

Article

A Study on Advancing Runoff Prediction Through Peak Protection Selective Learning in U.S. CAMELS Basins

Pengfei Xing, Kaiwei Zhang *, Qingtian Geng, Xiaoxiao Ma, Xiaochun Jin, Jing Wang, Yuguang Yan, Xiaoning Li and Qingliang Li 

School of Artificial Intelligence, Changchun Normal University, Changchun 130032, China; xingpengfei@ccsfu.edu.cn (P.X.); qtgeng@ccsfu.edu.cn (Q.G.); 18237991314@163.com (X.M.); debs222@163.com (X.J.); wangjing2738@ccsfu.edu.cn (J.W.); yanyuguang@ccsfu.edu.cn (Y.Y.); xiaonin4523@163.com (X.L.); liqingliang@ccsfu.edu.cn (Q.L.)

* Correspondence: zhangkw0022@163.com

Abstract

Accurate prediction of extreme runoff events is crucial for flood control and disaster mitigation and water resource risk management. In recent years, deep learning has emerged as a key method for runoff simulation. However, its training strategy applies uniform optimization across all time steps, which encourages the model to fit normal hydrological processes while reducing its sensitivity to extreme runoff events. To address this issue, this study introduces Selective Learning and proposes a runoff prediction model with a flood peak protection mechanism, namely PPSL-DI-LSTM. In this framework, sample importance is dynamically regulated through uncertainty filtering, anomaly filtering, and flood peak protection, thereby reducing the influence of unstable or non-generalizable samples while enhancing the learning of high-flow processes. Based on this strategy, a dynamic weighted loss function, termed Selective Weighted RMSE (SW-RMSE), is designed to optimize the training process. Experimental results on the CAMELS dataset show that, compared to the baseline DI-LSTM model, the proposed model achieves improvements of approximately 4% and 6% in median Nash-Sutcliffe Efficiency (NSE) and median Kring-Gupta Efficiency (KGE), respectively. Simultaneously, the Flow Duration Curve High Flow Variability (FHV) was significantly reduced by approximately 84% and the overall percent bias (PBIAS) improved from -3.80% to 1.15% . Additionally, when integrated into a standard LSTM, PPSL maintains comparable median NSE while improving median KGE (from 0.751 to 0.778) and substantially reducing FHV (from -12.86% to -3.34%). Our method overcomes the limitations of uniformly treating all training samples in complex hydrological simulations, and has strong potential to substantially improve the accuracy and reliability of future runoff modeling studies.



Academic Editor: Ataur Rahman

Received: 5 June 2026

Revised: 22 July 2026

Accepted: 24 July 2026

Published: 28 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: selective learning; runoff prediction; deep learning; Long Short-Term Memory (LSTM)

1. Introduction

Accurate prediction of peak runoff and other extreme hydrological events is essential for flood warning, reservoir operation, and water-resources risk management [1–3]. However, compared with ordinary flow conditions, extreme runoff events are sparse, rapidly varying, and strongly non-stationary, making them much harder to model. In practical datasets, normal-flow samples dominate the time series, whereas flood peaks and abrupt

rising limbs account for only a small fraction of observations but are far more important for decision making [4]. This imbalance often causes prediction models to perform reasonably well under average conditions while remaining biased or unstable during critical high-flow periods.

Traditional hydrological models and conceptual rainfall–runoff models—such as the physics-based SWAT model and conceptual models like HBV and SAC-SMA—offer strong physical interpretability. However, their performance often relies on basin-specific assumptions and extensive parameter calibration, which can limit their robustness in watersheds characterized by highly uneven spatial distributions and complexity, particularly under extreme event conditions [5]. Conventional machine learning models, such as Support Vector Machine (SVM) and Random Forest (RF), can capture nonlinear relationships to some extent but typically depend on manual feature engineering [6]. These models often remain inadequate in representing long-term temporal dependencies, catchment lag responses, and complex multivariate interactions. In contrast, various deep learning architectures can improve streamflow prediction performance from different perspectives [7]. For example, CNN models are effective at extracting local variation features from multivariate inputs, thereby enhancing the representation of short-term runoff responses [8]. The Transformer architecture captures long-term temporal dependencies through self-attention mechanisms [9]. In particular, recurrent neural networks such as LSTM are well-suited for rainfall–runoff prediction, as they can directly learn nonlinear input–output mappings from data and effectively capture the memory effects and time lags inherent in hydrological processes [10–13].

LSTM-based models have become one of the most widely used deep learning approaches in rainfall–runoff modelling because their gated memory structure can capture nonlinear catchment responses, cumulative effects, and delayed runoff generation. Kratzert et al. (2018) applied Long Short-Term Memory (LSTM) networks to rainfall–runoff modelling [14], utilising gating mechanisms to effectively capture long-term dependencies. On the CAMELS dataset (531 catchments), they achieved a median NSE of approximately 0.63, representing an improvement of around 8–10% over the traditional conceptual model SAC-SMA, indicating that LSTM possesses superior simulation accuracy and generalisation capabilities. Building on this, researchers have further enhanced the expressive power of RNNs through various improvements: Deng et al. (2022) proposed a CNN-LSTM hybrid model that uses CNN layers to extract local features and LSTM layers to capture temporal evolution [15]. Applied to the Feilaixia catchment for daily runoff forecasting, the model achieved an NSE of 0.838, outperforming standalone CNN (0.737) and LSTM (0.745), demonstrating the effectiveness of integrating spatial feature extraction with sequential modelling. Vaibhav Tripathi et al. developed a unified En-LSTM framework that couples conventional LSTM with TCN, attention mechanisms, and a peak-aware hybrid loss, incorporating multi-scale lagged and sliding window features. It attained KGE values of 0.84 and 0.71 during training and testing, respectively, while achieving $\text{NSE} \approx 0.79$ and $\text{KGE} \approx 0.76$ at completely unseen stations, indicating robust transferability. Although these LSTM variants improve overall simulation accuracy, their MSE-based optimization weights all time steps equally and lacks targeted constraints on peaks, leading to systematic underestimation of peak flow [16]. Hence, improved sequential modeling alone does not guarantee accurate peak flow prediction.

To improve the prediction of extreme runoff events, particularly peak flows and rapid rising processes, researchers have conducted targeted explorations in recent years from multiple perspectives, including model architecture, loss function design, sample balancing, and event screening. Chen et al. (2022) proposed the LSTM-ALSL model, which applies self-attention to both long-lag and short-lag inputs. Across four MOPEX basins in the

United States [17], it improved the 7-day NSE by 3.4–9.0% and reduced the average relative error (ARE) of peak flows by 16.9–49.5% compared with the standard LSTM, indicating that attention effectively enhances the extraction of critical information near the peak. Girihagama et al. (2022) developed an attention-based encoder–decoder LSTM for daily runoff forecasting in 10 basins of the Ottawa River, Canada, which raised the average test NSE from 0.668 to 0.985 and reduced RMSE from 40.39 to 8.2 m³/s [18]. However, slight underestimation persisted during spring snowmelt and some annual peak events, suggesting that structural improvements alone cannot fully eliminate biases in extreme flow prediction. In essence, these attention-based methods enhance feature extraction at the architectural level, but they do not explicitly differentiate the training importance of individual time steps within a runoff sequence.

To further address this challenge, Man et al. (2023) developed two peak-sensitive loss functions, Peak Error Tanh (PET) and Peak Error Swish (PES), for LSTM-based runoff prediction. The PES-based model achieved a comparable overall NSE (0.917) while improving flood-period NSE from 0.747 to 0.873 and flood peak qualification rate (QR) from 77.8% to 92.3%, demonstrating the effectiveness of peak-focused loss weighting for improving extreme runoff prediction [19]. PET and PES nonlinearly transform prediction residuals to penalize peak-flow errors more than ordinary-flow errors. However, they operate at the loss level and do not explicitly identify informative time steps. Similarly, Li et al. (2024) proposed an extreme-adaptive model (DAN) that integrates multi-loss weighting, probability modeling, and imbalance-aware sampling to improve extreme flow prediction. The model reduced relative RMSE by 19% overall and 30% on the UpperPen dataset, demonstrating the potential of extreme-adaptive strategies for characterizing extreme flows [20]. DAN enhances the learning of rare extreme-flow samples by jointly adjusting optimization targets and sample distributions. It therefore addresses imbalance at the sample/distribution level rather than the time-step level. Furthermore, Tofighi et al. (2025) combined event-based active learning using greedy selection with an asymmetric peak (AP) loss in LSTM-based streamflow forecasting. By selecting 40 representative rainfall events from 200 candidates, the model achieved NSE values of 0.93 and 0.88 for 1 h and 12 h lead times, respectively, with only marginal degradation compared with using all events. Incorporating AP loss further improved high-flow prediction, increasing high-flow NSE from 0.12 to 0.76 and from −0.90 to 0.62 for 1 h and 12 h forecasts, respectively [21]. Greedy selection is used to choose representative rainfall events, and AP loss emphasizes underestimation during high-flow periods. This improves training efficiency and peak fitting, but remains event-level selection. Overall, existing studies have shown that attention mechanisms, peak-sensitive losses, extreme-flow-aware optimization, and event-based sample selection can improve extreme runoff prediction.

However, these methods mainly focus on model structures, loss functions, or sample selection, while overlooking the unequal contributions of different time steps within a sequence. This motivates the development of time-step-based selective learning.

Selective learning offers a promising solution to the above challenge by identifying training samples that contribute most to model generalization [22], thereby reducing the influence of noisy or redundant information. This strategy has demonstrated its effectiveness in various domains. For example, in medical image analysis, a two-stage sample selection strategy effectively suppresses noise propagation and improves model robustness [23]. In time-series forecasting, Fu et al. proposed a time-step selection method based on predictive uncertainty, which significantly reduced forecasting errors [22]. This finding is particularly relevant to runoff forecasting, where key hydrological information is often concentrated in a limited number of critical time steps. These studies suggest that differentiating the contributions of training samples can improve both model stability

and generalization. Inspired by these advances, this study introduces selective learning into rainfall–runoff modeling at the time-step level to alleviate the learning bias caused by the scarcity of extreme runoff samples. However, hydrological processes are characterized by strong non-stationarity, seasonality, catchment heterogeneity [24], and the rarity of extreme events, making the identification of informative time steps considerably more challenging. Therefore, incorporating selective learning into runoff prediction has the potential to enhance the model’s ability to capture extreme runoff processes while simultaneously improving overall prediction accuracy and cross-catchment generalization.

Based on the above analysis, this paper introduces a selection learning strategy into the runoff forecasting task, with the aim of enhancing the generalisation ability and the ability to capture extreme events in deep learning models by improving the training paradigm. The core contributions are: (1) This study introduces a model-agnostic selective learning framework for rainfall–runoff prediction that improves how existing deep-learning models are optimized during training. By dynamically identifying informative time steps while reducing the influence of low-contribution samples, the proposed framework mitigates the optimization bias associated with conventional uniformly weighted loss functions and enables more effective learning of hydrologically important information, particularly during peak-flow periods. (2) The proposed PPSL framework consistently improves the simulation of extreme runoff events while maintaining or further enhancing overall predictive performance. Integrated with both LSTM and DI-LSTM, it significantly enhances the prediction of peak flows and rapid runoff variations without requiring architectural modifications. Extensive experiments on the CAMELS dataset demonstrate its robust performance and strong cross-basin generalization under diverse, non-stationary hydrological conditions.

2. Materials and Methods

2.1. Data Description

This study utilises the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS, USA) dataset, a large-scale catchment dataset originating from the USA and widely used in the field of hydrology [25]. This dataset encompasses a wealth of meteorologically driven data, hydrological observations and multi-dimensional catchment attribute information, all of which are publicly available. Given its extensive geographical coverage, long time series span and diverse hydro-climatic conditions, the CAMELS dataset has become the standard benchmark for evaluating the performance and generalisation ability of data-driven runoff prediction models [26]. Unlike many nationally developed hydrological datasets, which are typically designed for regional operational applications and therefore vary considerably in spatial coverage, data length, and processing standards, CAMELS provides a uniformly processed and openly accessible dataset across a large number of catchments. This level of standardization is essential for systematically evaluating the general applicability and cross-catchment generalization capability of the proposed framework, rather than focusing on region-specific operational accuracy. Accordingly, CAMELS is adopted in this study as a standardized benchmark dataset to evaluate the proposed framework for runoff and extreme runoff prediction, while facilitating fair comparisons with previous studies.

The CAMELS dataset comprises 671 small- to medium-sized catchments in the continental United States that have been minimally affected by human activity. These catchments exhibit significant variations in topography, climate and hydrological characteristics, with the study area encompassing a wide range of biogeographical types, from mountainous regions to plains, and from arid to humid zones. The dataset provides daily scale meteorological driver data and runoff observations for the period from 1980 to 2014, whilst also encompassing six major categories of static watershed attributes: topography, climate,

runoff, land cover, soil and geology. This study takes all 671 catchments as its research subjects, with model inputs comprising both meteorological variables and static catchment attributes [27]. The meteorological variables include precipitation, solar radiation, maximum temperature, minimum temperature, vapour pressure and duration of sunshine [28]; specific variable information is shown in Table 1 [29]. The model's prediction target is daily scale runoff.

Table 1. Overview of input data based on the CAMELS dataset [29].

	Variable Name	Description	Unit
Forcings	PRCP	Precipitation	mm/day
	SRAD	Solar radiation	W/m ²
	Tmax	Maximum temperature	°C
	Tmin	Minimum temperature	°C
	Vp	Vapor pressure	Pa
	Dayl	Day length	s
Attributes	Elev_mean	Catchment mean elevation	m
	Slope_mean	Catchment mean slope	m/km
	Area_gages2	Catchment area	km ²
	Frac_forest	Forest fraction	-
	Lai_max	Maximum monthly mean of the leaf area index	-
	Lai_diff	Monthly mean leaf area index amplitude	-
	Dom_land_cover_frac	Fraction of the catchment area	-
	Dom_land_cover	Dominant land cover type	-
	Root_depth_50	Root depth at 50th percentile	m
	Soil_depth_statsgo	Soil depth	m
	Soil_porosity	Volumetric soil porosity	-
	Soil_conductivity	Saturated hydraulic conductivity	cm/h
	Max_water_content	Maximum water content	m
	Geol_class_1st	Most common geologic class	-
	Geol_class_2nd	Second most common geologic class	-
	Geol_porosity	Subsurface porosity	-
	Geol_permeability	Subsurface permeability	m ²

Figure 1 and Appendix A Figure A1 show the spatial distributions of the major catchment attributes in the CAMELS dataset. In the main text, three representative catchment attributes—aridity, slope, and forest cover—are presented to characterize the climatic, topographic, and land-cover conditions of the study basins, while the spatial distributions of the remaining attributes are provided in Appendix A Figure A1. During both model training and prediction, the proposed model utilizes all 17 static catchment attributes provided by the CAMELS dataset together with meteorological forcings for both normal-flow and extreme-runoff prediction, rather than relying solely on these three representative attributes. To ensure consistency with previous CAMELS-based rainfall-runoff studies, particularly the DI-LSTM benchmark adopted in this work, and to enable a fair comparison between different methods, the available data were chronologically divided into training, validation, and testing periods. Specifically, the model was trained from 1 October 1985 to 30 September 1994, validated from 1 October 1994 to 30 September 1995, and finally evaluated on an independent testing period from 1 October 1995 to 30 September 2005. The validation dataset was used for hyperparameter tuning, model selection, and early stopping, while the testing dataset was used exclusively for the final performance evaluation.

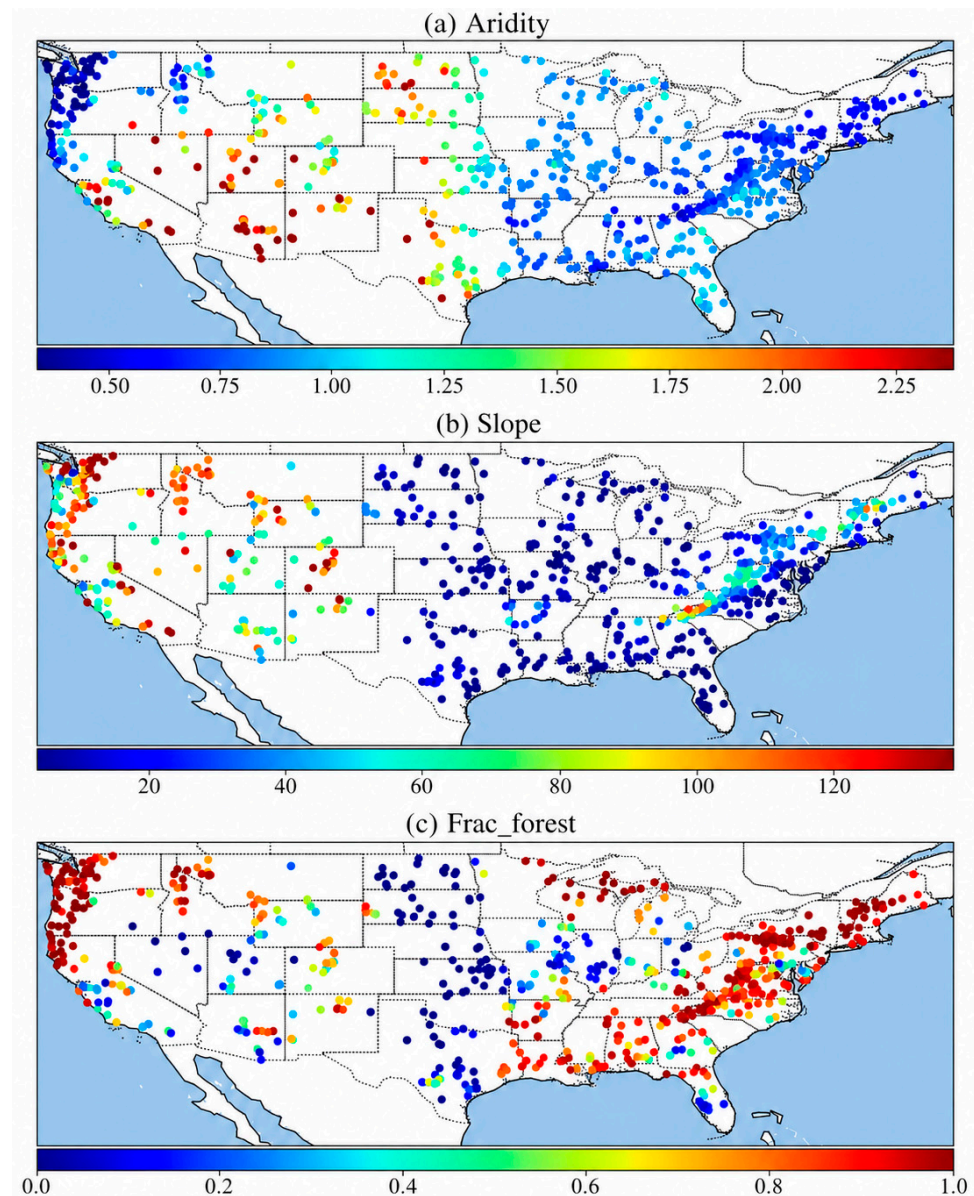


Figure 1. Spatial distribution of CAMELS catchment characteristics. **(a)** Aridity index (-): The ratio of the long-term average annual potential evapotranspiration to annual precipitation, blue denotes humid regions and red denotes arid regions. **(b)** Slope (m/km): The average slope of the catchment, where blue denotes gentle terrain and red denotes steep terrain. **(c)** Forest cover (-): The proportion of the catchment area covered by forest, where blue indicates low forest cover and red indicates high forest cover. Each point denotes the centroid of a catchment and the color represents the catchment-averaged value; therefore, adjacent or visually overlapping points may correspond to distinct catchments with different characteristics.

2.2. Data Preprocessing

When training on a catchment-based dataset, data from different catchments are aggregated in batches for loss function calculations. This process implicitly assumes that errors across catchments carry equal weight; however, without pre-processing, the model tends to favour fitting data from wet and large-scale catchments, whilst paying insufficient attention to arid or small-scale catchments. To mitigate this issue, this paper employs the following data pre-processing strategy [30,31].

Firstly, to reduce the influence of differences in catchment scale during regional training, daily streamflow was first normalized by catchment area and mean annual precipitation to obtain a dimensionless target variable. Subsequently, a logarithmic transformation was

applied to both streamflow and precipitation to reduce data skewness and compress the dynamic range, thereby improving numerical stability during model training [29]. The transformation is defined as follows [29]:

$$v^* = \log_{10}(\sqrt{v} + 0.1) \quad (1)$$

Here, v denotes the original variable, and v^* denotes the transformed variable. The square root transformation is used to reduce data skewness, whilst the logarithmic transformation further compresses the data range; the constant 0.1 is used to avoid numerical instability issues caused by zero values.

Finally, all input variables are standardised, that is,

$$x' = \frac{x - \mu}{\sigma} \quad (2)$$

Here, x represents the original input variable, whilst μ and σ denote the mean and standard deviation calculated within the study catchment, respectively. This step standardises the scale of different variables, thereby improving the stability and convergence efficiency of model training.

2.3. Methodology

The structural diagram of the LSTM cell is shown in Figure 2, while the overall workflow of the proposed PPSL framework for runoff simulation is shown in Figure 3. Our method combines the DI-LSTM base model described in Section 2.3.1 with the PPSL training strategy proposed in Section 2.3.2. Section 2.3.3 outlines the experimental setup and evaluation metrics.

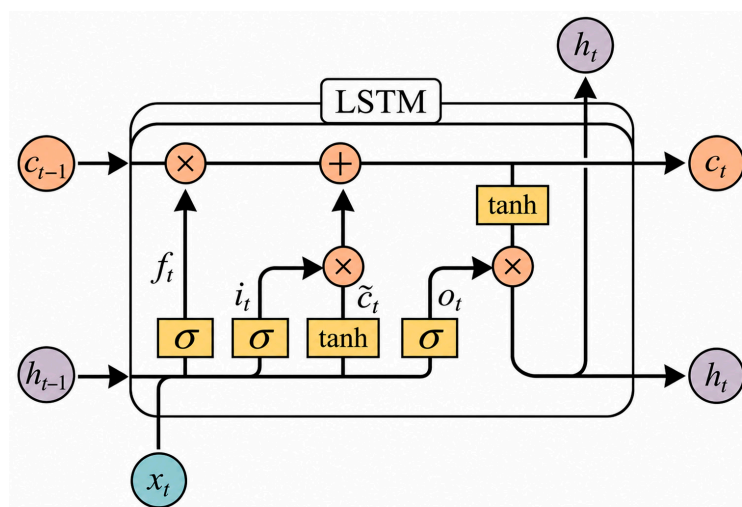


Figure 2. The architecture of an LSTM.

2.3.1. Base Model

Long Short-Term Memory (LSTM) is a recurrent neural network that captures long-term dependencies and mitigates gradient vanishing via gated memory cells. This structure is well-suited for modeling the lagged and nonlinear dynamics of rainfall-runoff processes, making it a common baseline in data-driven hydrology [32]. Its core gate equations at time step t are the following [32]:

Input gate:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (3)$$

Forget gate:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (4)$$

Output gate:

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (5)$$

Candidate cell state:

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (6)$$

Cell state update:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (7)$$

Hidden state update:

$$h_t = o_t \odot \tanh(c_t) \quad (8)$$

In Equations (3)–(8), i_t , f_t and o_t denote the input, forget, and output gates, respectively. c_t , $\tilde{c}_t$, and h_t represent the cell state, candidate cell state, and hidden state. x_t is the input at time step t , while h_{t-1} and c_{t-1} denote the hidden state and cell state from the previous time step. $\sigma(\cdot)$ and $\tanh(\cdot)$ denote the sigmoid and hyperbolic tangent activation functions, respectively, and $\odot$ represents element-wise multiplication. W , U and b are the learnable weight matrices and bias vectors [32].

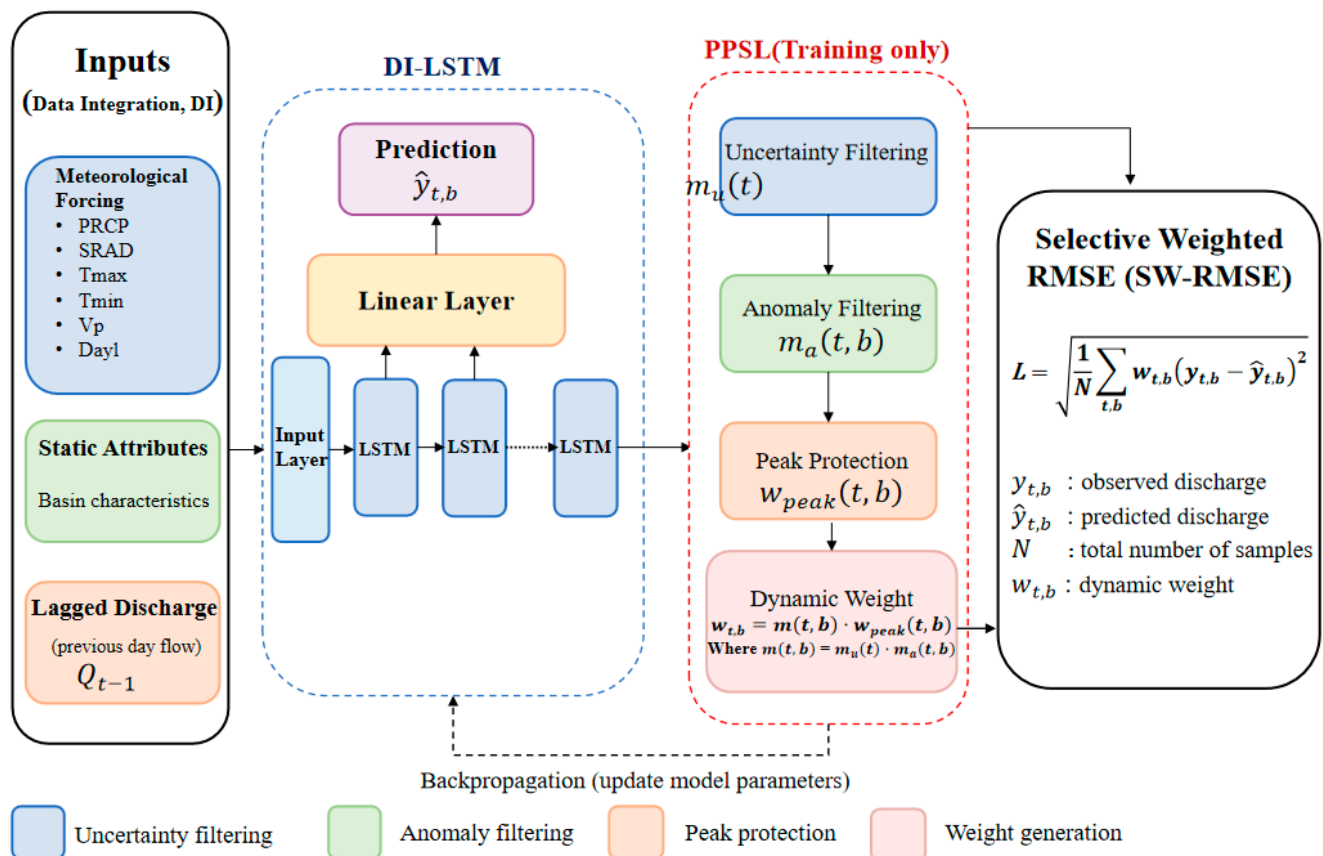


Figure 3. Flowchart of runoff simulation using PPSL-DI-LSTM.

However, standard LSTMs often struggle to capture prior catchment conditions as they typically rely only on meteorological forcings, leading to biases in flood peak simulation.

To address this, DI-LSTM extends the input vector to include lagged observed runoff (DI(1)) [29], enhancing the model's representation of these prior states:

$$\tilde{x}_t = [x_t; q_{t-1}^{obs}] \quad (9)$$

where x_t denotes the meteorological forcing vector at the current time, and q_{t-1}^{obs} represents the observed runoff from the previous day. The extended input $\tilde{x}_t$ together with the basin's static attributes c drives the LSTM cells [29]; the recursive process can be expressed as

$$h_t = LSTM(h_{t-1}, [\tilde{x}_t; c]) \quad (10)$$

$$\hat{q}_t = g(h_t) \quad (11)$$

where h_t denotes the current hidden state, $g(\cdot)$ denotes the output layer mapping function, and $\hat{q}_t$ denotes the predicted runoff.

Although the use of observed runoff lagged by one day can cause the model to be highly autoregressive and underrepresent meteorological-based prediction capabilities, this input is retained following the original DI-LSTM framework to represent catchment memory and antecedent hydrological conditions.

It should be noted that the proposed PPSL framework is not intended to compensate for the inclusion of lagged runoff itself, but rather to improve the model training process by dynamically emphasizing informative time steps. Moreover, the effectiveness of PPSL has also been verified on the standard LSTM model, demonstrating that its performance improvement is not limited to the DI-LSTM architecture.

To address these issues, this paper proposes a Peak Protection Selective Learning (PPSL) strategy based on DI-LSTM. By adaptively weighting training samples, PPSL mitigates sample imbalance and optimization bias, improving both overall runoff prediction and the simulation of extreme runoff events.

2.3.2. A Forecasting Model Based on a Flood Peak Protection Selection Learning Framework

Selective Learning is a model-agnostic training strategy for time-series forecasting that alleviates overfitting on complex data. Instead of optimising the loss uniformly over all time steps, it selects time steps with better generalisation so that the model focuses on stable temporal patterns. It mainly targets two problematic cases: highly uncertain time steps dominated by noise and anomalous time steps caused by abrupt disturbances [22]. Accordingly, a dual-mask scheme is used to filter these time steps and compute the loss only on the selected ones.

Different from the hard masking in [22], we tailor Selective Learning to runoff forecasting, where non-stationarity is strong and flood peaks are rare and easily misclassified. Specifically, we replace time-step removal with a continuous, weighted loss to suppress noise while retaining useful information. Meanwhile, we redesign the anomaly criterion by treating rainfall-driven high flows as physically meaningful extremes rather than anomalies, and introduce a peak-protection mechanism to strengthen learning on flood peaks.

The proposed strategy comprises three collaborative modules: uncertainty filtering, anomaly filtering, and peak protection. The first two suppress unstable and high-bias samples via weighted down-weighting, while peak protection ensures that extreme runoff (especially flood peaks) receives sufficient attention through weight amplification and lower-bound constraints.

The uncertainty filtering module identifies time steps with unstable predictions by measuring the dispersion of prediction residuals across samples within each mini-batch [22].

The resulting uncertainty is used for adaptive sample weighting during training. The residual of the b -th sample at time step t is defined as

$$\epsilon_{t,b} = \hat{y}_{t,b} - y_{t,b} \quad (12)$$

Calculate the residual variance at the batch level to measure the degree of fluctuation in the prediction results for that time step:

$$\text{Var}_t = \frac{1}{N_t} \sum_{b \in \mathcal{V}_t} (\epsilon_{t,b} - \mu_t)^2 \quad (13)$$

Among these,

$$\mu_t = \frac{1}{N_t} \sum_{b \in \mathcal{V}_t} \epsilon_{t,b} \quad (14)$$

To reduce the fluctuations caused by small-batch training, an exponential moving average (EMA) with a smoothing coefficient of $\beta = 0.95$ is adopted to stabilize the uncertainty estimates:

$$\tilde{V}_t = \beta \cdot \tilde{V}_t + (1 - \beta) \cdot \text{Var}_t \quad (15)$$

Based on the estimated uncertainty, a quantile threshold θ_u is determined using the filtering intensity parameter r_u , which is set to 0.1 to identify high-uncertainty samples:

$$m_u(t) = \sigma\left(\frac{\theta_u - \tilde{V}_t}{\tau_u}\right) \quad (16)$$

where $\sigma(\cdot)$ denotes the sigmoid function. A larger $\tilde{V}_t$ indicates higher prediction uncertainty, resulting in a smaller uncertainty weight $m_u(t)$.

Anomaly filtering suppresses high-bias samples caused by sudden disturbances [22]. We define both model prediction error and persistence baseline error:

$$e_f(t, b) = |\hat{y}_{t,b} - y_{t,b}| \quad (17)$$

$$e_g(t, b) = |y_{t,b} - y_{t-1,b}| \quad (18)$$

where $y_{t-1,b}$ denotes the observed value at time step $t - 1$, which is used to construct a persistence baseline.

Based on the errors defined above, the anomaly scoring function is defined as

$$S_{t,b} = e_f(t, b) - e_g(t, b) \quad (19)$$

A larger $S_{t,b}$ indicates a higher likelihood that the sample is an outlier, whereas a smaller value suggests better predictability.

Within each mini-batch, threshold θ_a is determined by quantile ($r_a = 0.07$), and the sample weight is

$$m_a(t, b) = \sigma\left(\frac{S_{t,b} - \theta_a}{\tau_a}\right) \quad (20)$$

Higher anomaly scores result in smaller anomaly weights $m_a(t, b)$, thereby reducing their contribution to the loss function.

Peak protection targets rainfall-driven high-flow processes to prevent over-suppression and enhance learning of flood peak samples. The peak threshold is defined as $\theta_p = \text{Quantile}(y, q_p)$ with $q_p = 0.95$. Following the commonly adopted percentile-based definition of high-flow events in hydrological studies (e.g., van Vliet et al., 2013) [33], the

95th percentile of observed runoff was used to identify peak-flow samples. The probability of belonging to a peak process is

$$p_{t,b} = \sigma\left(\frac{y_{t,b} - \theta_p}{\tau_p}\right) \quad (21)$$

where $\sigma(\cdot)$ is the Sigmoid function and τ_p is the temperature parameter. When $y_{t,b}$ is significantly higher than the threshold θ_p , $p_{t,b}$ approaches 1, indicating that the time step belongs to a peak flow process.

Furthermore, a peak weighting function is constructed to enhance the model's focus on high-flow samples:

$$w_{peak}(t, b) = 1 + \alpha_p \cdot p_{t,b} \quad (22)$$

where α_p denotes the peak weighting intensity parameter, which is set to 1.75 in this study. Consistent with previous peak-aware weighting strategies for runoff prediction (e.g., Man et al., 2023) [19] When a sample falls within the peak flow regime, its loss weight is significantly amplified, thereby guiding the model to prioritize learning from extreme runoff processes.

To prevent peak flow samples from being excessively down-weighted during the selective learning process, a weight lower-bound constraint is further introduced:

$$m(t, b) = \max(m_u(t) \cdot m_a(t, b), w_{\min} \cdot p_{t,b}) \quad (23)$$

where $m_u(t)$ and $m_a(t, b)$ represent the uncertainty weight and anomaly weight, and $w_{\min}$ is the minimum weight coefficient for peak protection, which is set to 1.50 in this study. The design is consistent with previous peak-aware and imbalance-aware weighting strategies for runoff prediction [19]. Its value is determined using the same validation procedure as $w_{\min}$ and is set to 1.50 (Appendix A, Table A1). This mechanism ensures that peak samples retain a sufficient weight to participate in model training, even when they have high uncertainty or anomaly scores, thereby preventing key extreme events from being diminished.

In summary, the proposed PPSL framework combines uncertainty filtering, anomaly filtering, and peak protection to suppress noisy samples while emphasizing informative and extreme-runoff samples. This adaptive weighting strategy improves the representation of extreme runoff events without compromising overall runoff prediction performance.

2.3.3. Models Setting and Evaluation Metrics

To enhance the model's characterization ability for extreme runoff events while maintaining overall predictive accuracy, we introduce uncertainty filtering, anomaly filtering, and peak protection mechanisms based on the traditional Root Mean Square Error (RMSE). This leads to the construction of a dynamically weighted RMSE loss function, termed Selective Weighted RMSE (SW-RMSE), which is formulated as follows:

$$\text{SW-RMSE} = \sqrt{\frac{1}{N} \sum_{t,b} w_{t,b} (y_{t,b} - \hat{y}_{t,b})^2} \quad (24)$$

where $y_{t,b}$ and $\hat{y}_{t,b}$ represent the observed and predicted values for the b -th basin at time step t ; N is the total number of samples, and $w_{t,b}$ denotes the dynamic weighting coefficient.

The weight $w_{t,b}$ is determined by the selective learning mechanism and is expressed as

$$w_{t,b} = m(t, b) \cdot w_{peak}(t, b) \quad (25)$$

where $m(t, b)$ is the composite weight derived from the joint action of the uncertainty filtering factor $m_u(t)$ and the anomaly filtering factor $m_a(t, b)$ defined in Section 2.3.2; $w_{peak}(t, b)$ is the peak-aware weighting function designed to enhance the importance of high-flow and extreme runoff samples.

To ensure comparability with existing studies, the model configuration on the CAMELS dataset is kept consistent with the study by Feng et al. (2020) [29]. The model takes a continuous sequence of 365 days as input, including 6 meteorological variables and 17 catchment static attributes, for runoff prediction. The hidden layer size is set to 256. During the training process, the batch size is 100, and the number of epochs is 300. Meteorological forcing data are obtained from the Daymet dataset. To ensure the reproducibility of the experiments, the random seed is set to 888. All experiments were conducted on a server equipped with an Intel Core™ i9-10980XE CPU (3.00 GHz $\times$ 36), 128 GB of RAM, and an NVIDIA RTX A5000 GPU.

Model performance was evaluated using six widely adopted hydrological metrics: Percent Bias (PBIAS) [34], Kling–Gupta Efficiency (KGE) [35], Nash–Sutcliffe Efficiency (NSE) [36], Coefficient of Determination (R^2) [37], the percent bias of the top 2% peak flow range (FHV) and the percent bias of the bottom 30% low flow range (FLV) [29,38]. FHV and FLV are flow duration curve based metrics that evaluate model performance under peak-flow and base-flow conditions, respectively. These metrics assess overall prediction accuracy, bias, correlation, variability, and model performance under low-flow and high-flow conditions. The corresponding formulations are given as follows:

$$NSE = 1 - \frac{\sum_{t=1}^T (y_t - \hat{y}_t)^2}{\sum_{t=1}^T (y_t - \bar{y})^2} \quad (26)$$

$$KGE = 1 - \sqrt{(r - 1)^2 + \left(\frac{\sigma_{\hat{y}}}{\sigma_y} - 1\right)^2 + \left(\frac{\bar{\hat{y}}}{\bar{y}} - 1\right)^2} \quad (27)$$

$$R^2 = 1 - \frac{\sum_{t=1}^T (y_t - \hat{y}_t)^2}{\sum_{t=1}^T (y_t - \bar{y})^2} \quad (28)$$

$$PBIAS = \frac{\sum_{t=1}^T (\hat{y}_t - y_t)}{\sum_{t=1}^T y_t} \times 100 \quad (29)$$

$$FLV = \frac{\sum_{i=1}^{\lfloor 0.3T \rfloor} (\hat{y}_{(i)} - y_{(i)})}{\sum_{i=1}^{\lfloor 0.3T \rfloor} y_{(i)}} \times 100 \quad (30)$$

$$FHV = \frac{\sum_{i=\lfloor 0.98T \rfloor}^T (\hat{y}_{(i)} - y_{(i)})}{\sum_{i=\lfloor 0.98T \rfloor}^T y_{(i)}} \times 100 \quad (31)$$

where y_t and $\hat{y}_t$ denote the observed and predicted runoff at time t , respectively; $\bar{y}$ and $\bar{\hat{y}}$ are their corresponding means, σ_y and $\sigma_{\hat{y}}$ are the standard deviations, r is the Pearson correlation coefficient, and T is the time-series length. $y_{(i)}$ and $\hat{y}_{(i)}$ denote the i -th sorted observed and predicted runoff values, respectively. Following Refs. [29,38], $0.3T$ and $0.98T$ correspond to the bottom 30% and top 2% of the sorted flow sequence, respectively.

To evaluate the sensitivity of the PPSL model to different configurations of selective learning parameters, we respectively adjust the high-flow threshold, peak weighting intensity, and sample filtering parameters, conducting comparative experiments under unified experimental conditions. The parameter configurations for each experimental group are listed in Table A1 of Appendix A, and the model evaluation results are provided in Table A2 of Appendix A.

3. Results

This study incorporates the selective learning framework into a deep learning model, conducting runoff simulation experiments and performance evaluations using the CAMELS dataset. First, we evaluate the overall performance of the PPSL-LSTM and PPSL-DI-LSTM model across the CAMELS catchments (Section 3.1). Next, a local evaluation of the model's ability to simulate hydrological processes under different scenarios is performed through time-series comparisons in representative basins (Section 3.2). Subsequently, we analyze the spatial distribution of Δ NNSE improvements and explore the underlying factors affecting PPSL performance across different catchments using SHAP analysis (Section 3.3). Finally, the model's characterization capability and robustness under high-flow extreme scenarios are examined through the reproduction of typical extreme flood events (Section 3.4).

3.1. Overall Performance Comparison

To comprehensively evaluate the effectiveness and generality of the proposed selective learning strategy, Figure 4 compares the performance of four models (LSTM, PPSL-LSTM, DI-LSTM, and PPSL-DI-LSTM) using six evaluation metrics across all CAMELS catchments, and Table 2 reports the corresponding median values. Compared with the conventional LSTM, PPSL-LSTM achieved consistent improvements in hydrological performance despite using only meteorological forcing variables as inputs. In particular, median KGE increased from 0.751 to 0.778, while FHV improved substantially from -12.86% to -3.34% , indicating that PPSL effectively enhances the learning of critical hydrological processes and improves peak-flow simulation. In contrast, the median NSE and R^2 values remained essentially unchanged ($0.741 \rightarrow 0.740$ for median NSE), suggesting that the selective learning strategy does not significantly affect overall variance-based accuracy in the standard LSTM configuration. To verify if these improvements are robust against random initialization noise, we evaluated the models across three random seeds (Seeds 42, 66, and 123; Table 3). The results show that while NSE gains remain marginal across different seeds, the improvements in KGE and FHV are robust. PPSL-LSTM consistently outperformed the baseline in KGE ($0.767\text{--}0.776$ vs. 0.751) and substantially reduced peak-flow bias (FHV) from -12.86% to as low as -3.38% . This confirms that selective learning reliably improves high-flow modeling, even when overall NSE changes are small.

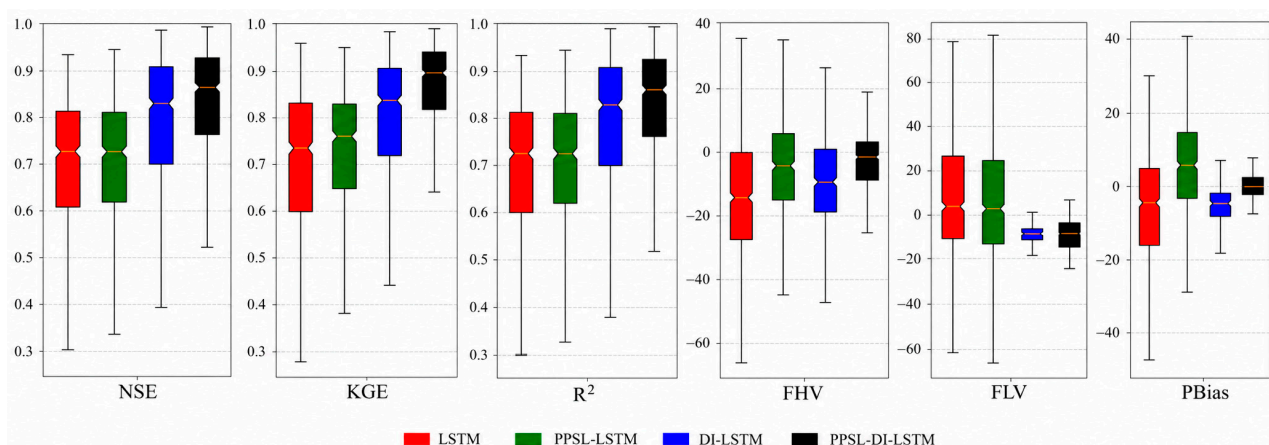


Figure 4. Boxplot comparison of the performance of the four models across six evaluation metrics over all CAMELS catchments. Red, green, blue, and black denote the distributions of the LSTM, PPSL-LSTM, DI-LSTM, and PPSL-DI-LSTM models, respectively. The central line in each box represents the median value of the corresponding metric.

Table 2. Comparison of the hydrological simulation performance of the four models across six evaluation metrics during the testing period. Bold values indicate the best performance.

Model	NSE	KGE	R ²	FHV	FLV	PBias
LSTM	0.741	0.751	0.741	−12.863	6.457	−3.705
PPSL-LSTM	0.740	0.778	0.740	−3.3438	5.999	6.611
DI-LSTM	0.835	0.847	0.835	−8.202	−4.721	−3.808
PPSL-DI-LSTM	0.870	0.901	0.870	−1.298	−4.540	1.153

Table 3. Median basin test performance under different random seeds.

Seed	Model	NSE	KGE	FHV	FLV	PBias
42	LSTM	0.7418	0.7516	−12.8634	6.4572	−3.7055
	PPSL-LSTM	0.7504	0.7712	−8.4508	5.9921	1.9454
66	LSTM	0.7418	0.7516	−12.8634	6.4572	−3.7055
	PPSL-LSTM	0.7496	0.7678	−3.3824	5.8993	6.2119
123	LSTM	0.7418	0.7516	−12.8634	6.4572	−3.7055
	PPSL-LSTM	0.7464	0.7761	5.4352	8.8683	5.2707

When integrated into DI-LSTM, PPSL further improved the overall prediction performance. Compared with DI-LSTM, PPSL-DI-LSTM increased the median NSE, KGE, and R² by approximately 4%, 6%, and 4%, respectively, while reducing PBIAS and further improving both FHV and FLV. These results indicate that PPSL effectively enhances both the overall runoff simulation accuracy and the representation of extreme flow conditions. Overall, the proposed PPSL strategy consistently improves runoff prediction under both input configurations. Even when using only meteorological forcing data, PPSL can still effectively enhance the model's ability to learn critical hydrological processes and improve runoff prediction performance.

Figure 5 presents the spatial distributions of NSE across the CAMELS catchments for both the DI framework and the standard LSTM framework, together with their corresponding improvements in normalized NSE (Δ NNSE). Δ NNSE is adopted instead of relative NSE gain because its bounded formulation allows more robust cross-catchment comparisons, particularly when baseline model accuracies vary substantially.

As shown in Figure 5a,b, both DI-LSTM and PPSL-DI-LSTM achieve NSE values above 0.6 for most catchments. The PPSL-DI-LSTM model exhibits a broader distribution of high NSE values, particularly in the western United States, the northern Midwest, and parts of the Southeast. Similarly, in the standard LSTM framework (Figure 5d,e), PPSL-LSTM consistently outperforms the baseline LSTM, with noticeable expansion of high-performance regions.

The improvement patterns are summarized in Figure 5c,f. For both model structures, positive Δ NNSE values dominate across the vast majority of catchments, indicating that PPSL provides systematic rather than localized improvements. The magnitude of improvement is generally larger in the DI framework, suggesting that PPSL interacts more effectively with the DI structure. Spatial clustering of larger gains can be observed in several hydro-climatic regions, implying coherent regional benefits of the PPSL strategy. Negative Δ NNSE values occur only in a limited number of isolated catchments and are typically small in magnitude, indicating that the proposed strategy rarely causes meaningful performance degradation.

Overall, these results demonstrate that PPSL not only improves average predictive performance but also enhances spatial robustness and transferability across heterogeneous hydro-climatic conditions, with consistent benefits observed in both model architectures.

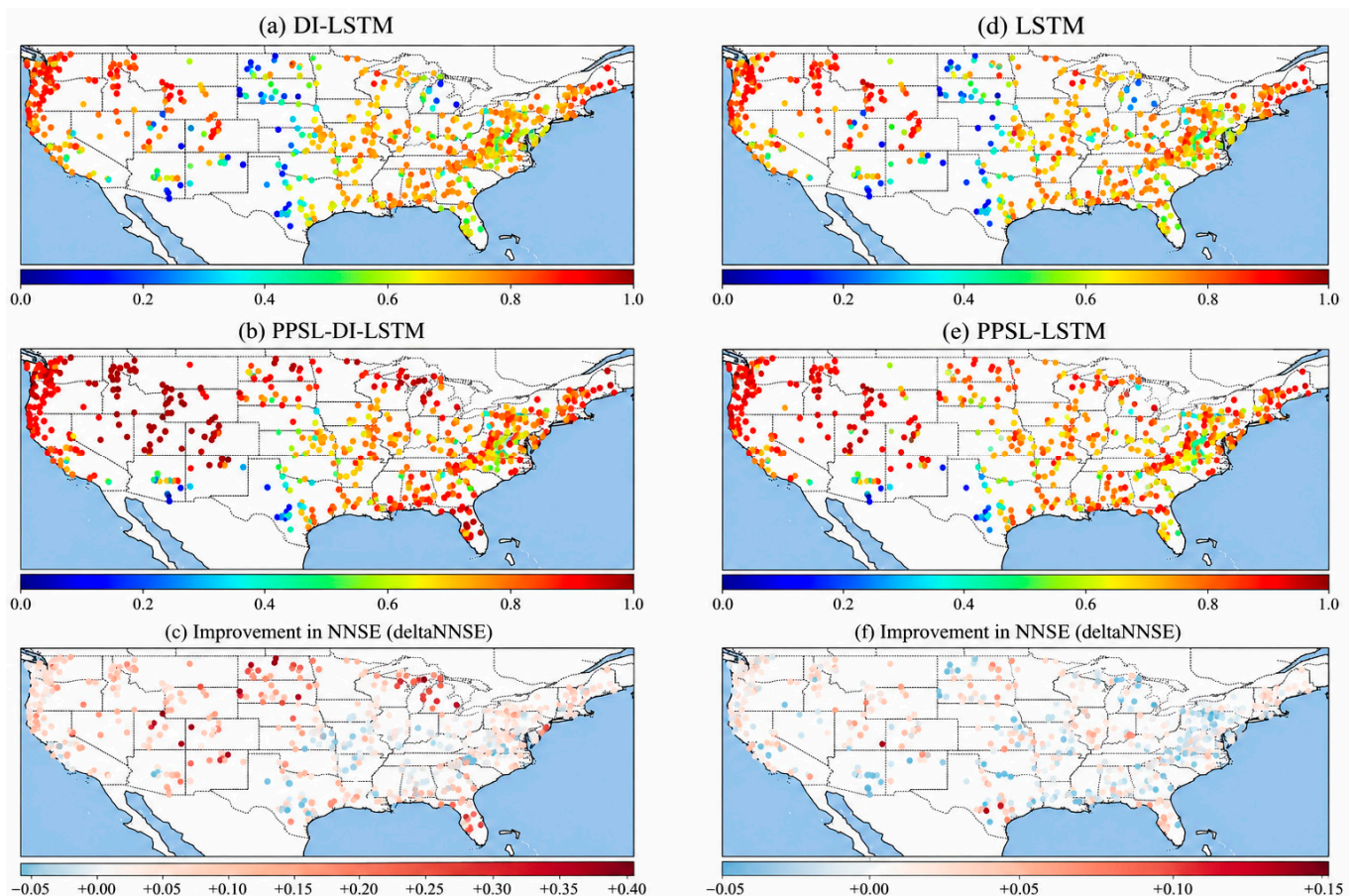


Figure 5. Spatial distribution of NSE and model improvement across the CAMELS catchments for two model structures. Panels (a–c) show results for the DI framework: (a) DI-LSTM, (b) PPSL-DI-LSTM, and (c) the corresponding improvement in NNSE [39] (Δ NNSE). Panels (d–f) show results for the standard LSTM framework: (d) LSTM, (e) PPSL-LSTM, and (f) the corresponding Δ NNSE. Δ NNSE calculated as Δ NNSE = $\frac{1}{2 - \text{NSE}_{\text{PPSL}}} - \frac{1}{2 - \text{NSE}_{\text{Baseline}}}$, where Baseline denotes the corresponding baseline model (i.e., DI-LSTM in (a) and LSTM in (d)). Red and blue dots represent catchments with positive and negative relative improvements, respectively, with darker colors indicating larger magnitudes.

3.2. Time-Series Evaluation of Predictions

Figure 6 presents the daily runoff comparisons for seven representative CAMELS catchments during the testing period. To clearly evaluate the models' capability in simulating extreme hydrological processes, we focused on 2–3 year windows containing the highest peak flow events for each catchment, comparing USGS observations with DI-LSTM and PPSL-DI-LSTM simulations. The selected catchments cover diverse hydrological response types, including event-driven basins dominated by prominent peak pulses (Figure 6a,b,d), basins with high low-flow proportions (Figure 6c,f), and basins with frequent fluctuations (Figure 6e,g).

Results indicate that PPSL-DI-LSTM achieves the highest NSE in most catchments and delivers consistent cross-basin improvements, demonstrating robust generalization. In event-driven basins with significant peak pulses, DI-LSTM captures general flood trends but commonly suffers from peak underestimation and response lag. By contrast, PPSL-DI-LSTM markedly improves constraints on both peak magnitude and timing (e.g., NSE increased from 0.46 to 0.67 in Figure 6f, from 0.88 to 0.91 in Figure 6b, and from 0.85 to 0.88 in Figure 6c).

Notably, improvements from PPSL-DI-LSTM are not confined to catchments with weak baseline performance; further gains are also achieved where DI-LSTM already attains

high NSE. This suggests that the selective learning strategy functions not merely as an error-correction mechanism, but enhances the model's unified capacity to characterize complex hydrological processes through more effective sample organization and training constraints.

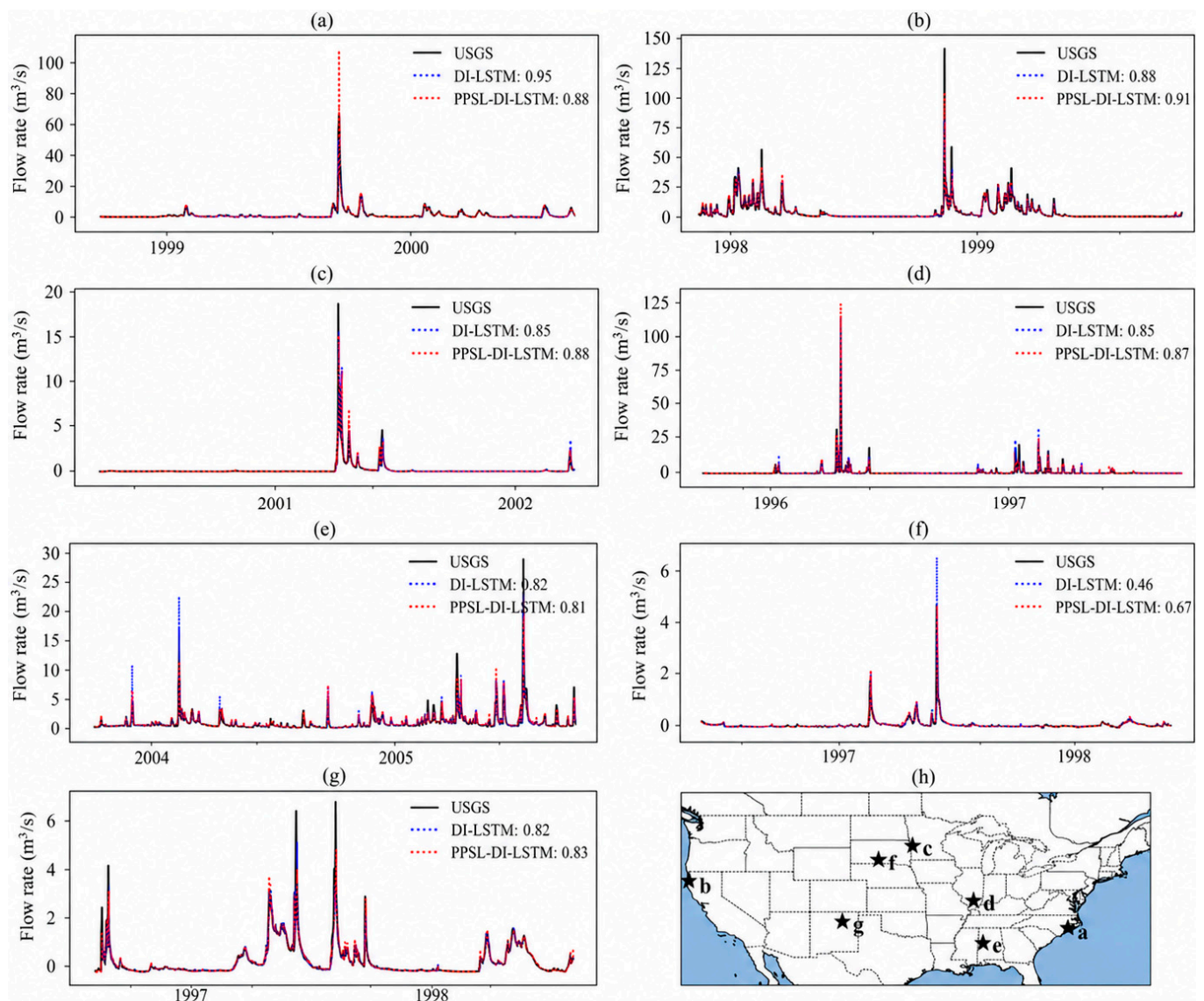


Figure 6. Daily runoff comparisons for seven representative CAMELS catchments during the two-year window containing the highest observed flood peak in the testing period. USGS observations are compared against DI-LSTM and PPSL-DI-LSTM simulations, with NSE values indicated in the legends. Panels (a–g) show the time series for individual catchments, and panel (h) shows their spatial locations.

In contrast to the overall comparison over multi-year peak windows in Figure 6, Figure 7 focuses on short-term hydrological transition processes for the same set of catchments—encompassing sharp rising limbs, flood peaks, and rapid recessions, indicating an improved ability to capture abrupt runoff dynamics.

Representative improvements are observed in catchments Figure 7b,d,f, where PPSL-DI-LSTM effectively suppresses flood-peak overestimation while preserving the overall hydrograph evolution. Similar improvements in peak timing and hydrograph shape are also evident in catchments Figure 7c,g. In catchment Figure 7e, both models perform similarly, suggesting that the proposed strategy provides limited additional benefit when runoff variability is relatively moderate.

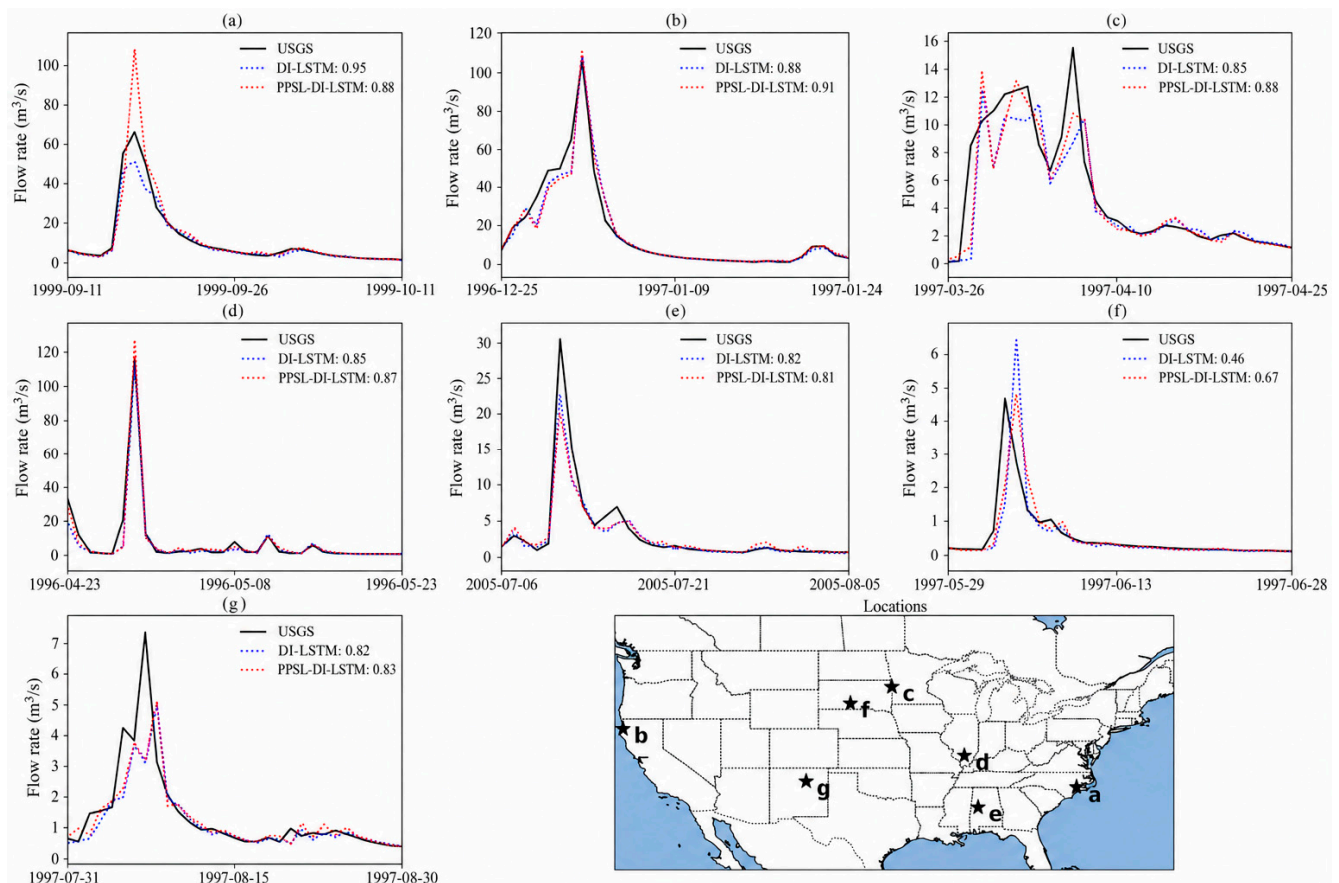


Figure 7. Zoomed-in views of runoff periods with abrupt hydrograph changes for the seven representative catchments. (a–g) comparisons of observed USGS flow rates with DI-LSTM and PPSL-DI-LSTM model simulations across different geographical locations as shown in the map.

Overall, these comparisons demonstrate that PPSL-DI-LSTM more effectively captures rapid hydrological transitions and improves the representation of flood processes without compromising the overall hydrograph characteristics.

3.3. Analysis of Model Performance Distribution Characteristics

To further investigate how the proposed PPSL strategy affects prediction performance across different catchments, Figure 5 presents the spatial distribution of the NNSE [39] improvement (Δ NNSE) for both the DI and standard LSTM frameworks, comparing PPSL-enhanced models with their respective baselines across the CAMELS catchments. Positive Δ NNSE values dominate most regions of the continental United States in both model structures, demonstrating that PPSL consistently improves streamflow prediction under diverse hydro-climatic conditions. The largest improvements are mainly observed in the western United States, the northern Midwest, and parts of the southeastern United States, whereas only a limited number of catchments exhibit negligible or negative improvements, with generally larger magnitudes occurring under the DI framework.

These spatial differences are closely related to the selective learning mechanism of PPSL. By assigning greater weights to difficult samples, particularly peak-flow events, PPSL enables the model to focus on hydrologically informative runoff processes instead of treating all samples equally. Through its time-step weighting and peak-preservation mechanisms, PPSL adaptively strengthens the learning of critical moments and prevents extreme peaks from being underrepresented, thus improving the representation of extreme hydrological behaviors. Consequently, catchments characterized by rapid hydrological responses benefit most from the proposed strategy. For example, many western moun-

tainous catchments exhibit steep terrain and flashy runoff responses, while southeastern catchments frequently experience intense precipitation and highly dynamic flood events. In these regions, emphasizing peak-flow samples improves the model's ability to capture nonlinear rainfall–runoff processes, resulting in larger performance gains. In contrast, catchments in the central Great Plains and some eastern basins show relatively limited improvements. These basins are more strongly influenced by groundwater storage, baseflow recession, or other slow hydrological processes, where runoff variability is less dependent on individual rainfall events. Therefore, increasing the learning emphasis on peak-flow samples alone provides relatively limited benefits [31,40].

To clarify the underlying mechanism, a Δ SHAP analysis was performed to quantify the per-feature change in importance between paired models (Figure 8). Unlike independent SHAP rankings, Δ SHAP directly captures how PPSL redistributes feature contributions after introducing its selective weighting strategy.

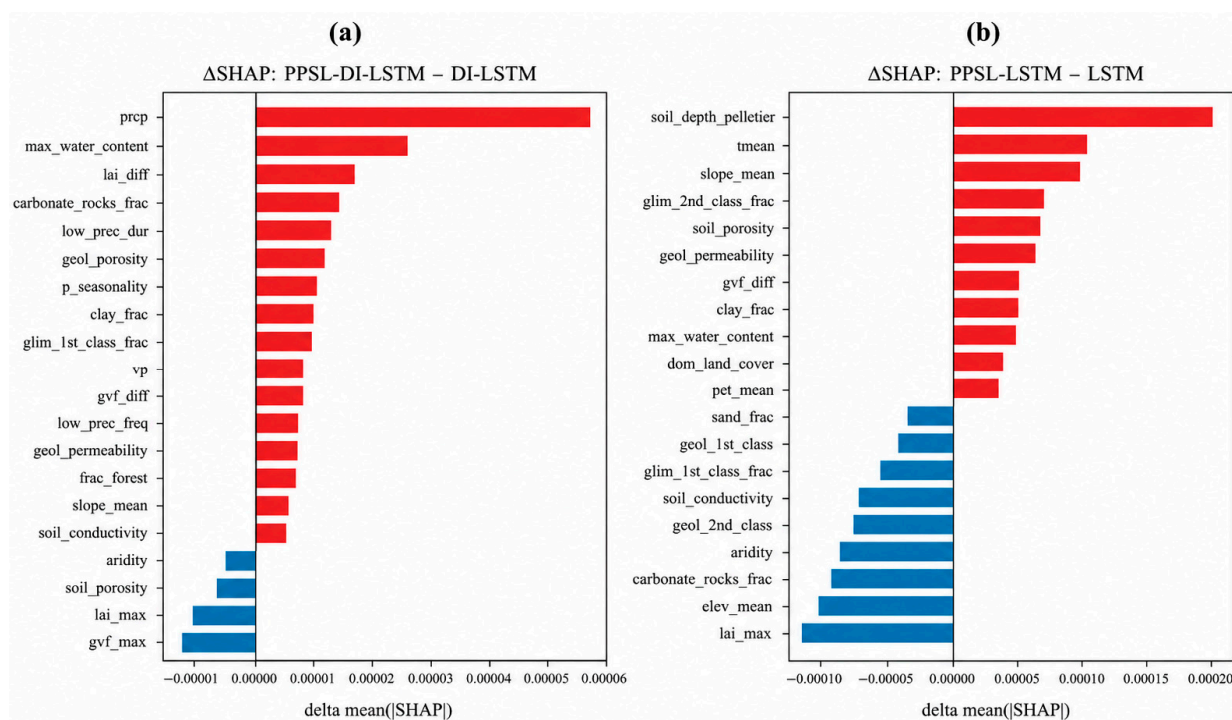


Figure 8. Per-feature change in mean absolute SHAP values ($\Delta \text{mean} | \text{SHAP} |$) after introducing PPSL. (a) Δ SHAP for the DI framework (PPSL-DI-LSTM – DI-LSTM). (b) Δ SHAP for the standard LSTM framework (PPSL-LSTM – LSTM). Positive values indicate increased feature importance after applying PPSL, whereas negative values indicate reduced importance. Features are ranked by the combined magnitude of importance shifts across both model pairs, and only the top 20 features with the largest overall Δ SHAP values are displayed. The figure illustrates how selective learning redistributes model attribution across hydrometeorological and catchment descriptors.

Under the DI framework (Figure 8a), precipitation (prcp) shows the largest positive Δ SHAP, followed by variables associated with soil water storage (e.g., max_water_content) and vegetation dynamics (e.g., lai_diff), while static vegetation indicators (e.g., gvf_max and lai_max) exhibit negative Δ SHAP values. A similar but weaker redistribution is observed in the standard LSTM framework (Figure 8b), where the shifts in importance are more moderate. This milder adjustment suggests that, without antecedent streamflow input, PPSL can still enhance sensitivity to dynamic predictors.

These systematic shifts are consistent with the design of PPSL. By assigning higher weights to difficult samples—particularly peak-flow events—and dynamically adjusting time-step importance, PPSL encourages the model to prioritize precipitation forcing

and storage-sensitive processes that dominate rapid runoff responses. At the same time, the reduced relative importance of static vegetation indicators suggests that PPSL alleviates excessive reliance on persistent background signals or long-term memory effects. The stronger Δ SHAP responses within the DI framework further indicate that selective learning interacts synergistically with antecedent streamflow input, amplifying the reallocation of importance toward event-driven hydrological drivers.

Overall, the spatial Δ NNSE patterns and Δ SHAP shifts jointly indicate that PPSL primarily enhances event-driven hydrological processes. While rapid-response catchments benefit substantially, storage-dominated systems remain comparatively insensitive, highlighting the potential value of hydrological-regime-adaptive learning mechanisms in future work.

3.4. Analysis of Extreme Runoff Events

To facilitate a detailed analysis of different extreme flood processes, four typical flood events are classified into two categories based on their formation mechanisms. Figure 9a,b represent rainfall-driven flood events, including the 1995 Louisiana heavy rainfall flood and the 2001 Texas Tropical Storm Allison flood. In contrast, Figure 10a,b depicts hurricane-driven flood events, comprising the 1999 North Carolina hurricane flood and the 2005 Hurricane Katrina flood. These two categories correspond to sudden convective rainfall processes and large-scale persistent storm processes, respectively, providing a rigorous test of the model's simulation capabilities under diverse extreme hydrological scenarios.

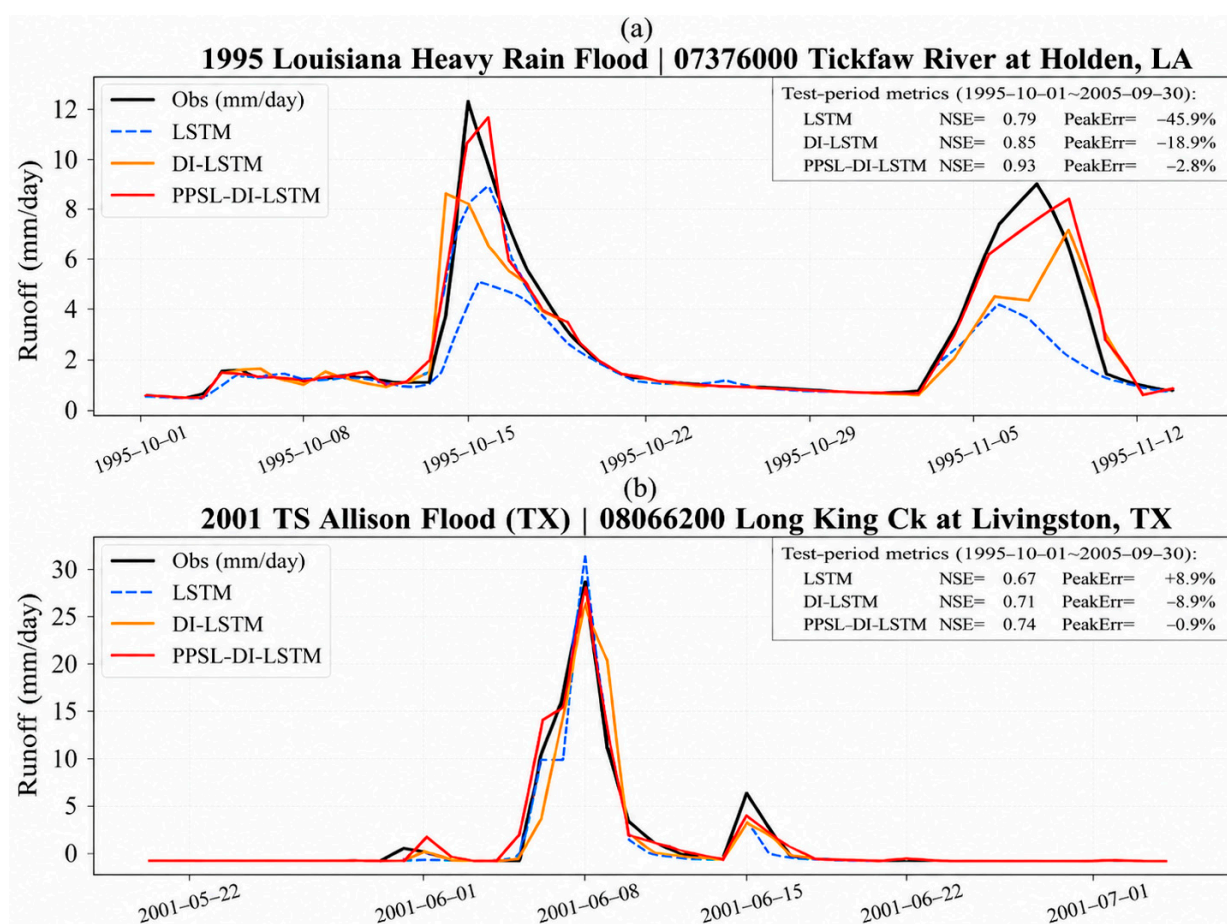


Figure 9. Comparison of runoff simulations during rainfall-driven flood events. (a) The 1995 Louisiana heavy rainfall flood event; (b) the 2001 Tropical Storm Allison heavy rainfall event. NSE and PeakErr for each model are reported in the upper-right corner of each subplot. LA and TX represent Louisiana and Texas, respectively.

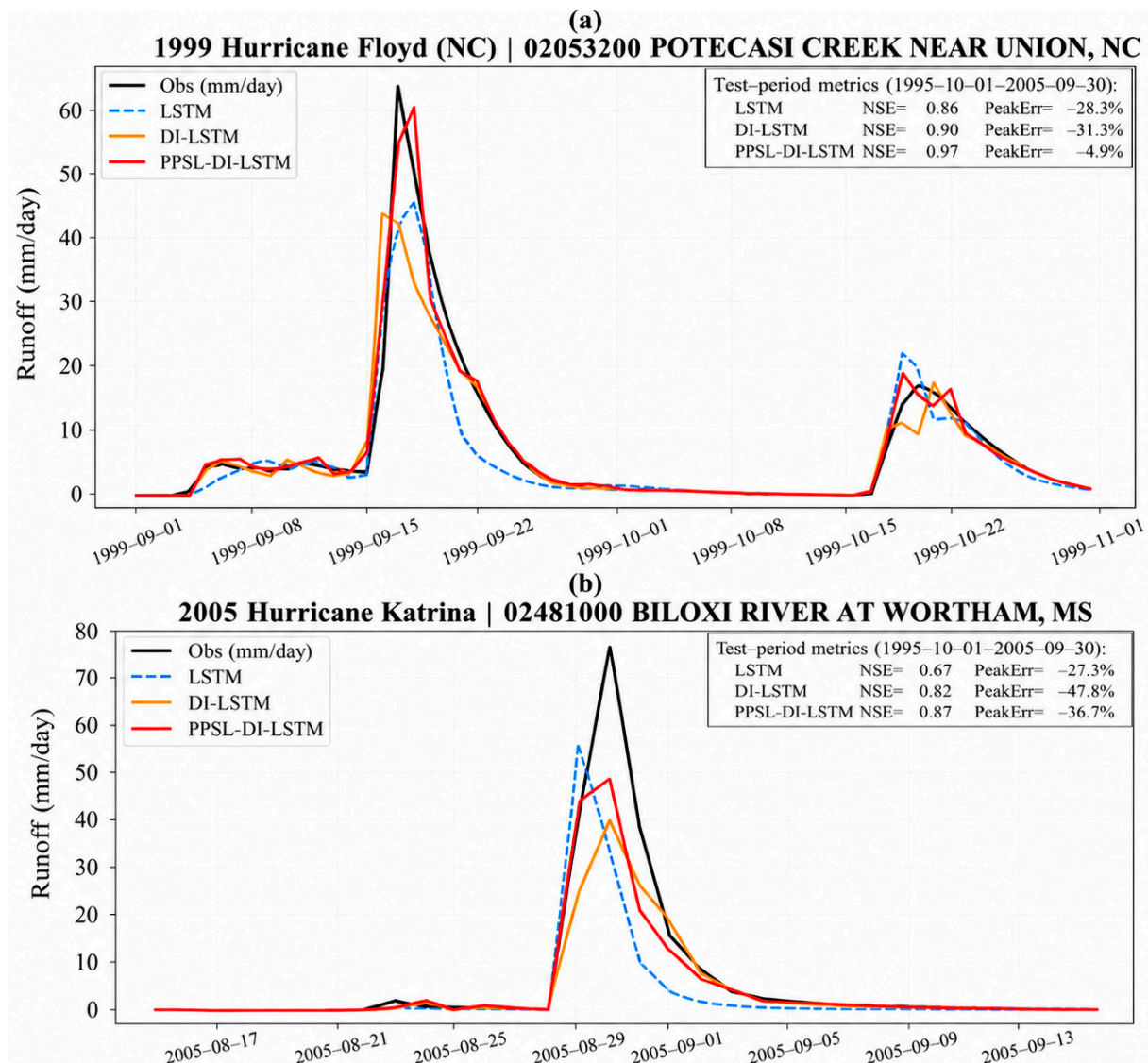


Figure 10. Comparison of runoff simulations during hurricane-driven flood events. (a) 1999 Hurricane Floyd heavy rainfall flood event; (b) 2005 Hurricane Katrina heavy rainfall flood event. NSE and PeakErr for each model are reported in the upper-right corner of each subplot. NC and MS represent North Carolina and Mississippi, respectively.

For clarity, the representative flood events analyzed in this study are further summarized in Appendix A (Table A4), including the event name, occurrence date, basin ID, rainfall, observed peak runoff, simulated peak runoff, peak error, timing error, and event-based NSE for both the DI-LSTM and PPSL-DI-LSTM. Models. PeakErr denotes the relative error of the peak runoff, indicating flood peak overestimation or underestimation within the selected event window [41]. $\text{PeakErr}(\%) = (Q_{\text{sim}}^{\text{max}} - Q_{\text{obs}}^{\text{max}}) / Q_{\text{obs}}^{\text{max}} \times 100$, where $Q_{\text{obs}}^{\text{max}}$ and $Q_{\text{sim}}^{\text{max}}$ are the maximum observed and simulated streamflow values, respectively.

For rainfall-driven flood events (Figure 9a,b), the hydrographs typically exhibit characteristics such as rapid rising limbs, sharp peaks, and abrupt fluctuations, serving as a rigorous test of the model's capability to simulate rapid runoff responses. Notably, the Tropical Storm Allison flood event was characterized by a short-duration runoff response, with substantial runoff occurring within only a few days and remaining close to zero before and after the event. The results indicate that the traditional LSTM exhibits widespread peak underestimation and delayed peak timing, particularly struggling to accurately track runoff variations during the rapid rising stage. Although DI-LSTM improves peak simulation

to some extent, limitations persist regarding peak magnitude and shape fidelity. In contrast, the simulations from PPSL-DI-LSTM align more closely with observed hydrographs, capturing peak magnitudes with higher accuracy and achieving superior NSE scores.

For hurricane-driven flood events (Figure 10a,b), the hydrographs typically exhibit prolonged durations, distinct recession limbs, and occasionally multiple peaks. Consequently, they serve as a rigorous test of the model's overall simulation capability for complex flood processes. In these scenarios, the traditional LSTM tends to suffer from issues such as overly rapid recession or significant overall bias, leading to discrepancies between the simulated and observed hydrographs. Although DI-LSTM improves the general trend, it remains susceptible to error accumulation during prolonged flood periods. In contrast, PPSL-DI-LSTM demonstrates closer alignment with observations across the primary and secondary peaks as well as the recession phase. The resulting hydrographs are smoother and more stable, indicating that the model not only enhances peak flow simulation but also strengthens the stability of simulating prolonged complex flood processes.

4. Discussion

The proposed PPSL framework consistently improves both overall runoff prediction and extreme-flow simulation. Compared with DI-LSTM, PPSL-DI-LSTM achieves notable reductions in FHV while also improving FLV, indicating a better balance between high-flow and low-flow prediction. These improvements can be attributed to the adaptive sample-weighting strategy, which encourages the model to focus on hydrologically informative and difficult samples during training. Furthermore, the spatial distribution of Δ NNSE together with the SHAP analysis demonstrates that PPSL improves prediction performance across most catchments by reducing the model's reliance on antecedent runoff while increasing the contributions of precipitation and catchment characteristics, thereby enhancing the physical representation of rainfall–runoff processes.

Compared with previous studies, the improvements achieved by the proposed PPSL framework are consistent with the general trend that emphasizing informative hydrological samples can enhance extreme runoff prediction. Existing approaches have primarily focused on improving model performance through attention mechanisms (Chen et al., 2022; Giriagama et al., 2022) [17,18], peak-aware loss functions (Man et al., 2023) [19], imbalance-aware optimization strategies, or event-based sample selection (Tofighi et al., 2025) [21]. While these methods have demonstrated notable improvements in peak-flow simulation, they mainly rely on modifications to the network architecture, loss function, or event selection strategy. In contrast, PPSL addresses the problem from the perspective of the training process by dynamically identifying and emphasizing informative time steps within runoff sequences. This complementary learning strategy not only improves the representation of extreme runoff events but also enhances overall model generalization without requiring changes to the underlying network architecture. These findings suggest that selective learning provides an effective and flexible framework that can complement existing deep-learning-based rainfall–runoff models.

Although PPSL-DI-LSTM consistently improves prediction performance, several limitations should be acknowledged. First, the dynamic sample-weighting strategy introduces additional computations during training, increasing the training time from approximately 69 min for DI-LSTM to about 90 min for PPSL-DI-LSTM. However, this additional cost is incurred only during model training, while the inference efficiency remains essentially unchanged. Second, the effectiveness of PPSL depends on dynamic sample difficulty estimation, and prediction errors during the early stages of training may not fully reflect the intrinsic learning difficulty of runoff samples, potentially affecting the accuracy of sample weighting.

In addition, PPSL provides relatively limited improvements in some catchments, suggesting that sample difficulty alone may not fully capture complex hydrological processes. Future work will focus on developing hydrological-regime-aware adaptive weighting strategies to further improve the robustness and generalization capability of the proposed framework. Another limitation is that the proposed framework was validated only using historical hydrological records from the CAMELS dataset. Therefore, its robustness and applicability under current hydrological conditions, as well as its transferability to other large-sample hydrological datasets and geographical regions, remain to be further evaluated using more recent observations and additional datasets.

Peak-emphasis learning entails an inherent trade-off between improved peak representation and magnitude calibration. The slight peak overestimation observed in catchment in Figure 7a illustrates this structural characteristic, as emphasizing extreme events may occasionally amplify peak magnitudes. This behavior reflects a consequence of event-focused optimization rather than model instability, suggesting that future work may refine adaptive weighting to better balance peak preservation and magnitude control. It should also be noted that the static catchment attributes used in this study are inherited from the CAMELS dataset. Consequently, temporal variations in catchment characteristics (e.g., land cover or vegetation changes) are not explicitly represented, which may limit the model's ability to capture long-term watershed evolution. Future work could explore incorporating time-varying catchment descriptors to address this limitation.

5. Conclusions

This study proposed a runoff prediction method that integrates the selective learning mechanism (PPSL) into the DI-LSTM framework, and further examined its effectiveness when integrated into a standard LSTM model, and systematically evaluated the model performance across multiple basins in the CAMELS dataset, with a particular focus on overall prediction accuracy and extreme runoff simulation performance. The main conclusions are as follows:

- (1). By introducing the selective learning mechanism into the DI-LSTM framework, the proposed PPSL-DI-LSTM model achieved consistent improvements in overall prediction accuracy. Compared with DI-LSTM, the median NSE and KGE values increased by approximately 4% and 6%, respectively, while PBias was improved from -3.81% to 1.15% , indicating better runoff volume representation.
- (2). PPSL significantly improved the simulation of extreme runoff processes. The FHV metric improved substantially from -8.20 to -1.29 , demonstrating a clear reduction in flood peak underestimation. In addition, the model achieved more stable low-flow simulation and better reproduced flood peak magnitude and timing during typical extreme flood events.
- (3). When applied to a standard LSTM, PPSL maintained comparable median NSE (from 0.741 to 0.740) while improving median KGE (from 0.751 to 0.778) and substantially reducing FHV (from -12.86% to -3.34%), indicating enhanced representation of peak-flow processes without degrading overall predictive accuracy.

These findings suggest that, in the absence of antecedent streamflow input ($q(t-1)$), PPSL mainly improves peak-flow simulation rather than overall NSE, whereas larger overall accuracy gains emerge when PPSL is integrated with the DI-LSTM framework.

Overall, the proposed PPSL framework improves how deep-learning-based hydrological models are optimized during training by adaptively adjusting the learning order and importance distribution of training samples. The proposed framework effectively alleviates the optimization bias caused by uniformly weighted training, thereby improving overall runoff prediction while enhancing the prediction of hydrologically important ex-

treme runoff events. Furthermore, its model-independent design enables straightforward integration into different deep-learning-based hydrological models without modifying their architectures, highlighting its potential applicability to diverse hydrological datasets and geographical regions.

Author Contributions: Methodology, P.X. and K.Z.; formal analysis, Q.G.; data curation, J.W.; writing—original draft preparation, P.X., K.Z., Q.G., X.M., X.J., J.W., Y.Y., X.L. and Q.L.; writing—review and editing, P.X., K.Z., Q.G., X.M., X.J., J.W., Y.Y., X.L. and Q.L.; visualization, K.Z.; All authors have read and agreed to the published version of the manuscript.

Funding: This research funded by the Science and Technology Development Plan of Jilin Province under Grant 20260102273JC.

Data Availability Statement: The CAMELS data set can be obtained from the U.S. Geological Survey National Water Information System website <https://doi.org/10.5066/F7P55KJN> (accessed on 5 July 2026). The code used in this paper is available at <https://github.com/kwkw-z/PPSL-DI-LSTM> (accessed on 5 July 2026).

Acknowledgments: We thank the School of Artificial Intelligence at Changchun Normal University for providing computing equipment and computer maintenance.

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

Abbreviations

The following abbreviations are used in this manuscript:

PPSL	Peak Protection Selective Learning
LSTM	Long Short-Term Memory
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Network
DI	Data Integration
CAMELS	Catchment Attributes and Meteorology for Large-sample Studies
NSE	Nash-Sutcliffe Efficiency
KGE	Kling-Gupta Efficiency
PBIAS	Percent Bias
R^2	Coefficient of Determination
FHV	Flow Duration Curve High Flow Variability
FLV	Flow Duration Curve Low Flow Variability
RMSE	Root Mean Square Error
SW-RMSE	Selective Weighted Root Mean Square Error
CDF	cumulative density functions
FDC	Flow Duration Curve
PeakErr	Peak Error

Appendix A

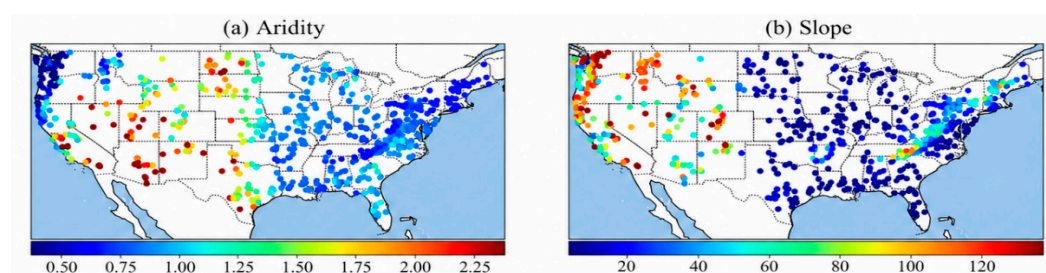


Figure A1. Cont.

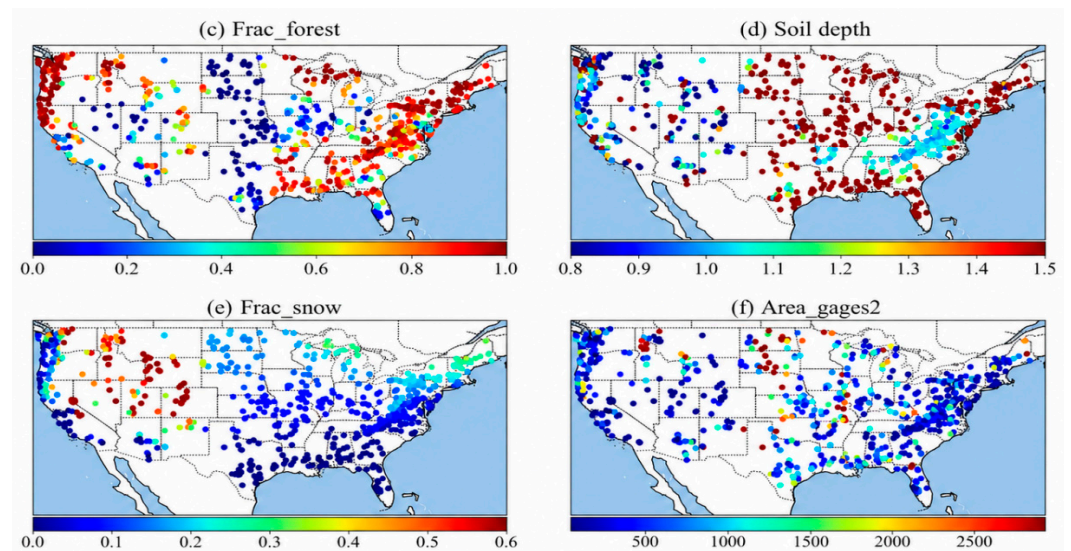


Figure A1. Spatial distribution of CAMELS catchment characteristics. (a) Aridity index: the ratio of the long-term average annual potential evapotranspiration to annual precipitation. (b) Slope: the average slope of the catchment. (c) Forest cover: the proportion of the catchment area covered by forest. (d) Soil depth: the average soil depth of the catchment. (e) Snow cover fraction: the proportion of the catchment area covered by snow. (f) Drainage area: the size of the area controlled by the gauging station. The colorbar below each subplot represents the magnitude of the corresponding catchment characteristic, with units of (a) dimensionless, (b) m/km, (c) dimensionless, (d) m, (e) dimensionless, and (f) km². Colors ranging from blue to red represent increasing values of each characteristic, reflecting the spatial variation in catchments from humid to arid, gentle to steep, sparsely to densely forested, shallow to deep soils, snow-poor to snow-rich, and small to large in drainage area.

Table A1. Parameter settings of different selective learning experiments.

Group	r_u	r_a	q_p	α_p	$w_{\min}$
SL1	0.10	0.10	0.85	1.50	1.50
SL2	0.10	0.05	0.85	2.00	1.50
SL3	0.08	0.05	0.70	2.50	2.00
SL4	0.10	0.10	0.80	1.50	1.50
SL5	0.10	0.07	0.95	1.75	1.50
SL6	0.10	0.10	0.85	1.65	1.50

Table A2. Performance comparison under different selective learning parameter settings.

Group	NSE	KGE	R ²	PBIAS	FLV	FHV
SL1	0.867	0.897	0.867	2.465	2.436	1.817
SL2	0.860	0.893	0.860	2.690	2.9457	2.278
SL3	0.864	0.885	0.864	2.112	3.1045	−7.279
SL4	0.866	0.892	0.866	1.692	−2.925	−1.654
SL5	0.870	0.901	0.870	1.153	−4.540	−1.298
SL6	0.867	0.892	0.867	2.078	−3.229	−1.496

Table A3. Candidate hyperparameter values used for tuning and the selected optimal configuration of the proposed mode.

Hyperparameter	Candidate Values	Selected Value
Training sequence length	150, 200, 365	365
Mini-batch size	50, 100, 150	100
LSTM hidden size	64, 128, 256	256
Dropout rate	0, 0.3, 0.5	0.5
Maximum training epochs	100–500	300

Table A4. Summary of representative flood events and event-based performance comparison between the DI-LSTM and PPSL-DI-LSTM models.

Event Name	Date	Basin ID	Model	Rainfall (mm)	Observed Peak'(mm/day)	Simulated Peak (mm/day)	Peak Error (%)	Timing Error (Days)	Event NSE
1995 Louisiana Heavy Rain Flood	15 October 1995	07376000	DI-LSTM	46.37	12.4279	10.0736	−18.9435	−1	0.8512
	15 October 1995	07376000	PPSL-DI-LSTM	46.37	12.4279	12.0765	−2.8276	0	0.9313
2001 TS Allison Flood (TX)	8 June 2001	08066200	DI-LSTM	62.41	29.5642	26.9468	−8.8534	0	0.7211
	8 June 2001	08066200	PPSL-DI-LSTM	62.41	29.5642	29.2959	−0.9075	0	0.7402
1999 Hurricane Floyd (NC)	18 September 1999	02053200	DI-LSTM	11.26	63.7149	43.7581	31.3219	+1	0.9016
	18 September 1999	02053200	PPSL-DI-LSTM	11.26	63.7149	60.5816	−4.9177	0	0.9718
2005 Hurricane Katrina	30 August 2005	02481000	DI-LSTM	170.03	76.6670	40.0362	−47.7790	0	0.8206
	30 August 2005	02481000	PPSL-DI-LSTM	170.03	76.6670	48.5417	−36.6850	0	0.8739

References

- Wang, X.; Xie, H. A review on applications of remote sensing and geographic information systems (GIS) in water resources and flood risk management. *Water* **2018**, *10*, 608. [\[CrossRef\]](#)
- Jia, W.; Chen, M.; Yao, H.; Wang, Y.; Wang, S.; Ni, X. Improving sub-daily runoff forecast based on the multi-objective optimized extreme learning machine for reservoir operation. *Water Resour. Manag.* **2024**, *38*, 6173–6189. [\[CrossRef\]](#)
- Patel, A.; Yadav, S.; Teegavarapu, R. Enhancing real-time flood forecasting and warning system by integrating ensemble techniques and hydrologic model simulations. *J. Water Clim. Change* **2024**, *15*, 4307–4327. [\[CrossRef\]](#)
- Merz, B.; Basso, S.; Fischer, S.; Lun, D.; Blöschl, G.; Merz, R.; Guse, B.; Viglione, A.; Vorogushyn, S.; Macdonald, E. Understanding heavy tails of flood peak distributions. *Water Resour. Res.* **2022**, *58*, e2021WR030506. [\[CrossRef\]](#)
- Coulthard, T.J.; Skinner, C.J. The sensitivity of landscape evolution models to spatial and temporal rainfall resolution. *Earth Surf. Dyn.* **2016**, *4*, 757–771. [\[CrossRef\]](#)
- Adamczyk, J.; Malawski, F. Comparison of manual and automated feature engineering for daily activity classification in mental disorder diagnosis. *Comput. Inform.* **2021**, *40*, 850. [\[CrossRef\]](#)
- Kraft, B.; Jung, M.; Körner, M.; Koirala, S.; Reichstein, M. Towards hybrid modeling of the global hydrological cycle. *Hydrol. Earth Syst. Sci.* **2022**, *26*, 1579–1614. [\[CrossRef\]](#)
- Jogin, M.; Madhulika, M.; Divya, G.; Meghana, R.; Apoorva, S. Feature extraction using convolution neural networks (CNN) and deep learning. In *Proceedings of the 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 18–19 May 2018*; IEEE: New York, NY, USA, 2018.
- Su, L.; Zuo, X.; Li, R.; Wang, X.; Zhao, H.; Huang, B. A systematic review for transformer-based long-term series forecasting. *Artif. Intell. Rev.* **2025**, *58*, 80. [\[CrossRef\]](#)
- Arsenault, R.; Martel, J.-L.; Brunet, F.; Brissette, F.; Mai, J. Continuous streamflow prediction in ungauged basins: Long Short-Term Memory Neural Networks clearly outperform traditional hydrological models. *Hydrol. Earth Syst. Sci. Discuss.* **2023**, *27*, 139–157. [\[CrossRef\]](#)
- Shekar, P.R.; Mathew, A.; Sharma, K.V. A hybrid CNN–RNN model for rainfall–runoff modeling in the Potteruvagu watershed of India. *CLEAN–Soil Air Water* **2025**, *53*, 2300341. [\[CrossRef\]](#)
- Hashemi, R.; Brigode, P.; Garambois, P.-A.; Javelle, P. How can we benefit from regime information to make more effective use of long short-term memory (LSTM) runoff models? *Hydrol. Earth Syst. Sci.* **2022**, *26*, 5793–5816. [\[CrossRef\]](#)
- Calvette, T.; Reis, M.; Paz, I. Hybrid Deep Learning Time-Lagged Precipitation-Runoff Model. *J. Hydrol. Eng.* **2024**, *29*, 05024020. [\[CrossRef\]](#)
- Kratzert, F.; Klotz, D.; Brenner, C.; Schulz, K.; Herrnegger, M. Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 6005–6022. [\[CrossRef\]](#)
- Deng, H.; Chen, W.; Huang, G. Deep insight into daily runoff forecasting based on a CNN-LSTM model. *Nat. Hazard.* **2022**, *113*, 1675–1696. [\[CrossRef\]](#)
- Sharma, P.; Tripathi, V.; Mohanty, M.P. Improving Daily Streamflow Predictions over Large Watersheds: Introducing a Novel Enhanced Long Short-Term Memory (En-LSTM) Model. *Water Resour. Manag.* **2026**, *40*, 92. [\[CrossRef\]](#)
- Chen, X.; Huang, J.; Wang, S.; Zhou, G.; Gao, H.; Liu, M.; Yuan, Y.; Zheng, L.; Li, Q.; Qi, H. A new rainfall-runoff model using improved LSTM with attentive long and short lag-time. *Water* **2022**, *14*, 697. [\[CrossRef\]](#)
- Girihagama, L.; Naveed Khaliq, M.; Lamontagne, P.; Perdikaris, J.; Roy, R.; Sushama, L.; Elshorbagy, A. Streamflow modelling and forecasting for Canadian watersheds using LSTM networks with attention mechanism. *Neural Comput. Appl.* **2022**, *34*, 19995–20015. [\[CrossRef\]](#)
- Man, Y.; Yang, Q.; Shao, J.; Wang, G.; Bai, L.; Xue, Y. Enhanced LSTM model for daily runoff prediction in the upper Huai river basin, China. *Engineering* **2023**, *24*, 229–238. [\[CrossRef\]](#)
- Li, Y.; Xu, J.; Anastasiu, D.C. An extreme-adaptive time series prediction model based on probability-enhanced lstm neural networks. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 8684–8691. [\[CrossRef\]](#)
- Tofighi, S.; Gurbuz, F.; Mantilla, R.; Xiao, S. Advancing Machine Learning-Based Streamflow Prediction Through Event Greedy Selection, Asymmetric Loss Function, and Rainfall Forecasting Uncertainty. *Appl. Sci.* **2025**, *15*, 11656. [\[CrossRef\]](#)
- Fu, Y.; Shao, Z.; Yu, C.; Li, Y.; An, Z.; Wang, Q.; Xu, Y.; Wang, F. Selective learning for deep time series forecasting. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 98084–98115. [\[CrossRef\]](#)
- Li, J.; Ma, X.; Shi, Y. Dual-Stage Clean-Sample Selection for Incremental Noisy Label Learning. *Bioengineering* **2025**, *12*, 743. [\[CrossRef\]](#) [\[PubMed\]](#)
- Slater, L.J.; Anderson, B.; Buechel, M.; Dadson, S.; Han, S.; Harrigan, S.; Kelder, T.; Kowal, K.; Lees, T.; Matthews, T. Nonstationary weather and water extremes: A review of methods for their detection, attribution, and management. *Hydrol. Earth Syst. Sci. Discuss.* **2020**, *25*, 3897–3935. [\[CrossRef\]](#)
- Mangunkiya, N.K.; Kumar, K.B.; Dey, P.; Sharma, S.; Bejagam, V.; Mujumdar, P.P.; Sharma, A. CAMELS-IND: Hydrometeorological time series and catchment attributes for 228 catchments in Peninsular India. *Earth Syst. Sci. Data* **2025**, *17*, 461–491. [\[CrossRef\]](#)

26. Lees, T.; Buechel, M.; Anderson, B.; Slater, L.; Reece, S.; Coxon, G.; Dadson, S.J. Benchmarking Data-Driven Rainfall-Runoff Models in Great Britain: A comparison of LSTM-based models with four lumped conceptual models. *Hydrol. Earth Syst. Sci. Discuss.* **2021**, *25*, 5517–5534. [\[CrossRef\]](#)
27. Yang, Q.; Zhang, H.; Wang, G.; Luo, S.; Chen, D.; Peng, W.; Shao, J. Dynamic runoff simulation in a changing environment: A data stream approach. *Environ. Model. Softw.* **2019**, *112*, 157–165. [\[CrossRef\]](#)
28. Newman, A.J.; Clark, M.P.; Sampson, K.; Wood, A.; Hay, L.E.; Bock, A.; Viger, R.J.; Blodgett, D.; Brekke, L.; Arnold, J. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrol. Earth Syst. Sci.* **2015**, *19*, 209–223. [\[CrossRef\]](#)
29. Feng, D.; Fang, K.; Shen, C. Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resour. Res.* **2020**, *56*, e2019WR026793. [\[CrossRef\]](#)
30. Addor, N.; Newman, A.J.; Mizukami, N.; Clark, M.P. The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrol. Earth Syst. Sci.* **2017**, *21*, 5293–5313. [\[CrossRef\]](#)
31. Kratzert, F.; Klotz, D.; Shalev, G.; Klambauer, G.; Hochreiter, S.; Nearing, G. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrol. Earth Syst. Sci.* **2019**, *23*, 5089–5110. [\[CrossRef\]](#)
32. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Van Vliet, M.T.; Franssen, W.H.; Yearsley, J.R.; Ludwig, F.; Haddeland, I.; Lettenmaier, D.P.; Kabat, P. Global river discharge and water temperature under climate change. *Glob. Environ. Change* **2013**, *23*, 450–464. [\[CrossRef\]](#)
34. Gupta, H.V.; Sorooshian, S.; Yapo, P.O. Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. *J. Hydrol. Eng.* **1999**, *4*, 135–143. [\[CrossRef\]](#)
35. Gupta, H.V.; Kling, H.; Yilmaz, K.K.; Martinez, G.F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.* **2009**, *377*, 80–91. [\[CrossRef\]](#)
36. Nash, J.E.; Sutcliffe, J.V. River flow forecasting through conceptual models part I—A discussion of principles. *J. Hydrol.* **1970**, *10*, 282–290. [\[CrossRef\]](#)
37. Note, M.L. A note on a general definition of the coefficient of determination. *Biometrika* **1991**, *78*, 691–692. [\[CrossRef\]](#)
38. Yilmaz, K.K.; Gupta, H.V.; Wagener, T. A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resour. Res.* **2008**, *44*, W09417. [\[CrossRef\]](#)
39. Nossent, J.; Bauwens, W. Application of a normalized Nash-Sutcliffe efficiency to improve the accuracy of the Sobol’ sensitivity analysis of a hydrological model. In *EGU General Assembly Conference Abstracts*; EGU: Vienna, Austria, 2012.
40. Shen, C. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res.* **2018**, *54*, 8558–8593. [\[CrossRef\]](#)
41. Moriasi, D.N.; Arnold, J.G.; Van Liew, M.W.; Bingner, R.L.; Harmel, R.D.; Veith, T.L. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* **2007**, *50*, 885–900. [\[CrossRef\]](#)

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.