geographies

MDPI

Article

Application of Machine Learning Models for Improving Discharge Prediction in Ungauged Watershed: A Case Study in East DuPage, Illinois

Amin Asadollahi , Binod Ale Magar, Bishal Poudel , Asyeh Sohrabifar and Ajay Kalra *

Department of Civil, Environmental and Infrastructure Engineering, Southern Illinois University, Carbondale, IL 62901, USA; amin.asadollahi@siu.edu (A.A.); binod.alemagar@siu.edu (B.A.M.); bishal.poudel@siu.edu (B.P.); asyeh.sohrabifar@siu.edu (A.S.)
* Correspondence: kalraa@siu.edu; Tel.: +1-618-453-7008

Abstract: Accurate flood prediction models and effective flood preparedness rely on thoroughly understanding rainfall-runoff dynamics. Similarly, effective rainfall–runoff models account for multiple interrelated parameters for robust runoff prediction. Process-based physical models offer valuable insights into hydrological processes, but their effectiveness can be hindered by data limitations or difficulties in acquiring specific data. Motivated by the frequent flooding events and limited data availability in the East Branch DuPage watershed, Illinois, this study addresses a critical gap in research by investigating effective discharge prediction methods. In this study, two significant machine learning (ML) models, artificial neural network (ANN) and support vector machine (SVM), were employed for discharge prediction. Historical data spanning from 2006 to 2021 were utilized to assess the performance of the models. Hyperparameter tuning was performed on the models to optimize their performance, and root mean square error (RMSE), Nash-Sutcliffe efficiency (NSE), percent bias (PBIAS), coefficient of determination (R2), and the normalized root mean squared error (NRMSE) were used as evaluation metrics. Although both machine learning models demonstrated strong performance, the analysis revealed that the ANN model emerged as the more reliable option for predicting discharge in the watershed. Crucially, the ANN model surpassed the SVM model's performance, achieving superior accuracy in predicting peak discharge events within the study area. Our findings have the potential to assist decision-makers and communities in implementing more dependable flood mitigation strategies, particularly in regions where hydrology data are limited.

Keywords: discharge; SVM; ANN; forecast; machine learning

check for updates

Citation: Asadollahi, A.; Magar, B.A.; Poudel, B.; Sohrabifar, A.; Kalra, A. Application of Machine Learning Models for Improving Discharge Prediction in Ungauged Watershed: A Case Study in East DuPage, Illinois. *Geographies* **2024**, *4*, 363–377. https://doi.org/10.3390/geographies4020021

Academic Editor: Xu Chen

Received: 10 May 2024
Revised: 25 May 2024
Accepted: 4 June 2024
Published: 6 June 2024

1. Introduction

Flooding is a natural phenomenon that can have devastating effects on communities and ecosystems, making it a significant concern for disaster preparedness and management [1]. It can cause significant harm to both the environment and human life, resulting in damage to property and infrastructure. It can occur gradually or suddenly, leading to flash floods [2]. Various factors such as global warming, changes in land use and land cover, and urbanization can exacerbate the impact and frequency of flooding events [2]. According to estimates, the mean annual probability of high property damage in certain regions of the US could vary from 38% to 80% under the RCP4.5 scenario and from 46% to 95% under the RCP8.5 scenario by the end of the century (2090s) [3]. The increasing severity of flooding highlights the need for reliable and effective prediction and mitigation methods [4]. Therefore, it is crucial to develop accurate methods for predicting and managing flooding, particularly in urban areas.

An important aspect of understanding and managing floods is capturing the dynamics of runoff, which is one of the primary contributors to flooding events [5]. Accurate flood risk assessment relies on precise peak runoff estimation, which is determined through rainfall–runoff simulation [2]. Predicting accurate discharge is a crucial factor in flood control and reducing damage to the environment and infrastructure [6]. Over time, researchers have explored various methods for simulating runoff in watersheds. These include empirical models, such as machine learning algorithms (e.g., ANN and SVM) that utilize historical data for predictions and physical-based models, specifically distributed hydrological models, which also possess conceptual characteristics and are used to predict runoff [3]. By leveraging these diverse approaches, researchers aim to improve our understanding of hydrological processes and enhance flood prediction and management capabilities.

Over the years, traditional hydrological models such as the Hydrological Engineering Centre Hydrological Modeling System (HEC-HMS) have played a crucial role in modeling the interactions between precipitation, land surface, and river systems to calculate the runoff of the catchment area [7]. Physically based hydrological models have greatly contributed to the calculation of complex hydrological processes, providing valuable understanding for flood risk assessment and hydrological management [7]. However, their application faces several hurdles. These include the need for diverse, high-quality datasets encompassing various hydrological parameters and often require significant computational resources and expertise in interpreting and selecting appropriate hydrological parameters, especially in short-term predictions [8]. On the other hand, empirical models are used as an approach for predicting hydrological events, particularly in flood modeling. These models leverage historical data analysis to establish statistical relationships between observed variables, offering a valuable tool for flood risk assessment and mitigation strategies [9]. While empirical models offer several advantages, including the ability to handle and learn from large datasets, flexibility in adapting to new data, and reduced reliance on in-depth knowledge of the physical system, they also come with limitations. These limitations can include potential accuracy issues, a steeper learning curve for users compared to traditional methods, and higher computational demands [4].

In hydrology research, ML has gained significant attention because of its ability to facilitate accurate predictions by training and testing datasets. ML has been developed over the past few years to better understand real-world problems by using current and observed data and experiences to generate accurate predictive outputs [10]. These approaches are advantageous, particularly in scenarios where process-based models may be constrained by high modeling costs, data scarcity, or the need for supplementary analytical capabilities to interpret complex datasets [11]. Several water and hydrology studies have already used ML for research applications such as sediment transport, rainfall-runoff simulation, water distribution networks, water quality analysis, and flood inundation mapping [12–14]. The use of machine learning models like ANN, SVM, and Random Forests (RFs) has increased in hydrology and water resource modeling. Many researchers use these models to forecast water levels, streamflow, and rainfall–runoff in time series problems [15–17].

ANNs are artificial intelligence models that mimic the information-processing capabilities of biological neural networks. These networks are trained using algorithms to recognize patterns and relationships in data, making them useful for various tasks [15]. Bekele and Nicklow [18] explore the application of ANNs, trained using a hybrid evolutionary approach, to simulate various hydrological responses such as flow, sediment, and nutrient dynamics. Ghorbani et al. [19] compared ANNs and SVMs to model the discharge of the Big Cypress River in Texas, USA, utilizing time series data. They found that both ANN and SVM models were more reliable than conventional models for modeling the river discharge. Tamiru et al. [20] employed an ANN model for flood inundation mapping in the lower Baro Akobo River Basin, Ethiopia. Their study revealed that the ANN model exhibited good performance during both training and testing periods. Another study conducted in two reservoirs in Illinois, USA, aimed to compare the performance of various machine learning models in predicting reservoir outflow. The research demonstrated that ML models are an effective approach for reservoir management. Additionally, the study found that the ANN model outperformed RF and SVM models in predicting reservoir outflow [21]. Riad et al. [22] modelled the rainfall–runoff relationship in a semiarid area in

Morocco using an ANN. They found that this method predicted more reliable data than classical regression models. Comparing ANNs with other ML models, especially SVMs, provides valuable insights into their relative performance and informs the selection of the most effective modeling technique for a particular problem domain.

SVMs offer a flexible and powerful approach to modeling hydrological processes and can complement other techniques such as empirical models, distributed hydrological models, and conceptual models in watershed management studies [23]. They utilized the SVM method to forecast streamflow in the Western United States, and they found the model to be highly accurate. Asefa et al. [6] utilized SVM models for multi-time scale streamflow predictions, which yielded promising results. By leveraging local climatological data, the SVM model effectively forecasted streamflow with minimal input requirements compared to traditional physical models. In a study focused on predicting the discharge of the Mahanadi River in India, researchers applied SVM models and found them to outperform the traditional Box-Jenkins approach [24]. Lin et al. [25] conducted a study where SVM was employed to analyze long-term observations of monthly river flow discharges in a watershed in China. This highlights the SVM's effectiveness as a viable option for long-term discharge prediction in hydrological studies. Guo et al. [26] endeavoured to enhance the performance of the SVM model for predicting monthly streamflow by eliminating noise from runoff time series. The study aimed to verify whether the refined SVM model could effectively handle complex hydrological data series. Finding the best ML method for each basin based on available data enables tailored and optimized flood forecasting models, leading to more accurate and reliable predictions.

While ML offers a promising path for discharge prediction, a key challenge lies in identifying the most effective model for specific watersheds. The novelty of this study is to evaluate the efficacy of ANN and SVM models for rainfall-runoff simulation in a region with limited hydrological data availability. These models can predict runoff from rainfall by identifying complex patterns and correlations within limited and diverse datasets, without relying on detailed physical processes like antecedent soil moisture and infiltration rate [10]. This study also leveraged global precipitation measurement (GPM) data from Climate Engine since the East Branch DuPage watershed lacks precipitation gauging stations. The subsequent sections of this paper delve into the following aspects. The Materials and Methods section encompasses the study area description, data preprocessing techniques, and performance evaluation methods employed. The Results section then presents the findings obtained from the model application. A comprehensive discussion follows, comparing these results with existing research in this field. Finally, the concluding section summarizes the key takeaways and potential implications of this study.

2. Materials and Methods

2.1. Study Area

The East Branch DUPAGE watershed is located on the northern side of Illinois, USA. The region's watershed area is approximately 62.2 square kilometres and is located at an elevation of 204 to 250 m above sea level (Figure 1). The study area extends from latitude 41°50′ N to 41°57′ N and longitude 87°59′ W to 88°6′ W. The average soil permeability of the study area is 6.24 cm/h [27]. The outlet of the watershed consists of USGS gauge station 05540160 downstream of the study area (https://earthexplorer.usgs.gov/, accessed on 25 October 2023). In 1996, 2008, 2013, and 2020, the area experienced major flooding events [3]. This study area does not possess any rainfall gauging stations; therefore, for the current research purpose, gridded precipitation data were used.

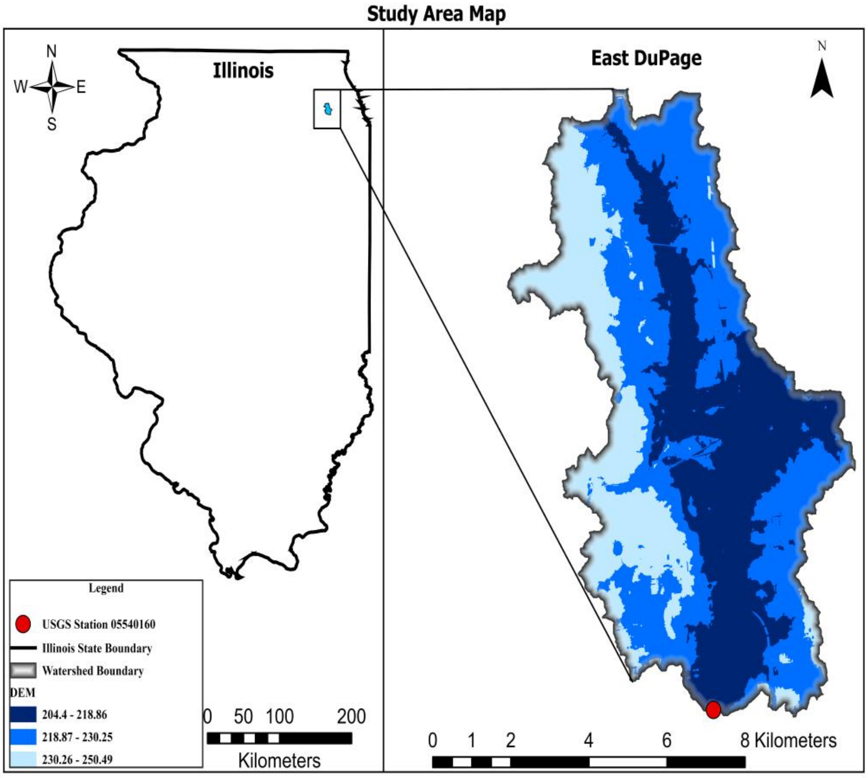


Figure 1. The East Branch DuPage watershed, Illinois.

2.2. Precipitation Data

Precipitation is an important dataset used in the current study. As the study watershed does not have any gauging stations to measure the precipitation data, global precipitation measurements (GPMs) from Climate Engine were used. The GPM is a next-generation system that measures snow and rain using satellite technology. The high spatial resolution of the GPM method, typically 0.1° by 0.1°, enables detailed and precise monitoring of precipitation patterns and distribution worldwide [28]. Gridded precipitation data for 2006–2021 were converted into daily time series data using Python.

2.3. Data

Selecting significant input variables is crucial in developing time series forecasting models as it enhances model performance by eliminating irrelevant and redundant variables that introduce noise and diminish accuracy and speed [29]. This study employed a brute force feature selection method, similar to previous studies, to identify the most significant input variables from the available meteorological data for the region [3,30]. This study uses gauge height (https://earthexplorer.usgs.gov/, accessed on 25 October 2023), meteorological model data such as precipitation (https://www.climateengine.org/, accessed on 26 October 2023), and climatic data such as humidity, temperature, and evapotranspiration for machine learning simulation (https://power.larc.nasa.gov/, accessed on 26 October 2023) inputs. In addition, as the seasonal variations play a vital role in rainfall–runoff dynamics, seasonality is also incorporated as a feature. This study utilized a 15-year data series from 2006 to 2021 obtained from measurement gauges to investigate rainfall–runoff dynamics within the East Branch DuPage watershed. This timeframe was chosen considering the recommended minimum of a decade of data for reliable hydrological predictions [10]. Additionally, the period encompasses a history of significant flooding events in the region potentially offering valuable data points for model training. While the analysis suggests a possible biennial pattern of high rainfall–runoff events, the data also reveal inconsistencies [3]. Years with substantial precipitation and runoff are interspersed with periods of drought characterized by minimal rainfall and low streamflow. This variability in the data series presents a potential challenge for machine learning models [25]. Models primarily trained on data rich in high-flow events might struggle to accurately predict discharge during low-flow periods. However, the extensive 15-year daily data allow the models to benefit from a large and diverse dataset, potentially improving their ability to learn complex hydrological behavior and ultimately predict flood events [25]. These data were used to calculate and predict the hydrology and hydraulic parameters for the ANN and SVM models. Based on previous research, the data were split into training and testing sets [31,32]. In the study period from 2006 to 2021, 80% of the entire daily stage discharge dataset was randomly selected for training purposes. The other 20% of the dataset was reserved specifically for evaluating the models' performance.

2.4. Pre-Processing Data

Choosing the best inputs is crucial in developing time series models. This step enhances model performance by eliminating irrelevant and redundant variables that introduce noise, thereby improving accuracy and computational efficiency [31,33]. Correlated input variables can obscure the true relationships between important variables, which can negatively impact the prediction ability of the model [34]. For this purpose, the identification and management of outliers play a crucial role in ensuring the integrity and accuracy of statistical inferences [35,36]. Outliers, defined as data points significantly deviating from the average, can distort the results of analyses and compromise the robustness of models [37]. Addressing outliers requires careful consideration and a systematic approach. Several strategies exist for replacing outliers, each catering to different data characteristics and the underlying reasons behind the outlier presence [36,37]. The methodical approach utilized in this study involved identifying outliers based on a criterion related to their deviation from the mean. Specifically, values that exceeded three standard deviations were flagged for correction [38]. Subsequently, a strategy of imputation through interpolation was employed to replace these outlier values with contextually inferred data points, preserving the overall trend of the dataset [38]. This approach seeks to mitigate the impact of extreme observations while maintaining the continuity and coherence of the dataset. The resulting corrected dataset, represented through a box plot, reflects the application of a judicious outlier-handling technique [37]. Such systematic outlier correction processes are essential in research and analysis, contributing to the refinement of datasets for more accurate and reliable statistical interpretations [38]. This study employs the corrected data derived through this method. Figure 2a,b present a box plot depicting the flood events of the DuPage River dataset before and after the removal of outliers, respectively.

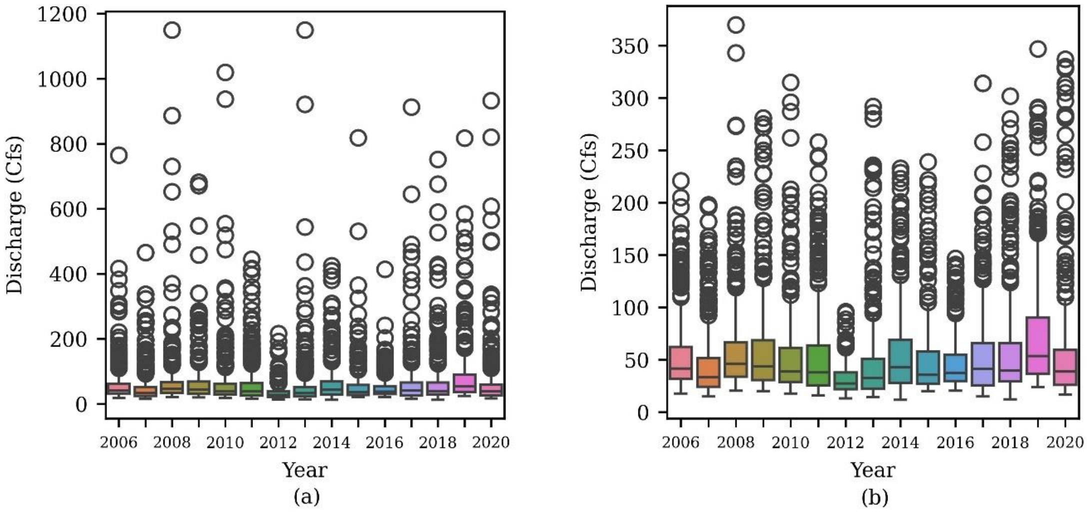


Figure 2. Discharge box plot. (**a**) Before removing outlier. (**b**) After removing outlier.

2.5. ANN Model

The development of the ANN algorithm and its interconnected nodes is influenced by the intricate neural network present in the human brain and body [39,40]. ANNs with similar network structures and interconnected neuron units enable efficient modeling of mathematical systems. The accuracy of ANN algorithms in modeling flood prediction has been proven [41–43]. ANNs are recognized as effective machine learning models for capturing complex relationships between variables, particularly in the context of predicting rainfall, flood occurrences, and discharge levels [44,45]. ANNs are also noted for their computational efficiency compared to other models, allowing for faster processing of data and predictions [46]. Previous researchers have also utilized this model for predicting streamflow, rainfall, and runoff, demonstrating its versatility and effectiveness across various hydrological applications [47–49]. This study employed a feedforward neural network for regression tasks, leveraging the TensorFlow and Keras libraries. This architecture is commonly referred to as a "Multilayer Perceptron" (MLP), which serves as a fundamental model in artificial neural networks [50]. The MLP architecture comprises input layers, hidden layers, and an output layer. Based on the selected input parameters (gauge height, precipitation, temperatures, humidity, and evapotranspiration) and the output (discharge), a neural network configuration was chosen with 5 input nodes, 3 hidden layers, and 1 output node. The selection of the number of hidden nodes was guided by insights from the research conducted by Lv et al. [30]. Three hidden layers with 100, 50, and 20 neurons, respectively, are added with the ReLU activation function. The neural network architecture includes an output layer with a neuron optimized for regression analysis. The model is constructed utilizing the RMSprop optimizer with a rate of learning 0.001 and employs the mean squared error loss function [30]. To prevent overfitting, the training procedure includes the EarlyStopping callback, which evaluates the validation error and stops training if no enhancement is observed for 20 consecutive epochs. The training dataset was used to train and validate the model on a separate validation dataset, with a maximum of 5000 epochs specified for the training process. Equation (1) was used to prepare the inputs and normalized to a range between 0 and 1 [51]. Random initial weights were established for the connections between input, hidden, and output nodes within the ANN networks [52]. Commencing the modeling process requires the meticulous collection of essential input parameters, documented daily and spanning the temporal domain from 2006 to 2021.

$$X\_normalized = \frac{x - min\_val}{max\_val - min\_val} \tag{1}$$

2.6. SVM Model

Established on the Structural Risk Minimization (SRM) principle, the SVM uses a training set of sample objects to find a hyperplane in the data space with the largest minimum distance between the sample objects [26]. This hyperplane separates the two different classes of sample objects with the largest minimum distance [53]. The sample objects on the edges of the hyperplane, called support vectors, are used to separate the objects into different classes. The object samples (support vectors) provide the basis for the hyperplane, and thus the algorithm is named support vector machines [44]. One of the supervised learning methods, SVR predicts the future data based on the underlying dependency of the known observations (training samples) [53].

Various kernel functions (linear kernel, polynomial kernel, RBF kernel, sigmoid kernel) in the SVM operate in their own way depending on kernel parameters. These parameters define the regression model complexity, and the kernel functions give the option of input space dimensions. Noise in input data is also significantly supressed by parameter settings. Here are some popular kernel functions [24].

1. Linear kernel: $k(x_i, x_j) = x_i^T x_j$;
2. Polynomial kernel: $k(x_i, x_j) = \left(\gamma x_i^T x_j + r\right)^d, \gamma > 0$;
3. RBF kernel: $k(x_i, x_j) = \exp\left(-\gamma \parallel x_i - x_j \parallel\right)^2, \gamma > 0$;

4. Sigmoid kernel: $k(x_i, x_j) = \tanh\prime(-\gamma x_i^T x_j + r)\gamma > 0$.

C, γ, r, and d are kernel parameters. These parameters define the regression model complexity, and the kernel functions give us the option of input space dimensions.

The linear kernel function serves as the foundation of the SVM method, primarily because the target variable (discharge) exhibits a linear relationship with the features. The dataset used comprises discharge (Q), gauge height (f_{gh}), precipitation (f_p), humidity (f_h), temperature (f_t), and evapotranspiration (f_e) of the years 2006 to 2020. Data preprocessing steps included handling missing values, removing outliers, and scaling features. The resources for the data were derived from creditable sources, which are mentioned above in the materials section. Based on the parameters that have influences on the discharge, Equation (2) is modeled. To ensure consistent and effective model training, the features were standardized using the Standard Scaler.

$$Q = f\left(f_{gh}, f_p, f_h, f_t, f_e\right), \tag{2}$$

where f_{gh}, f_p, f_h, f_t, f_e are the input parameters and f_d is the output. The SVR model with a linear kernel was initialized. In this study, we employed the grid search method to estimate the hyperparameters in the SVM model, as performed in previous research [23]. The hyperparameters C (regularization parameter) and epsilon (insensitivity parameter) were set based on a preliminary hyperparameter tuning process, Grid Search, where the values [C = 100 and epsilon = 1.0] were found to yield optimal performance. Figure 3a,b illustrate the architecture of both the ANN and SVM models, respectively.

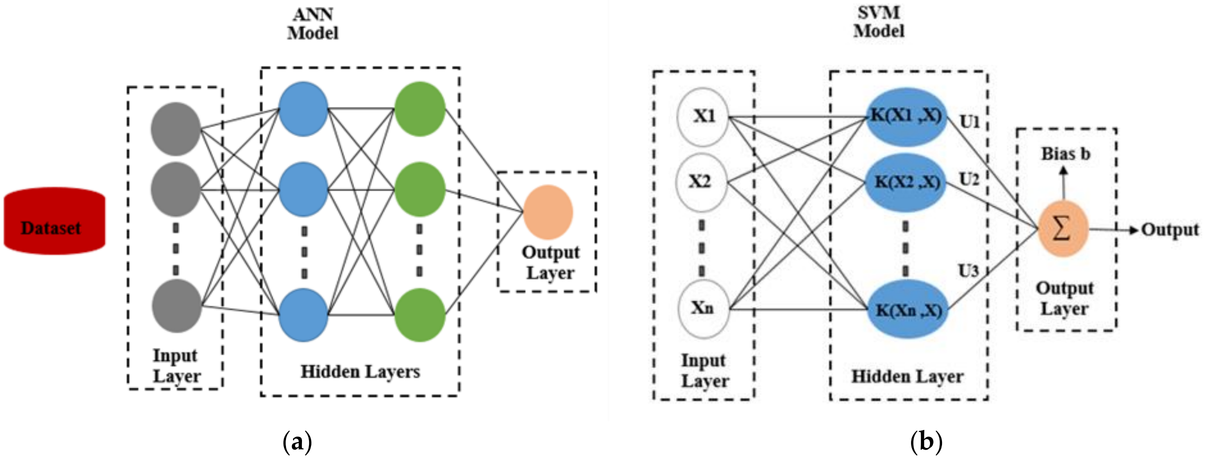


Figure 3. The architecture of (**a**) ANN model and (**b**) SVM model.

2.7. Performance Evaluation

The evaluation of the models was conducted using various evaluation methods. These metrics provide a detailed evaluation of the models and can help determine which models perform better in terms of accuracy, reliability, and predictive power. In this study, the five evaluation metrics root mean square error (RMSE), Nash-Sutcliffe efficiency (NSE), percent bias (PBIAS), coefficient of determination (R^2), and the normalized root mean squared error (NRMSE) were utilized to evaluate the performance of the model [3,54]. Table 1 displays all the evaluation methods utilized. $Q_{o,i}$ = observed data, $Q_{s,i}$ = simulated data, $\overline{Q}_{o,i}$ = mean value of real data, and N = total data.

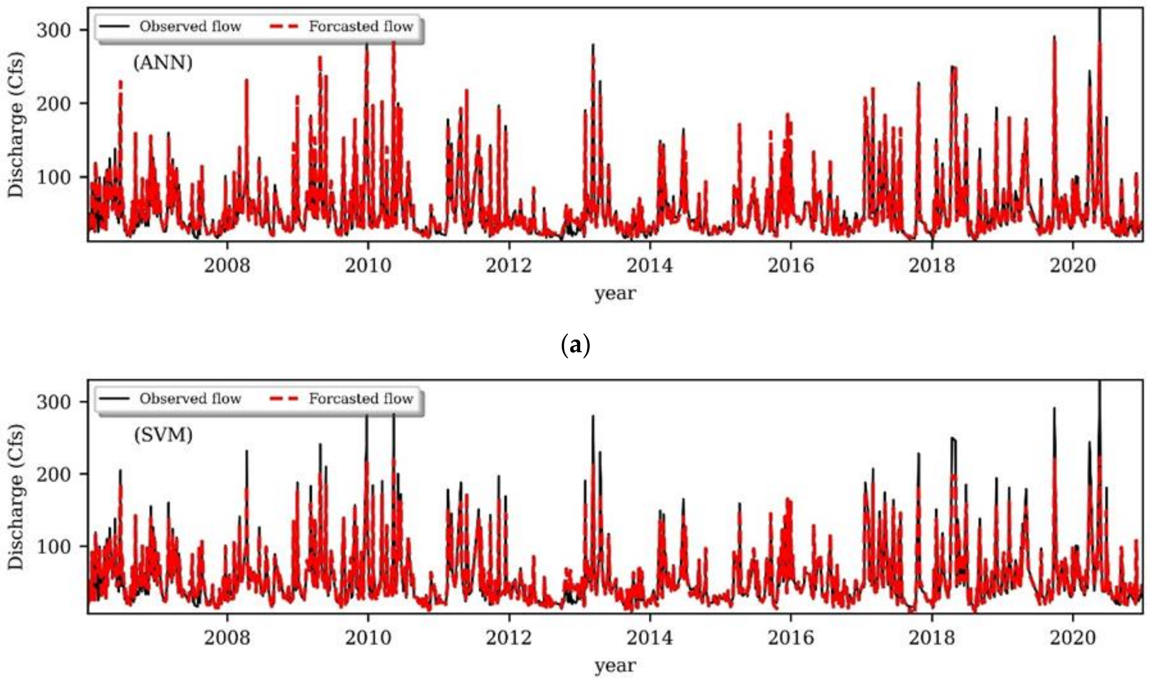
Table 1. List of evaluation methods for assessing the performance of models.

Indices	Mathematical Expression	Satisfactory Range
RMSE	$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(Q_{s,i} - Q_{o,i}\right)^2}{N}}$	
NSE	$\text{NSE} = 1 - \left[\frac{\sum_{i=1}^{N}\left(Q_{O,i} - Q_{S,i}\right)^2}{\sum_{i=1}^{N}\left(Q_{o,i} - \overline{Q}_{o}\right)^2}\right]$	$0.5 < \text{NSE} \leq 1$
R^2	$R^2 = \frac{\left(\sum_{i=1}^{N}\left(Q_{o,i} - \overline{Q}_{o,i}\right) * \left(Q_{s,i} - \overline{Q}_{o,i}\right)\right)^2}{\sum_{i=1}^{N}\left(Q_{a,i} - \overline{Q}_{o,i}\right)^2 * \sum_{i=1}^{N}\left(Q_{s,i} - \overline{Q}_{o,i}\right)^2}$	>0.5
PBIAS	$\text{PBIAS} = \frac{\sum_{i=1}^{N}(Q_{o,i} - Q_{s,i})}{\sum_{i=1}^{N} Q_{o,i}} * 100$	$-25\% < \text{PBIAS} < +25\%$
NRMSE	$\text{NRMSE} = \frac{\frac{1}{N}\sum_{i=1}^{N}(Q_{s,i} - Q_{o,i})^2}{Mean}$	$0 \leq$

3. Results and Discussion

This study focuses on evaluating the performance of runoff modeling using ANN and SVM models in the East Branch DUPAGE watershed, Illinois, USA. As mentioned earlier, the five key inputs selected for this study include gauge height, precipitation, humidity, temperature, and evapotranspiration for both ML models, based on insights drawn from previous research [3,21]. These variables include past and current information from 2006 to 2021. The irrelevant and redundant data, which reduce the accuracy and speed of the model, were not included in the model. The precipitation data utilized in this study were derived from satellite-based rainfall products spanning from 2006 to 2021, similar to the approach employed in Bhusal et al.'s research [3]. They showed that the model matches the observed daily discharge data and can be considered completely reliable. Figure 4a,b display the observed runoff plotted against the calculated runoff of the watershed for testing the ANN and SVM models, respectively. For both models, the data indicate a strong correlation between predicted values and the corresponding daily discharge, demonstrating a good match across the board. The evaluation metrics for the ANN model are as follows: PBIAS = 0.15%, RMSE (Cfs) = 7.01, NSE = 0.97, and NRMSE = 0.91. In comparison, the evaluation metrics for the SVM model are PBIAS = −0.39%, RMSE (Cfs) = 10.41, NSE = 0.94, and NRMSE = 0.0328. These results confirm that the ANN model outperforms the SVM model in predicting runoff for the watershed, consistent with findings from previous studies [19,21]. Additionally, Bafitlhile et al. [31] compared the performance of these two models for flood forecasting in humid, semi-humid, and semi-arid basins in China. They found that SVM generally outperformed ANN in streamflow simulation. This difference in results could be attributed to the shorter time-series data and using only 60% of the data for training the model.

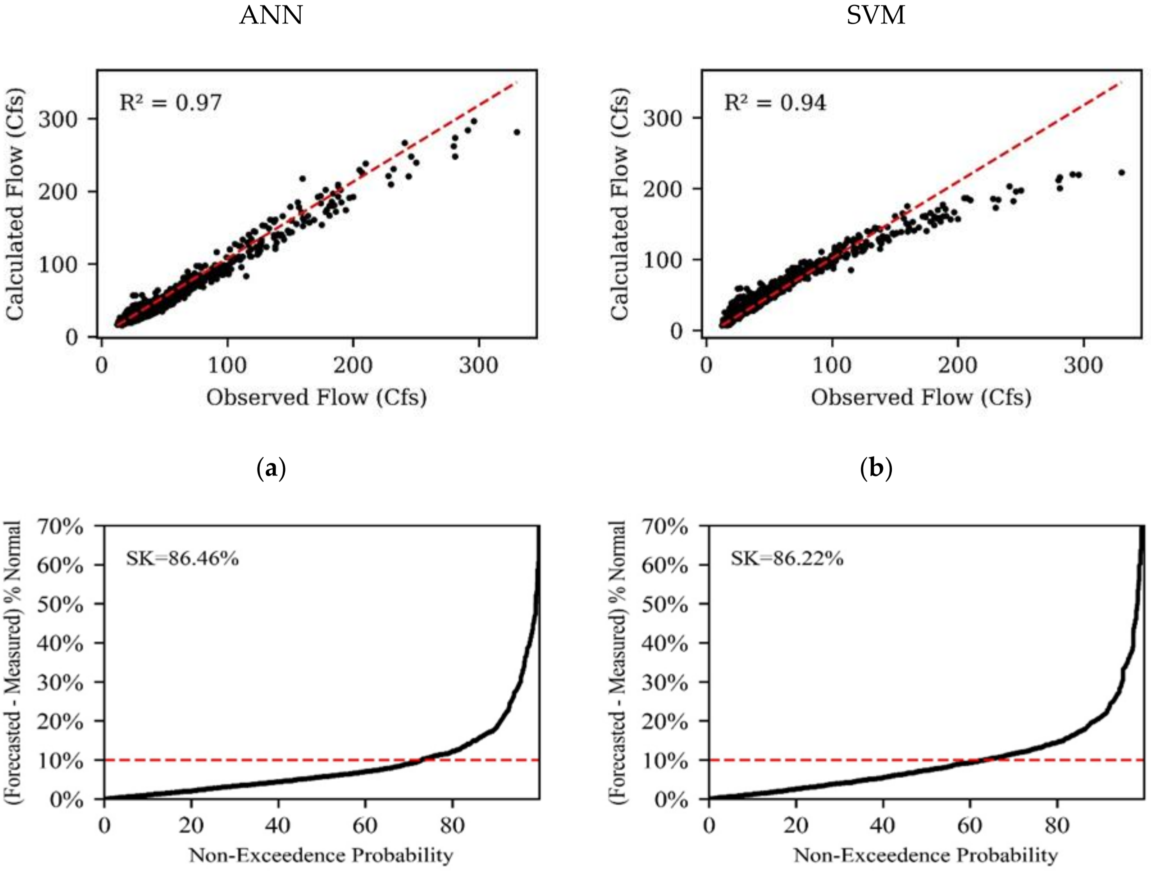
The scatter plots for both ANN and SVM models are depicted in Figure 5a,b, respectively, showcasing the goodness of fit and performance. The coefficient of determination (R^2) remains unchanged under linear transformations of the independent variables' distribution, making it a robust metric in regression analysis, and provides more informative insights compared to other metrics [55]. In this context, R^2 ratings were calculated for both models in training and testing data, and it indicates excellent performance, with R^2 values 0.94 and 0.97 for testing SVM and ANN results, respectively. In a study by Bhusal [3], they employed RFs in the watershed and reported an R^2 value of 0.72 for their model. Additionally, they utilized the HEC-HMS model, known as one of the top-performing physical methods for discharge prediction, yielding an R^2 value of 0.99. This observation underscores the ANN model's superior performance compared to the SVM and RF models in predicting runoff data, aligning closely with the results obtained from HEC-HMS.



(**a**)

(**b**)

Figure 4. Observed and calculated runoff time series during testing set for (**a**) ANN and (**b**) SVM models.



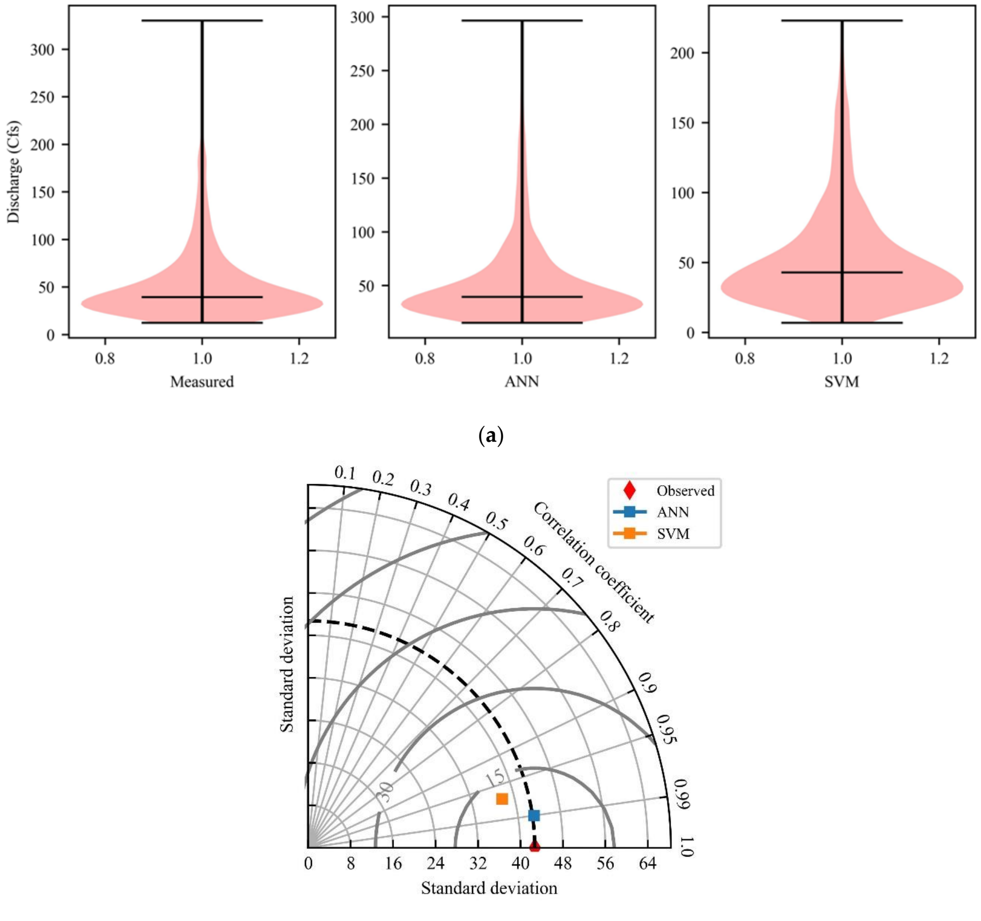
(**a**) (**b**)

(**c**) (**d**)

Figure 5. The East Branch DUPAGE watershed runoff′s scatter plots for (**a**) ANN model and (**b**) SVM model and non-exceedance probability plots for (**c**) ANN and (**d**) SVM models.

This study evaluated the accuracy of SVM and ANN models for streamflow forecasting in the basin using the cumulative non-exceedance probability error, specifically calculated as the LEPS SK score. A perfect forecast would result in an SK score of 100%, while a random forecast would have an SK score of 0. Higher SK scores indicate better model performance [56]. Figure 5c,d illustrates the probabilistic cumulative error between measured and predicted flow rates. Both models achieved high SK scores exceeding 80%, suggesting good overall performance. However, the ANN model exhibited a slight advantage, with scores consistently exceeding those of the SVM model. Furthermore, the ANN model results achieved a higher proportion of predictions with low errors (around 10%). This means a potentially more reliable forecasting tool for water managers in the basin.

Figure 6a shows the violin plot for the distribution of measured and forecasted discharge values obtained with SVM and ANN models in the watershed. The wider, violin-shaped areas depict the probability density of the discharge values [57]. The SVM model exhibits a broader range of predicted discharge values, whereas the ANN model demonstrates a distribution more closely aligned with the measured values. The Taylor diagram (Figure 6b) [58] summarizes the performance of two machine learning models, ANN and SVM, in predicting streamflow discharge. Ideal model performance is represented by a point at the centre of the diagram, where the correlation coefficient is 1 and the standard deviation matches the observed data [58]. The plot depicts the output of both models for a randomly selected 20% of the entire dataset as discharge. Both models exhibit a high correlation coefficient. However, the ANN model notably surpasses the SVM model, indicating a superior alignment between predicted and observed discharge values. Overall, these plots suggest that the ANN model provides more precise discharge forecasts for the basin.



(**a**)

(**b**)

Figure 6. The observed and forecasted flowrate values: (**a**) violin plots; (**b**) Taylor plot.

Table 2 presents the results of the evaluation of the ANN and SVM models using various statistical indices. The RMSE is commonly employed to assess the accuracy of predictions [59]. During the ANN training phase, the RMSE was found to be 9.56 Cfs, while during the testing phase, it was 7.01 Cfs. These values indicate a satisfactory level of accuracy in the model's predictions. The NSE is a widely used metric for evaluating hydrological models with values between 0.5 and 1, typically considered optimal. These values indicate a good agreement between observed and simulated values, suggesting that the model performs well in replicating the observed data [54]. For the ANN model, the NSE was found to be 0.95 during training and 0.97 during testing. These values are close to 1, indicating a high level of agreement between the predictions and the real data. The PBIAS indicates the typical trend of the predicted data. In a model, PBIAS values should be zero or fall within the range of ±25% [60]. The ANN model exhibited a slight underestimation of peak discharge, with underestimation percentages of 1.34% during training and 0.15% during testing. NRMSE serves as a robust standardization technique to assess the relationship between RMSE and a baseline data range. It normalizes RMSEs across various magnitudes in time series, thereby generating a standardized value for comparison [61]. According to Table 2, the NRMSE values for training and testing results in the ANN model are 1.68 and 0.91, respectively. These values show that the model's predictions are relatively accurate, especially during the testing phase when the NRMSE value is lower than 1. Furthermore, the statistical index analyses consistently indicated that the SVM performed well in predicting daily discharge data. This clustering indicates the model's ability to accurately capture and predict discharge under these specific flow conditions. Nonetheless, it is worth mentioning that the SVM model, like the ANN model, exhibited an underestimation of high discharge values. These underestimations are particularly notable during extreme events. During the training period, the RMSE, NSE, PBIAS, R^2, and NRMSE values were 13.15 m^3/s, 0.92, −1.49%, 0.91, and 0.036, respectively. For the testing period, the corresponding values were 10.41 m^3/s, 0.94, −0.39%, 0.94, and 0.033, respectively. It was noted that the SVM model exhibited more noticeable deviations, especially for higher discharge values. This suggests that the SVM model's effectiveness in estimating peak discharge was relatively lower compared to its performance with other discharge values.

Table 2. Comparison between ANN and SVM models for runoff estimation.

Statistical Index	ANN Model		SVM Model	
	Training	Testing	Training	Testing
RMSE (Cfs)	9.56	7.01	13.15	10.41
NSE	0.95	0.97	0.92	0.94
PBIAS (%)	1.34	0.15	−1.49	−0.39
R^2	0.96	0.97	0.92	0.94
NRMSE	1.68	0.91	0.036	0.0328

This study shows that both the ANN and SVM models demonstrated accurate recreation of discharge characteristics, including flood peaks and time, throughout the study area. The evaluation indices, encompassing both training and testing outcomes of the models, reinforce the successful utilization of machine learning models in predicting discharge at the watershed outlet. ML models, particularly ANN and SVM, can be highly dependable for predicting runoff, especially in areas where limited data are available. These findings are consistent with a prior study that similarly highlighted the effectiveness of ANN and SVM models as prediction methods in hydrology projects [3,31,40]. ANNs are particularly adept at capturing non-linear relationships between various hydrological factors influencing discharge due to their multi-layered structure. Unlike the SVM model, which often relies on a linear kernel for classification or regression, ANNs have a flexible architecture that includes activation functions. These activation functions enable the model to learn and represent complex, non-linear interactions within the data, enhancing their predictive capabilities [10]. This study suggests that these models exhibit satisfactory performance in predicting discharge during non-flooding periods but tend to underestimate discharge when flooding events are substantial. Tis indicates that both models may have limitations in accurately predicting discharge during extreme flood conditions. However, considering the effective performance of the machine learning models in generating discharge data under non-flooding conditions, integrating these models with HEC-RAS can be valuable for water resource planning and flood control in the study area. To enhance reliability, it is essential to evaluate dependent factors such as precision, robustness, and applicability to different site conditions. It is recommended to test and analyze the proposed approaches across various locations to ensure their effectiveness in different contexts. Overall, the ANN model outperformed the predictions made by the SVM model. The comprehensive findings of this research work strongly support the effectiveness of machine learning models in accurately predicting rainfall–runoff and floods in regions where data availability is limited. These models have demonstrated their potential to overcome data scarcity challenges and provide valuable insights for effective water resource management and flood mitigation strategies in such areas.

4. Conclusions

This study aimed to compare the accuracy of ANN and SVM methods in predicting the discharge of a watershed that is prone to significant flooding events. To achieve the most precise runoff prediction, the models utilized a combination of daily gauge height, precipitation, humidity, temperature, and evapotranspiration inputs. The study used a long time series of data from 2006 to 2021; 80% of the data were used to train the models, while the remaining 20% were reserved for testing purposes. The results revealed that both the ANN and SVM models were highly effective in estimating daily discharge at watershed outlets. However, it is worth noting that the accuracy of the ANN model surpassed that of the SVM model, particularly during extreme flood conditions. This study demonstrated that ML models, unlike the HEC-HMS model, do not require a large number of input variables for flood prediction. This is particularly beneficial in data-scarce regions where comprehensive input data may not always be available. This dataset served as the primary source of information for analyzing and simulating the relationship between precipitation and runoff in the study area. According to the reasonably strong performance of the models, it can be concluded that the precipitation, temperature, humidity, and gauge height data used in this study are reliable for discharge prediction. The data utilized in this study have demonstrated their dependability and accuracy, making them suitable for hydrology investigations. In both the ANN and SVM models, the peak flows were found to be underestimated. However, when an integrated approach combining physical-based models and ML models was employed, it yielded promising results in accurately predicting the runoff flood depth downstream of the watershed. To enhance the accuracy models in predicting discharge, it is recommended to remove outliers from the dataset.

This study demonstrates that ML methods, particularly the ANN, offer a reliable and cost-effective approach for discharge prediction in regions with limited data availability, with high accuracy. Also, ML models, when trained on remotely sensed data, can significantly improve the accuracy of rainfall–runoff simulations compared to traditional methods.

In future research, there are several potential areas that researchers could explore. Firstly, to enhance the accuracy of precipitation data, incorporating precipitation stations alongside the PERSIANN precipitation product utilized in this study would be beneficial. Additionally, exploring the use of other precipitation products and considering additional input variables in machine learning models, such as curve number, infiltration, land use/land cover, and radiation, may contribute to improved predictions. Conducting feature selection to identify the most influential input variables would also make more accurate models. Several considerations must be addressed when applying the ANN model to other hydrological contexts with similar data limitations, including environmental parameters like anthropogenic land cover changes and climate conditions. Relevant feature selection, hyperparameter optimization, cross-validation, and hydrological patterns must be checked to ensure its successful application. Our models were not completely able to capture the behavior of some of the extreme events; this can be the result of the data imbalance. Various sampling techniques can be explored to increase the efficiency of the models. Future research should target larger and more diverse basins with unique meteorological patterns to broaden our understanding of runoff prediction in different environmental contexts.

Author Contributions: Conceptualization, A.K. and A.A.; Methodology, A.A. and B.P.; Software, A.A. and B.A.M.; Validation, A.A., A.S. and B.A.M.; Formal analysis, A.A. and A.K.; Writing—original draft, A.A., B.A.M., and B.P.; Writing—review and editing, A.A., A.K., B.A.M., B.P., and A.S.; Supervision, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data generated or analyzed during this study are included in this published article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Jonkman, S. Global Perspectives on Loss of Human Life Caused by Floods. *Nat. Hazards* **2005**, *34*, 151–175. [CrossRef]
2. Ke, Q.; Tian, X.; Bricker, J.; Tian, Z.; Guan, G.; Cai, H.; Huang, X.; Yang, H.; Liu, J. Urban pluvial flooding prediction by machine learning approaches—A case study of Shenzhen city, China. *Adv. Water Resour.* **2020**, *145*, 103719. [CrossRef]
3. Bhusal, A.; Parajuli, U.; Regmi, S.; Kalra, A. Application of Machine Learning and Process-Based Models for Rainfall-Runoff Simulation in DuPage River Basin, Illinois. *Hydrology* **2022**, *9*, 117. [CrossRef]
4. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]
5. Wilkinson, M.E.; Quinn, P.F.; Barber, N.J.; Jonczyk, J. A framework for managing runoff and pollution in the rural landscape using a Catchment Systems Engineering approach. *Sci. Total Environ.* **2014**, *468–469*, 1245–1254. [CrossRef] [PubMed]
6. Asefa, T.; Kemblowski, M.; McKee, M.; Khalil, A. Multi-Time Scale Stream Flow Predictions: The Support Vector Machines Approach. *Hydrology* **2006**, *318*, 7–16. [CrossRef]
7. Halwatura, D.; Najim, M.M.M. Application of the HEC-HMS model for runoff simulation in a tropical catchment. *Environ. Model. Softw.* **2013**, *46*, 155–162. [CrossRef]
8. Sahoo, S.; Russo, T.A.; Elliott, J.; Foster, I. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. *Water Resour. Res.* **2017**, *53*, 3878–3895. [CrossRef]
9. Brunner, M.I.; Slater, L.J.; Tallaksen, L.M.; Clark, M.P. Challenges in modeling and predicting floods and droughts: A review. *Environ. Sci.* **2021**, *8*, e1520. [CrossRef]
10. Mosavi, A.; Ozturk, P.; Chau, K.W. Flood Prediction Using Machine Learning Models: Literature Review. *Water* **2018**, *10*, 1536. [CrossRef]
11. Hosseiny, H.; Nazari, F.; Smith, V.; Nataraj, C. A Framework for Modeling Flood Depth Using a Hybrid of Hydraulics and Machine Learning. *Sci. Rep.* **2020**, *10*, 8222. [CrossRef] [PubMed]
12. Zahura, F.T.; Goodall, J. Predicting combined tidal and pluvial flood inundation using a machine learning surrogate model. *J. Hydrol. Reg. Stud.* **2022**, *41*, 101087. [CrossRef]
13. Yang, S.N.; Chang, L.C. Regional Inundation Forecasting Using Machine Learning Techniques with the Internet of Things. *Water* **2020**, *12*, 1578. [CrossRef]
14. Chen, G.; Hou, J.; Liu, Y.; Xue, S.; Wu, H.; Wang, T.; Lv, J.; Jing, J.; Yang, S.N. Urban inundation rapid prediction method based on multi-machine learning algorithm and rain pattern analysis. *J. Hydrol.* **2024**, *633*, 131059. [CrossRef]
15. Erdal, H.; Karakurt, O. Advancing Monthly Streamflow Prediction Accuracy of CART Models Using Ensemble Learning Paradigms. *Hydrology* **2013**, *447*, 119–128. [CrossRef]
16. Schnier, S.; Cai, X. Prediction of Regional Streamflow Frequency Using Model Tree Ensembles. *Hydrology* **2014**, *517*, 298–309. [CrossRef]
17. Yaseen, Z.M.; El-Shafie, A.; Jaafar, O.; Afan, H.A.; Sayl, K.N. Artificial Intelligence Based Models for Stream-Flow Forecasting: 2000–2015. *Hydrology* **2015**, *530*, 829–844. [CrossRef]
18. Bekele, E.G.; Nicklow, J.W. Hybrid Evolutionary Search Methods for Training an Artificial Neural Network. *Impacts Glob. Clim. Change* **2012**. [CrossRef]
19. Ghorbani, M.A.; Khatibi, R.; Goel, A.; FazeliFard, M.H.; Azani, A. Modeling river discharge time series using support vector machine and artificial neural networks. *Environ. Earth Sci.* **2016**, *75*, 685. [CrossRef]
20. Tamiru, H.; Dinka, M.O. Application of ANN and HEC-RAS model for flood inundation mapping in lower Baro Akobo River Basin, Ethiopia. *J. Hydrol. Reg. Stud.* **2021**, *36*, 100855. [CrossRef]

21. Qie, G.; Zhang, Z.; Getahun, E.; Allen Mamer, E. Comparison of Machine Learning Models Performance on Simulating Reservoir Outflow: A Case Study of Two Reservoirs in Illinois, U.S.A. *Am. Water Resour. Assoc.* **2022**, *59*, 554–570. [CrossRef]
22. Riad, S.; Jacky, M.; Bouchaou, L.; Najjar, Y. Rainfall-runoff model using an artificial neural network approach. *Math. Comput. Model.* **2004**, *40*, 839–846. [CrossRef]
23. Carrier, C.; Kalra, A.; Ahmad, S. Using Paleo Reconstructions to Improve Streamflow Forecast Lead Time in the Western United States. *Am. Water Resour. Assoc.* **2013**, *49*, 1351–1366. [CrossRef]
24. Maity, R.; Bhagwat, P.; Bhatnagar, A. Potential of Support Vector Regression for Prediction of Monthly Streamflow Using Endogenous Property. *Hydrol. Process.* **2010**, *24*, 917–923. [CrossRef]
25. Lin, J.Y.; Cheng, C.T.; Chau, K.W. Using Support Vector Machines for Long-Term Discharge Prediction. *Hydrol. Sci.* **2006**, *51*, 599–612. [CrossRef]
26. Guo, J.; Zhou, J.; Qin, H.; Zou, Q.; Li, Q. Monthly Streamflow Forecasting Based on Improved Support Vector Machine Model. *Expert Syst. Appl.* **2011**, *38*, 13073–13081. [CrossRef]
27. Battelle. *Dupage River, Illinois Feasibility Report and Integrated Environmental Assessment*; W912HQ-15-D-0001; US Army Corps of Engineers: Columbus, OH, USA, 2018.
28. Dubey, S.; Gupta, H.; Goyal, M.K.; Joshi, N. Evaluation of precipitation datasets available on Google earth engine over India. *Int. J. Climatol.* **2021**, *41*, 4844–4863. [CrossRef]
29. Rajaee, T.; Khani, S.; Ravansalar, M. Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: A review. *Chemom. Intell. Lab. Syst.* **2020**, *200*, 103978. [CrossRef]
30. LV, Z.; Zuo, J.; Rodriguez, D. Predicting of runoff using an optimized SWAT-ANN: A case study. *J. Hydrol. Reg. Stud.* **2020**, *29*, 100688. [CrossRef]
31. Bafitlhile, T.M.; Li, Z. Applicability of Support Vector Machine and Artificial Neural Network for Flood Forecasting in Humid, Semi-Humid and Semi-Arid Basins in China. *Water* **2019**, *11*, 85. [CrossRef]
32. Sharifi, A.; Dinpashoh, Y.; Mirabbasi, R. Daily runoff prediction using the linear and non-linear models. *Water Sci. Technol.* **2017**, *76*, 793–805. [CrossRef] [PubMed]
33. Alexandridis, A.; Patrinos, P.; Sarimveis, H.; Tsekouras, G. A two-stage evolutionary algorithm for variable selection in the development of RBF neural network models. *Chemom. Intell. Lab. Syst.* **2005**, *75*, 149–162. [CrossRef]
34. Tran, H.D.; Muttil, N.; Perera, B.J.C. Selection of significant input variables for time series forecasting. *Environ. Model. Softw.* **2015**, *64*, 156–163. [CrossRef]
35. Silva, A.P.D.; Filzmoser, P.; Brito, P. Outlier detection in interval data. *Adv. Data Anal. Classif.* **2018**, *12*, 785–822. [CrossRef]
36. Ramsay, J.O.; Silverman, B.W. *Functional Data Analysis*; Springer: New York, NY, USA, 2005.
37. Chebana, F.; Dabo-Niang, S.; Ouarda, T.B. Exploratory functional flood frequency analysis and outlier detection. *Water Resour. Res.* **2012**, *48*, 1–20. [CrossRef]
38. Sunderland, K.M.; Beaton, D.; Fraser, J.; Kwan, D.; McLaughlin, P.M.; Montero-Odasso, M.; Peltsch, A.J.; Pieruccini-Faria, F.; Sahlas, D.J.; Swartz, R.H.; et al. The utility of multivariate outlier detection techniques for data quality evaluation in large studies: An application within the ONDRI project. *BMC Med. Res. Methodol.* **2019**, *19*, 102. [CrossRef] [PubMed]
39. Kumar, V.; Ashu, V.; Shikha, J. Modeling rainfall—Runoff process using artificial neural network with emphasis on parameter sensitivity. *Model. Earth Syst. Environ.* **2020**, *6*, 2177–2188. [CrossRef]
40. Daliakopoulos, I.N.; Tsanis, I.K. Comparison of an artificial neural network and a conceptual rainfall–runoff model in the simulation of ephemeral streamflow. *Hydrol. Sci.* **2016**, *61*, 2763–2774. [CrossRef]
41. Abbot, J.; Marohasy, J. Input selection and optimisation for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks. *Atmos. Res.* **2014**, *138*, 166–178. [CrossRef]
42. Li, L.; Xu, H.; Chen, X.; Simonovic, S. Streamflow forecast and reservoir operation performance assessment under climate change. *Water Resour. Manag.* **2010**, *24*, 83–104. [CrossRef]
43. Wu, C.; Chau, K.W. Data-driven models for monthly streamflow time series prediction. *Eng. Appl. Artif. Intell.* **2010**, *23*, 1350–1367. [CrossRef]
44. Kar, A.K.; Lohani, A.K.; Goel, N.K.; Roy, G.P. Development of flood forecasting system using statistical and ANN techniques in the downstream catchment of mahanadi basin, india. *Water Resour. Prot.* **2010**, *2*, 880. [CrossRef]
45. Sulaiman, J.; Wahab, S.H. Heavy rainfall forecasting model using artificial neural network for flood prone area. In *It Convergence and Security 2017*; Springer: Singapore, 2018; Volume 449, pp. 68–76. [CrossRef]
46. Jain, A.; Indurthy, S.K.V.P. Closure to "comparative analysis of event-based rainfall-runoff modeling techniques—Deterministic, statistical, and artificial neural networks" by ASHU JAIN and SKV prasad indurthy. *Hydrolic Eng.* **2004**, *9*, 551–553. [CrossRef]
47. Tanty, R.; Desmukh, T.S. Application of artificial neural network in hydrology—A review. *Int. J. Eng. Res. Technol.* **2015**, *4*, 184–188. [CrossRef]
48. Badrzadeh, H.; Sarukkalige, R.; Jayawardena, A.W. Impact of multi-resolution analysis of artificial intelligence models inputs on multi-step ahead river flow forecasting. *Hydrology* **2013**, *507*, 75–85. [CrossRef]
49. Taormina, R.; Chau, K.W.; Sethi, R. Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the Venice Lagoon. *Eng. Appl. Artif. Intell.* **2012**, *25*, 1670–1676. [CrossRef]
50. Chen, C.; Hui, Q.; Xie, W.; Wan, S.; Zhou, Y.; Pei, Q. Convolutional Neural Networks for forecasting flood process in Internet-of-Things enabled smart city. *Comput. Netw.* **2021**, *186*, 107744. [CrossRef]

51. Chan, V.K.H.; Chan, C.W. Towards explicit representation of an artificial neural network model: Comparison of two artificial neural network rule extraction approaches. *Petroleum* **2020**, *6*, 329–339. [CrossRef]
52. Pham, S.; Le, H.M.; Thanh, D.V.; Dang, T.D.; Loc, H.H.; Anh, D.T. Deep learning convolutional neural network in rainfall-runoff modelling. *Hydroinformatics* **2020**, *22*, 541–561. [CrossRef]
53. Kadam, V.; Kumar, S.; Bongale, A.; SeemaWazarkar, S.; Kamat, P.; Patil, S. Enhancing Surface Fault Detection Using Machine Learning for 3D Printed Products. *Appl. Syst. Innov.* **2021**, *4*, 34. [CrossRef]
54. Asadollahi, A.; Sohrabifar, A.; Ghimire, A.B.; Poudel, B.; Shin, S. The Impact of Climate Change and Urbanization on Groundwater Levels: A System Dynamics Model Analysis. *Environ. Prot. Res.* **2024**, *4*, 1–15. [CrossRef]
55. Chicco, D.; Warrens, M.J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [CrossRef] [PubMed]
56. Thakur, B.; Kalra, A.; Ahmad, S.; Lamb, K.W.; Lakshmi, V. Bringing statistical learning machines together for hydroclimatological predictions—Case study for Sacramento San joaquin River Basin, California. *Hydrol. Reg. Stud.* **2020**, *27*, 100651. [CrossRef]
57. Yaseen, Z.M.; Awadh, S.M.; Sharafati, A.; Shahid, S. Complementary data-intelligence model for river flow simulation. *J. Hydrol.* **2018**, *567*, 180–190. [CrossRef]
58. Taylor, K.E. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res. Atmos.* **2001**, *106*, 7183–7192. [CrossRef]
59. Kumar, N.; Singh, S.K.; Srivastava, P.K.; Narsimlu, B. SWAT Model calibration and uncertainty analysis for streamflow prediction of the Tons River Basin, India, using Sequential Uncertainty Fitting (SUFI-2) algorithm. *Model. Earth Syst. Environ.* **2017**, *3*. [CrossRef]
60. Abbaspour, K.C.; Rouholahnejad, E.; Vaghefi, S.; Srinivasan, R.; Yang, H.; Kløve, B. A Continental-Scale Hydrology and Water Quality Model for Europe: Calibration and Uncertainty of a High-Resolution Large-Scale SWAT Model. *Hydrology* **2015**, *524*, 733–752. [CrossRef]
61. Ranatunga, T.; Tong, S.T.; Yang, Y.J. An approach to measure parameter sensitivity in watershed hydrological modelling. *Hydrol. Sci. J.* **2017**, *62*, 76–92. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.