML model can capture the signals for most of the large watersheds (>1,000 km^2).

Dam density shows a preferred range of best simulations compared to the overall performance (Fig. 13b). More than 90 % of the reaches within the best group have dam density from 0.004 to 0.02 dams/km^2. However, only 40 % of overall reaches have dam density within this range. About 20 % of overall reaches have no dam within the upstream watershed. In comparison, only one reach (around 2.3 %) within the best group has no dam influence. The best group tends to have a narrower range of dam density than the overall reaches. According to NID (2022), the average dam density within the ORB is around 0.01 dams/km^2. The streamflow generation process learned by the RF model carries the influence of dams and fits better for watersheds with a moderate density of reservoirs (0.004 to 0.02 dams/km^2).

Rainfall and snowfall are the major meteorological forcing and the most important variables in the model building process here. The best performing stations show a moderate range of annual rainfall from 1 m to 1.5 m compared to the entire distribution of annual rainfall from 0.7 m to 2.5 m (Fig. 14a). The top thirty percent (70 %-100 % CP) of the stations within the best group have annual snowfall greater than 1.5 m, which is much higher than the ones in overall stations with top ten percent (90 %-100 % CP) of the stations. As a result, the streamflow generation regime learned by RF involves the meteorological transformation process, which can manage a moderate range of rainfall and relatively high snowfall. RF models show high efficiency and accuracy in dealing with watersheds with meteorological forcing in the form of rainfall and snowfall. On the other hand, 50 % (50 %-100 % CP) of the stations within the best group have annual snowfall of more than 0.5 m, which is much higher than the worst group with about 15 % (85 %-100 % CP) of the stations. RF fails to simulate the streamflow generation processes at pseudo ungauged reaches with relatively low snowfall due to the relatively high feature importance of long-term snowfall in the model building process of RF. The absence of the variation of snowfall inputs may lead to the malfunctioning of some splits of the nodes related to the long-term snowfall within a RF model, such as melting of large snowpacks. These dominant processes of RF models cannot participate in predicting streamflow in the watersheds with low snowfall, and the streamflow predicted at such watersheds tend to be more inaccurate than the ones with the participation of all the learned processes in RF (Sivakumar, 2000; Sivakumar, 2004).

5. Conclusions

ML methods provide an appealing alternative to traditional hydrologic models for PUB. A better understanding of the sources of uncertainty in ML model performance can enable them to be implemented more widely across large scales with greater confidence. This study quantifies the effect of data splitting process and the resulting covariate shift on the performance of ML models for PUB. Specifically, it implements RF and ANN to analyze (i) how to assess the problem of data splitting process of ML models, (ii) how covariate shifts of variables link to the high uncertainty incorporated in the performance of ML models, and (iii) what are the preferred ranges of watershed characteristics for ML models.

In order to assess the effect of data splitting in ML models, a Monte Carlo analysis is implemented to estimate the predictive distribution of

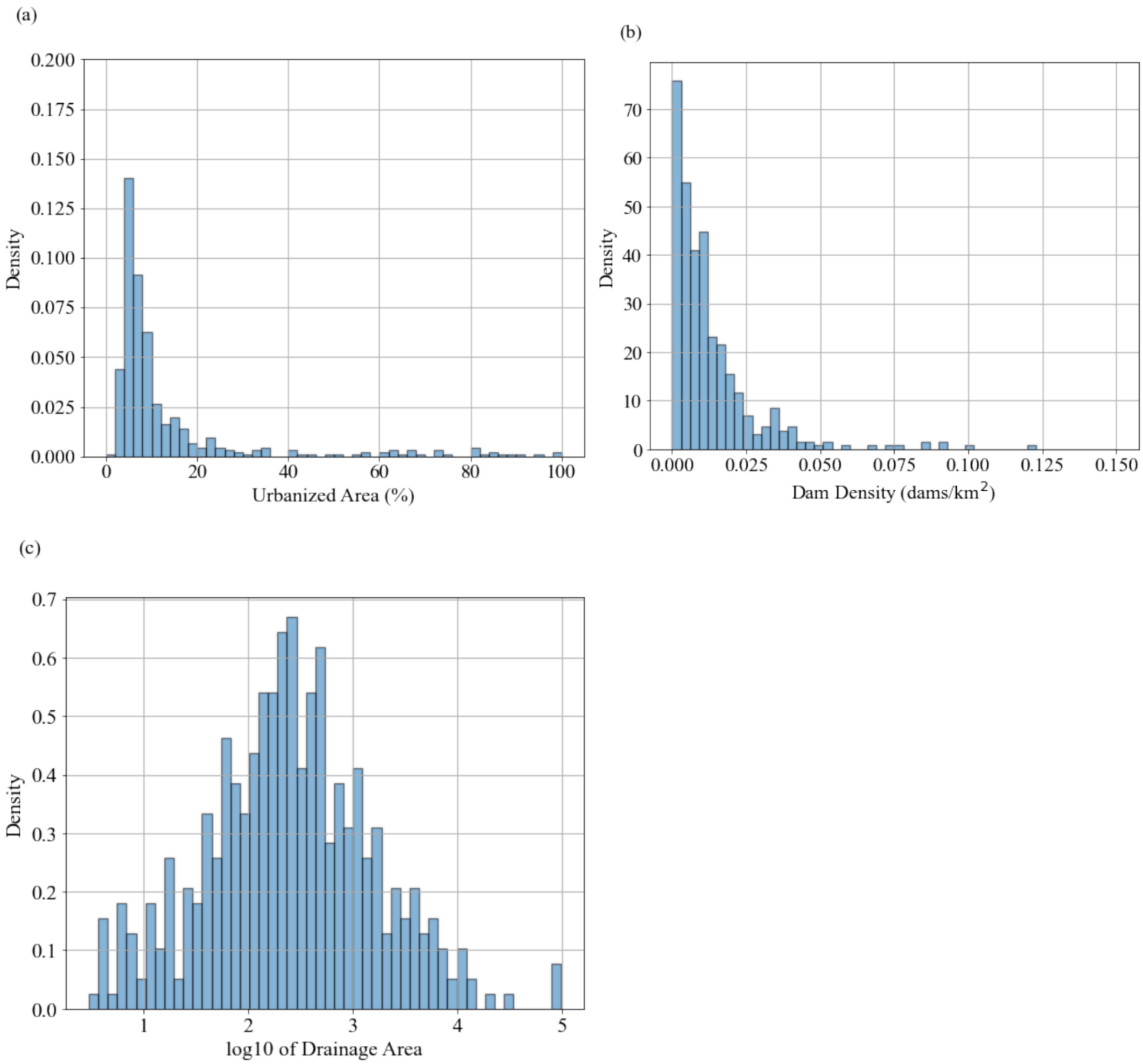

Fig. 10. The density plots of (a) urbanized area, (b) dam density, and (c) $\log_{10}$ of drainage area in km^2 showing the heterogeneity of distributions.

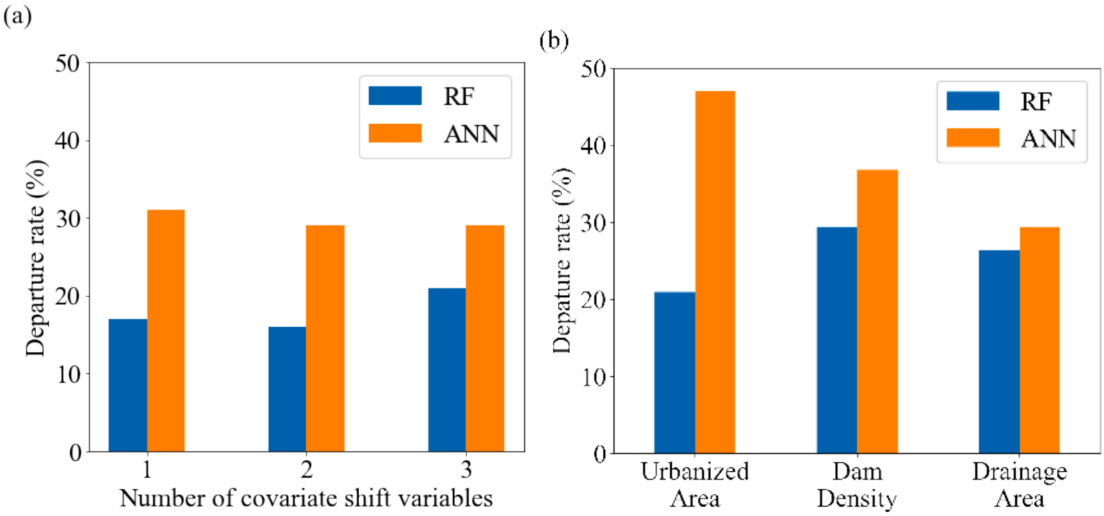

Fig. 11. Plots showing departure rates of ML models related to (a) global heterogeneity and (b) individual variable heterogeneity of urbanized area, dam density, and drainage area.

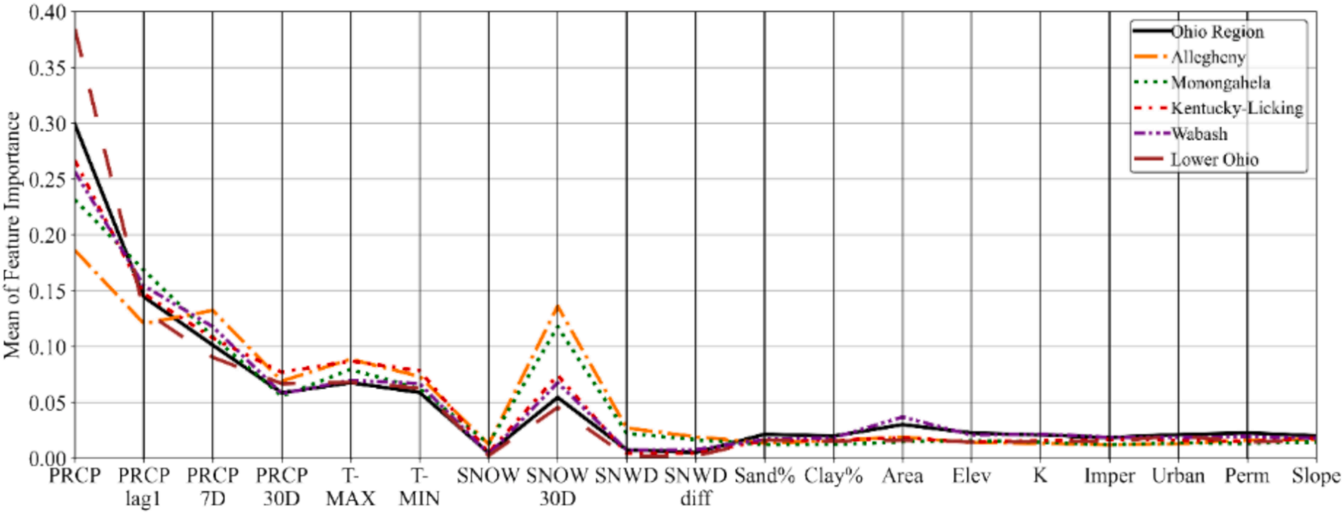

Fig. 12. Parallel coordinate plots of the feature importance in RF models for different study areas showing the mean of feature importance of variables. (The standard deviation of feature importance is orders of magnitude smaller and varies similar to the feature importance).

Table 7
KS test p-value of variables between the overall simulations and simulations from the best/worst 10 % group in the ORB. The vertical red line represents the p = 0.05 significance level, which is the criteria to reject/ or fail to reject the null hypothesis of KS test. (The distributions of all variables can be found in Fig. S1).

Variables	Code	Details	kS test p value for worst case	kS test p value for best case
Annual Rainfall	–	Annual averaged precipitation	0.15	0.04
Annual Snowfall	–	Annual averaged snowfall	0.00	0.02
Annual Snow Depth	–	Annual averaged snow depth	0.00	0.04
Maximum temperature	TMAX	Maximum daily temperature	0.04	0.08
Minimum temperature	TMIN	Minimum daily temperature	0.00	0.01
Drainage area	Area	Drainage area	0.26	0.00
Averaged elevation	Elev	Averaged elevation	0.00	0.01
Slope	Slope	Averaged Slope	0.11	0.36
Impervious%	Imper	Averaged impervious percentage	0.95	0.04
Urbanized area (%)	Urban	Averaged urbanized area	0.83	0.10
Clay%	Clay%	Averaged clay percentage	0.08	0.10
Sand%	Sand %	Averaged sand percentage	0.00	0.09
Permeability	Perm	Averaged permeability	0.06	0.00
Hydraulic conductivity	K	Averaged hydraulic conductivity		
Dam density	–	Averaged dam density	0.20	0.00

NSE for PUB. Further, the analysis is also implemented on different subbasins of ORB to study the differences in predictive distribution with respect to spatial scale of the watersheds. The predictive distributions are relatively consistent in terms of mean and have low variance indicating that RF performs satisfactorily at most stations across different data splits which indicates that ML has the potential for satisfactory performance in PUB. However, the negative skewness and high kurtosis of the predictive distributions indicate the likelihood of outliers in single RF simulations which can lead to erroneous predictions when ML is applied with random data splitting for PUB.

To overcome the limitations of the random data splitting, this study analyzed the effect of heterogeneity of the input variables, resulting in covariate shifts, on the performance of ML modeling for PUB. Covariate shift causes departures from normality for both parametric and non-parametric ML models in the applications of PUB. Under the influence of covariate shifts, ML may learn biased meteorological transformation processes from input variables to streamflow. Data splitting processes for training/testing ML models should avoid covariate shifts to alleviate the uncertainty in PUB. Global heterogeneity and heterogeneity of specific watershed characteristics in data splitting processes add uncertainty to ML modeling results. Specifically, the distribution of watershed characteristics like drainage area, urbanized area, and dam influence should be consistent between training and testing inputs so that ML models can learn and assess the appropriate meteorological transformation process.

Preferred ranges of variables, including drainage area, dam influence, and meteorological variables in the ORB for RF can be determined. RF can successfully learn the streamflow generation process from mesoscale watersheds with areas from 1,000 to 10,000 km^2 with a moderate range of dam density (4 to 20 dams/1,000 km^2). Meteorological forcings have a higher impact on the performance of RF for the ORB, as evidenced by the fact that the model performs well at stations with a moderate range of annual rainfall (1 to 1.5 m/year) and relatively high snowfall (0.5 to 1.5 m/year). RF works well in watersheds with the composition of rainfall and snowfall. On the contrary, watersheds with low snowfall tend to show poorer performance from RF due to the malfunctions of the splits created in RF for long-term snowfall. RF fails to predict the streamflow in regions with relatively low annual snowfall ($<$0.5 m) in the ORB.

Taken together, the three objectives can provide important guidelines regarding the implementation of ML for PUB. ML implementations need to incorporate targeted data splitting where the distribution of input variables are consistent between training and testing sets instead of using random sampling. Input variables should also be checked to ensure that they fall within the preferred range for the adopted ML technique. When implementing a trained ML model to an ungauged site, it is not possible to check its performance because of the lack of streamflow data at the site. However, checking the distribution and range of input variables can provide important indications to the robustness of the ML model's performance.

The impact of covariate shift can be alleviated using validation

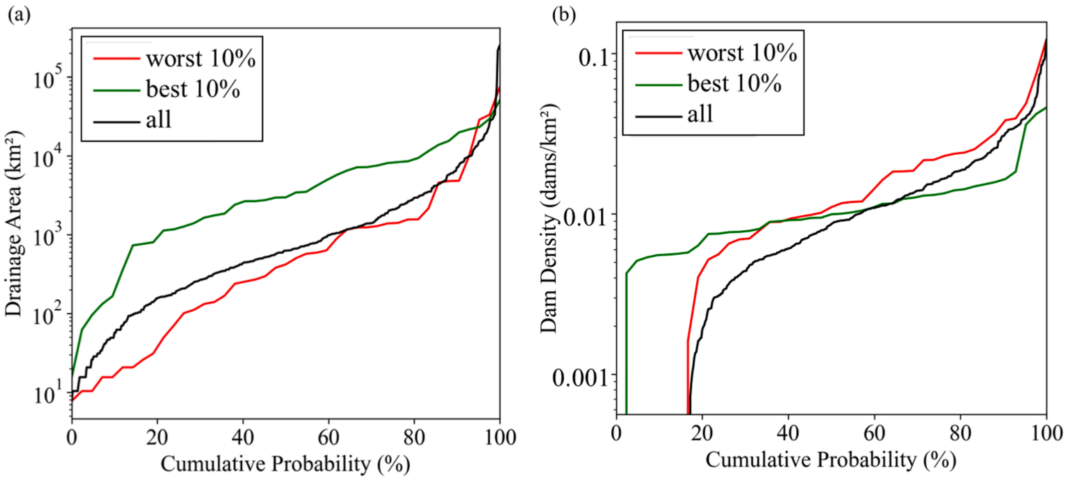

Fig. 13. CDF plots of (a) Drainage area and (b) dam density distribution of the best/worst 10% group and the overall performance.

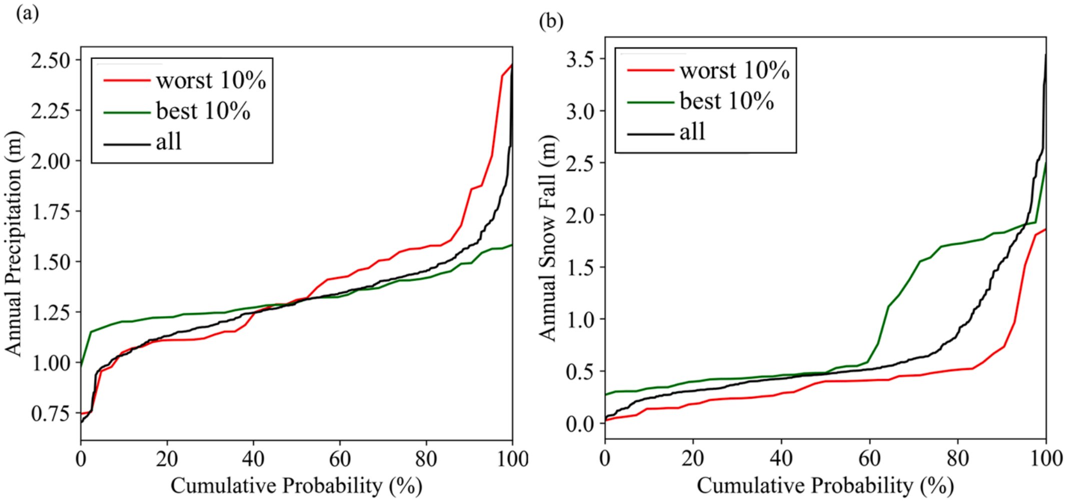

Fig. 14. CDF plots of (a) annual precipitation and (b) annual snowfall distribution of the best/worst 10% group and the overall performance.

techniques. These techniques include applying weights to training data or testing data (Reddi et al., 2015; Segovia-Martín et al., 2023), adjusting loss function and optimization process (Bickel et al. 2009; Liu and Ziebart, 2017), and splitting training/ testing data with pre-defined conditions (López et al., 2014; Rezaei et al., 2021). Future forays in this direction should analyze the effect of different validation techniques and data split ratios for ML in PUB. Also, this study was implemented on ORB which has gauges on streams with a wide range of drainage areas. However, smaller basins are generally less likely to be gauged and are therefore often underrepresented in training/testing dataset which can lead to more unpredictable behavior in prediction of streamflow generation process. This calls for a more in-depth analysis on the effect of covariate shift at different scales with a particular focus on smaller basins.

CRediT authorship contribution statement

Pin-Ching Li: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sayan Dey:** Writing – review & editing, Writing – original draft, Validation, Conceptualization. **Venkatesh Merwade:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was performed with funding from the U.S. National Science Foundation (OAC:1835822, OAC:1829764, and OAC:2118329). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) Bridges-2 at the service-provider through allocation TG-ASC170016. Authors thank Roberta Lynn Vance from the Lyles School of Civil Engineering at Purdue University for proofreading the paper.

Open Research

The datasets generated in this study are made available in the Purdue University Research Repository (PURR) at Li et al. (2023a), and the Python codes are made available in the Purdue University Research Repository (PURR) at Li et al. (2023b). In addition, the list of all stations in the ORB (1284 stations), including stage and streamflow stations, the list of daily streamflow stations in the ORB (598 stations), and the final list of selected stations (431 stations) are available in the Purdue University Research Repository (PURR) at Li et al. (2023a).

Data availability

Data will be made available on request.

References

Addor, N., Nearing, G., Prieto, C., Newman, A.J., Le Vine, N., Clark, M.P., 2018. A ranking of hydrological signatures based on their predictability in space. *Water Resour. Res.* 54 (11), 8792–8812. https://doi.org/10.1029/2018WR022606.

Adnan, R.M., Petroselli, A., Heddam, S., Santos, C.A.G., Kisi, O., 2021. Comparison of different methodologies for rainfall–runoff modeling: machine learning vs conceptual approach. *Nat. Hazards* 105 (3), 2987–3011. https://doi.org/10.1007/s11069-020-04438-2.

Adombi, A.V.D.P., Chesnaux, R., Boucher, M.A., 2021. Theory-guided machine learning applied to hydrogeology—state of the art, opportunities and future challenges. *Hydrgeol. J.* 29 (8), 2671–2683. https://doi.org/10.1007/s10040-021-02403-2.

AghaKouchak, A., Habib, E., 2010. Application of a conceptual hydrologic model in teaching hydrologic processes. *Int. J. Eng. Educ.* 26 (4 (S1)), 963.

Anifowose, F., Khoukhi, A., Abdulraheem, A., 2017. Investigating the effect of training–testing data stratification on the performance of soft computing techniques: an experimental study. *J. Exp. Theor. Artif. Intell.* 29 (3), 517–535. https://doi.org/10.1080/0952813X.2016.1198936.

Araza, A., Hein, L., Duku, C., Rawlins, M. A., & Lomboy, R. (2020). Data-driven streamflow modelling in ungauged basins: regionalizing random forest (RF) models. *bioRxiv*. doi:10.1101/2020.11.14.382598.

Athira, P., Sudheer, K.P., Cibin, R., Chaubey, I., 2016. Predictions in ungauged basins: an approach for regionalization of hydrological models considering the probability distribution of model parameters. *Stoch. Env. Res. Risk A.* 30 (4), 1131–1149. https://doi.org/10.1007/s00477-015-1190-6.

Balogun, I., Attoh-Okine, N., 2021. Random Forest–based covariate shift in addressing nonstationarity of railway track data. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part a: Civil Engineering* 7 (3), 04021028. https://doi.org/10.1061/AJRUA6.0001141.

Beck, H.E., van Dijk, A.I., Roo, A.D., Dutra, E., Fink, G., Orth, R., Schellekens, J., 2017. Global evaluation of runoff from 10 state-of-the-art hydrological models. *Hydrol. Earth Syst. Sci.* 21 (6), 2881–2903. https://doi.org/10.5194/hess-21-2881-2017.

Bergeron, J.M., Trudel, M., Leconte, R., 2016. Combined assimilation of streamflow and snow water equivalent for mid-term ensemble streamflow forecasts in snow-dominated regions. *Hydrol. Earth Syst. Sci.* 20 (10), 4375–4389. https://doi.org/10.5194/hess-20-4375-2016.

Besaw, L.E., Rizzo, D.M., Bierman, P.R., Hackett, W.R., 2010. Advances in ungauged streamflow prediction using artificial neural networks. *J. Hydrol.* 386 (1–4), 27–37. https://doi.org/10.1016/j.jhydrol.2010.02.037.

Betson, R.P., 1964. What is watershed runoff? *J. Geophys. Res.* 69 (8), 1541–1552. https://doi.org/10.1029/JZ069i008p01541.

Beven, K., 2006. A manifesto for the equifinality thesis. *J. Hydrol.* 320 (1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007.

Biau, G., Scornet, E., 2016. A random forest guided tour. *TEST* 25 (2), 197–227. https://doi.org/10.1007/s11749-016-0481-7.

Bickel, S., Brückner, M., Scheffer, T., 2009. Discriminative learning under covariate shift. J. Mach. Learn. Res. 10 (9).

Breiman, L., 2001. Random Forests. *Machine Learning* 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 2017. Classification and regression trees. *Routledge*. https://doi.org/10.1201/9781315139470.

Breuer, L., Huisman, J.A., Frede, H.G., 2006. Monte Carlo assessment of uncertainty in the simulated hydrological response to land use change. *Environ. Model. Assess.* 11 (3), 209–218. https://doi.org/10.1007/s10666-006-9051-9.

Carlisle, D.M., Falcone, J., Wolock, D.M., Meador, M.R., Norris, R.H., 2010. Predicting the natural flow regime: models for assessing hydrological alteration in streams. *River Research and Applications* 26 (2), 118–136. https://doi.org/10.1002/rra.1247.

Catal, C., Diri, B., 2009. Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem. *Inf. Sci.* 179 (8), 1040–1058. https://doi.org/10.1016/j.ins.2008.12.001.

Chang, Y., Wu, J., Jiang, G., Kang, Z., 2017. Identification of the dominant hydrological process and appropriate model structure of a karst catchment through stepwise simplification of a complex conceptual model. *J. Hydrol.* 548, 75–87. https://doi.org/10.1016/j.jhydrol.2017.02.050.

Chen, C.S., Chou, F.N.F., Chen, B.P.T., 2010. Spatial information-based back-propagation neural network modeling for outflow estimation of ungauged catchment. *Water Resour. Manag.* 24, 4175–4197. https://doi.org/10.1007/s11269-010-9652-6.

Choubin, B., Solaimani, K., Rezanezhad, F., Roshan, M.H., Malekian, A., Shamshirband, S., 2019. Streamflow regionalization using a similarity approach in ungauged basins: Application of the geo-environmental signatures in the Karkheh River Basin. Iran*Catena* 182, 104128. https://doi.org/10.1016/j.catena.2019.104128.

Cibin, R., Athira, P., Sudheer, K.P., Chaubey, I., 2014. Application of distributed hydrological models for predictions in ungauged basins: a method to quantify predictive uncertainty. *Hydrol. Process.* 28 (4), 2033–2045. https://doi.org/10.1002/hyp.9721.

Clark, M.P., Vogel, R.M., Lamontagne, J.R., Mizukami, N., Knoben, W.J., Tang, G., Papalexiou, S.M., 2021. The abuse of popular performance metrics in hydrologic modeling. *Water Resour. Res.* 57 (9), e2020WR029001. https://doi.org/10.1029/2020WR029001.

R.A.L.P.H. D'Agostino E.S. Pearson Tests for departure from normality. Empirical results for the distributions of b2 and $\sqrt{}$b1 Biometrika 60 3 1973 613 622 10.1093/biomet/60.3.613.

Darbandsari, P., Coulibaly, P., 2020. Inter-comparison of lumped hydrological models in data-scarce watersheds using different precipitation forcing data sets: Case study of Northern Ontario. *Canada. journal of Hydrology: Regional Studies* 31, 100730. https://doi.org/10.1016/j.ejrh.2020.100730.

Das, K.R., Imon, A.H.M.R., 2016. A brief review of tests for normality. *Am. J. Theor. Appl. Stat.* 5 (1), 5–12. https://doi.org/10.11648/j.ajtas.20160501.12.

Das, T., Bárdossy, A., Zehe, E., He, Y., 2008. Comparison of conceptual model performance using different representations of spatial variability. *J. Hydrol.* 356 (1–2), 106–118. https://doi.org/10.1016/j.jhydrol.2008.04.008.

Desai, S., Ouarda, T.B., 2021. Regional hydrological frequency analysis at ungauged sites with random forest regression. *J. Hydrol.* 594, 125861. https://doi.org/10.1016/j.jhydrol.2020.125861.

Electric Power Research Institute (EPRI) (2010). Program on technology innovation: Ohio River water quality trading pilot program—business case for power company participation, 2008. Palo Alto, CA: Electric Power Research Institute. Technical Report 1018861. Retrieved from http://kieser-associates.com/uploaded/epri_business_case_report.pdf [Accessed on 3rd August 2020].

Electric Power Research Institute (EPRI) Ohio River Basin water quality trading project http://wqt.epri.com/pdf/3002001739_WQT-Program-Summary_2014-03.pdf 2014 Electric Power Research Institute Palo Alto, CA Retrieved from [Accessed on 3rd August 2020].

Electric Power Research Institute (EPRI) Ohio River Basin water quality trading project http://wqt.epri.com/pdf/NEWSLETTER%20Jan%202016.pdf 2016 Electric Power Research Institute Palo Alto, CA Retrieved from [Accessed on 3rd August 2020].

Esmaeili-Gisavandani, H., Zarei, H., Fadaei Tehrani, M.R., 2023. Regional flood frequency analysis using data-driven models (M5, random forest, and ANFIS) and a multivariate regression method in ungauged catchments. *Appl Water Sci* 13 (6), 139. https://doi.org/10.1007/s13201-023-01940-3.

A. Fisher C. Rudin F. Dominici All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously J. Mach. Learn. Res. 20 177 2019 1 81 PMCID: PMC8323609.

Fleming, B.J., Archfield, S.A., Hirsch, R.M., Kiang, J.E., Wolock, D.M., 2021a. Spatial and temporal patterns of low streamflow and precipitation changes in the Chesapeake Bay Watershed. *JAWRA Journal of the American Water Resources Association* 57 (1), 96–108. https://doi.org/10.1111/1752-1688.12892.

Fleming, S.W., Watson, J.R., Ellenson, A., Cannon, A.J., Vesselinov, V.C., 2021b. Machine learning in Earth and environmental science requires education and research policy reforms. *Nat. Geosci.* 14 (12), 878–880. https://doi.org/10.1038/s41561-021-00865-3.

Gholamy, A., Kreinovich, V., Kosheleva, O., 2018. Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. *Int. J. Intell. Technol. Appl. Stat.* 11 (2), 105–111. https://doi.org/10.6148/IJITAS.201806_11(2).0003.

Gibson, L. (2020). 113,000 more properties may be at risk of flooding in Indiana than previously thought, report says. Indianapolis Star. Retrieved from https://www.indystar.com/story/news/environment/2020/08/17/113-000-more-homes-risk-flooding-indiana-report-says/5571932002/. [Accessed on 12th December 2020].

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press. ISBN: 0262337371.

Hauser, C. (2020, March 20). Heavy Rains Flood Parts of Ohio, Stranding Residents. *The New York Times*. Retrieved from https://www.nytimes.com/2020/03/20/us/ohio-flooding.html [Accessed on 12th December 2020].

Her, Y., Chaubey, I., 2015. Impact of the numbers of observations and calibration parameters on equifinality, model performance, and output and parameter uncertainty. *Hydrol. Process.* 29 (19), 4220–4237. https://doi.org/10.1002/hyp.10487.

Her, Y., Yoo, S.H., Cho, J., Hwang, S., Jeong, J., Seong, C., 2019. Uncertainty in hydrological analysis of climate change: multi-parameter vs. multi-GCM ensemble predictions. *Sci. Rep.* 9 (1), 1–22. https://doi.org/10.1038/s41598-019-41334-7.

Hill, R.A., Weber, M.H., Leibowitz, S.G., Olsen, A.R., Thornbrugh, D.J., 2016. The Stream-Catchment (StreamCat) Dataset: A database of watershed metrics for the conterminous United States. *JAWRA Journal of the American Water Resources Association* 52 (1), 120–128. https://doi.org/10.1111/1752-1688.12372.

Hinton, G.E., 2012. A practical guide to training restricted Boltzmann machines. *Neural Networks: Tricks of the Trade: Second Edition* 599–619. https://doi.org/10.1007/978-3-642-35289-8_32.

Hodgkins, G.A., Over, T.M., Dudley, R.W., Russell, A.M., LaFontaine, J.H., 2023. The consequences of neglecting reservoir storage in national-scale hydrologic models: An appraisal of key streamflow statistics. *JAWRA Journal of the American Water Resources Association*. https://doi.org/10.1111/1752-1688.13161.

Hrachowitz, M., Savenije, H.H.G., Blöschl, G., McDonnell, J.J., Sivapalan, M., Pomeroy, J.W., Cudennec, C., 2013. A decade of Predictions in Ungauged Basins (PUB)—a review. *Hydrol. Sci. J.* 58 (6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183.

Istok, J.D., Boersma, L., 1986. Effect of antecedent rainfall on runoff during low-intensity rainfall. *J. Hydrol.* 88 (3–4), 329–342. https://doi.org/10.1016/0022-1694(86)90098-3.

Jha, M., Gassman, P.W., Secchi, S., Gu, R., Arnold, J., 2004. Effect Of watershed subdivision on SWAT flow, sediment, and nutrient predictions. *JAWRA Journal of the American Water Resources Association* 40 (3), 811–825. https://doi.org/10.1111/j.1752-1688.2004.tb04460.x.

Janjić, J., Tadić, L., 2023. Fields of Application of SWAT Hydrological Model—A Review. *Earth* 4 (2), 331–344. https://doi.org/10.3390/earth4020018.

Khandelwal, A., Xu, S., Li, X., Jia, X., Stienbach, M., Duffy, C., Nieber, J., & Kumar, V. (2020). Physics Guided Machine Learning Methods for Hydrology. *arXiv preprint*. http://arxiv.org/abs/2012.02854.

Krajewski, A., Sikorska-Senoner, A.E., Hejduk, A., Hejduk, L., 2020. Variability of the initial abstraction ratio in an urban and an agroforested catchment. *Water* 12 (2), 415. https://doi.org/10.3390/w12020415.

Krajewski, A., Sikorska-Senoner, A.E., 2021. Suspended sediment routing through a small on-stream reservoir based on particle properties. *J. Soil. Sediment.* 21, 1523–1538. https://doi.org/10.1007/s11368-020-02872-0.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G., 2019. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrol. Earth Syst. Sci.* 23 (12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019.

Kuchment, L. S., & Gelfan, A. N. (2009). Assessing parameters of physically-based models for poorly gauged basins. New approaches to hydrological prediction in data sparse regions. *Wallingford: IAHS Press, IAHS Publ*, *333*, 3-10. ISSN: 0144-7815.

Li, P., Dey, S., Merwade, V.M., 2023a. Data for analyzing the effect of data splitting and covariate shift on machine learning based streamflow prediction in ungauged basins. Purdue University Research Repository.

P. Li S. Dey V.M. Merwade Codes for analyzing the effect of data splitting and covariate shift on machine learning based streamflow prediction in ungauged basins 2023 [Software]. Purdue University Research Repository 10.4231/B783-2C47.

Liu, A., & Ziebart, B. D. (2017). Robust covariate shift prediction with general losses and feature views. arXiv preprint arXiv:1712.10043. doi: 10.48550/arXiv.1712.10043.

López, V., Fernández, A., Herrera, F., 2014. On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed. Inf. Sci. 257, 1–13. https://doi.org/10.1016/j.ins.2013.09.038.

Lucas, Y., Portier, P.E., Laporte, L., Calabretto, S., He-Guelton, L., Oblé, F., Granitzer, M., 2019. Dataset shift quantification for credit card fraud detection. In: In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pp. 97–100. https://doi.org/10.1109/AIKE.2019.00024.

McCuen, R.H., 2004. Hydrologic analysis and design. *Journal of the American Water Resources Association (JASWR)* 40(3), 838, ISBN:0-13-142424-6.

McGaughey, G., Walters, W.P., Goldman, B., 2016. Understanding Covariate Shift in Model Performance. F1000Research, 5. https://doi.org/10.12688/f1000research.8317.3.

McGovern, A., Lagerquist, R., Gagne, D.J., Jergensen, G.E., Elmore, K.L., Homeyer, C.R., Smith, T., 2019. Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Am. Meteorol. Soc.* 100 (11), 2175–2199. https://doi.org/10.1175/BAMS-D-18-0195.1.

Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E., Houston, T.G., 2012. An overview of the global historical climatology network-daily database. *J. Atmos. Oceanic Tech.* 29 (7), 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1 [Accessed on 1st May 2020].

Miller, M.P., Carlisle, D.M., Wolock, D.M., Wieczorek, M., 2018. A database of natural monthly streamflow estimates from 1950 to 2015 for the conterminous United States. *JAWRA Journal of the American Water Resources Association* 54 (6), 1258–1269. https://doi.org/10.1111/1752-1688.12685.

Milly, P.C., Dunne, K.A., Vecchia, A.V., 2005. Global pattern of trends in streamflow and water availability in a changing climate. *Nature* 438 (7066), 347–350. https://doi.org/10.1038/nature04312.

Moges, E., Demissie, Y., Larsen, L., Yassin, F., 2021. Sources of hydrological model uncertainties and advances in their analysis. *Water* 13 (1), 28. https://doi.org/10.3390/w13010028.

Moorthy, K., Saberi Mohamad, M., Deris, S., 2014. A review on missing value imputation algorithms for microarray gene expression data. *Curr. Bioinform.* 9 (1), 18–22. https://doi.org/10.2174/1574893608999140109120957.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* 50 (3), 885–900. https://doi.org/10.13031/2013.23153.

Mosavi, A., Ozturk, P., Chau, K.W., 2018. Flood prediction using machine learning models: Literature review. *Water* 10 (11), 1536. https://doi.org/10.3390/w10111536.

Multi-Resolution Land Characteristics Consortium, 2022. National Land Cover Database (NLCD) 2016. Retrieved from https://www.mrlc.gov/national-land-cover-database-nlcd-2016 [Accessed on 1st June. 2022].

Nathan, R., Lowe, L., 2012. The hydrologic impacts of farm dams. *Australasian Journal of Water Resources* 16 (1), 75–83. https://doi.org/10.7158/13241583.2012.11465405.

National Inventory of Dams, 2022. NID Data Downloads. Retrieved from https://nid.sec.usace.army.mil/#/downloads [Accessed on 20th June. 2022].

Neal, R. M. (1992). Bayesian training of backpropagation networks by the hybrid Monte Carlo method. Technical Report CRG-TR-92-1. Dept. of Computer Science, University of Toronto. Retrieved from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=38b7a4d7d9646c4474c893fc53a606dee3264fec [Accessed on 1st May 2020].

Nearing, G.S., Kratzert, F., Sampson, A.K., Pelissier, C.S., Klotz, D., Frame, J.M., Gupta, H.V., 2021. What role does hydrological science play in the age of machine learning? *Water Resour. Res.* 57 (3), e2020WR028091. https://doi.org/10.1029/2020WR028091.

National Weather Service (NWS), 2017. Flooding in Kentucky. Retrieved from http://www.floodsafety.noaa.gov/states/ky-flood.shtml [Accessed on 13th June. 2023].

Pechlivanidis, I.G., Arheimer, B., 2015. Large-scale hydrological modelling by using modified PUB recommendations: the India-HYPE case. *Hydrol. Earth Syst. Sci.* 19 (11), 4559–4579. https://doi.org/10.5194/hess-19-4559-2015.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine Learning research, 12*, 2825-2830. HAL Id: hal-00650905v2.

Petty, T.R., Dhingra, P., 2018. Streamflow hydrology estimate using machine learning (SHEM). *JAWRA Journal of the American Water Resources Association* 54 (1), 55–68. https://doi.org/10.1111/1752-1688.12555.

Piotrowski, A.P., Napiorkowski, J.J., Piotrowska, A.E., 2020. Impact of deep learning-based dropout on shallow neural networks applied to stream temperature modelling. *Earth Sci. Rev.* 201, 103076. https://doi.org/10.1016/j.earscirev.2019.103076.

Prieto, C., Le Vine, N., Kavetski, D., García, E., Medina, R., 2019. Flow prediction in ungauged catchments using probabilistic random forests regionalization and new statistical adequacy tests. *Water Resour. Res.* 55 (5), 4364–4392. https://doi.org/10.1029/2018WR023254.

Prieto, C., Le Vine, N., Kavetski, D., Fenicia, F., Scheidegger, A., Vitolo, C., 2022. An exploration of Bayesian identification of dominant hydrological mechanisms in ungauged catchments. *Water Resour. Res.* 58 (3), e2021WR030705. https://doi.org/10.1029/2021WR030705.

Ramchandran, M., & Mukherjee, R. (2021). On ensembling vs merging: least squares and Random Forests under covariate shift. *arXiv preprint.* doi:10.48550/arXiv.2106.02589.

Raza, H., Prasad, G., Li, Y., 2015. EWMA model based shift-detection methods for detecting covariate shifts in non-stationary environments. *Pattern Recogn.* 48 (3), 659–669. https://doi.org/10.1016/j.patcog.2014.07.028.

Razavi, T., Coulibaly, P., 2013. Streamflow prediction in ungauged basins: review of regionalization methods. *J. Hydrol. Eng.* 18 (8), 958–975. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690.

Razavi, T., Coulibaly, P., 2017. An evaluation of regionalization and watershed classification schemes for continuous daily streamflow prediction in ungauged watersheds. *Canadian Water Resources Journal/revue Canadienne Des Ressources Hydriques* 42 (1), 2–20. https://doi.org/10.1080/07011784.2016.1184590.

Reddi, S., Poczos, B., & Smola, A. (2015, February). Doubly robust covariate shift correction. In Proceedings of the AAAI conference on artificial intelligence (Vol. 29, No. 1). Doi: 10.1609/aaai.v29i1.9576.

Reitermanova, Z. (2010). Data splitting. *In WDS'10 Proceedings of Contributed Papers*, Part I (pp. 31–36). ISBN 978-80-7378-139-2.

Rezaei, A., Liu, A., Memarrast, O., & Ziebart, B. D. (2021, May). Robust fairness under covariate shift. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 11, pp. 9419-9427). Doi: 10.1609/aaai.v35i11.17135.

Robert, C., Casella, G., 2013. Monte Carlo statistical methods. *Springer Science & Business Media.* https://doi.org/10.1007/978-1-4757-4145-2.

Saadi, M., Oudin, L., Ribstein, P., 2019. Random Forest ability in regionalizing hourly hydrological model parameters. *Water* 11 (8), 1540. https://doi.org/10.3390/w11081540.

Saksena, S., Merwade, V., Singhofen, P.J., 2019. Flood inundation modeling and mapping by integrating surface and subsurface hydrology with river hydrodynamics. *J. Hydrol.* 575, 1155–1177. https://doi.org/10.1016/j.jhydrol.2019.06.024.

Samadi, V.S., Tabas, S.S., Wilson, C.A., Hitchcock, D.R., 2024. Regression-Based Machine Learning Approaches for Daily Streamflow Modeling. *Advanced Hydroinformatics: Machine Learning and Optimization for Water Resources* 129–147. https://doi.org/10.1002/9781119639268.ch5.

Schmidt, L., Heße, F., Attinger, S., Kumar, R., 2020. Challenges in applying machine learning models for hydrological inference: A case study for flooding events across Germany. Water Resour. Res. 56 (5), e2019WR025924. https://doi.org/10.1029/2019WR025924.

Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., Bethge, M., 2020. Improving robustness against common corruptions by covariate shift adaptation. *Adv. Neural Inf. Proces. Syst.* 33, 11539–11551.

Schoppa, L., Disse, M., Bachmair, S., 2020. Evaluating the performance of random forest for large-scale flood discharge simulation. *J. Hydrol.* 590, 125531. https://doi.org/10.1016/j.jhydrol.2020.125531.

Schlef, K.E., Francois, B., Robertson, A.W., Brown, C., 2018. A general methodology for climate-informed approaches to long-term flood projection—Illustrated with the Ohio river basin. *Water Resour. Res.* 54 (11), 9321–9341. https://doi.org/10.1029/2018WR023209.

Schwarz, G.E., Alexander, R.B., 1995. *State Soil Geographic (STATSGO) Data Base for the Conterminous United States* No. 95–449. https://doi.org/10.3133/ofr95449.

J.I. Segovia-Martín S. Mazuelas A. Liu July). Double-Weighting for Covariate Shift Adaptation 2023 PMLR 30439 30457.

Seibert, J., Staudinger, M., van Meerveld, H.J., 2019. Validation and over-parameterization—experiences from hydrological modeling. *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives* 811–834. https://doi.org/10.1007/978-3-319-70766-2_33.

Singh, N.K., Basu, N.B., 2022. The human factor in seasonal streamflows across natural and managed watersheds of North America. *Nat. Sustainability* 5 (5), 397–405. https://doi.org/10.1038/s41893-022-00848-1.

Sivakumar, B., 2000. Chaos theory in hydrology: important issues and interpretations. *J. Hydrol.* 227 (1–4), 1–20. https://doi.org/10.1016/S0022-1694(99)00186-9.

Sivakumar, B., 2004. Dominant processes concept in hydrology: moving forward. *Hydrol. Process.* 18 (12), 2349–2353. https://doi.org/10.1002/hyp.5606.

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshmi, V., Zehe, E., 2003. IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrol. Sci. J.* 48 (6), 857–880. https://doi.org/10.1623/hysj.48.6.857.51421.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929-1958. ISSN: 1532-4435.

M. Sugiyama S. Nakajima H. Kashima P. Buenau M. Kawanabe Direct importance estimation with model selection and its application to covariate shift adaptation Adv. Neural Inf. Proces. Syst. 20 2007 ISBN: 9781605603520.

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 43 (6), 1947–1958. https://doi.org/10.1021/ci034160g.

Thiessen, A.H., 1911. Precipitation averages for large areas. *Mon. Weather Rev.* 39 (7), 1082–1089. https://doi.org/10.1175/1520-0493(1911)39<1082b:PAFLA>2.0.CO;2.

Thomas, T., Rajabi, E., 2021. A systematic review of machine learning-based missing value imputation techniques. *Data Technol. Appl.* 55 (4), 558.

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Wilkins-Diehr, N., 2014. XSEDE: accelerating scientific discovery. *Comput. Sci. Eng.* 16 (5), 62–74. https://doi.org/10.1109/MCSE.2014.80.

Tyralis, H., Papacharalampous, G., Langousis, A., 2019. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water* 11 (5), 910. https://doi.org/10.3390/w11050910.

Underwood, K.L., Rizzo, D.M., Hanley, J.P., Sterle, G., Harpold, A., Adler, T., Perdrial, J. N., 2023. Machine-learning reveals equifinality in drivers of stream DOC concentration at continental scales. *Water Resour. Res.* 59 (3), e2021WR030551. https://doi.org/10.1029/2021WR030551.

United States Department of Agriculture, 2022. United_States_General_Soil_Map_STATSGO2. Retrieved from https://agdatacommons.nal.usda.gov/articles/model/United_States_General_Soil_Map_STATSGO2_/24660345 [Accessed on 1st June. 2022].

United States Environmental Protection Agency, 2022. StreamCat Dataset. Retrieved from https://www.epa.gov/national-aquatic-resource-surveys/streamcat-dataset [Accessed on 1st June. 2022].

United States Geological Survey, 2022. USGS 3D Elevation Program. Retrieved from https://www.usgs.gov/3d-elevation-program [Accessed on 1st June. 2022].

United States Geological Survey, 2023. USGS current water data for the nation. Retrieved from https://waterdata.usgs.gov/nwis/rt [Accessed on 20th June. 2023].

Upreti, P., Ojha, C.S.P., 2021. Comparison of antecedent precipitation based rainfall-runoff models. *Water Supply* 21 (5), 2122–2138. https://doi.org/10.2166/ws.2020.315.

Valizadeh, N., Mirzaei, M., Allawi, M.F., Afan, H.A., Mohd, N.S., Hussain, A., El-Shafie, A., 2017. Artificial intelligence and geo-statistical models for stream-flow forecasting in ungauged stations: state of the art. *Nat. Hazards* 86 (3), 1377–1392. https://doi.org/10.1007/s11069-017-2740-7.

Wang, W., Chen, L., Lin, C., Liu, Y., Dong, X., Xiong, J., Shen, Z., 2023. Source appointment at large-scale and ungauged catchment using physically-based model and dynamic export coefficient. *J. Environ. Manage.* 326, 116842. https://doi.org/10.1016/j.jenvman.2022.116842.

Westfall, P.H., 2014. Kurtosis as peakedness, 1905–2014. RIP. *the American Statistician* 68 (3), 191–195. https://doi.org/10.1080/00031305.2014.917055.

Wickham, J.D., Stehman, S.V., Gass, L., Dewitz, J., Fry, J.A., Wade, T.G., 2013. Accuracy assessment of NLCD 2006 land cover and impervious surface. *Remote Sens. Environ.* 130, 294–304. https://doi.org/10.1016/j.rse.2012.12.001.

Winkler, M., Schellander, H., Gruber, S., 2020. May). Snow water equivalents exclusively from snow heights and their temporal Changes: the ΔSNOW. In: MODEL. *EGU General Assembly Conference Abstracts*. https://doi.org/10.5194/egusphere-egu2020-11298.

Worland, S.C., Farmer, W.H., Kiang, J.E., 2018. Improving predictions of hydrological low-flow indices in ungaged basins using machine learning. *Environ. Model. Softw.* 101, 169–182. https://doi.org/10.1016/j.envsoft.2017.12.021.

Xiang, Z., Yan, J., Demir, I., 2020. A rainfall-runoff model with LSTM-based sequence-to-sequence learning. *Water Resour. Res.* 56 (1), e2019WR025326. https://doi.org/10.1029/2019WR025326.

Yihdego, Y., Webb, J., 2013. An empirical water budget model as a tool to identify the impact of land-use change in stream flow in southeastern Australia. *Water Resour. Manag.* 27, 4941–4958. https://doi.org/10.1007/s11269-013-0449-2.

Yilmaz, M.U., Bihrat, Ö.N.Ö.Z., 2019. Evaluation of statistical methods for estimating missing daily streamflow data. *Teknik Dergi* 30 (6), 9597–9620. https://doi.org/10.18400/tekderg.421091.

Zhang, Y., Chiew, F.H., Li, M., Post, D., 2018. Predicting runoff signatures using regression and hydrological modeling approaches. *Water Resour. Res.* 54 (10), 7859–7878. https://doi.org/10.1029/2018WR023325.

Zhang, M., Li, X., Wang, L., 2019. An adaptive outlier detection and processing approach towards time series sensor data. *IEEE Access* 7, 175192–175212. https://doi.org/10.1109/ACCESS.2019.2957602.

Ziegler, A., König, I.R., 2014. Mining data with random forests: current options for real-world applications. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery* 4 (1), 55–63. https://doi.org/10.1002/widm.1114.