



Enhancing rainfall–runoff model accuracy with machine learning models by using soil water index to reflect runoff characteristics

Sarunphas lamampai , Yutthana Talaluxmana, Jirawat Kanasut and Prem Rangsiwanichpong *

Department of Water Resources Engineering, Faculty of Engineering, Kasetsart University, Bangkok, Thailand

*Corresponding author. E-mail: prem.r@ku.th

 SI, 0000-0002-0613-5665; PR, 0000-0002-1663-3537

ABSTRACT

The advancement of data-driven models contributes to the improvement of estimating rainfall–runoff models due to their advantages in terms of data requirements and high performance. However, data-driven models that rely solely on rainfall data have limitations in responding to the impact of soil moisture changes and runoff characteristics. To address these limitations, a method was developed for selecting predictor variables that utilize the accumulation of rainfall at various time intervals to represent soil moisture, the changes in the runoff coefficient, and runoff characteristics. Furthermore, this study investigated the utility of rainfall products [such as climate hazards group infrared precipitation with station data (CHIRPS) and global precipitation measurement (GPM)] for representing rainfall data, while also using the soil water index (SWI) to enhance runoff estimation. To assess these methods, the random forest (RF) and artificial neural network (ANN) models were utilized to simulate daily runoff. Incorporating both the rainfall and SWI data led to improved outcomes. The RF demonstrated superior performance compared with the ANN and the conceptual model, without the need for baseflow separation or antecedent runoff. Furthermore, accumulated rainfall was shown to be a valuable input for the models. These findings should facilitate the estimation of runoff in locations with limited measurement data on rainfall and soil moisture by utilizing remote sensing data.

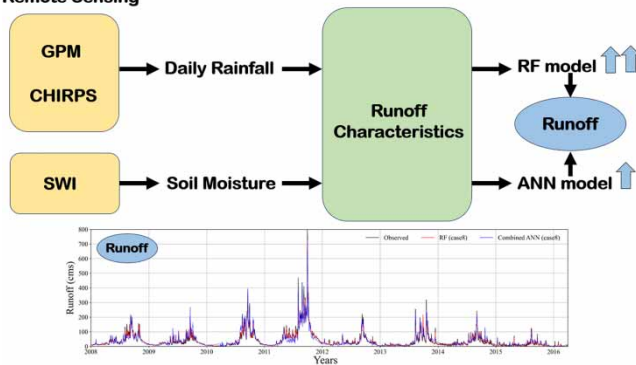
Key words: ANN, CHIRPS, GPM, rainfall–runoff model, RF, SWI

HIGHLIGHTS

- This work proposes a new method for selecting predictors from accumulated rainfall to reflect the features of runoff.
- This method is efficient enough for runoff estimation without needing to use rain gauge station data.
- RF model outperforms the conceptual model and ANN model in all seasons.
- SWI can enhance the accuracy of the model.
- GPM and CHIRPS yield results similar to those obtained using rainfall data from rain gauge stations.

GRAPHICAL ABSTRACT

Remote Sensing



This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY 4.0), which permits copying, adaptation and redistribution, provided the original work is properly cited (<http://creativecommons.org/licenses/by/4.0/>).

INTRODUCTION

Rainfall and runoff data are essential components of hydrology, offering crucial insights that help determine effective water resource management strategies. The weather patterns in the Ping River Basin are affected by the Northeast Monsoon during the dry season (from November to mid-March) and the Southwest Monsoon throughout the wet season (from mid-May to September) (Rangsiwanichpong & Melesse 2022). In the Upper Ping River (UPR) basin in Thailand, heavy rainfall and rapid changes in rainfall patterns are particularly concentrated between August and October. Consequently, the basin frequently experiences flash floods and subsequent floods within a few days after heavy rain. Given these challenges, accurately predicting daily streamflow in the basin becomes crucial for mitigating flood damage and enabling early warnings.

Rainfall–runoff models are hydrological models used to simulate the transformation of rainfall into runoff within a specific area. These models imitate the phenomenon under various conditions, including factors such as soil type, land cover, topography, and meteorological conditions. The complex processes of the FLEX (Gao *et al.* 2014) and NAM (Bao *et al.* 2011) models primarily involve infiltration, evapotranspiration, and runoff generation. Furthermore, the physically based and conceptual models require many input variables that represent various physical characteristics of a basin to achieve accurate estimation (Maier *et al.* 2010). However, many areas in Thailand face a major lack of measurement data such as soil moisture, ground water, and rainfall station measurements.

In recent decades, data-driven models have gained widespread popularity in the field of water resource management (Khan *et al.* 2021; Mao *et al.* 2021; Mohammadi 2021). The primary advantages of these models are their simplicity and speed. The calculation process relies on establishing relationships between inputs and outputs through mathematical functions, without requiring a detailed consideration of the underlying physical characteristics (Patel & Joshi 2017). Many studies demonstrated the capacity of artificial neural network (ANN) models to estimate runoff based on nonlinear relationships, often resulting in superior outcomes compared with the conceptual models (Shoab *et al.* 2016; Huo *et al.* 2017; Mishra & Karmakar 2019). Similarly, the random forest (RF) model has the potential to generate more precise and stable predictions through its ensemble approach (Schoppa *et al.* 2020; Qiao *et al.* 2023). However, it is well known that these models are insufficient at accurately predicting runoff when using only meteorological variables and not considering antecedent runoff (Tongal & Booj 2018).

Some studies have made attempts to improve the accuracy of runoff prediction by using the antecedent runoff and upstream runoff as predictor variables (Hosseini & Mahjouri 2016; Mei & Smith 2021; Qiao *et al.* 2023; Sayed *et al.* 2023). Although these approaches have demonstrated enhanced accuracy, their utilization may be limited in some locations. For example, these methods are impractical in upstream basins where inflow data are unavailable or in areas characterized by discontinuous or missing runoff data. Furthermore, the accuracy of runoff prediction can be improved by separating the baseflow components. Lakshmi *et al.* (2022) separated rainy days and non-rainy days for runoff simulation using machine models. Tongal & Booj (2018) demonstrated that the one-parameter, recursive, single-pass, digital filter method can enhance the simulation performance of machine learning models to a certain extent; however, it still requires the use of antecedent runoff data. Song (2022) used groundwater data as a substitute for baseflow, which resulted in very good performance. However, notably, the method has the limitation that groundwater data may not be accessible in all areas, especially within the UPR.

Remote sensing data can serve as substitutes for traditional ground gauges in areas where there is a deficiency in rainfall and soil moisture data. Many gridded precipitation products at the global or regional scale are now publicly available, such as global precipitation measurement (GPM) (Hou *et al.* 2014), tropical rainfall measurement mission (TRMM) (Huffman *et al.* 2007), Climate Prediction Centre Morphing Algorithm (CMORPH) (Joyce *et al.* 2004), and Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) (Funk *et al.* 2015). Zhang *et al.* (2022) investigated precipitation products using the variable infiltration capacity (VIC) model across various regions. The outcomes revealed that streamflow simulations based on GPM produced the best performance compared with GPM, TRMM, and APHRODITE. Furthermore, these satellite products were utilized to simulate runoff using the conceptual model GR2M on both daily and monthly scales (Ye *et al.* 2022). The results illustrated that CHIRPS had the best performance during the validation period in the middle of the TRMM, CHIRPS, and CMORPH. Furthermore, the soil water index (SWI) has effectively represented soil moisture levels (Paulik *et al.* 2014; Bouaziz *et al.* 2021). Sriwongsitanon *et al.* (2023) provided evidence that SWI at a 40-day interval was significantly correlated with soil moisture, as obtained from a conceptual model. Therefore, the current study was interested in the performance of rainfall products (GPM and CHIRPS) and SWI.

Although many studies have developed rainfall–runoff models using machine learning, much of the data used for enhancement is either limited or lacks measurements. Additionally, some techniques for separating baseflow are difficult to apply in a

given area, such as the UPR. However, to the best of our knowledge, there has been no publication describing the process of selecting accumulated rainfall to effectively represent runoff characteristics and how this may enhance the accuracy of runoff estimation models using SWI.

This research marks a departure from previous studies in various significant aspects. First, it introduces the utilization of cumulative rainfall data as a surrogate for the essential components influencing the occurrence of surface water. Second, it substantiates that antecedent surface water data from northern water measurement stations are not imperative; instead, rainfall data serve as a valuable aid in assessment. Third, this research diverges from prior methodologies by not employing the technique of segregating surface water components into baseflow and fast flow. Fourth, it demonstrates the efficacy of the SWI as a tool for enhancing the accuracy of surface water assessments. Lastly, it highlights the effectiveness of solely using rainfall data for surface water assessment, eliminating the dependency on other factors such as land use and soil type.

To overcome the limitations of data, the current study proposed a novel and easily accessible methodology for creating predictors that can be applied to all areas using satellite data. Thus, we identified a suitable period for accumulating rainfall to effectively capture runoff characteristics, encompassing aspects such as soil moisture and runoff patterns. Unlike processes in some reports that relied on randomness or trial approaches, the current approach applied a principled selection process. Moreover, ANN and RF models were selected to evaluate the performance of the predictor variables.

MATERIALS AND METHODS

Study area

The UPR, situated in the northern region of Thailand, covers an area of approximately 6,355 km² (Figure 1). This catchment area is characterized by diverse landscapes, including mountains, forests, and agricultural land. The basin's geographical features, coupled with its fertile soil and abundant water resources, establish it as a crucial hub for agricultural activities. Notably, rice cultivation is the predominant agricultural activity in the area, playing a crucial role in the country's food production. The primary river of the basin originates from the northern mountains and flows in a southerly direction, eventually

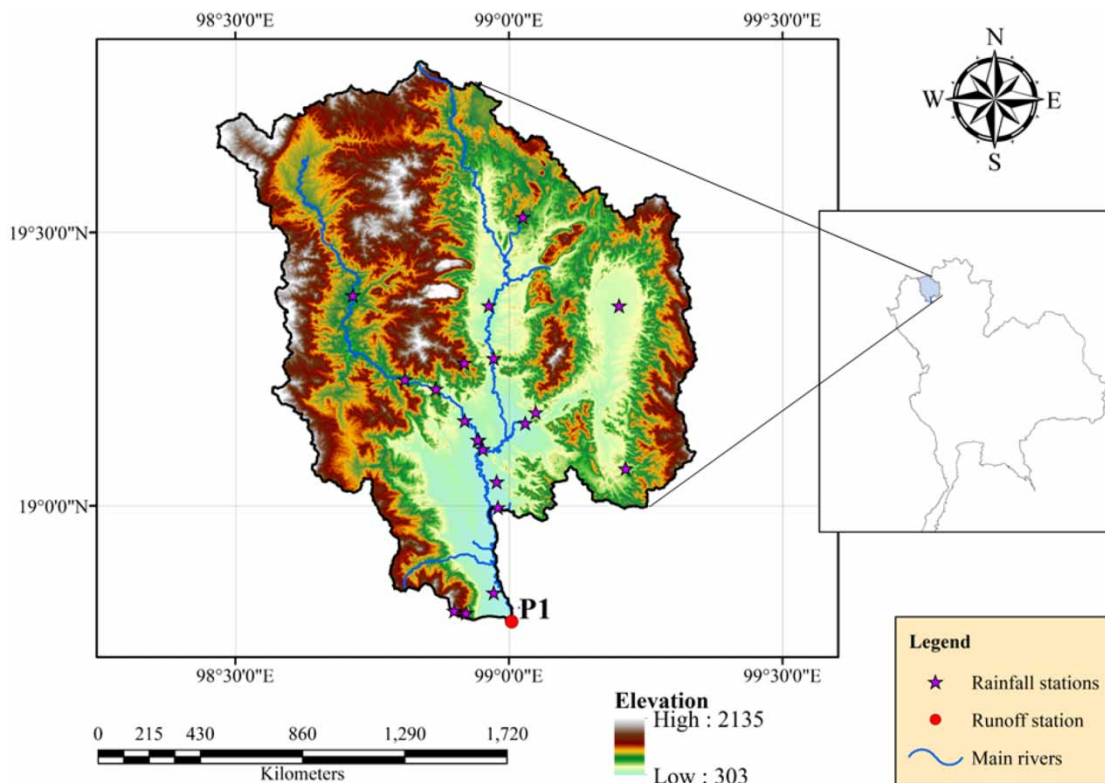


Figure 1 | The topography of Upper Ping River basin and the location of the rainfall and runoff stations.

into the Bhumibol Dam. This river system is a vital resource for numerous water-dependent activities within the region. The basin's climate is significantly influenced by the northeast and southeast monsoons, giving rise to well-defined dry and wet seasons. The dry season spans from November to April, while the wet season prevails from May to October. Because of the hydrological characteristics of the basin, it typically requires around 2–3 days for water to travel from the furthest point within the catchment area to the outlet.

Measurement data

This study required two essential datasets: rainfall data and runoff data. They were sourced from the Royal Irrigation Department (RID) based on daily data. To represent average rainfall, the Thiessen Polygon method was utilized to calculate the average rainfall from 19 rain gauge stations situated within the study area. Furthermore, the runoff data at the station at P.1, located at the basin outlet, was used for model calibration. However, notably, data availability was influenced by concerns related to data integrity and limited online access. Consequently, the dataset covering the period from 2008 to 2016 was selected for this study.

Soil water index

The SWI provides a measure of soil moisture levels at various soil depths, primarily influenced by precipitation-induced infiltration processes. The calculation of SWI relies on surface soil moisture (SSM) data, extracted from the Advanced Scatterometer (ASCAT) sensor, as specified in studies by *Albergel et al. (2008)* and *Paulik et al. (2014)*. The available data cover the period since 2008. To capture soil moisture dynamics at different depths, the SWI uses characteristic temporal length (T) parameters. These parameters include intervals of 1, 5, 10, 15, 20, 40, 60, and 100 days. Each T -length corresponds to a specific depth within the soil profile, offering insights into the moisture conditions at distinct levels. Notably, research conducted in the Ping River basin has indicated that the 40-day T -length is well suited for representing soil moisture in the upper region of the basin (*Sriwongsitanon et al. 2023*). Therefore, SWI-40 days may enhance runoff forecasting accuracy and contribute to more informed water resource management strategies.

Rainfall products

For this study, two rainfall product datasets were selected as key inputs: the GPM from Integrated Multisatellite Retrievals for GPM (GPM-IMERG) and CHIRPS. These datasets offer a valuable solution for estimating precipitation in areas where ground-based data might be scarce or limited. The GPM datasets, developed by NASA, are designed to provide estimates of surface precipitation across a substantial portion of the globe. These datasets offer a high spatial resolution of 0.1 km. GPM consists of three distinct products, each tailored to address varying user needs concerning accuracy and timeliness. These products consist of early, late, and final versions. Among these, the final version stands out by having the highest accuracy, even though it has a longer latency of three months (*Foelsche et al. 2017*; *Ramadhan et al. 2022*). The CHIRPS datasets, established through collaboration between the United States Geological Survey (USGS) and the Climate Hazards Group at the University of California, utilize infrared precipitation data in conjunction with station measurements. CHIRPS cover a substantial area from 50° S to 50° N with a fine spatial resolution of 0.05 km. They have two distinctive products: the preliminary rainfall product with a two-day latency and the final product with a three-week latency (*Funk et al. 2015*). To ensure the accuracy and quality of the input data, as well as to achieve robust and reliable outcomes for runoff estimation, the current study utilized the final versions of both GPM and CHIRPS.

Models

The ANN model is a complicated computational process inspired by the neural networks observed in the human brain. It exhibits a remarkable ability to handle intricate, nonlinear relationships within data. ANNs excel at discerning complex patterns and resolving multifaceted problems, ranging from classification tasks to regression analysis. An ANN comprises three fundamental layers: the input layer, hidden layers, and the output layer. The input layer is the initial point for introducing input data into the model and relaying the raw information to the network. The hidden layers are responsible for processing the input dataset and transforming it to make it suitable for the final output. The output layer determines the ultimate target output of the model based on the processed information from the hidden layers.

An RF model is a powerful ensemble learning technique used in machine learning for both classification and regression. It combines multiple decision trees during the training process and aggregates their predictions to produce the best results. These individual decision trees are created using a process called bootstrapped sampling, where each tree is trained on a

different subset of the data with replacement. The ensemble model can help improve model accuracy, reduce overfitting, and increase robustness against variance in the data.

Methods

The methodology of this study can be summarized as follows: 1. Presenting a method for selecting predictor variables from accumulated rainfall data to reflect the runoff characteristics. 2. Evaluating the predictor variables through simulation using the RF and ANN models. 3. Assessing the performance of the gridded rainfall products as input data.

Selection of predictor variables

The primary components of runoff are surface flow and base flow. Surface flow reacts promptly to rainfall events, exhibiting a rapid increase and then a decrease within a few days when it rains. Conversely, base flow changes gradually and provides a stable water supply to streamflow during dry periods without precipitation. The natural factors influencing runoff change comprise the intensity of rainfall, soil moisture, land use, and soil type. In this study, both rainfall and soil moisture were chosen as predictors for runoff forecasting, under the assumption that there are no changes in land use and soil type that could influence the runoff. Soil moisture plays a crucial role in the variation of runoff as it indicates the water that can be absorbed or retained in the soil. When the soil is already saturated from previous rainfall, its ability to store additional water is diminished, resulting in an increase in runoff. Due to the absence of soil moisture measurement data in the region, we opted to use the SWI as a substitute for soil moisture. Additionally, accumulated rainfall at various periods was compared with the SWI to find the most suitable correlation. To analyze the trend of runoff, the runoff and accumulated rainfall were compared using Pearson correlations to determine the suitability of the time interval of the accumulated rainfall as a predictor. Changes in the runoff coefficient (RC) were calculated on a monthly scale by dividing the monthly average runoff by the monthly average rainfall. This calculation helped us to understand the runoff pattern occurring during various periods. Typically, the RC is relatively low during the beginning of the rainy season due to the arid conditions in the area caused by persistent drought and low groundwater levels. Then, the RC starts to steadily increase toward the end of the rainy season. Even after the rainfall has decreased, the RC continues to rise. These findings illustrate that utilizing data spanning an extended period in the past aids in identifying trends and comprehending fluctuations in runoff over time in a specific area.

Therefore, a few days of prior rainfall cannot fully capture the comprehensive condition of runoff changes, because even with the same rainfall intensity at different times, it may not result in identical runoff outcomes. To overcome this limitation, we proposed integrating the SWI, accumulated rainfall data, and prior rainfall data to enhance runoff forecasting. The accumulated rainfall data had the highest correlation when the SWI and runoff were used as predictors, offering valuable insights for forecasting. Finally, we attempted to utilize all accumulated rainfall data at intervals of 10, 30, 60, 90, 120, and 150 days as predictors, to highlight the importance of using determined accumulated rainfall values.

Evaluation models and predictors

In this study, runoff forecasting was conducted using ANN and RF models, which were implemented using TensorFlow libraries in the Python language and system. For predicting 1-day runoff, the predictor variables obtained in the section “Study area” were used as input features for both the ANN and RF models. The dataset was randomly split into training and testing sets, with a ratio of 70:30, respectively. Furthermore, cross-validation was also incorporated to ensure the models’ robustness and evaluate their generalization capabilities. The RF model was configured with default hyperparameters, where `n_estimators`, `min_sample_leaf`, `max_features`, and `max_depth` were set to 10, 1, ‘auto’, and ‘None’, respectively. The ANN model was configured using the Rectified Linear Unit (ReLU) function as the activation function and consisted of two hidden layers with 128 and 256 nodes, respectively. Generally, the ANN model shows the capability to effectively capture the relationships between high rainfall and runoff during the rainy season. However, it struggles in the dry season because of the distinct hydrological conditions. To address this issue, we proposed a combined-ANN model, consisting of two sub-models tailored for the rainy and dry seasons. The first sub-model operates during the dry season when there is either no rain or less than 1 mm rain/day persisting for over five consecutive days, as such conditions have minor impact on runoff (Figure 2). Conversely, the second sub-model is used for the rainy season, encompassing the remaining timeframe. The combined-ANN model used the best predictor obtained from Section ‘Selection of predictor variables’.

Furthermore, we selected the RF model for runoff forecasting because of its ability to handle various conditions well. The RF naturally divides the dataset into subgroups and performs regression within each subgroup. The inherent characteristic of the RF enables it to adapt and capture the different relationships between rainfall and runoff during the rainy and dry seasons

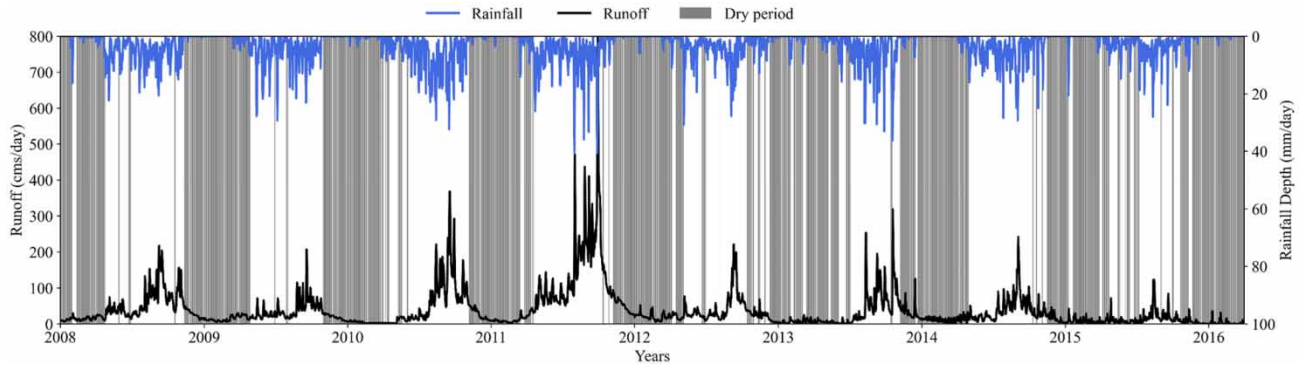


Figure 2 | Separation runoff into dry period with duration rainfall less than 1 mm/day and continuous for more than 5 days.

without the need for explicit data splitting or the creation of separate sub-models. As the RF can automatically detect and accommodate these fluctuations, it is a more convenient and efficient option for capturing the diverse hydrological conditions spanning the entire year.

To assess the effectiveness and suitability of models for real-world use, the Nash–Sutcliffe efficiency (NSE) (Nash & Sutcliffe 1970) and root mean square error (RMSE) statistics were applied to assess the runoff forecasting outcomes. NSE is an essential performance metric for evaluating hydrograph accuracy, with higher values indicating superior model performance (Equation (1)). Conversely, RMSE is a commonly applied metric to quantify the differences between observed and simulated data, where lower values indicate better model performance (Equation (2)).

$$NSE = 1 - \frac{\sum_{i=1}^N (Q_{obs,i} - Q_{sim,i})^2}{\sum_{i=1}^N (Q_{obs,i} - \bar{Q}_{obs})^2} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Q_{sim} - Q_{obs})^2}{N}} \tag{2}$$

where $Q_{obs,i}$ and $Q_{sim,i}$ are the observed and simulated runoff data at time i , respectively, \bar{Q}_{obs} is the mean observed runoff data, and N is the period length of the runoff data.

Performance evaluation of rainfall products

Rainfall products obtained through remote sensing offer a notable advantage by providing spatial information that covers the entire area of Thailand. This is particularly beneficial in regions where ground-based measurement data are unavailable or sparse. To use the overall rainfall conditions as predictors for the model, they were averaged across all pixels within the UPR. This study selected CHIRPS and GPM products as predictors for the RF and ANN models to forecast runoff. The runoff results based on the rainfall products were evaluated using NSE and RMSE to determine the suitable product for the Ping River basin. Considering the efficiency of the products, both products were assessed using four metrics: probability of detection (POD), false alarm ratio (FAR), critical success index (CSI), and RMSE (Kim & Han 2021; Ramadhan *et al.* 2022). These metrics illustrate the variations in efficiency among the different products, which in turn affect the accuracy of runoff estimation.

The POD index reflects the accuracy ratio of precipitation events in these products, indicating how well the products identify actual rain events (Equation (3)). The FAR index represents the ratio of precipitation detected by these products but does not correspond to actual rain events (Equation (4)). The CSI reflects the score ratio of precipitation events detected by these products compared with the overall events, providing a measure of the products’ overall accuracy (Equation (5)). To calculate these metrics, many studies have defined the occurrence of an event based on rainfall exceeding 1 mm/day (Yong *et al.* 2010; Tang *et al.* 2020).

$$POD = H / (H + M) \tag{3}$$

$$FAR = F(H + F) \tag{4}$$

$$CSI = H / (H + F + M) \tag{5}$$

where H is the frequency of observed precipitation detected by satellite, M is the frequency of observed precipitation not detected by satellite, and F is the frequency of false precipitation detected by satellite.

A summary of all the processes for estimating runoff is presented in Figure 3.

RESULTS AND DISCUSSION

Selection predictors

In this study, the basin had a time concentration of three days, representing the period required for runoff to travel from the farthest point of the basin to the outlet. Consequently, the antecedent rainfall from the preceding three days was selected for predicting runoff. Figures 4 and 5 illustrate the time series of SWI and runoff, respectively, for accumulated rainfall data covering 2008–2016. The 10- and 30-day accumulated rainfall data had more pronounced fluctuations, primarily influenced by rainfall patterns. By contrast, the accumulated rainfall data over longer periods appeared smoother and more effectively highlighted seasonal variations. According to the Pearson correlation analysis, the 10-day accumulated rainfall data displayed a substantial correspondence with runoff variations, with a correlation coefficient (r) of 0.71. The 120-day accumulated rainfall data had the highest correlation with the SWI ($r = 0.88$).

In Figure 6, the RC displays a rising trend from May until the end of the rainy season. These findings underscored the variations in runoff at different times, even when the amount of rainfall remained the same. However, after October, the RC was not considered because the increasing runoff was mainly influenced by baseflow, while the rainfall decreased. This pattern extended over five months, highlighting the importance of selecting the 150-day accumulated rainfall as a predictor in the model to effectively capture changes in the RC.

Thus, to reveal the performance of selecting predictor variables, they were categorized into eight cases, as shown in Table 1 (Equations (6)–(13)). Case 1 represents surface flow based solely on three-days antecedent rainfall; Cases 2 and 3 represent

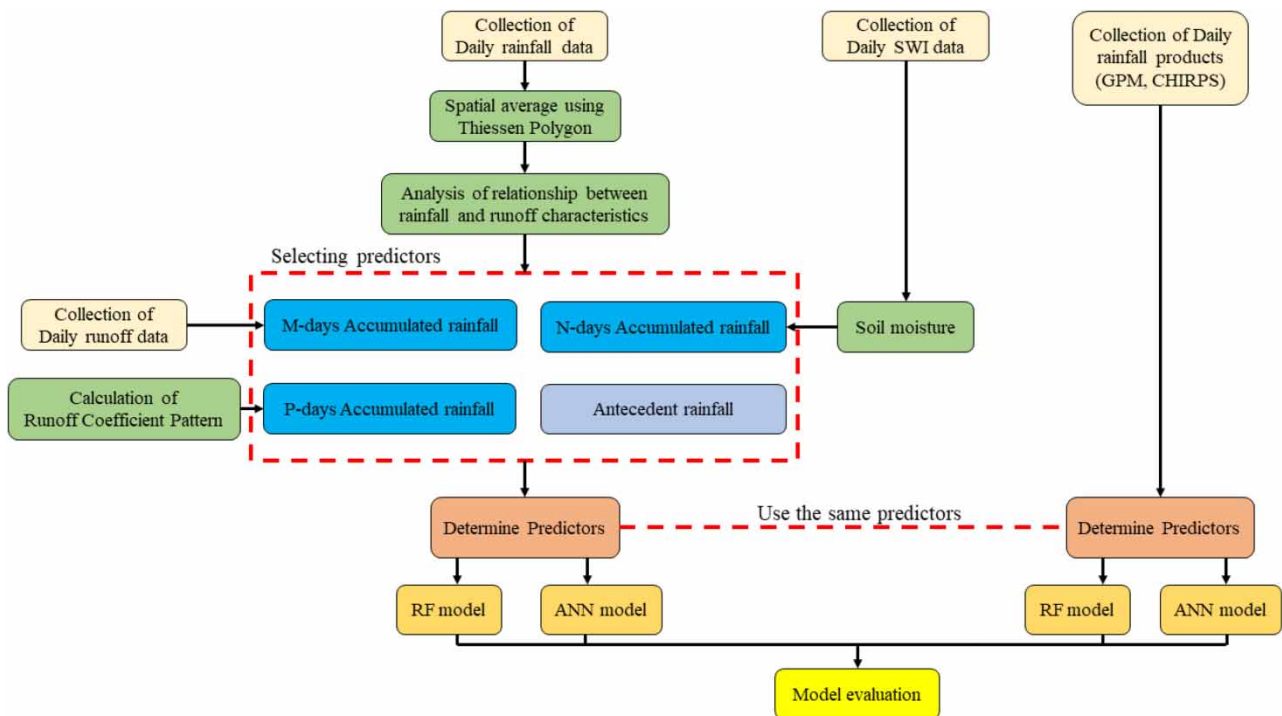


Figure 3 | Flowchart of selecting predictors from rainfall data and estimating runoff through RF and ANN models.

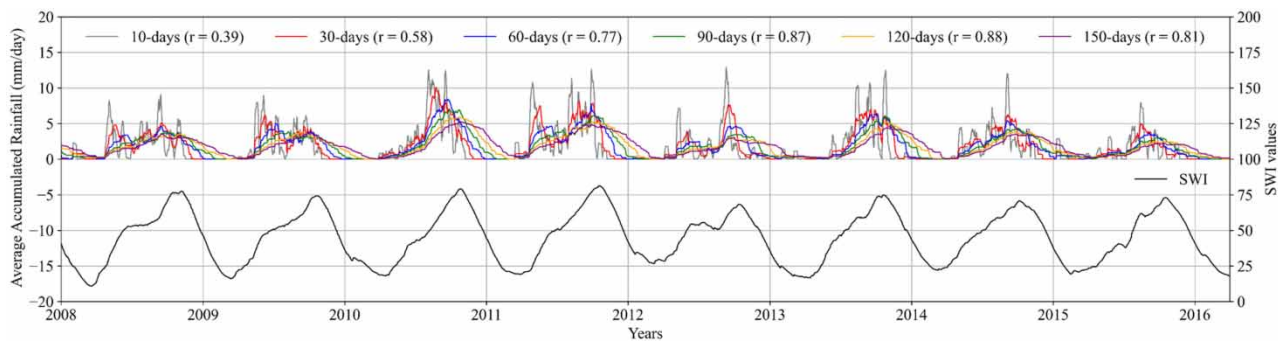


Figure 4 | Relationship between time series of SWI and accumulated rainfall with Pearson correlation (r).

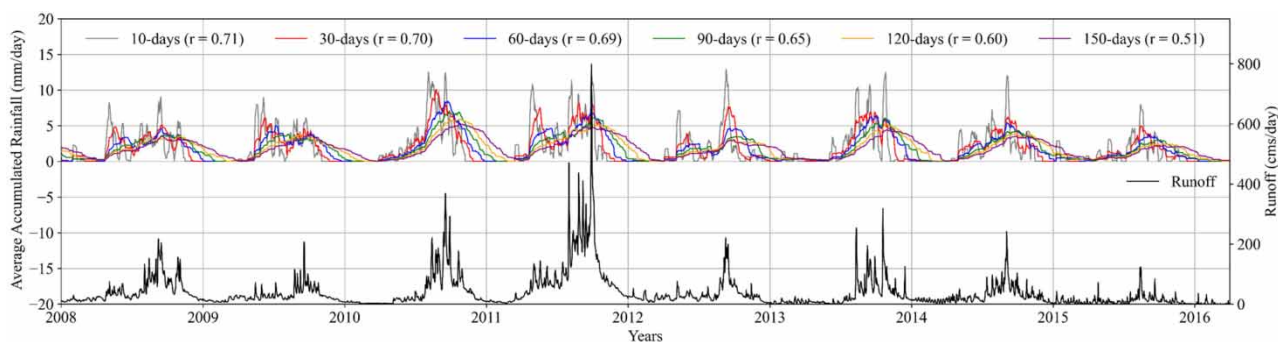


Figure 5 | Relationship between time series of runoff and accumulated rainfall with Pearson correlation (r).

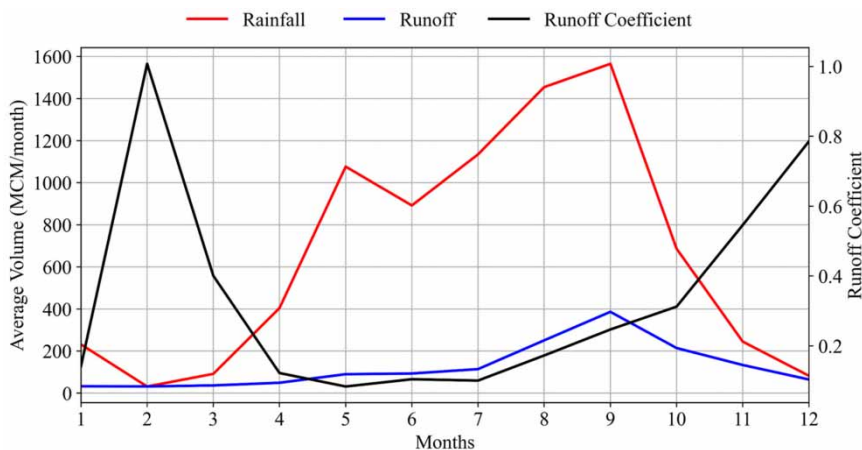


Figure 6 | Fluctuations in monthly runoff coefficient, rainfall, and runoff.

surface flow and soil moisture, with the addition of SWI data and 120-day accumulated rainfall; Cases 4 and 5 represent surface flow, soil moisture, and RC changes, with the addition of 150-day accumulated rainfall; Cases 6 and 7 represent surface flow, soil moisture, RC changes, and trend analysis, with the addition of 10-day accumulated rainfall; and Case 8 includes all components and incorporates accumulated rainfall at all time intervals.

Model performance for each predictor variable

The runoff forecasting models, utilizing the predictor variables from the nine cases were evaluated using the performance metrics NSE and RMSE with three rounds of cross-validation. The results of the training and testing are summarized in Tables 2 and 3.

Table 1 | All predictor variables for runoff simulation with machine learning models

Case	Predictors	Equations
1	$Q = f(P_{t-1}, P_{t-2}, P_{t-3})$	(6)
2	$Q = f(P_{t-1}, P_{t-2}, P_{t-3}, SWI)$	(7)
3	$Q = f(P_{t-1}, P_{t-2}, P_{t-3}, P_{acc120d})$	(8)
4	$Q = f(P_{t-1}, P_{t-2}, P_{t-3}, SWI, P_{acc150})$	(9)
5	$Q = f(P_{t-1}, P_{t-2}, P_{t-3}, P_{acc120d}, P_{acc150})$	(10)
6	$Q = f(P_{t-1}, P_{t-2}, P_{t-3}, SWI, P_{acc150}, P_{acc10})$	(11)
7	$Q = f(P_{t-1}, P_{t-2}, P_{t-3}, P_{acc120d}, P_{acc150}, P_{acc10})$	(12)
8	$Q = f(P_{t-1}, P_{t-2}, P_{t-3}, P_{acc120d}, P_{acc150}, P_{acc10}, P_{acc30}, P_{acc60}, P_{acc90})$	(13)

Q is the daily runoff; P_t is the rainfall data at t day. P_{accN} is the accumulated rainfall at N day.

Table 2 | NSE metric results for all models

Rainfall station	GPM				CHIRPS							
	Training		Testing		Training		Testing					
	RF	ANN	RF	ANN	RF	ANN	RF	ANN				
Case 1	0.83	0.43	0.25	0.35	0.89	0.30	0.08	0.20	0.90	0.38	0.31	0.33
Case 2	0.95	0.69	0.66	0.60	0.95	0.49	0.57	0.49	0.95	0.66	0.65	0.54
Case 3	0.96	0.69	0.71	0.67	0.96	0.72	0.64	0.70	0.95	0.68	0.64	0.61
Case 4	0.97	0.80	0.81	0.71	0.96	0.68	0.78	0.68	0.96	0.74	0.74	0.62
Case 5	0.96	0.81	0.78	0.71	0.97	0.72	0.72	0.70	0.95	0.72	0.69	0.62
Case 6	0.98	0.83	0.88	0.76	0.97	0.75	0.84	0.77	0.97	0.79	0.83	0.67
Case 7	0.98	0.88	0.85	0.75	0.97	0.78	0.80	0.77	0.96	0.82	0.77	0.66
Case 8	0.99	0.95	0.91	0.82	0.98	0.82	0.87	0.82	0.98	0.90	0.86	0.70
Case 8*		0.91		0.85		0.85		0.77		0.93		0.73

Predictor in Case 8* was used only for the combined-ANN model.

Table 3 | RMSE metric results for all models

Rainfall station	GPM				CHIRPS							
	Training		Testing		Training		Testing					
	RF	ANN	RF	ANN	RF	ANN	RF	ANN				
Case 1	23.91	43.10	43.24	40.35	18.63	48.00	47.99	44.60	18.23	44.97	41.63	41.02
Case 2	12.27	31.97	28.97	31.72	13.11	40.78	32.87	35.65	12.27	33.46	29.56	33.97
Case 3	11.82	31.80	27.00	28.89	11.20	30.34	29.87	27.59	13.18	32.61	29.88	31.40
Case 4	9.27	25.44	21.79	26.86	10.81	32.50	23.65	28.48	11.19	29.34	25.55	30.63
Case 5	10.80	25.21	23.59	27.02	10.36	30.15	26.28	27.43	12.86	30.32	28.03	30.88
Case 6	8.26	23.35	17.03	24.51	9.11	28.48	19.98	24.11	9.71	26.15	20.50	28.69
Case 7	8.53	19.57	19.53	24.85	9.36	27.04	22.19	23.79	10.80	24.46	23.80	28.95
Case 8	6.99	13.36	15.09	21.06	8.48	24.05	18.26	21.48	8.49	18.04	18.48	27.33
Case 8*		16.64		22.39		22.93		22.29		16.54		26.08

Predictor in Case 8* was used only for the combined-ANN model.

Case 1, based on the preceding 3-day rainfall, was inadequate for both the RF and ANN models. Utilizing solely antecedent rainfall data did not comprehensively represent the intricate runoff generation process. Although these outcomes might have corresponded to peak flows during rainy periods, they displayed excessive fluctuations and failed to accurately capture the base flow response, as shown in Figure 7.

Therefore, more comprehensive predictor variables were required to enhance the accuracy and performance of the models. Case 2 showed an improvement in the accuracy for estimating runoff in both rainy and dry periods. In this case, the RF model performed slightly better than the ANN model (NSE values of 0.66 and 0.60, respectively). Likewise, in Case 3, the RF model outperformed the ANN model, obtaining NSE values of 0.71 and 0.67, respectively. These findings underscored the importance of both the SWI and the 120-day accumulated rainfall as they contributed to improving the efficiency of estimating runoff, including baseflow during dry periods. Cases 4 and 5 provided additional improvements by incorporating the 150-day accumulated rainfall, representing changes in the RC. In the RF model, Cases 4 and 5 had higher NSE values of 0.81 and 0.78, respectively. Conversely, the ANN model had NSE values of 0.71 for both candidates. Furthermore, Cases 6 and 7 demonstrated enhanced efficiency with the incorporation of the 10-day accumulated rainfall. The RF model for Cases 6 and 7 achieved notable NSE values of 0.88 and 0.85, respectively, while the ANN model achieved NSE values of 0.76 and 0.75, respectively. Finally, Case 8 outperformed all the case scenarios above, producing NSE values for the RF and ANN models of 0.91 and 0.82, respectively, and RMSE values of 15.09 and 21.06 mm-day⁻¹, respectively. Case 8 was selected as a predictor for the combined-ANN model, resulting in an enhanced NSE of 0.85 but a slightly higher RMSE of 22.39 compared with the single ANN model. Overall, SWI plays a crucial role in improving accuracy, particularly in the RF model, but not as much in the ANN model. However, SWI can be replaced by 120-day accumulated rainfall as an input, resulting in slightly lower accuracy than the SWI.

Furthermore, Figures 8 and 9 illustrate the runoff results based on Cases 6 and 8, respectively. The RF model provided robust runoff estimations for baseflow compared with the ANN model. Due to the RF model's proficiency in separating

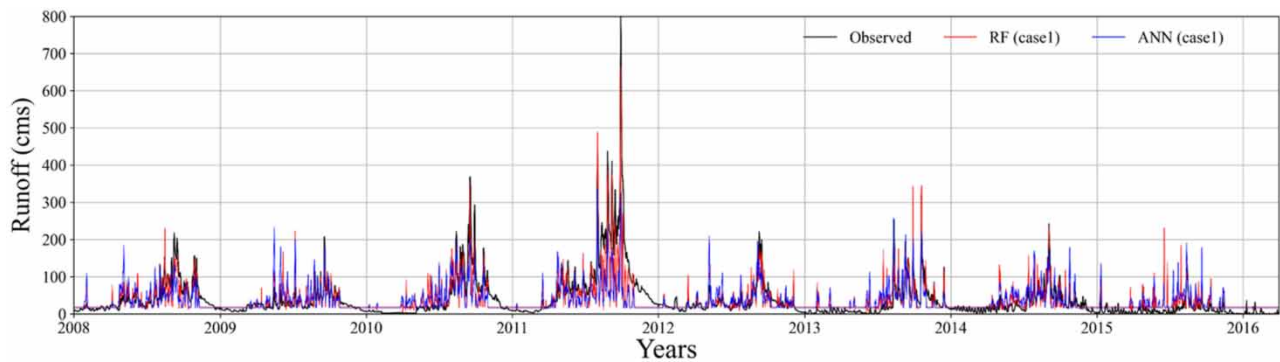


Figure 7 | Comparison of daily runoff between observed data and the results from the ANN and RF models using Case 1 as input.

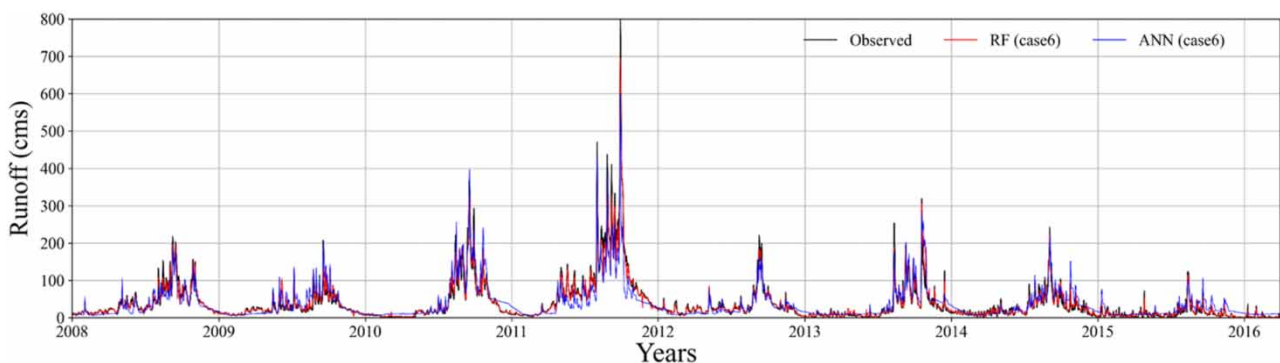


Figure 8 | Comparison of daily runoff between observed data and results from the ANN and RF models using Case 6 as input.

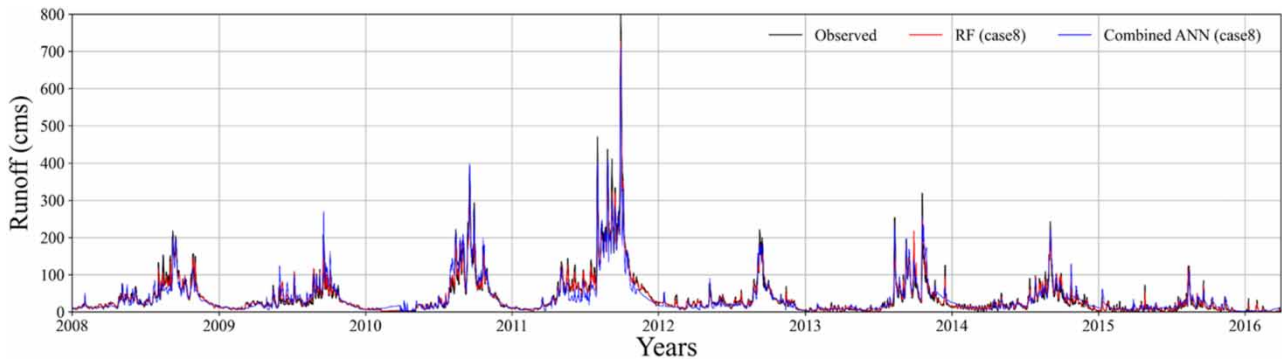


Figure 9 | Comparison of daily runoff between observed data and results from the combined-ANN and RF models using Case 8 as input.

the variations and patterns of runoff, it produced precise runoff forecasts, especially for low flow conditions. However, notably, both models underestimated the runoff during peak flow periods.

Efficiency of rainfall products

Table 4 presents the performance of the rainfall outputs based on four metrics. The POD and CSI for GPM were slightly higher than those for CHIRPS, with POD values of 0.93 and 0.9, and CSI values of 0.9 and 0.87 for GPM and CHIRPS, respectively. The CSI for GPM was slightly lower than for CHIRPS, with CSI values of 0.03 and 0.04, respectively. The RMSE values for GPM and CHIRPS were 2.65 and 3.43 mm·day⁻¹, respectively. These results suggested that the GPM estimates were closer to the observed rainfall and provided better detection than CHIRPS.

In Tables 2 and 3, the testing phase for Case 8 produced similar runoff forecasting results using the RF model with GPM and CHIRPS data, with the highest NSE values of 0.87 for GPM and 0.86 for CHIRPS, respectively. By contrast, in the case using the ANN model, the GPM data outperformed CHIRPS with NSE values of 0.82 (GPM) and 0.70 (CHIRPS). Notably, the combined-ANN model based on RF-GPM had slightly poorer performance with an NSE of 0.77 compared with the single ANN. The combined-ANN model based on RF-CHIRPS with an NSE value of 0.73 had better performance than the single ANN. Furthermore, Case 6 was a potent predictor, with the RF model having NSE values of 0.84 (GPM) and 0.83 (CHIRPS). However, Case 6 did not produce robust outcomes using the ANN model, with NSE values of 0.77 (GPM) and 0.67 (CHIRPS). Figure 10 illustrates the runoff results from GPM based on Case 8, providing similar observed data but slightly underestimating during peak flows. Conversely, CHIRPS produced more fluctuations compared with GPM, as shown in Figure 11.

DISCUSSION

The novel selection of predictors that we proposed seemed to offer a more systematic and principled approach, whereas some studies have applied trial-and-error or random methods for predictor selection (Ali & Shahbaz 2020; Mei & Smith 2021). The selection of predictors based on specific criteria aimed to enhance the accuracy and applicability of the model. These methods do not require data about previous runoff, soil moisture, groundwater, land use, or soil type. Integrating the SWI as a predictor in machine learning models for runoff prediction represents an innovative approach. The SWI provided a substantial contribution in enhancing prediction accuracy. Similarly, the 10- and 150-day accumulated rainfall data effectively represented the runoff tendency. However, notably, these time intervals may not represent exact patterns for every location. Nevertheless, this approach could be applied to identify the predictor for each specific location.

Table 4 | Performance of the rainfall products

CHIRPS				GPM			
POD	FAR	CSI	RMSE	POD	FAR	CSI	RMSE
0.90	0.04	0.87	3.43	0.93	0.03	0.90	2.65

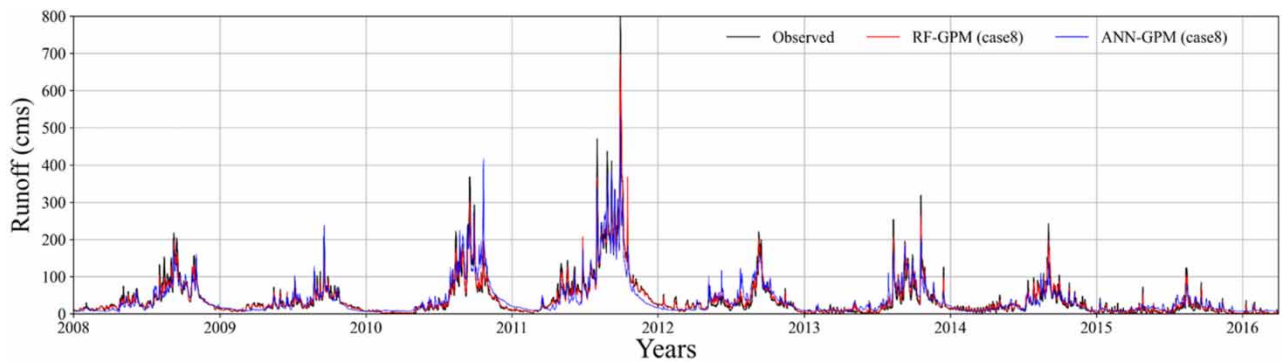


Figure 10 | Comparison of daily runoff between observed data and results from the RF and ANN models based on GPM using Case 8.

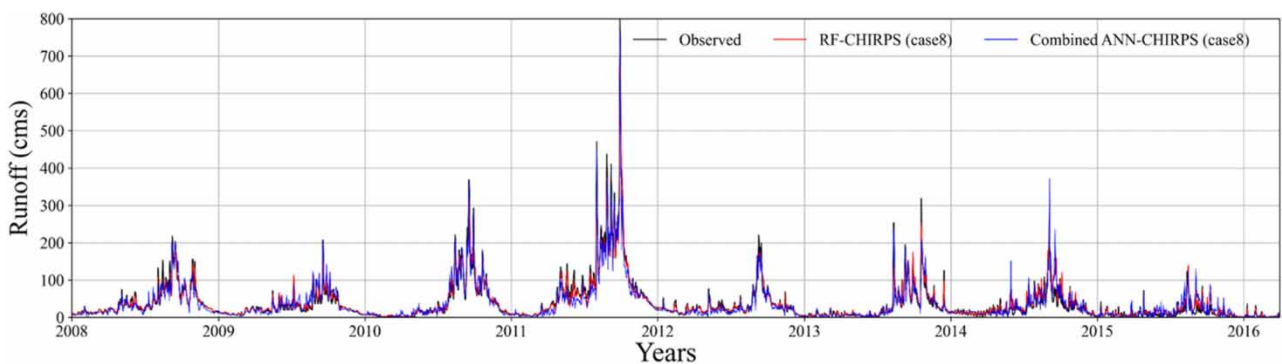


Figure 11 | Comparison of daily runoff between observed data and results from RF and ANN models based on CHIRPS using Case 8.

The success of the proposed method also opens up new possibilities for its application in other areas where soil moisture or groundwater measurement data are scarce or not readily available. The novel approach of incorporating the SWI with accumulated rainfall, or using accumulated rainfall alone, appeared to produce better performance compared with the previously used conceptual model in the region. The FlexL model produced an NSE of about 0.8 for testing at P.1, while the best current findings of this study produced an NSE of 0.91 for testing (Sriwongsitanon *et al.* 2023). Furthermore, the performance of these methods was comparable to that of other models that required multiple inputs and techniques for splitting baseflow. For example, the convolutional neural networks (CNN) model used additional inputs, such as groundwater, land cover, and soil maps (Song 2022). The ANN and RF models utilized baseflow separation techniques (Tongal & Booij 2018; Song 2022). Some studies may require antecedent runoff data to improve accuracy (Qiao *et al.* 2023; Vaheddoost *et al.* 2023). Therefore, the current findings suggest that the proposed methods are particularly valuable, especially in areas where comprehensive data sources are limited and rely solely on rainfall data.

The RF model accurately estimated the runoff in both the dry and rainy seasons from internal features, separating characteristics of runoff occurrence. In addition, the RF model demonstrated that using Case 6 as a predictor was sufficient to achieve a prediction with high accuracy. By contrast, the ANN model required Case 8 as a predictor to attain comparable levels of accuracy. Additionally, the combined-ANN model developed by segregating the runoff components enhanced accuracy compared with the single ANN model. Nonetheless, even with this improvement, the combined-ANN model still fell behind the performance of the RF model.

Considering the overfitting problem, the RF model tended to more readily overfit than the ANN and combined-ANN models, since the RF model often produced very high NSE values of above 0.98 in the training phase, while in the testing phase, it had an NSE value in the range 0.25–0.91. Conversely, the ANN model typically produced closer performance results between the training and testing data, with NSE values of 0.91 and 0.85, respectively. However, the RF model still provided better results in the testing phase than the ANN model even though the latter model generalized unseen data better.

Based on the GPM and CHIRPS datasets, the RF model demonstrated a remarkable degree of flexibility with input data. This was evident even though the RF used distinct rainfall products, with the results producing close values for both NSE and

RMSE. By contrast, the ANN produced substantial differences in performance when utilizing GPM data compared with CHIRPS data. The results based on the GPM data indicated comparable efficiency to those obtained using rainfall station data.

CONCLUSIONS

A methodology was successfully developed to select predictors to represent the characteristics of runoff, addressing the limitation of estimating runoff solely based on rainfall data in a data-driven model. The accumulation of rainfall can be used to represent runoff characteristics. The utilization of accumulated rainfall results shows a significant improvement in the accuracy of the models. Meanwhile, SWI demonstrates its importance as a factor that can enhance the accuracy of both models. Using the GPM and CHIRPS datasets instead of rainfall station data enhanced the applicability of the models in areas without any rain gauge data. The daily runoff predicted by the RF model using GPM was more accurate than the others. The runoff results from the RF and ANN models slightly underestimated peak flow, suggesting a limitation in accurately predicting extreme runoff events. However, the limitation of this rainfall–runoff model is that it does not account for changes in land use, soil type, or the impact of human activities such as agriculture. Therefore, incorporating these factors into the model may enhance the accuracy of runoff estimation.

ACKNOWLEDGMENTS

The authors heartily thank the Land Development Department for the topography map and the Royal Irrigation Department for the rainfall and runoff data.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

- Albergel, C., Rüdiger, C., Pellarin, T., Calvet, J. C., Fritz, N., Froissard, F., Suquia, D., Petitpa, A., Pignatelli, B. & Martin, E. 2008 From near-surface to root-zone soil moisture using an exponential filter: An assessment of the method based on in-situ observations and model simulations. *Hydrology and Earth System Sciences* **12** (6), 1323–1337.
- Ali, S. & Shahbaz, M. 2020 Streamflow forecasting by modeling the rainfall–streamflow relationship using artificial neural networks. *Modeling Earth Systems and Environment* **6**, 1645–1656.
- Bao, A. M., Liu, H. L., Chen, X. & Pan, X. L. 2011 The effect of estimating areal rainfall using self-similarity topography method on the simulation accuracy of runoff prediction. *Hydrological Processes* **25** (22), 3506–3512.
- Bouaziz, L. J., Fenicia, F., Thirel, G., de Boer-Euser, T., Buitink, J., Brauer, C. C., Niel, J. D., Dewals, B. J., De, G., Grelier, B., Melsen, L. A., Moustakas, S., Nossent, J., Pereira, F., Sprokkereef, E., Stam, J., Weerts, A. H., Willems, P., Savenije, H. G. & Hrachowitzrogu, M. 2021 Behind the scenes of streamflow model performance. *Hydrology and Earth System Sciences* **25** (2), 1069–1095.
- Foelsche, U., Kirchengast, G., Fuchsberger, J., Tan, J. & Petersen, W. A. 2017 Evaluation of GPM IMERG early, late, and final rainfall estimates using WegenerNet gauge data in southeastern Austria. *Hydrology and Earth System Sciences* **21** (12), 6559–6572.
- Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A. & Michaelsen, J. 2015 The climate hazards infrared precipitation with stations – a new environmental record for monitoring extremes. *Scientific Data* **2** (1), 1–21.
- Gao, H., Hrachowitz, M., Fenicia, F., Gharari, S. & Savenije, H. H. G. 2014 Testing the realism of a topography-driven model (FLEX-Topo) in the nested catchments of the Upper Heihe, China. *Hydrology and Earth System Sciences* **18** (5), 1895–1915.
- Hosseini, S. M. & Mahjouri, N. 2016 Integrating support vector regression and a geomorphologic artificial neural network for daily rainfall-runoff modeling. *Applied Soft Computing* **38**, 329–345.
- Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M. & Iguchi, T. 2014 The global precipitation measurement mission. *Bulletin of the American Meteorological Society* **95** (5), 701–722.
- Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., Hong, Y., Bowman, K. P. & Stocker, E. F. 2007 The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology* **8** (1), 38–55.
- Huo, J., Zhang, Y., Luo, L., Long, Y., He, Z. & Liu, L. 2017 Model parameter optimization method research in Heihe River open modeling environment (HOME). *International Journal of Pattern Recognition and Artificial Intelligence* **31** (09), 1759017.

- Joyce, R. J., Janowiak, J. E., Arkin, P. A. & Xie, P. 2004 CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *Journal of Hydrometeorology* **5** (3), 487–503.
- Khan, M. T., Shoaib, M., Hammad, M., Salahudin, H., Ahmad, F. & Ahmad, S. 2021 Application of machine learning techniques in rainfall-runoff modelling of the Soan River basin, Pakistan. *Water* **13** (24), 3528.
- Kim, J. & Han, H. 2021 Evaluation of the CMORPH high-resolution precipitation product for hydrological applications over South Korea. *Atmospheric Research* **258**, 105650.
- Lakshmi, P. J., Apaza, R. A., Alkhayyat, A., Marhoon, H. A. & Alameri, A. A. 2022 Hybrid wavelet-gene expression programming and wavelet-support vector machine models for rainfall-runoff modeling. *Water Science & Technology* **86** (12), 3205–3222.
- Maier, H. R., Jain, A., Dandy, G. C. & Sudheer, K. P. 2010 Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environmental Modelling & Software* **25** (8), 891–909.
- Mao, G., Wang, M., Liu, J., Wang, Z., Wang, K., Meng, Y., Zhong, R., Wang, H. & Li, Y. 2021 Comprehensive comparison of artificial neural networks and long short-term memory networks for rainfall-runoff simulation. *Physics and Chemistry of the Earth, Parts A/B/C* **123**, 103026.
- Mei, X. & Smith, P. K. 2021 A comparison of in-sample and out-of-sample model selection approaches for artificial neural network (ANN) daily streamflow simulation. *Water* **13** (18), 2525.
- Mishra, P. K. & Karmakar, S. 2019 Performance of optimum neural network in rainfall-runoff modeling over a river basin. *International Journal of Environmental Science and Technology* **16**, 1289–1302.
- Mohammadi, B. 2021 A review on the applications of machine learning for runoff modeling. *Sustainable Water Resources Management* **7** (6), 98.
- Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models part I – a discussion of principles. *Journal of Hydrology* **10** (3), 282–290.
- Patel, A. B. & Joshi, G. S. 2017 Modeling of rainfall-runoff correlations using artificial neural network – a case study of Dharoi Watershed of a Sabarmati river basin, India. *Civil Engineering Journal* **3** (2), 78–87.
- Paulik, C., Dorigo, W., Wagner, W. & Kidd, R. 2014 Validation of the ASCAT soil water index using in situ data from the international soil moisture network. *International Journal of Applied Earth Observation and Geoinformation* **30**, 1–8.
- Qiao, X., Peng, T., Sun, N., Zhang, C., Liu, Q., Zhang, Y., Liu, Q., Zhang, Y., Wang, Y. & Nazir, M. S. 2023 Metaheuristic evolutionary deep learning model based on temporal convolutional network, improved aquila optimizer and random forest for rainfall-runoff simulation and multi-step runoff prediction. *Expert Systems with Applications* **229** (Part B), 120616.
- Ramadhan, R., Yusnaini, H., Marzuki, M., Muharsyah, R., Suryanto, W., Sholihun, S., Vonnisa, M., Harmadi, H., Ningsih, A. P., Battaglia, A., Hashiguchi, H. & Tokay, A. 2022 Evaluation of GPM IMERG performance using gauge data over Indonesian maritime continent at different time scales. *Remote Sensing* **14** (5), 1172.
- Rangsiwanichpong, P. & Melesse, A. M. 2022 Analyzing the benefit-cost ratio of sediment resources by remote sensing data in the Ping River Basin, Thailand. *Water* **14** (13), 2071.
- Sayed, B. T., Al-Mohair, H. K., Alkhayyat, A., Ramírez-Coronel, A. A. & Elshahabi, M. 2023 Comparing machine-learning-based black box techniques and white box models to predict rainfall-runoff in a northern area of Iraq, the Little Khabor River. *Water Science and Technology* **87** (3), 812–822.
- Schoppa, L., Disse, M. & Bachmair, S. 2020 Evaluating the performance of random forest for large-scale flood discharge simulation. *Journal of Hydrology* **590**, 125531.
- Shoaib, M., Shamseldin, A. Y., Melville, B. W. & Khan, M. M. 2016 A comparison between wavelet based static and dynamic neural network approaches for runoff prediction. *Journal of Hydrology* **535**, 211–225.
- Song, C. M. 2022 Data construction methodology for convolution neural network based daily runoff prediction and assessment of its applicability. *Journal of Hydrology* **605**, 127324.
- Sriwongsitanon, N., Jandang, W., Williams, J., Suwawong, T., Maekan, E. & Savenije, H. H. 2023 Using normalised difference infrared index patterns to constrain semi-distributed rainfall-runoff models in tropical nested catchments. *Hydrology and Earth System Sciences* **27** (11), 2149–2171.
- Tang, G., Clark, M. P., Papalexiou, S. M., Ma, Z. & Hong, Y. 2020 Have satellite precipitation products improved over last two decades? A comprehensive comparison of GPM IMERG with nine satellite and reanalysis datasets. *Remote Sensing of Environment* **240**, 111697.
- Tongal, H. & Booij, M. J. 2018 Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. *Journal of Hydrology* **564**, 266–282.
- Vaheddoost, B., Safari, M. J. S. & Yilmaz, M. U. 2023 Rainfall-runoff simulation in ungauged tributary streams using drainage area ratio-based multivariate adaptive regression spline and random forest hybrid models. *Pure and Applied Geophysics* **180** (1), 365–382.
- Ye, X., Guo, Y., Wang, Z., Liang, L. & Tian, J. 2022 Extensive evaluation of four satellite precipitation products and their hydrologic applications over the Yarlung Zangbo River. *Remote Sensing* **14** (14), 3350.
- Yong, B., Ren, L. L., Hong, Y., Wang, J. H., Gourley, J. J., Jiang, S. H., Chen, X. & Wang, W. 2010 Hydrologic evaluation of multisatellite precipitation analysis standard precipitation products in basins beyond its inclined latitude band: A case study in Laohahe basin, China. *Water Resources Research* **46** (7), W07542.
- Zhang, Y., Ju, Q., Zhang, L., Xu, C. Y. & Lai, X. 2022 Evaluation and hydrological application of four gridded precipitation datasets over a large Southeastern Tibetan Plateau Basin. *Remote Sensing* **14** (12), 2936.