

Article

Inter-Comparison of Deep Learning Models for Flood Forecasting in Ethiopia's Upper Awash Basin

Girma Moges Mengistu ¹, Addisu G. Semie ^{2,3,4,*}, Gulilat T. Diro ⁵, Natei Ermias Benti ⁶,
Emiola O. Gbobaniyi ⁷ and Yonas Mersha ⁸

- ¹ Computational Data Science Program, College of Natural and Computational Sciences, Addis Ababa University, Addis Ababa P.O. Box 1176, Ethiopia
² LEAP NSF Science and Technology Center, Columbia University, New York, NY 10027, USA
³ Department of Earth and Environmental Engineering, Columbia University, New York, NY 10027, USA
⁴ NSF National Center for Atmospheric Research, Boulder, CO 80301, USA
⁵ ESCER Center, University of Quebec at Montreal, Montreal, QC H2X 3Y7, Canada
⁶ Center for Environmental Science, College of Natural and Computational Sciences, Addis Ababa University, Addis Ababa P.O. Box 1176, Ethiopia
⁷ Swedish Meteorological and Hydrological Institute (SMHI), 60176 Norrköping, Sweden
⁸ International Livestock Research Institute (ILRI), Addis Ababa P.O. Box 5689, Ethiopia
* Correspondence: as7690@columbia.edu

Abstract

Flood events driven by climate variability and change pose significant risks for socio-economic activities in the Awash Basin, necessitating advanced forecasting tools. This study benchmarks five deep learning (DL) architectures, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM (BiLSTM), and a Hybrid CNN–LSTM, for daily discharge forecasting for the Hombole catchment in the Upper Awash Basin (UAB) using 40 years of hydrometeorological observations (1981–2020). Rainfall, lagged discharge, and seasonal indicators were used as predictors. Model performance was evaluated against two baseline approaches, a conceptual HBV rainfall–runoff model as well as a climatology, using standard and hydrological metrics. Of the two baselines (climatology and HBV), the climatology showed limited skill with large bias and negative NSE, whereas the HBV model achieved moderate skill (NSE = 0.64 and KGE = 0.82). In contrast, all DL models substantially improved predictive performance, achieving test NSE values above 0.83 and low overall bias. Among them, the Hybrid CNN–LSTM provided the most balanced performance, combining local temporal feature extraction with long-term memory and yielding stable efficiency (NSE $\approx$ 0.84, KGE $\approx$ 0.90, and PBIAS $\approx$ −2%) across flow regimes. The LSTM and GRU models performed comparably, offering strong temporal learning and robust daily predictions, while BiLSTM improved flood timing through bidirectional sequence modeling. The CNN captured short-term variability effectively but showed weaker representation of extreme peaks. Analysis of peak-flow metrics revealed systematic underestimation of extreme discharge magnitudes across all models. However, a post-processing flow-regime classification based on discharge quantiles demonstrated high extreme-event detection skill, with deep learning models exceeding 89% accuracy in identifying extreme-flow occurrences on the test set. These findings indicate that, while magnitude errors remain for rare floods, DL models reliably discriminate flood regimes relevant for early warning. Overall, the results show that deep learning models provide clear improvements over climatology and conceptual baselines for daily streamflow forecasting in the UAB, while highlighting remaining challenges in peak-flow magnitude prediction. The study indicates promising results for the integration of deep learning methods into flood early-warning workflows; however, these



Academic Editor: Fang Yenn Teo

Received: 9 November 2025

Revised: 25 January 2026

Accepted: 28 January 2026

Published: 3 February 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.

results could be further improved by adopting a probabilistic forecasting framework that accounts for model uncertainty.

Keywords: flood forecasting; deep learning; long short-term memory; gated recurrent unit; hydrological modeling; Upper Awash Basin; artificial intelligence

1. Introduction

Floods are among the most devastating and recurrent natural disasters, causing widespread socio-economic and environmental disruption worldwide. Between 1998 and 2017, over 2 billion people globally were affected by floods, according to the World Health Organization [1]. Their impacts, ranging from displacement of people and infrastructural damage to agricultural losses and loss of life, are especially pronounced in developing regions [2]. Climate change has intensified the hydrological cycle, increasing the frequency and magnitude of flood events [1]. Compounding this are human-induced factors such as deforestation, urban encroachment on wetlands, and unregulated land-use practices that increase surface runoff and weaken natural flood mitigation systems.

In Ethiopia, the risks associated with flooding are magnified by limited infrastructure and adaptive capacity, weak early-warning mechanisms, and the reliance of a large portion of the population on climate-sensitive livelihoods. Recent floods have affected over 590,000 people nationwide, displacing around 95,000 [3]. The most impacted regions include Somali, Oromia, Sidama, southern and central Ethiopia, Amhara, and Tigray. In the Somali region, 247,000 individuals have been affected, with 51,000 displaced and approximately 18,000 hectares of cropland lost. Oromia has experienced impacts on more than 285,000 people, with 38,300 displaced and 34,700 hectares of cropland destroyed [3]. Nationwide, floods have caused over 2900 livestock deaths. These disasters have severely damaged homes, infrastructure, and farmland, exacerbating vulnerabilities in areas already affected by conflict and drought [3,4]. The Upper Awash Basin (UAB), located within one of Ethiopia's most economically vital and densely populated regions, exemplifies this vulnerability. With rapid urbanization, intensified agriculture, and high population density, the UAB frequently experiences severe flooding, leading to damage to housing, transportation infrastructure, and cropland, worsening food insecurity and poverty among vulnerable communities.

Flood forecasting systems are key to reducing disaster risks by providing early warnings and enabling proactive responses. These systems have evolved from simple statistical approaches to increasingly complex physical models, which can be broadly categorized into statistical and process-based frameworks. Statistical models are computationally efficient but often lack robustness due to simplified assumptions and a limited ability to represent non-stationary and extreme events. Physical models simulate rainfall–runoff processes by integrating precipitation, catchment characteristics, and stream network properties and range from conceptual lumped models such as HBV to semi-distributed models like SWAT [5] and fully distributed systems including WRF-Hydro [6,7] and GeoSFM [8]. While lumped and conceptual models are attractive for their simplicity and operational efficiency, their oversimplified representation of spatial heterogeneity can lead to reduced forecast skill in complex terrains. Fully distributed models offer a more detailed and physically consistent representation of hydrological processes, but they are computationally demanding and often exhibit strong sensitivity to parameterization choices, making regional calibration challenging.

The emergence of artificial intelligence, especially deep learning (DL) techniques, offers new opportunities to enhance hydrological modeling by leveraging large volumes of historical data to learn intricate spatiotemporal relationships. Among these approaches, Convolutional Neural Networks (CNNs) [9], initially developed for image recognition, have also been applied to hydrological applications due to their strength in capturing spatial and temporal features. They have been successfully used to analyze gridded meteorological inputs and satellite data for flood modeling [10,11]. Research by [12,13] demonstrated CNNs' effectiveness in flood inundation mapping and forecasting using diverse datasets, ranging from IoT-enhanced hydrological observations to outputs from 2D hydraulic models. Their ability to generalize from spatially rich data makes CNNs particularly valuable for flood-prone regions with complex topographies.

For time series prediction, Recurrent Neural Networks (RNNs) and their variant, Long Short-Term Memory (LSTM) networks, have been widely recognized for their ability to capture temporal dependencies. Unlike basic RNNs that suffer from short-term memory limitations [14], LSTMs are designed to learn long-term dependencies, making them well-suited for hydrological forecasting. Developed by [15], LSTM networks use memory cells regulated by input, output, and forget gates to retain and update information over long sequences. Gate Recurrent Unit (GRU) is a simpler variant of LSTM, proposed by [16]. The key difference between GRU and LSTM is that GRU merges the input and forget gates into a single update gate, reducing the number of parameters and simplifying training. These models have proven especially useful for capturing delayed hydrological responses to rainfall events.

More recently, hybrid architectures like Convolutional LSTM (ConvLSTM) have integrated the spatial learning capabilities of CNNs with the temporal learning strengths of LSTMs [17]. This combination enables ConvLSTM models to simultaneously capture spatial features (e.g., rainfall distribution) and temporal dependencies (e.g., flood propagation), making them highly suitable for flood forecasting in complex environments such as the UAB.

Despite these global advances, deep learning applications for streamflow forecasting in Ethiopian basins remain extremely limited, with no prior work conducting a comprehensive, long-term, multi-model evaluation that enables direct comparison of deep learning approaches within a single basin. Existing studies rely on single-model experiments or limited event-focused analyses, providing neither a generalization assessment nor extreme-flow diagnostics. Yet the UAB possesses a uniquely rich 40-year hydrometeorological archive, offering a rare opportunity for rigorous data-driven hydrological evaluation.

This study addresses these gaps by formulating three tightly connected research questions. First, how reliably can different deep learning (DL) architectures learn, generalize, and reproduce extreme streamflow dynamics in the Upper Awash Basin (UAB) when evaluated using a consistent, multi-decade dataset? Second, which DL architecture demonstrates superior skill in streamflow forecasting for the UAB, particularly in capturing peak flows and hydrological extremes? Third, to what extent do DL-based forecasts improve predictive performance relative to a reference model, such as a traditional conceptual hydrological model, under the same data and evaluation framework? To answer this, we conduct the first 40-year benchmark of DL models in the UAB, evaluating CNN, LSTM, GRU, BiLSTM, and Hybrid CNN–LSTM architectures using historical rainfall and discharge data from 1981 to 2020. This long-term, multi-model benchmark represents a novel data-driven hydrological experiment for Ethiopia and for East African catchments more broadly. Beyond standard metrics such as Mean Absolute Error (MAE) and root mean squared error (RMSE), we integrate hydrology-specific skill scores (Nash–Sutcliffe Efficiency (NSE), Kling–Gupta Efficiency (KGE), and percent bias (PBIAS)), seasonal diagnostics, and peak-flow perfor-

mance evaluation. The outcome of this study represents an important step toward the development of robust early-warning systems and improved disaster risk reduction frameworks, contributing to Ethiopia's climate resilience and sustainable development goals. The framework is also adaptable for flood forecasting in other basins across East Africa.

2. Materials and Methods

2.1. Study Area

The UAB, located in central Ethiopia between latitudes $8^{\circ}16' \text{ N}$ and $9^{\circ}18' \text{ N}$ and longitudes $37^{\circ}57' \text{ E}$ and $39^{\circ}17' \text{ E}$, covers an area of approximately 7656 km^2 (see Figure 1). The basin's elevation ranges from 1580 m to 3396 m above sea level, with the Awash River, originating from the Ethiopian highlands, serving as its main watercourse and playing a central role in shaping the region's hydrology and socio-economic activities [17,18]. A key feature of the basin is the Hombole hydrological station, which acts as the primary outlet of the UAB and contributes about 67% of the total flow into the Koka Dam, a major hydropower reservoir [19]. This makes Hombole a critical point for monitoring and managing river flow. Its long-term hydrological data provide a strong basis for applying deep learning in flood forecasting. The basin experiences frequent flooding, mainly due to heavy rainfall and complex hydroclimatic conditions [5,20]. These factors make flood prediction challenging but also create opportunities to develop advanced forecasting models. This study focuses on Hombole due to its data availability, hydrological importance, and the need for reliable flood forecasting tools in the basin.

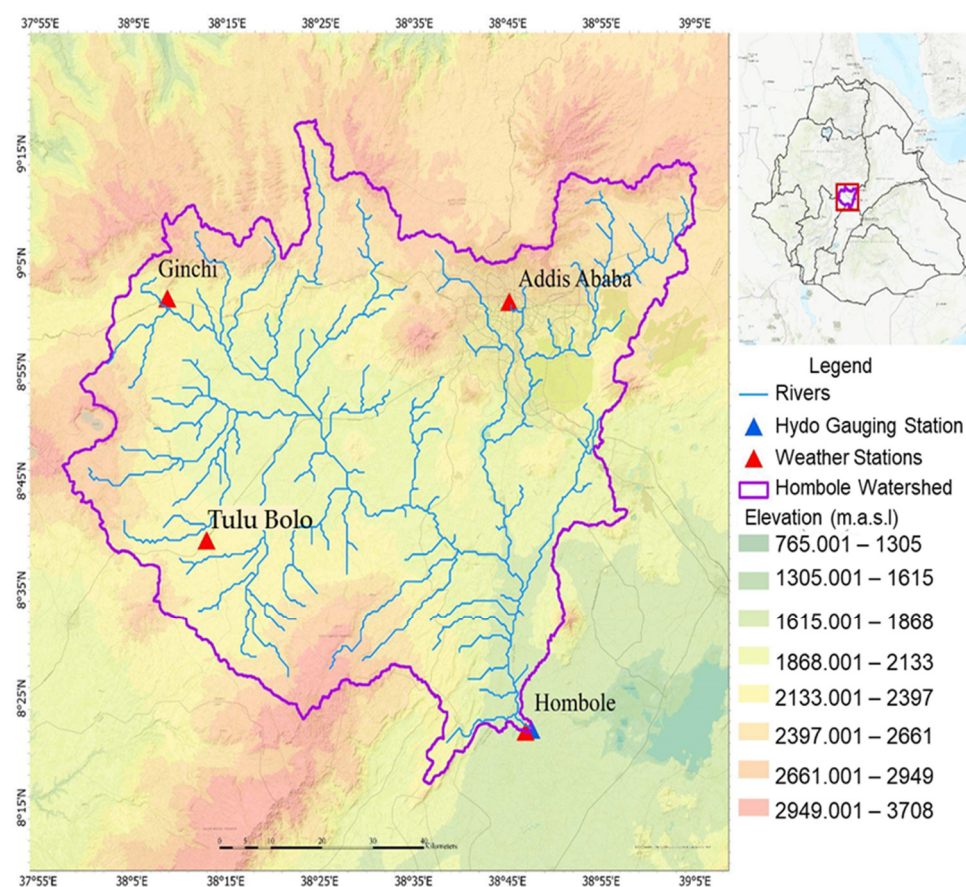


Figure 1. Map of the Hombole catchment located within the UAB, Ethiopia.

2.2. Dataset and Pre-Processing

2.2.1. Data Collection and Description

This study used daily hydrological and meteorological data from the UAB covering 1981–2020. The dataset includes 14,610 daily observations. Discharge data, measured in cubic meters per second (m^3/s), were obtained from the Ethiopian Ministry of Water and Energy (MoWE) at the Hombole gauge station. The discharge record is complete, with no missing values. Daily rainfall data were obtained from the Ethiopian Meteorological Institute (EMI). The gridded dataset has a spatial resolution of $4 \text{ km} \times 4 \text{ km}$ and covers the Hombole catchment and nearby stations: Hombole, Tulu Bolo, Ginchi, and Addis Ababa. The rainfall dataset (1981–2020) is also complete, with no gaps.

Hydrometeorological statistics show a mean, median, and maximum discharge of 44.69, 8.07, and $803.1 \text{ m}^3/\text{s}$, respectively. Rainfall means range from 2.14 to 3.08 mm , medians from 0.10 to 3.75 mm , and maxima from 66.54 to 86.81 mm across the four stations. The results confirm strong seasonality, with peak rainfall between June and September and corresponding discharge peaks in July and August. Table 1 presents the summary statistics for the key variables in the dataset.

Table 1. Summary statistics for discharge (m^3/s , at Hombole station) and rainfall (mm/day) at four locations.

Variables	Count	Mean	Std	Min	25%	50%	75%	Max
Discharge	14,610	44.69	79.27	0.40	4.22	8.074	40.37	803.10
Hombole	14,610	2.14	6.93	0.00	0.00	0.009	0.52	82.836
Tulu Bolo	14,610	2.93	6.25	0.00	0.00	0.082	2.57	83.248
Ginchi	14,610	3.08	5.91	0.00	0.00	0.176	3.75	66.537
Addis Ababa	14,610	2.77	5.58	0.00	0.00	0.103	3.00	86.808

Figure 2 shows a time series of daily discharge and rainfall for all four stations from 1981 to 2020. The plots highlight distinct wet and dry seasons, with low flows from January to May and October to December and sharp peaks during the main rainy season. Maximum rainfall occurs at Addis Ababa (86.81 mm) and Tulu Bolo, which strongly influences high-flow events in Hombole. A clear lag between rainfall and discharge peaks reflects runoff routing time, an important factor in model input design. The year-to-year variation in peak discharge also indicates climatic variability, supporting the inclusion of lagged rainfall, month, and seasonal indicators in the DL models. The completeness and higher temporal resolution of these datasets make them suitable for DL applications that depend on large, consistent records. In data-limited regions such as Ethiopia, this level of data integrity eliminates the need for imputation and improves the reliability and generalization of the model [21,22].

Figure 3 shows the monthly discharge distribution at Hombole station based on 14,610 daily records (1981–2020). The box plot highlights strong seasonal variation. Discharge is lowest from January to May, with medians near or below $50 \text{ m}^3/\text{s}$, corresponding to dry-season conditions. From June to September, discharge rises sharply, reaching medians and interquartile ranges between 300 and $600 \text{ m}^3/\text{s}$. July and August exhibit the highest peaks, with values exceeding $800 \text{ m}^3/\text{s}$ in extreme events. These results align with the UAB rainfall pattern, where the main rainy season occurs from June to September.

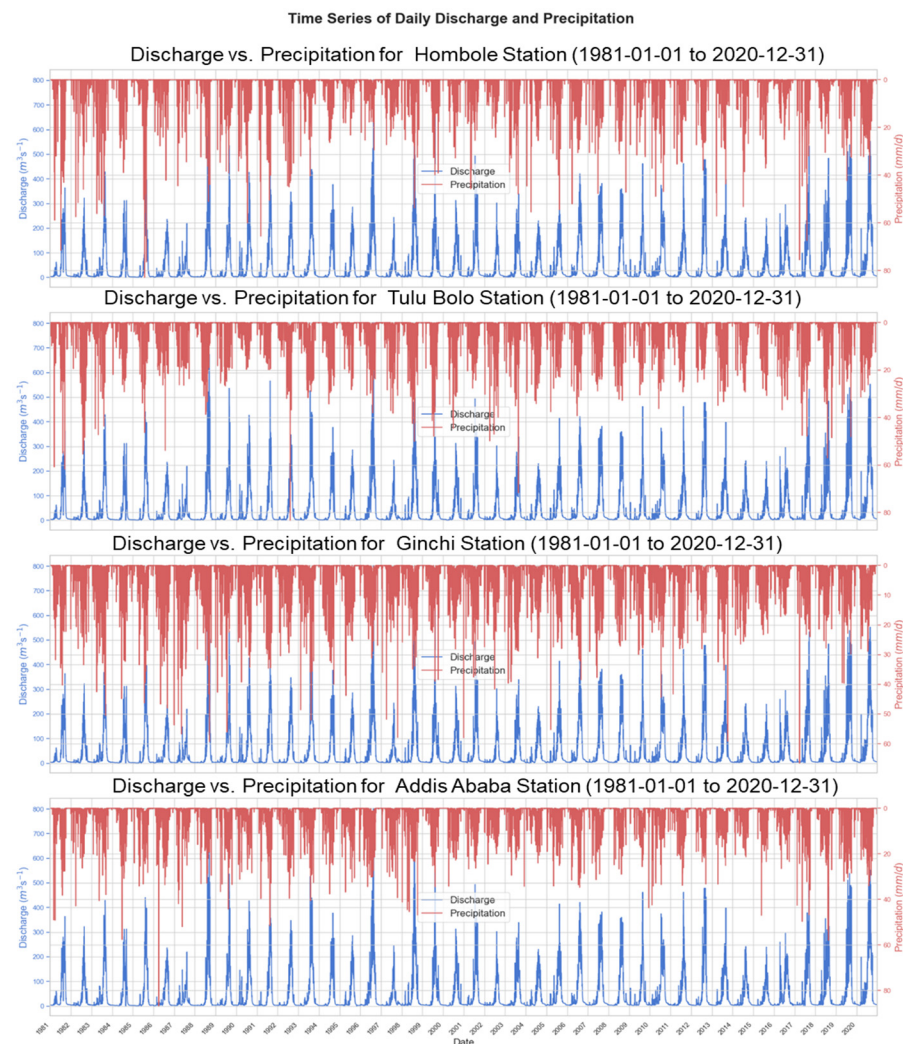


Figure 2. Daily rainfall and discharge time series for the Hombole, Tulu Bolo, Ginchi, and Addis Ababa stations from 1981 to 2020. The plot shows strong seasonal rainfall and discharge patterns with clear wet and dry periods.

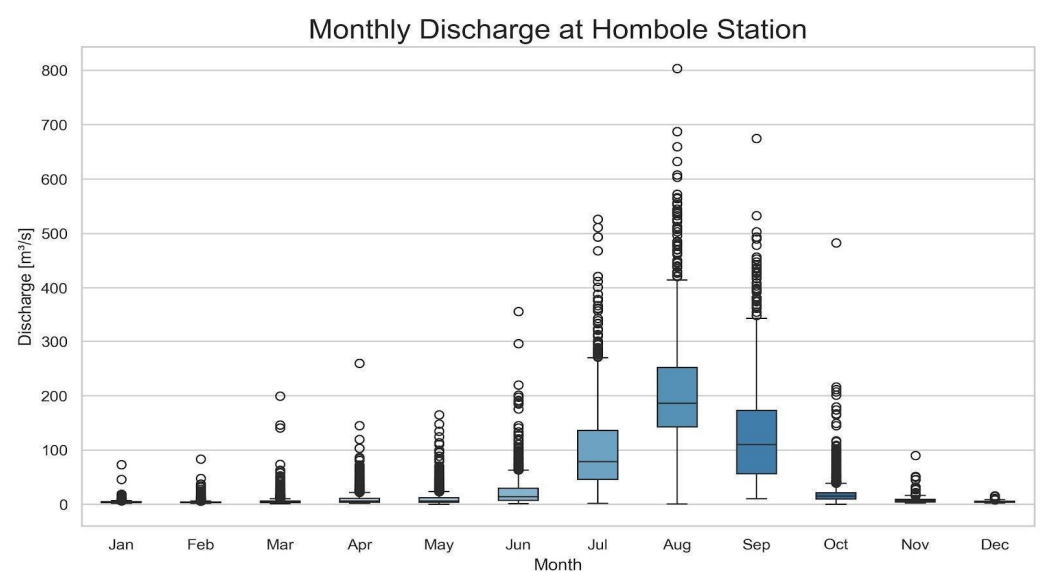


Figure 3. Monthly discharge distribution at Hombole station (1981–2020). The plot shows pronounced wet-season peaks and dry-season lows, consistent with rainfall patterns across the UAB.

2.2.2. Feature Engineering

Time series data like daily discharge and precipitation often exhibit cyclical patterns driven by natural rhythms, daily temperature shifts, weekly weather fluctuations, and seasonal rainfall (e.g., the June–September rainy season). Capturing these cycles is essential for accurate flood forecasting, especially using models such as RNNs and their variants, which benefit from temporally coherent inputs. Linear encoding of time features (e.g., months as 1–12) introduces discontinuities, such as between December (12) and January (1), that distort the true cyclical nature of the data and can hinder the model’s understanding of seasonal transitions. To address this, cyclical features were encoded using sine and cosine transformations, mapping time variables onto a unit circle and preserving their periodic continuity [23,24].

For monthly cycles,

$$mon_sin = \sin\left(\frac{2\pi \cdot month}{12}\right) \quad (1)$$

$$mon_cos = \cos\left(\frac{2\pi \cdot month}{12}\right) \quad (2)$$

where the month ranges from 1 to 12.

For weekly cycles,

$$Dayofweek_sin = \sin\left(\frac{2\pi \cdot Day\ of\ week}{7}\right) \quad (3)$$

$$Dayofweek_cos = \cos\left(\frac{2\pi \cdot Day\ of\ week}{7}\right) \quad (4)$$

where Dayofweek ranges from 0 to 6 (Monday to Sunday).

For daily cycles (normalized by month length),

$$Day_sin = \sin\left(\frac{2\pi \cdot Day}{Days\ in\ month}\right) \quad (5)$$

$$Day_cos = \cos\left(\frac{2\pi \cdot Day}{Days\ in\ month}\right) \quad (6)$$

This encoding is particularly important for the UAB, where flood patterns depend on seasonal rainfall and exhibit lagged responses between precipitation and discharge. Short-term lagged discharge predictors (2–7 days) were tested but found to be redundant, as the 14-day input window already captures temporal dependencies.

The UAB shows a strong seasonal hydrological pattern dominated by the JJA rainfall regime. Sinusoidal encodings of day-of-year and month were incorporated to explicitly represent this seasonality, enhancing the deep learning models’ ability to capture interannual variability and lagged hydrologic responses. While full deseasonalization was not applied, the results indicate that a substantial portion of predictive skill derives from the seasonal signal. This highlights both a strength and limitation of the modeling approach: deep learning effectively captures predictable seasonal dynamics, but forecasting anomalies or extreme events beyond this seasonal structure remains more challenging.

2.2.3. Correlation Analysis and Feature Selection

To improve the efficiency and accuracy of the model, Pearson’s correlation analysis was applied to identify key predictors of next-day discharge. Variables with strong or moderate correlations were retained, while those with negligible influence were excluded to reduce dimensionality and avoid overfitting. The correlation heatmap (Figure 4) highlights relationships between discharge in Hombole; precipitation from Hombole, Tulu Bolo,

Ginchi, and Addis Ababa stations; and time-based features. Discharge strongly correlates with target discharge ($r = 0.92$), confirming hydrological persistence and reasserting lagged discharge as a key predictor. Precipitation shows a moderate correlation with both discharge ($r = 0.29$ – 0.40) and target discharge ($r = 0.34$ – 0.48), peaking at Addis Ababa, indicating rainfall-driven discharge with a time delay. Spatial rainfall variability is evident, with higher peak values at Addis Ababa (86.81 mm) and Tulu Bolo (83.22 mm) compared to Ginchi (66.54 mm), underscoring their influence in flood dynamics. Month sin and month cos show moderate negative correlations with discharge and target discharge, capturing seasonal trends. Daily and weekly cyclical variables exhibit negligible correlations, suggesting minimal impact at the daily scale. Based on this analysis, the selected features, discharge, lagged discharge, precipitation (especially from Addis Ababa and Tulu Bolo), and monthly cycle indicators, provide a focused input set for the deep learning model, effectively capturing the seasonal, spatial, and temporal dynamics essential for flood forecasting.

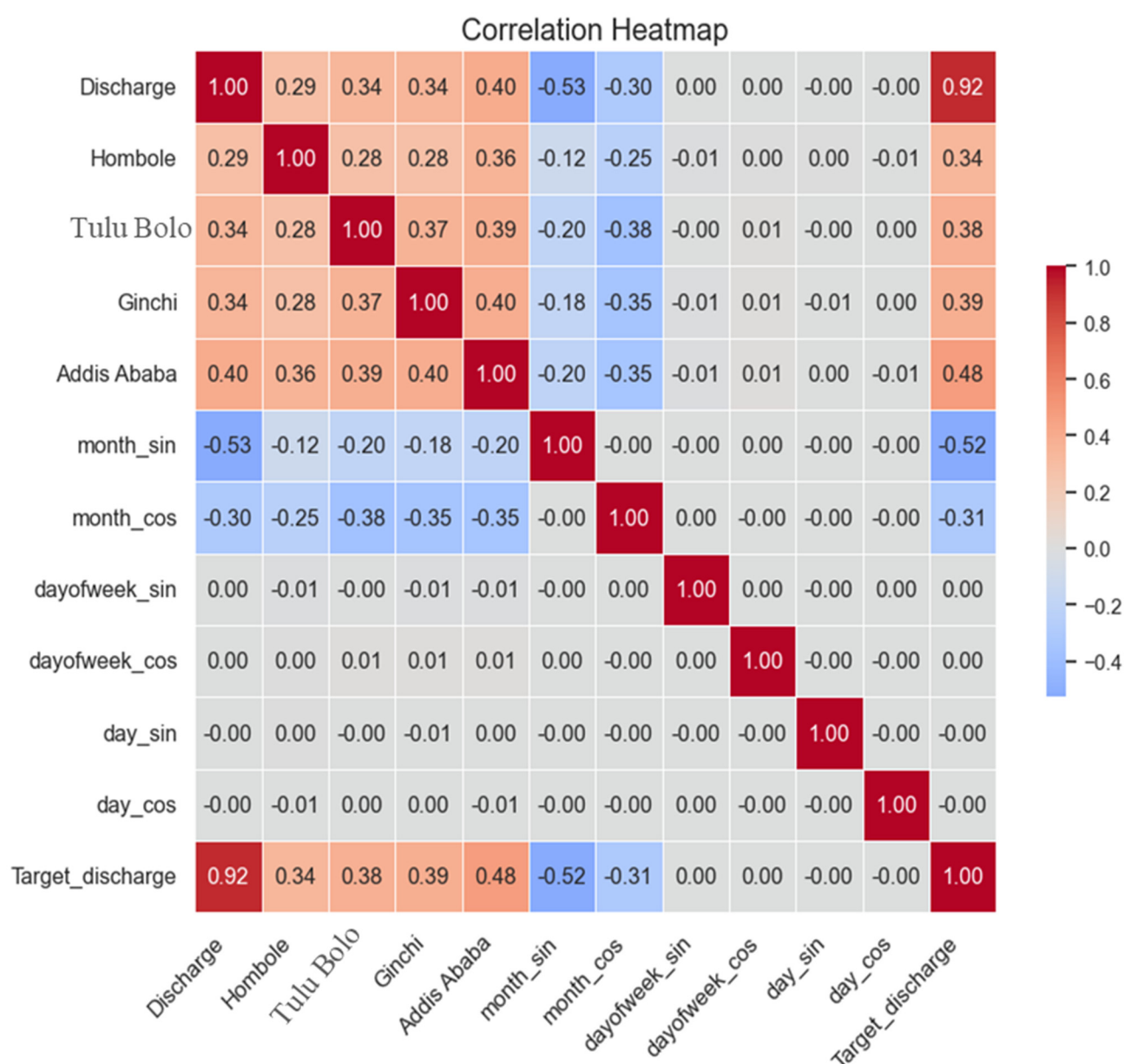


Figure 4. Heatmap of cross-correlations between discharge, lagged discharge, and precipitation at the Hombole, Tulu Bolo, Ginchi, and Addis Ababa stations from 1981 to 2020.

2.3. Deep Learning Models

2.3.1. Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) [25] were first developed for image classification but are now widely used in hydrological modeling, including flood forecasting with spatial and temporal inputs [10,26]. CNNs use convolutional layers to detect localized patterns within structured data, enabling the model to identify hierarchies and dependencies. This structure makes CNNs effective for feature extraction from satellite imagery, digital elevation models, rainfall sequences, and sensor-based hydrological records.

In this study, a one-dimensional temporal convolutional (Conv1D) network was implemented to extract short-term features from multivariate hydrological time series consisting of precipitation and lagged discharge. Unlike spatial CNNs that process two-dimensional image grids, the Conv1D filters slide along the time axis to identify rainfall–runoff patterns. The convolutional network comprised three sequential Conv1D layers. The first, second, and third convolutional layers employed 64, 128, and 256 filters, respectively, each with a kernel size of 2. Each Conv1D layer was followed by max pooling and batch normalization, with a dropout rate of 0.2 applied to mitigate overfitting. The resulting feature maps were flattened and passed through fully connected dense layers to produce the final discharge prediction. This hierarchical configuration enabled the model to progressively learn local temporal patterns at increasing levels of abstraction while maintaining relatively low model complexity.

2.3.2. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks [15] are a class of Recurrent Neural Networks (RNNs) designed to learn long-term dependencies in sequential data. They have been widely applied in hydrology for streamflow and flood forecasting due to their ability to retain temporal context and capture delayed rainfall–runoff responses [14,24,26–28]. LSTM cells use three internal gates, input, forget, and output, to regulate information flow and prevent gradient vanishing. This gating structure enables the network to remember relevant information from earlier time steps while discarding noise from less important inputs. In this study, the LSTM model was used on multivariate time series consisting of rainfall, lagged discharge, and seasonal indicators. The architecture included one to three stacked LSTM layers with 64 and 50 units, followed by dense layers (50, 1 neuron). Dropout layers with a rate of 0.2, batch normalization, and L2 regularization were employed to prevent overfitting.

2.3.3. Gated Recurrent Unit (GRU)

Gated Recurrent Unit (GRU) networks [25] are a simplified type of Recurrent Neural Network (RNN) designed to capture long-term dependencies in sequential data with fewer parameters than LSTM. GRUs have been widely used in hydrology for streamflow and flood forecasting due to their efficiency and ability to model nonlinear rainfall–runoff relationships [29,30]. GRU cells use two internal gates, update and reset, to control information flow and reduce computational complexity. The update gate determines how much past information is carried forward, while the reset gate adjusts the influence of previous hidden states. This compact gating structure enables GRUs to achieve similar accuracy to LSTM while improving training speed and stability. In this study, the GRU model was used on multivariate time series composed of rainfall, lagged discharge, and seasonal indicators. The architecture included one to three stacked GRU layers with 64 and 50 units, followed by dense layers (50, 1 neuron). Dropout layers with a rate of 0.2, batch normalization, and L2 regularization were used to improve generalization and prevent overfitting.

2.3.4. Bidirectional Long Short-Term Memory (BiLSTM)

Bidirectional Long Short-Term Memory (BiLSTM) networks [31] extend the conventional LSTM by processing input sequences in both forward and backward directions. This structure enables the model to learn dependencies from past and future time steps, improving its ability to represent temporal dynamics in hydrological data. BiLSTM networks have been applied successfully in streamflow and flood forecasting due to their enhanced sequence-learning capability [14,32]. In BiLSTM, two LSTM layers run in parallel. The forward layer reads the sequence in chronological order, while the backward layer processes it in reverse. The outputs of both layers are combined to form a richer temporal representation that captures cause-and-effect relationships across the entire rainfall–runoff sequence. This dual path structure helps the network better identify the timing and magnitude of flood peaks. In this study, the BiLSTM model was used on multivariate time series composed of rainfall, lagged discharge, and seasonal indicators. The architecture included two stacked BiLSTM layers with 64 and 50 units, followed by dense layers (50, 1 neuron). Dropout layers with a rate of 0.2, batch normalization, and L2 regularization were used to improve generalization and prevent overfitting.

2.3.5. Hybrid Convolutional Long Short-Term Memory (Hybrid CNN–LSTM)

The Hybrid CNN–LSTM model combines the feature extraction capability of CNNs with the sequence-learning strength of LSTM networks. This integration allows the model to capture both short-term local variations and long-term temporal dependencies in hydrological time series [26,33–35]. The CNN component identifies localized rainfall–runoff features, while the LSTM component models sequential relationships that influence streamflow generation. In the hybrid structure, convolutional layers act as a pre-processing stage that extracts temporal features from rainfall and discharge sequences. These extracted features are then passed to LSTM layers, which learn cumulative dependencies across time. This combined approach enhances flood forecasting performance by allowing the model to detect both event-driven peaks and persistent baseflow dynamics. In this study, the Hybrid CNN–LSTM architecture consisted of two Conv1D layers with 64 and 256 filters and kernel sizes of 3, followed by max pooling and batch normalization. The extracted feature maps were fed into LSTM layers with 64 and 50 units, followed by dense layers (50, 1 neuron). Dropout (0.2) and L2 regularization were applied to control overfitting. Finally, dense layers (50, 1 neuron) were used to produce discharge predictions. This configuration effectively integrated temporal feature extraction with memory-based learning for accurate discharge prediction.

2.4. Baseline Model for Benchmarking

To evaluate whether deep learning models provide meaningful improvements over simpler and physically based approaches, two baseline models were implemented: a climatology model and the Hydrologiska Byråns Vattenbalansavdelning (HBV) conceptual hydrological model [36].

Climatology (Long-Term Mean) Model: Forecasts use the long-term daily mean discharge for each calendar day, computed from the 40-year training dataset.

A climatology baseline model is included to provide a lower-bound reference against which the performance of the deep learning architectures can be meaningfully evaluated. The baseline represents the long-term daily mean discharge for each calendar day computed from the 40-year training record:

$$\hat{y}_t = \underline{y}_{train} \quad (7)$$

where $\hat{y}_t$ is the predicted discharge on day (t), and y_{train} is the corresponding climatological mean derived from the training period. This benchmark captures the inherent seasonal structure of the UAB and sets a necessary minimum skill threshold for all predictive models.

HBV Conceptual Hydrological Model: The HBV model [36,37] was implemented as a physically motivated baseline. HBV is a widely used conceptual rainfall–runoff model representing key hydrological processes including soil moisture dynamics, groundwater storage, and runoff generation through calibrated reservoirs. The model was calibrated using observed precipitation, temperature, and discharge data for the UAB and evaluated using the same training–validation–test splits as the deep learning models. Including HBV provides a critical benchmark for assessing whether data-driven approaches outperform traditional process-based hydrological models.

2.5. Training and Validation Procedure

The dataset was chronologically split into training (70%), validation (15%), and test (15%) sets to preserve temporal integrity for realistic flood forecasting scenarios. Data normalization was performed using StandardScaler, standardizing inputs to zero mean and unit variance to prevent bias toward high-magnitude variables such as discharge. Temporal sequences were generated using a sliding window of 14 time steps. This window size was selected after preliminary experiments with 7, 14, 21, and 30 day windows, where the 14-day configuration consistently yielded better performance across most architectures by balancing temporal context and responsiveness. The sequences were reshaped into *samples × timesteps × features* for deep learning model compatibility.

To ensure a fair and transparent comparison across architectures, a structured manual hyperparameter search was conducted on a single personal computer rather than an exhaustive grid or random search, which was computationally infeasible. Learning rate ({0.01, 0.001, 0.0001}), batch size ({16, 32, 64, 96}), and number of layers ({2, 3, 4}) were varied in preliminary runs, and the final configuration was selected based on validation performance and training stability.

Models were then trained using the Adam optimizer (learning rate = 0.001), mean squared error (MSE) as the loss function, a batch size of 64, and 100 epochs. Early stopping (patience = 15) and learning rate reduction (factor = 0.2 after 5 stagnant epochs) were applied to mitigate overfitting and improve convergence. Regularization techniques included 20% dropout, batch normalization, and L2 regularization ($\lambda = 0.01$) to enhance model stability and generalization. All models followed a consistent training pipeline to ensure fair comparison. Implementation was performed in Python 3.10 using TensorFlow 2.12 and the Keras API, executed on an Intel Core i7-10610U CPU with fixed random seeds (42) for reproducibility.

Figure 5 shows the workflow for the flood forecasting framework. Discharge and precipitation data go through pre-processing. This includes quality control, temporal alignment, and scaling. The dataset is split into training, validation, and test sets. The training set optimizes model parameters. The validation set tunes hyperparameters and controls early stopping to avoid overfitting. The test set evaluates final model performance without bias. After validation, the optimized model supports real-time flood forecasting.

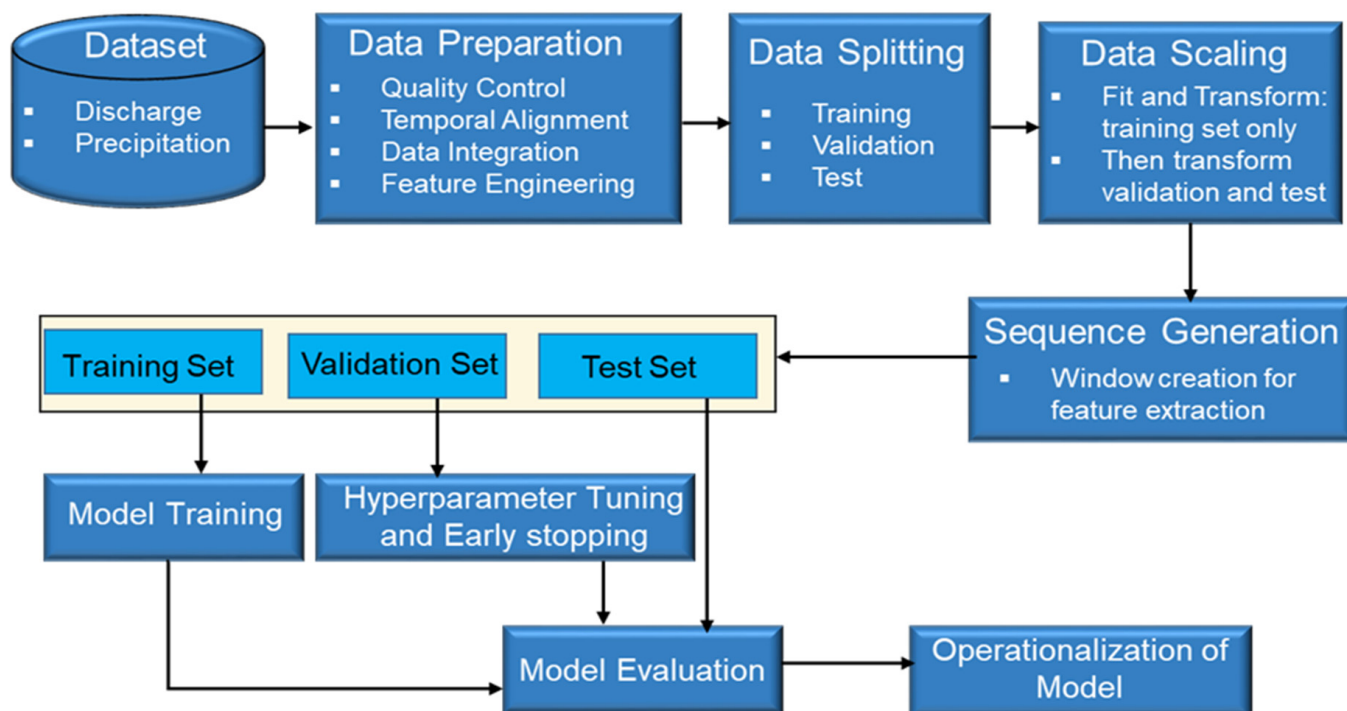


Figure 5. Methodological workflow for flood forecasting using deep learning.

3. Performance Evaluation Metrics

To evaluate model performance, we used both standard error-based measures and hydrology-specific scores. These metrics provide complementary perspectives on the accuracy, timing, and bias of the model in reproducing streamflow dynamics.

3.1. Standard Metrics

The standard error metrics used were the Mean Absolute Error (MAE) and root mean squared error (RMSE). MAE measures the average magnitude of errors without considering their direction, while RMSE penalizes larger errors more strongly, making it sensitive to high-flow extremes such as floods. Lower values of both metrics indicate better model performance.

3.2. Hydrology-Specific Metrics

Hydrology-specific metrics are more physically interpretable for hydrological predictions. These include the Nash–Sutcliffe Efficiency (NSE), Kling–Gupta Efficiency (KGE), and percent bias (PBIAS):

$$NSE = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (8)$$

$$KGE = 1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \quad (9)$$

$$PBIAS = 100 \times \frac{\sum_{i=1}^n (\hat{y}_i - y_i)}{\sum_{i=1}^n y_i} \quad (10)$$

where r is the Pearson correlation coefficient between observed and simulated discharges, $\alpha = \sigma_{\hat{y}}/\sigma_y$ represents the variability ratio, and $\beta = \mu_{\hat{y}}/\mu_y$ denotes the bias ratio. Here, σ and μ are the standard deviation and mean of the corresponding series, respectively.

The NSE quantifies how well the predicted hydrograph reproduces the magnitude and timing of the observed discharge; it ranges from $-\infty$ to 1, with values closer to 1 indicating excellent agreement. The KGE evaluates model performance as a combination of

correlation (timing), variability, and bias components, thus providing a balanced diagnostic of hydrological realism. PBIAS measures the general bias, indicating whether the model systematically overestimates (positive) or underestimates (negative) streamflow. To assess flood representation, PBIAS was also computed for the top- N peak-flow events. Following widely accepted hydrological performance criteria [36], the general interpretation ranges of these metrics are summarized in Table 2.

Table 2. Performance interpretation of model evaluation metrics used in this study.

Metric	Range	Ideal Value	Performance Criteria
MAE	$[0, \infty)$	0	Lower values indicate higher accuracy
RMSE	$[0, \infty)$	0	Sensitive to large errors; lower is better
NSE	$(-\infty, 1]$	1	>0.75 Excellent; 0.65–0.75 Good; 0.5–0.65 Satisfactory; <0.5 Poor
KGE	$(-\infty, 1]$	1	>0.75 Excellent; 0.65–0.75 Good; 0.5–0.65 Satisfactory; <0.5 Poor
PBIAS (%)	$(-\infty, \infty)$	0	PBIAS < 10 Excellent; 10–15 Good; 15–25 Satisfactory; >25 Poor

All metrics were calculated separately for the training, validation, and test datasets to evaluate both model learning and generalization capabilities.

Performance metrics consistently show that model skill in the test set is lower than in the validation period. This difference reflects a generalization gap, which arises when models trained on specific temporal data are applied to unseen periods. In the UAB, interannual variability, extreme events, and changing hydrological conditions contribute to reduced performance in the test set. Explicit reporting of training, validation, and test metrics allows quantification of this gap and provides a realistic assessment of predictive skill. Recognizing this generalization gap is essential for interpreting model performance and for guiding the design of operational flood forecasting systems in hydrologically complex and data-limited basins.

4. Results and Discussion

4.1. Model Performance and Interpretation

In this study, the performance of five deep learning models (CNN, LSTM, GRU, BiLSTM, and a Hybrid CNN–LSTM) was systematically analyzed to evaluate their effectiveness in daily flood forecasting within the UAB. The assessment employed visual diagnostics, including time series plots, scatter plots for training, validation, and test subsets, and learning curves that reflect model training behavior. Emphasis was placed on predictive accuracy, model fit, training dynamics, and generalization capacity. Additionally, the analysis incorporated seasonal hydrological characteristics, such as peak discharge events occurring during the June to September rainy season, to enhance contextual interpretation of model performance.

The learning curves in Figure 6 show the training and validation loss for all models across epochs. Both curves decline rapidly within the first 20–30 epochs and then stabilize, indicating efficient convergence. The small gap between training and validation loss (approximately 0.1) suggests good generalization and minimal overfitting. The convergence behavior highlights the success of early stopping (patience = 15), dropout (0.2), and L2 regularization ($\lambda = 0.01$) in managing model complexity.

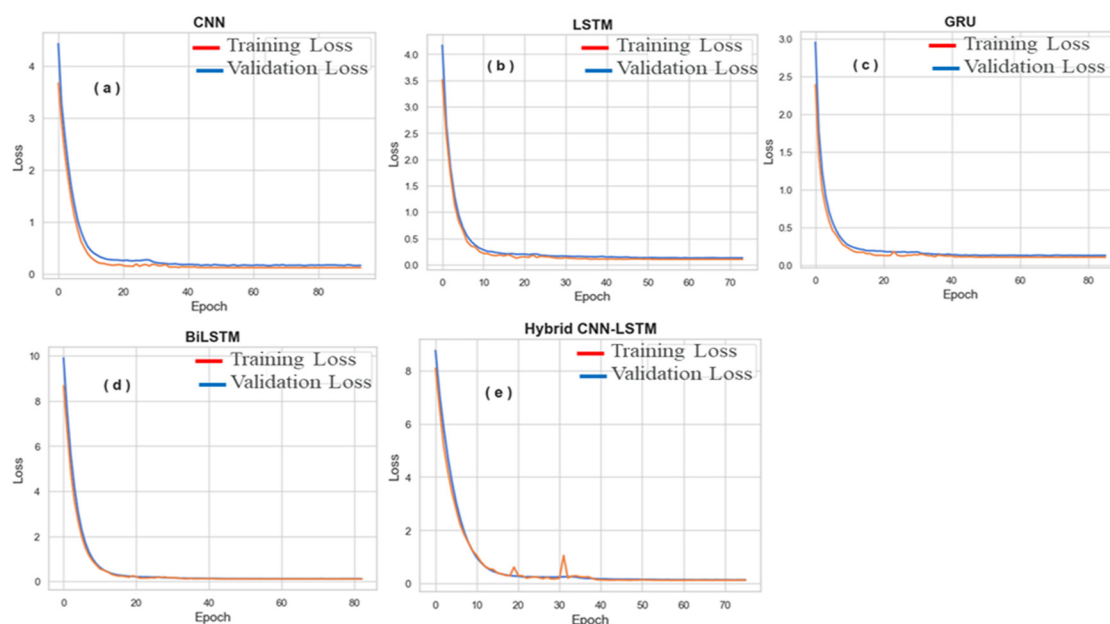


Figure 6. Learning curves for training and validation metrics of the (a) CNN, (b) LSTM, (c) GRU, (d) BiLSTM, and (e) Hybrid CNN–LSTM models.

4.1.1. Convolutional Neural Network (CNN)

Figure 7 compares observed and predicted discharge at the Hombole station during training, validation, and test periods using the CNN model. The observed discharge shows clear wet-season peaks in July and August, reaching up to $800 \text{ m}^3/\text{s}$, matching typical flood events in the UAB. The CNN model captures these seasonal patterns well during training, with $\text{MAE} = 9.54 \text{ m}^3/\text{s}$, $\text{RMSE} = 21.63 \text{ m}^3/\text{s}$, $\text{NSE} = 0.92$, $\text{KGE} = 0.90$, and near-zero bias ($\text{PBIAS} = 0.16\%$). The close fit between observed and predicted hydrographs demonstrates effective learning of rainfall–runoff relationships and stable convergence.

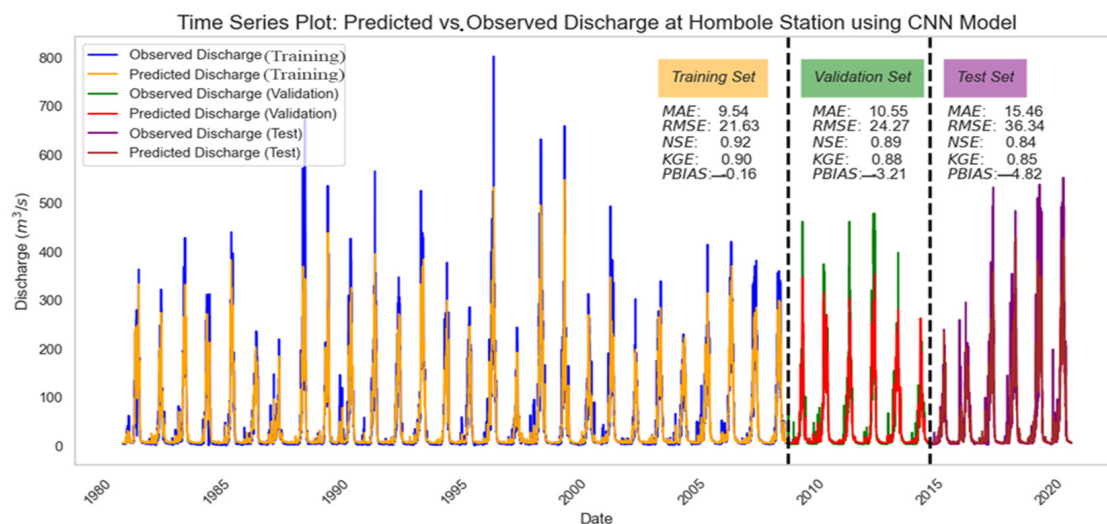


Figure 7. Time series plot of observed vs. predicted discharge at Hombole using the CNN model.

During validation, the model maintains good skill ($\text{MAE} = 10.55 \text{ m}^3/\text{s}$, $\text{RMSE} = 24.27 \text{ m}^3/\text{s}$, $\text{NSE} = 0.89$, $\text{KGE} = 0.88$, and $\text{PBIAS} = 3.21\%$), though errors increase slightly around discharge peaks. On the test set, performance drops somewhat ($\text{MAE} = 15.46 \text{ m}^3/\text{s}$, $\text{RMSE} = 36.34 \text{ m}^3/\text{s}$, $\text{NSE} = 0.84$, $\text{KGE} = 0.85$, $\text{PBIAS} = 4.82\%$), indicating mild underestimation during extreme flows but overall reasonable generalization.

Figure 8 shows scatter plots of observed versus predicted discharge for the three data splits. Predictions closely follow the 1:1 line in training but underestimate high flows. Validation shows increased scatter, especially between 200 and 400 m³/s. The test set exhibits the largest deviations at peak discharges, highlighting the CNN's limited ability to generalize extreme events.

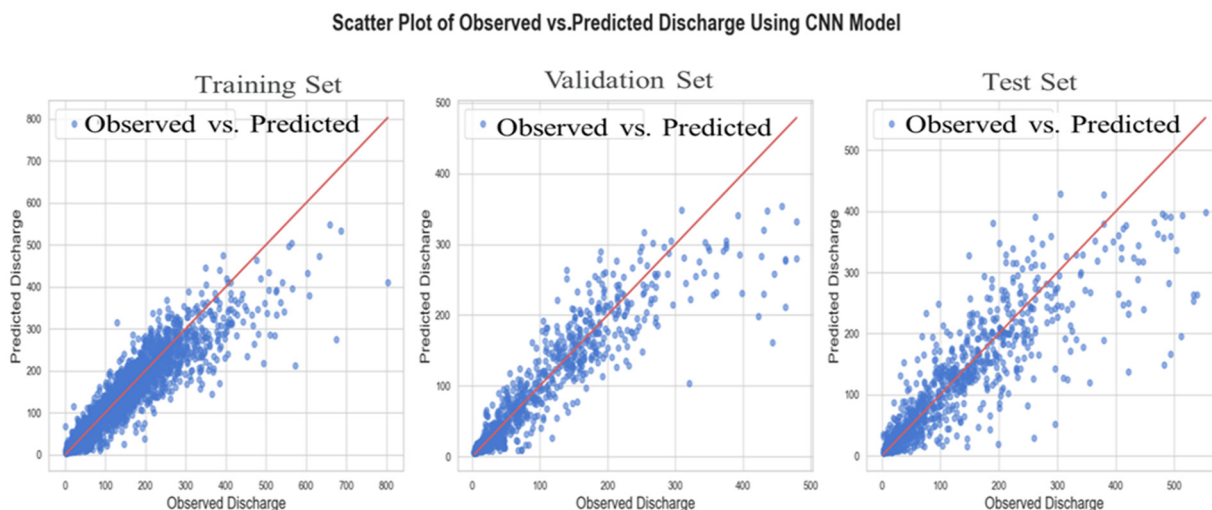


Figure 8. Scatter plots of observed vs. predicted discharge for training, validation, and test sets using CNN.

DL models were trained using standard MSE loss, which optimizes for average prediction accuracy. To evaluate performance on extreme events, the top 30 peak flows were analyzed. The results confirm underprediction of extreme discharges, a behavior explained by the emphasis of MSE on the bulk of the data rather than on rare, high-magnitude events. While peak-weighted or extreme-focused loss functions were not applied in this study, their potential to improve extreme-event forecasting is noted as a future methodological enhancement.

The seasonal metrics in Table 3 show stronger CNN performance in the wet season across all data splits. NSE ranges from 0.88 in training to 0.74 in testing, indicating good capture of rainfall–runoff dynamics and discharge peaks. KGE values between 0.86 and 0.79 confirm balanced correlation and hydrologic consistency. In the dry season, the model performs weaker. NSE drops to 0.71 (training) and 0.37 (testing), reflecting struggles with reproducing low-flow variability. High PBIAS values (up to 26.6%) show a tendency to overestimate low discharges, a common issue for models trained mostly on high-flow data.

Table 3. Seasonal performance metrics of the CNN model across training, validation, and test sets.

Dataset	Season	MAE	RMSE	NSE	KGE	PBIAS (%)
Training	Wet	21.188	36.459	0.875	0.859	−3.987
	Dry	3.684	5.925	0.708	0.645	26.592
Validation	Wet	24.876	40.34	0.809	0.83	−4.940
	Dry	3.301	5.632	0.520	0.511	8.195
Test	Wet	34.936	58.559	0.735	0.786	−6.070
	Dry	5.609	15.927	0.366	0.488	2.286

Table 4 shows CNN's poor performance on the top 30 peak flows. The model underestimates extreme floods, as indicated by the large negative PBIAS (31.68% to 35.27%) and negative NSE values, indicating poor predictive skill on the top 30 peak flows. High RMSE and low KGE confirm that the CNN struggles to capture nonlinear dynamics driving floods.

Table 4. Top 30 peak-flow performance metrics for the CNN model.

Dataset	MAE	RMSE	NSE	KGE	PBIAS (%)
Training	174.537	196.601	−5.702	0.294	−31.677
Validation	125.643	140.605	−7.409	0.294	−31.499
Test	165.536	185.624	−18.527	−0.277	−35.273

In summary, CNN captures seasonal and moderate flows well. However, it underestimates high-magnitude floods, due to smoothing effects of convolutional filters and limited extreme-event data in training. The results highlight the need for models with stronger temporal memory or attention mechanisms to improve flood peak prediction. CNN performs best in wet seasons, where rainfall–runoff signals dominate. Low dry season skill shows dependence on rainfall input and sensitivity to flow regime.

4.1.2. Long Short-Term Memory (LSTM)

Figure 9 shows observed and predicted discharge at the Hombole station during training, validation, and test periods. The LSTM model captures seasonal discharge patterns and peak timing accurately across all subsets. During training, the model shows strong performance with $NSE = 0.91$, $KGE = 0.93$, $RMSE = 23.20 \text{ m}^3/\text{s}$, and minimal bias ($PBIAS = 0.43\%$), indicating solid calibration and effective learning of rainfall–runoff dynamics. Validation results are consistent ($NSE = 0.90$, $KGE = 0.91$, and $RMSE = 23.52 \text{ m}^3/\text{s}$), demonstrating robust generalization. On the test set, performance remains high ($NSE = 0.85$ and $KGE = 0.90$), though $RMSE$ rises to $34.99 \text{ m}^3/\text{s}$. Small negative bias ($PBIAS = 2.94\%$) shows slight underestimation of peaks but no major systematic error. These metrics confirm LSTM’s stability in capturing baseflow and flood-season variability within the UAB.

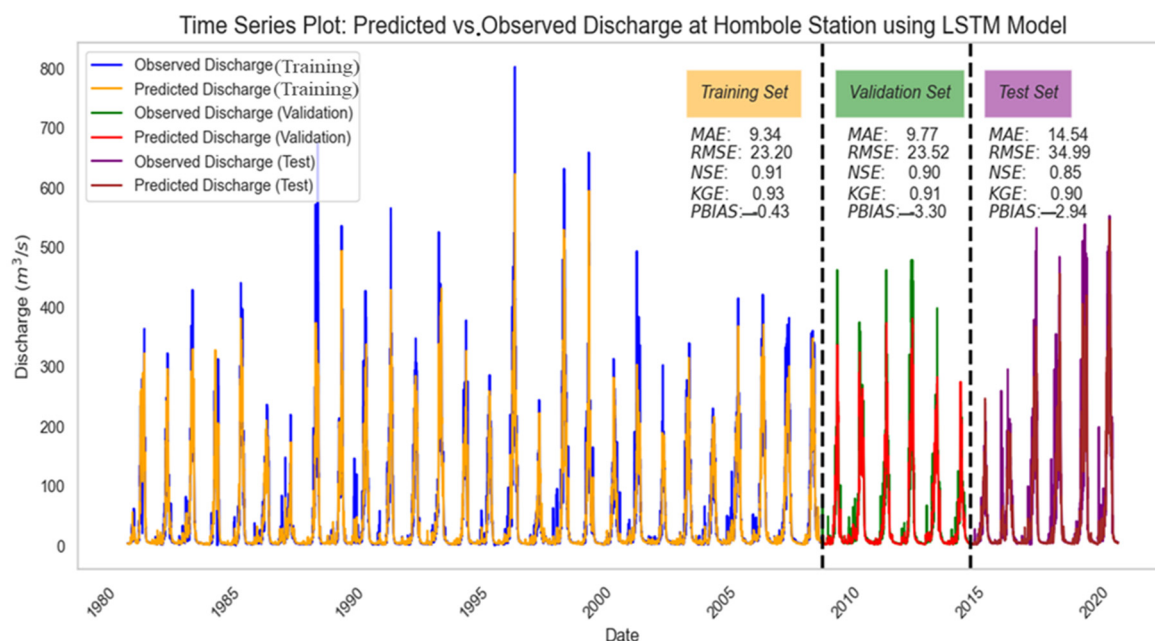
**Figure 9.** Time series plot of predicted vs. observed discharge using the LSTM model.

Figure 10 presents scatter plots for all data splits. Training data points tightly cluster near the 1:1 line, showing strong accuracy across flow ranges, including above $500 \text{ m}^3/\text{s}$. Validation shows similar clustering with slight spread at moderate flows ($200\text{--}400 \text{ m}^3/\text{s}$). Test data scatter increases, particularly for extreme flows ($600\text{--}800 \text{ m}^3/\text{s}$), indicating reduced generalization to unseen floods. Overall, the model maintains high calibration and validation fidelity, with minor expected performance drops under extreme events.

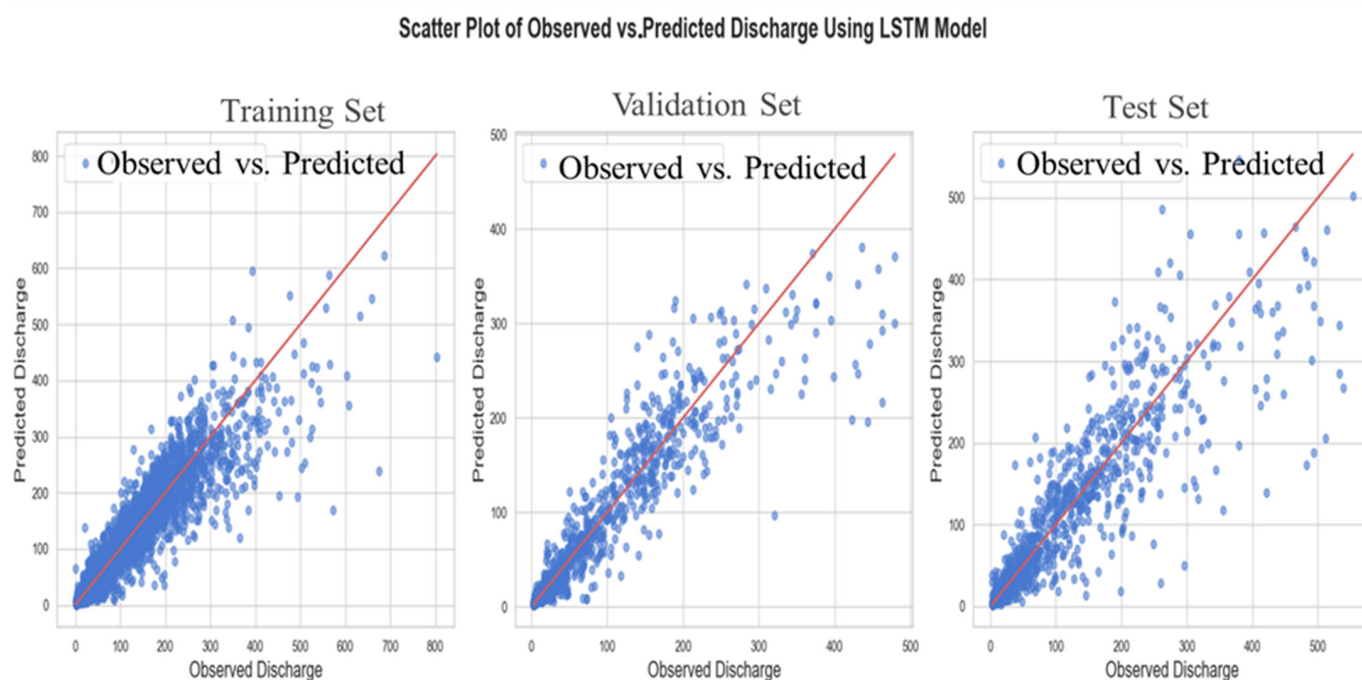


Figure 10. Scatter plots of observed vs. predicted discharge for training, validation, and test sets using the LSTM model.

Seasonal results (Table 5) show the LSTM performs strongly in wet seasons with NSE up to 0.86 and low bias, indicating balanced, accurate predictions. Dry season skill declines, reflecting variability in baseflow processes, with reduced NSE and KGE, and some over/underestimation error.

Table 5. Seasonal performance metrics of the LSTM model across data splits.

Dataset	Season	MAE	RMSE	NSE	KGE	PBIAS (%)
Training	Wet	23.03	39.28	0.86	0.89	−1.20
	Dry	2.45	5.78	0.72	0.79	4.92
Validation	Wet	24.26	39.89	0.82	0.86	−2.62
	Dry	2.43	5.28	0.58	0.62	−7.82
Test	Wet	34.35	56.57	0.75	0.84	−1.46
	Dry	4.52	14.99	0.44	0.49	−11.36

Evaluation of extreme peak flows (Table 6) reveals LSTM’s limitations in replicating the highest discharge values. Large RMSE and negative NSE show weak fit, while negative PBIAS indicates systematic underprediction of flood peaks. KGE near zero or below highlights poor consistency during extremes.

Table 6. Top 30 peak-flow performance metrics for the LSTM model.

Dataset	MAE	RMSE	NSE	KGE	PBIAS (%)
Training	171.25	200.85	−6.00	0.11	−29.90
Validation	107.52	126.97	−5.86	0.02	−26.90
Test	136.94	163.82	−14.21	−0.46	−28.62

To summarize, the LSTM consistently captures seasonal flow patterns with strength during wet periods but shows reduced skill at extreme peaks in nonlinear flood dynamics. The model excels in general discharge forecasting under normal/moderate flows but is less reliable for real-time extreme flood magnitude prediction.

4.1.3. Gate Recurrent Unit (GRU)

Figure 11 displays observed and predicted discharge at the Hombole station for the GRU model across training, validation, and test sets. Observed discharge shows strong seasonal peaks during July and August, consistent with flood periods in the UAB. The GRU model captures these temporal patterns well, with predicted discharges closely following observed values across all datasets.

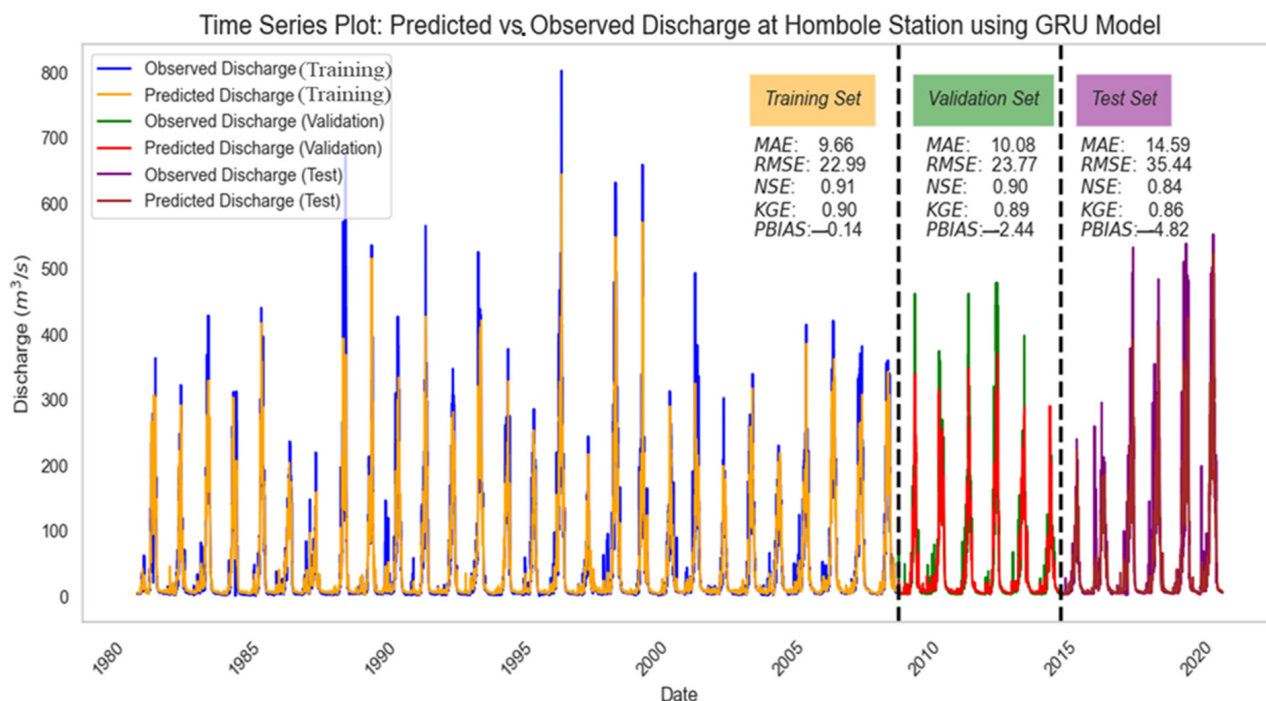


Figure 11. Time series plot of predicted vs. observed discharge using the GRU model.

During training, the model achieves high accuracy with $NSE = 0.91$, $KGE = 0.90$, and minimal bias ($PBIAS = 0.14\%$), indicating balanced performance and limited overfitting. Validation maintains stable results ($NSE = 0.90$ and $KGE = 0.89$), showing robust generalization. The test set retains strong skill ($NSE = 0.84$ and $KGE = 0.86$) with slightly increased RMSE ($35.44 \text{ m}^3/\text{s}$) and moderate underestimation bias ($PBIAS = 4.82\%$), especially during peak flows. The GRU's performance matches the LSTM but with faster convergence and fewer parameters. The simplified gating allows for efficient temporal learning while maintaining accuracy. Although it underpredicts extreme discharges, consistent NSE and KGE indicate reliable rainfall–runoff modeling and adaptation to varying flows.

Figure 12 shows scatter plots of observed versus predicted discharge across all datasets. Training predictions align closely with the 1:1 line, displaying minimal bias and strong agreement at moderate flows. Validation points cluster well but disperse slightly between 200 and $400 \text{ m}^3/\text{s}$. The test set exhibits broader scatter, especially above $600 \text{ m}^3/\text{s}$, reflecting reduced accuracy during high flows and limited precision in extreme floods.

Seasonal analysis (Table 7) shows strong wet-season performance with NSE between 0.86 and 0.75 and KGE from 0.87 to 0.81 , indicating consistent skill across data splits. Negative bias ($PBIAS$ from 3.5% to 5.6%) suggests mild underestimation of peak flows, typical in recurrent models. Dry-season skill decreases (NSE and KGE from 0.43 to 0.72), reflecting weaker prediction under low-flow conditions, influenced by near-zero discharges and low variability. Positive PBIAS during training and validation (up to 23.3%) indicates baseflow overestimation; near-zero testing bias shows improved balance.

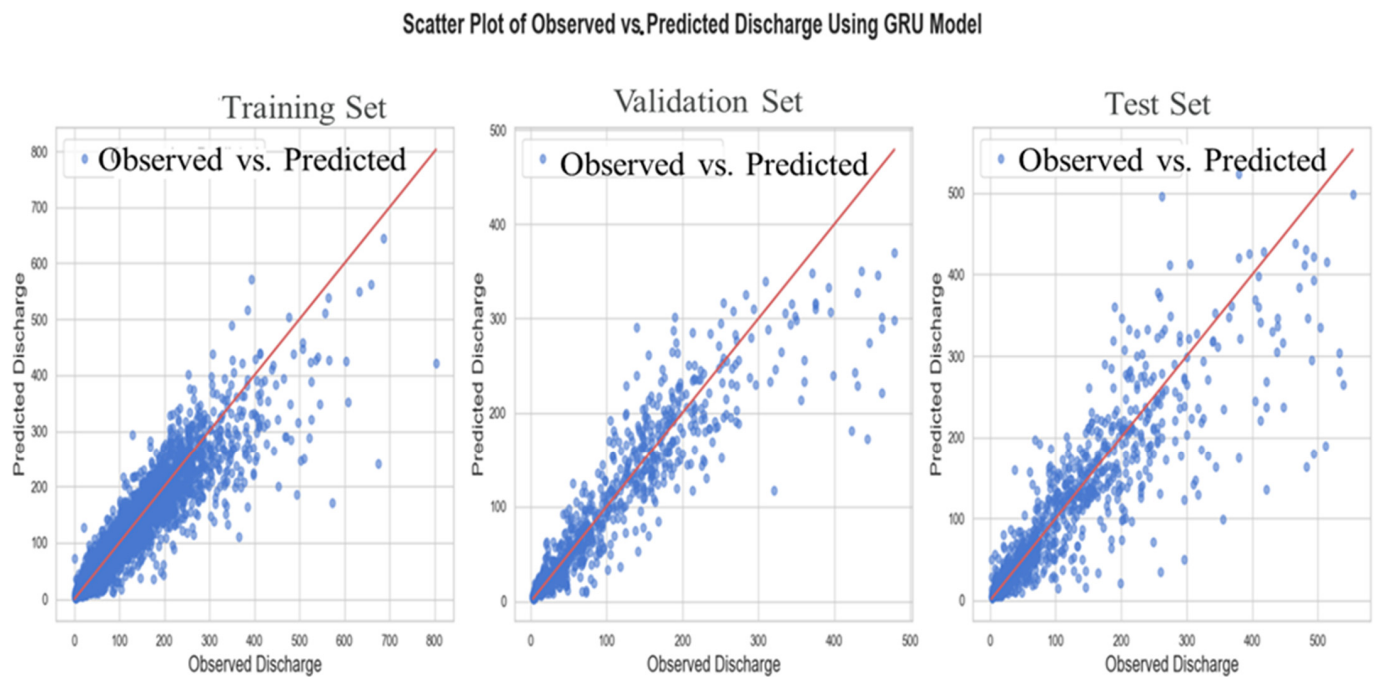


Figure 12. Scatter plots of observed vs. predicted discharge for training, validation, and test sets using the GRU model.

Table 7. Seasonal performance metrics of the GRU model across training, validation, and test sets.

Dataset	Season	MAE	RMSE	NSE	KGE	PBIAS (%)
Training	Wet	22.43	38.91	0.86	0.87	−3.49
	Dry	3.24	5.78	0.72	0.68	23.30
Validation	Wet	24.18	40.34	0.82	0.83	−4.29
	Dry	2.94	5.27	0.58	0.60	9.71
Test	Wet	33.95	57.40	0.75	0.81	−5.65
	Dry	4.80	15.01	0.44	0.46	−0.10

Performance on the top 30 peak flows (Table 8) is weak. Large RMSE (134–201 m³/s) and negative NSE reveal poor fit to observed floods. Negative PBIAS indicates systematic underprediction of extremes. KGE near or below zero confirms limited hydrological realism during floods.

Table 8. Top 30 peak-flow performance metrics for the GRU model.

Dataset	MAE	RMSE	NSE	KGE	PBIAS (%)
Training	168.74	201.28	−6.03	0.13	−30.29
Validation	115.45	133.99	−6.64	0.00	−28.94
Test	148.02	174.38	−16.23	−0.46	−31.40

Overall, GRU provides robust wet-season skill with better generalization than CNN, though it is slightly less stable than LSTM during low-flow periods. While it captures temporal continuity well, GRU lacks precision for extreme discharges. Enhancements like peak-weighted loss functions or hybridization with attention or convolutional layers may improve flood magnitude prediction.

4.1.4. Bidirectional Long Short-Term Memory (BiLSTM) Model

Figure 13 compares observed and predicted discharge at the Hombole station using the BiLSTM model across training, validation, and test sets. The BiLSTM architecture

processes input sequences in both forward and backward directions, enhancing temporal context capture compared to standard LSTM and GRU models.

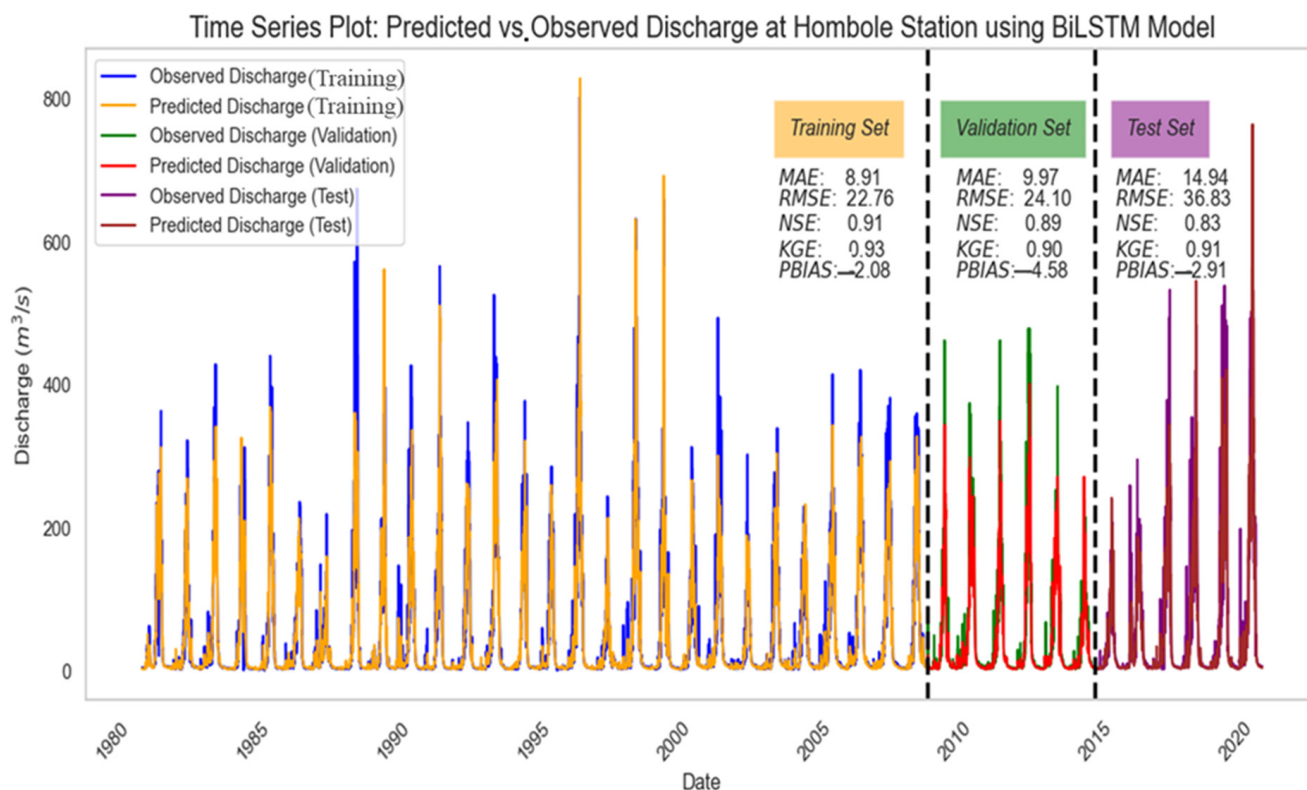


Figure 13. Time series of observed vs. predicted discharge using the BiLSTM model.

The model achieves strong performance across all phases, with NSE values of 0.91, 0.89, and 0.83 for training, validation, and test sets, respectively. KGE values (0.93–0.91) indicate high consistency and correlation between observed and simulated flows. Low PBIAS values (2.08% to 4.58%) show minimal bias, confirming balanced discharge estimation. Predicted discharge closely aligns with observed peaks in wet seasons, accurately capturing the timing and volume of major floods. Slight underestimation occurs at extreme peaks ($600 \text{ m}^3/\text{s}$), yet performance remains stable across data splits. Compared with LSTM and GRU, BiLSTM offers marginally better generalization and temporal precision through bidirectional memory, improving learning of both past and future flow dependencies.

Figure 14 shows scatter plots for training, validation, and test sets. Training predictions align tightly with the 1:1 line, with minimal deviation and slight underestimation beyond $500 \text{ m}^3/\text{s}$. Validation shows moderate scatter between 200 and $400 \text{ m}^3/\text{s}$, indicating minor precision loss. The test set presents wider dispersion beyond $600 \text{ m}^3/\text{s}$, reflecting reduced accuracy on extreme events.

Seasonal metrics (Table 9) show BiLSTM performs consistently during wet seasons, with NSE between 0.86 and 0.72 and KGE ranging 0.84–0.90, indicating reliable flood representation. Minor negative biases (0.9% to 3.3%) support accurate peak magnitude prediction. Dry-season skill is moderate, with NSE from 0.44 to 0.77 and larger negative biases (up to 13.9%), reflecting some underestimation of low flows. Despite this, BiLSTM achieves the best balance between high- and low-flow prediction among recurrent models thanks to its bidirectional processing.

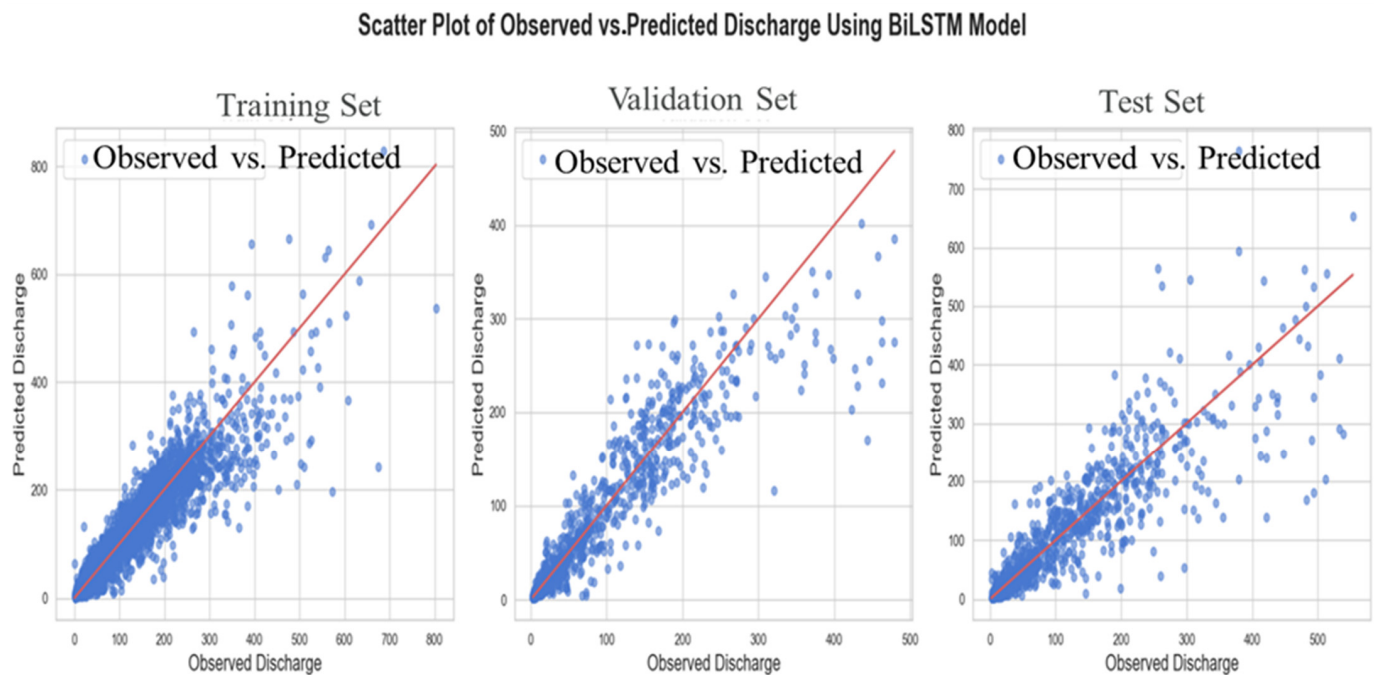


Figure 14. Scatter plots of observed vs. predicted discharge for training, validation, and test sets using BiLSTM.

Table 9. Seasonal performance metrics of the BiLSTM model across training, validation, and test sets.

Dataset	Season	MAE	RMSE	NSE	KGE	PBIAS (%)
Training	Wet	22.32	38.65	0.86	0.90	−2.23
	Dry	2.18	5.27	0.77	0.83	−1.02
Validation	Wet	25.03	40.90	0.81	0.84	−3.27
	Dry	2.34	5.29	0.58	0.61	−13.16
Test	Wet	35.05	59.96	0.72	0.86	−0.99
	Dry	4.76	14.95	0.44	0.54	−13.86

BiLSTM struggles to reproduce extreme peak discharges accurately, as seen in Table 10. Large RMSE and negative NSE values indicate a weak fit for flood peaks. Persistent negative PBIAS (20.6% to 28.9%) indicates systematic underestimation. Negative KGE values at high flows reveal low correlation.

Table 10. Top 30 peak-flow performance metrics for the BiLSTM model across all sets.

Dataset	MAE	RMSE	NSE	KGE	PBIAS (%)
Training	152.55	185.18	−4.95	−0.26	−20.63
Validation	115.48	134.15	−6.66	0.03	−28.95
Test	133.51	163.45	−14.14	−1.22	−22.20

Overall, BiLSTM demonstrates strong prediction consistency with minimal bias and reliable generalization across most flow conditions. Its bidirectional structure enhances flood timing and duration capture but shows limitations in accurately predicting peak flood magnitudes. These results highlight BiLSTM's improved temporal context learning, while also revealing sensitivity to data imbalance dominated by moderate flows. Incorporating weighted loss functions or event-focused training may improve its responsiveness to extreme flood events.

4.1.5. Hybrid Model (Hybrid CNN–LSTM)

Figure 15 shows observed and predicted discharge at the Hombole station using the Hybrid CNN–LSTM model across training, validation, and test sets. This hybrid architecture combines convolutional layers for short-term pattern extraction with LSTM layers for long-term temporal dependency learning, enabling joint representation of rainfall–runoff dynamics.

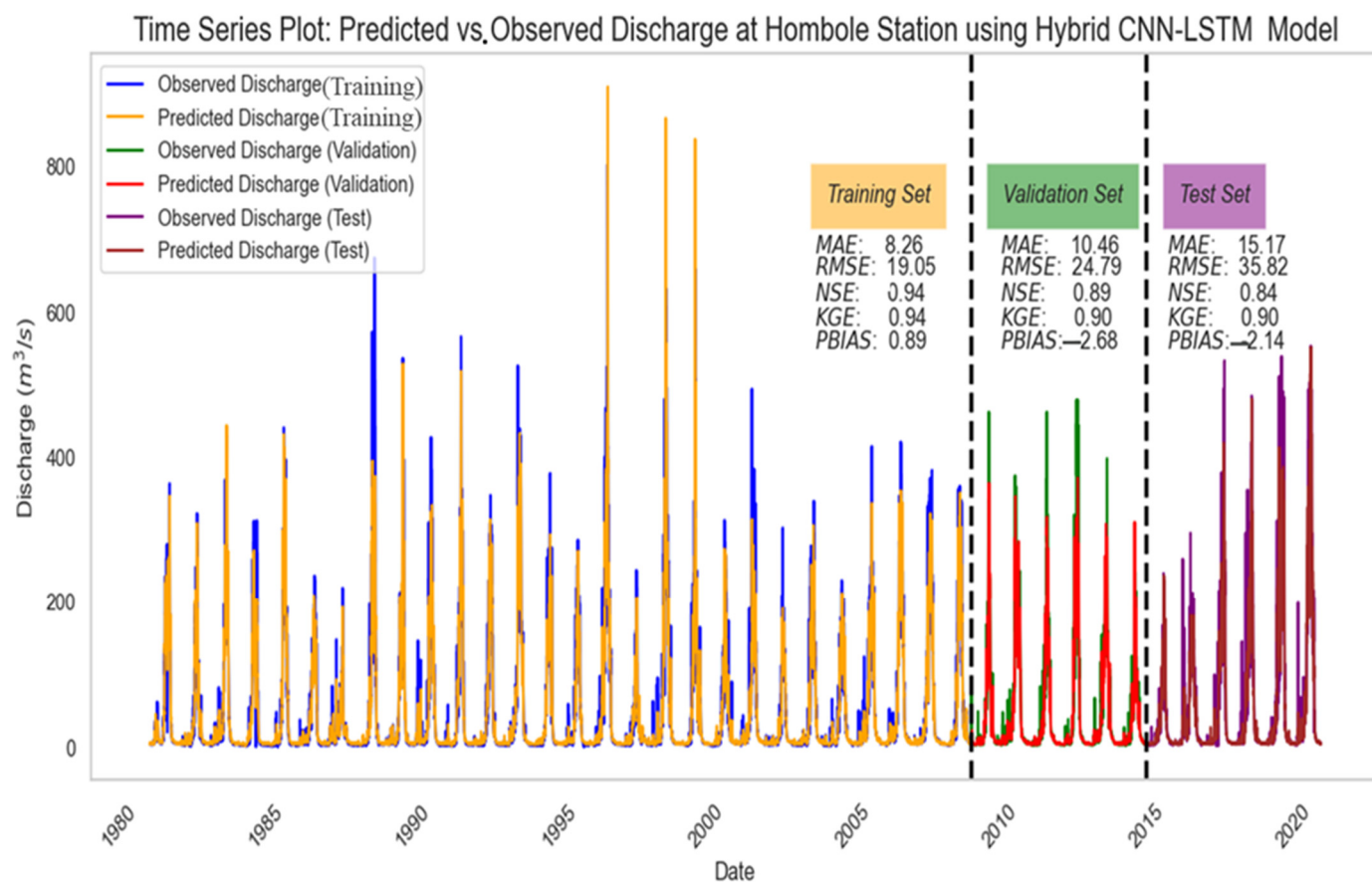


Figure 15. Time series of observed vs. predicted discharge using the Hybrid CNN–LSTM model.

The model shows superior predictive performance across all phases, achieving NSE values of 0.94, 0.89, and 0.84 and KGE values of 0.94, 0.90, and 0.90 for training, validation, and test sets, respectively. Low PBIAS values (2.14% to 0.89%) indicate balanced discharge estimation with minimal systematic error. RMSE values remain lower than standalone recurrent models, reflecting improved stability and convergence. Predicted discharge aligns closely with observed peaks, accurately capturing flood timing and amplitude during wet seasons while maintaining precision in dry periods. Minor underestimation occurs at extreme flows ($700 \text{ m}^3/\text{s}$), but overall consistency remains high. The hybrid outperforms individual CNN and LSTM models, confirming that combining spatial feature extraction with temporal memory enhances hydrological representation and forecasting reliability.

Figure 16 shows scatter plots of observed and predicted discharge for the hybrid model across all data splits. Training points cluster tightly along the 1:1 line, including high discharges above $500 \text{ m}^3/\text{s}$, indicating strong agreement. Validation shows moderate scatter around $200\text{--}400 \text{ m}^3/\text{s}$, with test data exhibiting a wider spread beyond $600 \text{ m}^3/\text{s}$. Overall, the hybrid model maintains high consistency and balanced prediction skill, outperforming single CNN or LSTM architectures in stability and accuracy.

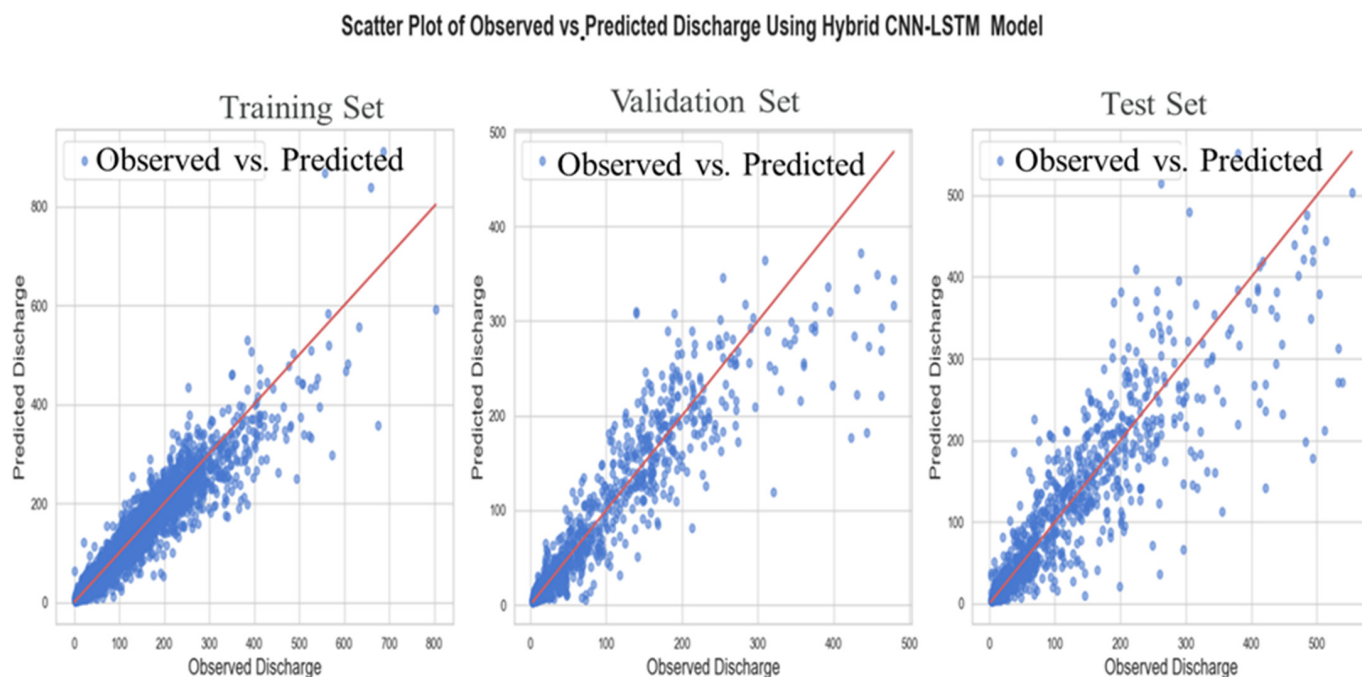


Figure 16. Scatter plots of observed vs. predicted discharge for training, validation, and test sets using the Hybrid CNN–LSTM model.

The hybrid model performs strongly in the wet season across datasets, with NSE values from 0.90 to 0.74 and KGE from 0.92 to 0.84. Low PBIAS values (1.3% to 3.2%) indicate minimal bias in flood magnitude prediction. It effectively captures rainfall–runoff relationships and peak timing in the main rainy season. Dry-season metrics show moderate skill (NSE = 0.45–0.79 and KGE = 0.51–0.76), reflecting stable low-flow representation but some overestimation during training (PBIAS = +16.4%) and slight underestimation during testing (PBIAS = 6.1%) (Table 11).

Table 11. Seasonal performance metrics of the Hybrid CNN–LSTM model across training, validation, and test sets.

Dataset	Season	MAE	RMSE	NSE	KGE	PBIAS (%)
Training	Wet	19.23	32.15	0.90	0.92	−1.33
	Dry	2.75	5.08	0.79	0.76	16.43
Validation	Wet	25.98	42.04	0.80	0.84	−3.22
	Dry	2.61	5.52	0.54	0.58	0.91
Test	Wet	35.98	58.15	0.74	0.84	−1.44
	Dry	4.64	14.89	0.45	0.51	−6.12

The hybrid model shows improved performance over standalone CNN and LSTM models but struggles to reproduce extreme peak flows accurately. High RMSE (136–162 m³/s) and negative NSE across datasets indicate difficulty matching observed flood magnitudes (Table 12). Consistently negative PBIAS values (15.2% to 30.3%) reveal systematic underestimation of flood peaks. KGE values near zero or negative suggest limited hydrological consistency under extreme conditions, though the hybrid architecture reduces error magnitude compared to LSTM and GRU. Convolutional layers enhance spatial feature extraction, while LSTM layers maintain temporal memory, resulting in smoother peak transitions and better timing.

Table 12. Top 30 peak-flow performance metrics for the Hybrid CNN–LSTM model across training, validation, and test sets.

Dataset	MAE	RMSE	NSE	KGE	PBIAS (%)
Training	133.34	157.59	−3.31	−0.19	−15.15
Validation	121.00	135.70	−6.83	0.12	−30.34
Test	128.70	161.58	−13.80	−0.55	−27.38

Overall, the Hybrid CNN–LSTM model delivers the most consistent seasonal performance, showing high hydrological fidelity during wet periods and improved dry-season stability. This highlights the advantage of combining the CNN’s ability to extract short-term temporal features with the LSTM’s capacity to capture long-term dependencies, resulting in more robust and adaptive discharge forecasts. Despite these strengths, the model remains sensitive to the rarity of extreme events in training data. Incorporating event-based training, data augmentation for rare floods, or attention mechanisms could enhance extreme-flow forecasting performance.

4.1.6. Baseline Models for Benchmarking

Climatology (Long-Term Mean) Model: Figure 17 compares observed and predicted discharge at the Hombole station using the climatology model for the training, validation, and test periods. This baseline predicts daily discharge from the long-term mean for each calendar day. Across all splits, the model shows weak skill, with high errors (MAE $\approx$ 50–58 m³/s and RMSE $\approx$ 74–91 m³/s) and near-zero or slightly negative NSE and KGE values, indicating failure to reproduce observed variability. The increasingly negative PBIAS in the test set (20.39%) reflects substantial underestimation of higher flows. These results confirm that the climatology model serves only as a conservative benchmark and that deep learning models add significant predictive value beyond a simple long-term mean.

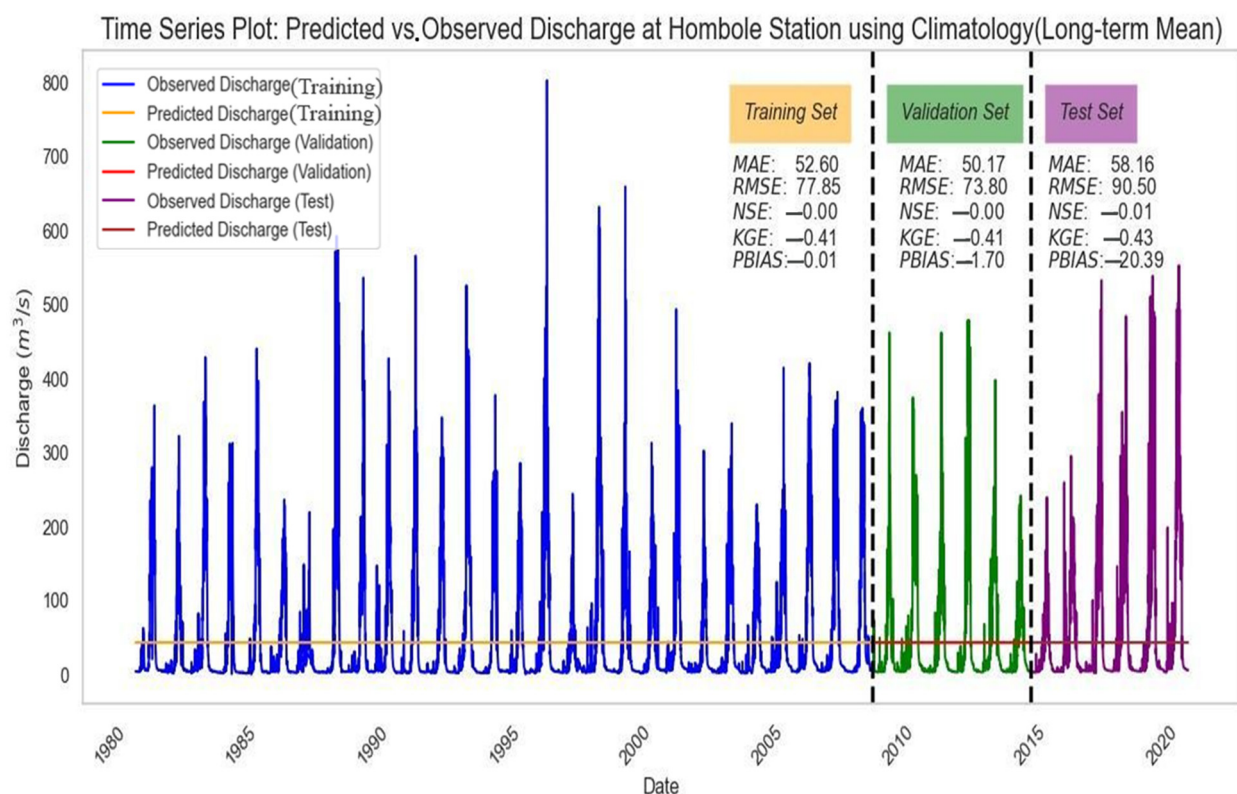
**Figure 17.** Time series of observed vs. predicted discharge using the climatology (long-term mean) model.

Figure 18 shows scatter plots of observed versus predicted discharge for the training, validation, and test sets using the climatology model. In all three subsets, points form nearly horizontal bands far from the 1:1 line, indicating that predictions cluster around the long-term mean and fail to track observed variability. Even at high flows, observed peaks correspond to almost constant predicted values, revealing very poor skill in representing both moderate and extreme discharges.

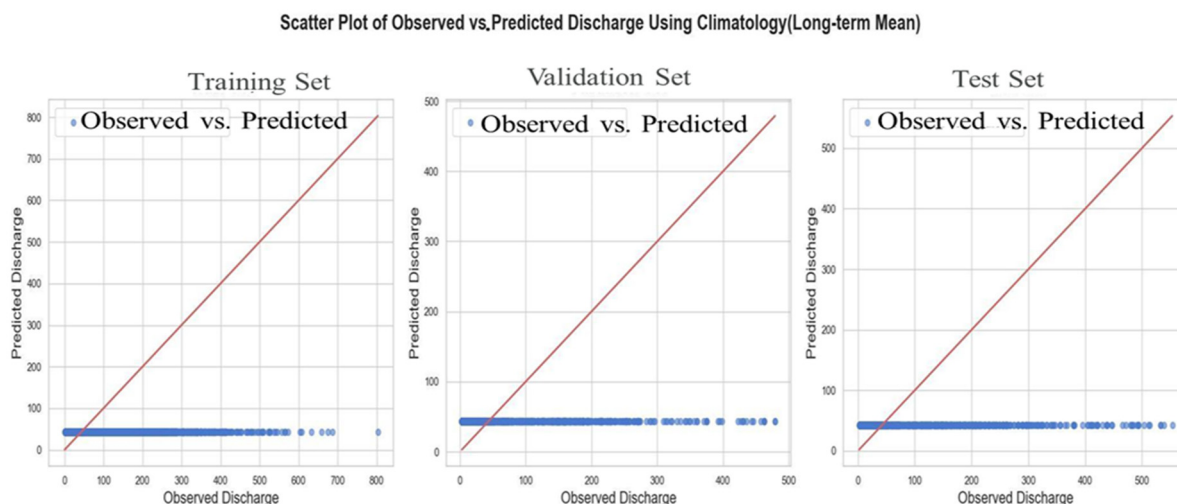


Figure 18. Scatter plots of observed vs. predicted discharge for training, validation, and test sets using the climatology (long-term mean) model.

The climatology model performs poorly in both seasons. Wet-season metrics show large errors (high MAE and RMSE), negative NSE, and negative KGE, indicating it cannot reproduce flood timing or magnitude and systematically underestimates high flows (PBIAS around 60% to 70%). Dry-season performance is even weaker: very negative NSE and KGE and extremely high positive PBIAS (over 250% and up to 440%) reveal severe overestimation of low flows and failure to represent baseflow dynamics (Table 13). Overall, the climatology model serves only as a crude benchmark, confirming the need for more advanced modeling approaches.

Table 13. Seasonal performance metrics for the climatology (long-term mean) model.

Dataset	Season	MAE	RMSE	NSE	KGE	PBIAS (%)
Training	Wet	86.01	124.40	−0.46	−0.54	−61.90
	Dry	35.85	36.57	−10.83	−3.62	440.06
Validation	Wet	82.54	117.83	−0.50	−0.54	−61.25
	Dry	33.92	34.81	−8.68	−2.59	330.44
Test	Wet	103.47	146.53	−0.71	−0.57	−68.76
	Dry	35.42	39.07	−1.67	−1.93	256.50

The climatology model performs very poorly for the top 30 peak flows. Very large MAE and RMSE values (MAE $\approx$ 356–508 m³/s; RMSE $\approx$ 359–514 m³/s), together with strongly negative NSE values (from −44.76 to −103.02), indicate almost no skill in reproducing flood peak magnitudes. Consistently negative KGE values around −0.7 and severe negative PBIAS (about −89% to −92%) show that the model massively underestimates extreme discharges across all data splits (Table 14). Overall, the climatology baseline fails to capture both the amplitude and variability of major flood events and provides only a crude lower bound for model performance. Its weak peak-flow skill highlights the substantial gains

achieved by the deep learning architectures, which add meaningful predictive value beyond a simple long-term mean assumption.

Table 14. Top 30 peak-flow performance metrics for the climatology (long-term mean) model.

Dataset	MAE	RMSE	NSE	KGE	PBIAS (%)
Training	508.05	513.69	−44.76	−0.69	−92.21
Validation	355.95	359.23	−53.89	−0.67	−89.24
Test	426.36	428.43	−103.02	−0.68	−90.85

HBV Conceptual Hydrological Model: Figure 19 compares observed and HBV-simulated discharge time series for the training, validation, and test periods at the Hombole station. Following calibration, the HBV model reproduces the general magnitude and seasonal evolution of daily discharge reasonably well across all data splits, as reflected by moderate NSE.

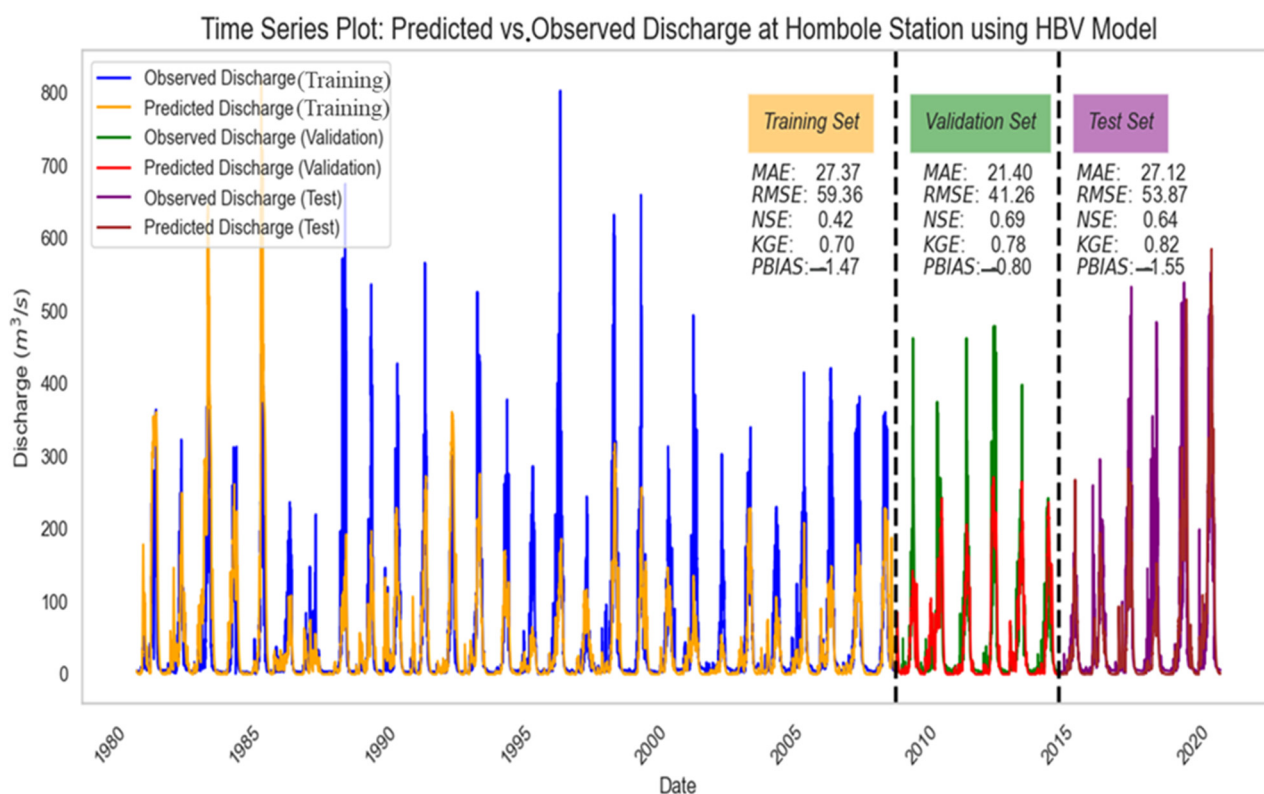


Figure 19. Time series of observed versus simulated discharge using the HBV conceptual hydrological model.

Figure 20 shows scatter plots of observed versus HBV-simulated discharge for the training, validation, and test periods. The calibrated HBV model exhibits a clear positive relationship between observed and simulated flows, particularly for low to moderate discharges, with points clustering closer to the 1:1 line. This indicates improved overall agreement after calibration. However, dispersion increases markedly at higher flows, and extreme discharges are consistently underestimated across all data splits. These results suggest that while calibration enhances average-flow simulation, the HBV model remains limited in capturing peak flood magnitudes in the UAB.

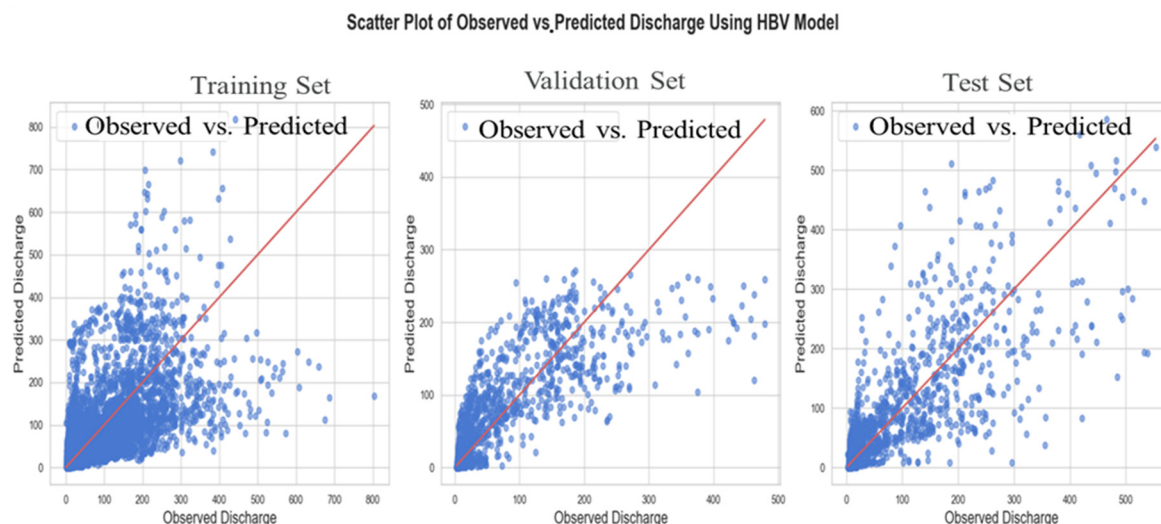


Figure 20. Scatter plots of observed versus simulated discharge for training, validation, and test sets using the HBV model.

The HBV model shows moderate skill across the training, validation, and test datasets. Positive NSE values (up to 0.69 in validation and 0.64 in testing) indicate that the model performs better than a mean-flow benchmark, while KGE values between 0.70 and 0.82 reflect reasonable agreement in correlation, variability, and bias. PBIAS values close to zero (within $\pm 2\%$) suggest good overall water-balance reproduction (Table 15). However, despite these improvements in average-flow simulation, HBV continues to underestimate extreme flood peaks, as evidenced by poor performance for the top 30 peak events and negative efficiency metrics during high-flow conditions.

Table 15. Seasonal performance metrics of the HBV model.

Dataset	Season	MAE	RMSE	NSE	KGE	PBIAS (%)
Training	Wet	61.35	97.03	0.11	−0.55	−11.22
	Dry	10.32	23.81	−4.01	−0.77	67.84
Validation	Wet	43.86	65.87	0.53	−0.38	−8.02
	Dry	10.12	19.47	−2.03	−2.03	39.49
Test	Wet	55.81	82.79	0.45	0.71	−8.684
	Dry	12.73	30.33	−0.61	0.12	39.1

Seasonal analysis indicates that the HBV model performs moderately during the wet season, with positive NSE values (up to 0.53 in validation and 0.45 in testing) and KGE values between 0.56 and 0.72, suggesting reasonable skill in reproducing seasonal runoff dynamics. Nevertheless, relatively large MAE and RMSE values and negative PBIAS (approximately -8% to -11%) indicate a tendency to underestimate high wet-season flows, particularly during flood events. In the dry season, model performance is weaker, with negative NSE values and positive PBIAS ($\approx 40\text{--}68\%$), reflecting difficulties in simulating low-flow conditions and baseflow dynamics (Table 16). Overall, HBV remains limited in accurately representing both extreme wet-season floods and dry-season low flows.

Table 16. Top 30 peak-flow performance metrics for the HBV model.

Dataset	MAE	RMSE	NSE	KGE	PBIAS (%)
Training	360.83	374.56	−23.32	−0.2	−65.5
Validation	195.76	204.94	−16.86	−0.06	−49.07
Test	158.14	188.52	−19.14	−1.45	−27.58

The HBV model shows limited skill in reproducing the top 30 flood peaks. Peak-flow errors remain large ($\text{MAE} \approx 158\text{--}361 \text{ m}^3/\text{s}$; $\text{RMSE} \approx 189\text{--}375 \text{ m}^3/\text{s}$), and NSE values are strongly negative (-16.9 to -23.3 across validation and test sets), indicating poor efficiency in capturing extreme flood magnitudes. Negative KGE values (-0.06 to -1.45) further reflect deficiencies in correlation and variability during peak events. Although PBIAS values ($\approx -28\%$ to -66%) indicate reduced bias relative to mean-flow conditions, the model still systematically underestimates extreme discharges. These results suggest that, while HBV reproduces average flow behavior reasonably well, it remains inadequate for reliable simulation of extreme flood peaks in the UAB.

The HBV model shows moderate overall discharge simulation ($\text{NSE} = 0.42\text{--}0.69$) but reveals clear limitations during wet seasons (negative NSE, $\text{PBIAS} \approx -99\%$) and for extreme peak flows (NSE from -52 to -123). These results highlight the challenges of lumped conceptual models for representing event-driven hydrology in the UAB using daily inputs. While HBV offers physical interpretability, deep learning models demonstrate improved flood representation compared to this process-based baseline in this data-limited setting.

4.2. Comparative Evaluation of Deep Learning Models

Table 17 summarizes the performance of the climatology and HBV baseline models alongside the five deep learning models (CNN, LSTM, GRU, BiLSTM, and Hybrid CNN–LSTM) using standard and hydrological evaluation metrics. The climatology model provides a conservative reference based on long-term mean discharge, while the HBV conceptual model shows limited skill and substantial bias, particularly under high-flow conditions. In contrast, the deep learning models more effectively captured the rainfall–runoff relationship, with only minor differences in error magnitude, bias, and stability across the training, validation, and test datasets.

Table 17. Quantitative performance metrics of deep learning models (Training/Validation/Test).

Model	MAE (m^3/s)	RMSE (m^3/s)	NSE	KGE	PBIAS (%)
Climatology	52.60/50.17/58.16	77.85/73.80/90.50	0.00/0.00/−0.01	−0.41/−0.41/−0.43	−0.01/−1.70/−20.39
HBV	27.37/21.4/27.12	59.36/41.26/27.12	0.42/0.69/0.64	0.70/0.78/0.82	−1.47/−0.8/−1.55
CNN	9.54/10.55/15.46	21.63/24.27/36.34	0.92/0.89/0.84	0.90/0.88/0.85	−0.16/−3.21/−4.82
LSTM	9.34/9.77/14.54	23.20/23.52/34.99	0.91/0.90/0.85	0.93/0.91/0.90	−0.43/−3.30/−2.94
GRU	9.66/10.08/14.59	22.99/23.77/35.44	0.91/0.90/0.84	0.90/0.89/0.86	−0.14/−2.44/−4.82
BiLSTM	8.91/9.97/14.94	22.76/24.10/36.83	0.91/0.89/0.83	0.93/0.90/0.91	−2.08/−4.58/−2.91
Hybrid CNN–LSTM	8.26/10.46/15.17	19.05/24.79/35.82	0.94/0.89/0.84	0.94/0.90/0.90	0.89/−2.68/−2.14

The climatology model shows weak skill, with high MAE and RMSE (e.g., 58.16 and $90.50 \text{ m}^3/\text{s}$ on the test set), near-zero or slightly negative NSE ($0.00/0.00/−0.01$), and negative KGE ($−0.41$ to $−0.43$), confirming that a long-term mean approximation cannot reproduce day-to-day variability or flood magnitudes. This baseline thus serves as a lower bound against which deep learning performance can be benchmarked.

The HBV model shows moderate overall predictive skill across data splits, with MAE ranging from 21.40 to $27.37 \text{ m}^3/\text{s}$ and RMSE from 41.26 to $59.36 \text{ m}^3/\text{s}$. It achieves reasonable efficiency scores ($\text{NSE} = 0.42\text{--}0.69$ and $\text{KGE} = 0.70\text{--}0.82$) and minimal bias ($\text{PBIAS} \approx -1\%$). However, performance degrades substantially during wet seasons (negative NSE and $\text{PBIAS} \approx -99\%$) and fails completely for extreme peak flows (NSE from -52 to -123). These results indicate that while HBV provides acceptable general discharge simulation, it cannot represent flood dynamics effectively at the Hombole station, confirming the advantage of deep learning models for capturing nonlinear rainfall–runoff relationships during high-flow events.

The CNN model performed well for short-term flow variations, maintaining an NSE of 0.84 and a KGE of 0.85 on the test set, but it underestimated flood peaks due to limited

temporal memory. The LSTM and GRU achieved higher temporal consistency, with NSE and KGE values above 0.84 across datasets. LSTM showed slightly lower error and bias, confirming its stability under both low- and high-flow conditions. GRU, while faster to train, tended to produce minor underpredictions during peak flows. The BiLSTM model improved flow timing and smoothness, leveraging its bidirectional processing to capture dependencies in both temporal directions. It showed a strong KGE of 0.91 on the test set, indicating balanced correlation, variability, and bias, though some dry-season bias persisted.

The Hybrid CNN–LSTM achieved the best overall results, integrating CNN’s local temporal feature extraction with LSTM’s long-term sequence learning. It recorded the lowest training error ($MAE = 8.26 \text{ m}^3/\text{s}$ and $RMSE = 19.05 \text{ m}^3/\text{s}$) and the most stable test performance ($NSE = 0.84$ and $PBIAS \approx -2\%$). Its balanced efficiency and bias metrics confirm its robustness and adaptability to both flood and dry conditions. Although performance differences were small across metrics, the hybrid and bidirectional models consistently provided better flood representation and generalization. The Hybrid CNN–LSTM stands out as the most reliable and operationally suitable model for real-time discharge forecasting in the UAB.

Operational flood forecasting with deep learning models must be interpreted cautiously, particularly in real-time settings where input data are subject to uncertainty. Non-linear architectures such as CNNs and LSTMs are sensitive to data quality issues, including missing observations, sensor errors, and abrupt changes in rainfall or discharge. While the models exhibit strong skill in capturing flood occurrence and timing on historical datasets, their ability to reliably predict peak magnitudes remains limited. Consequently, any application to early-warning or flood management should be viewed as conditional and primarily focused on event detection rather than fully operational deployment, pending further advances in data quality control and uncertainty-aware modeling frameworks.

4.3. Quartile-Based Post-Processing Classification of Discharge

To further assess model behavior under extreme flow conditions, a post-processing classification analysis was carried out using discharge quantiles derived from the training dataset. Instead of focusing solely on continuous error metrics, this approach evaluates whether models correctly distinguish hydrologically relevant flow regimes, including high and extreme flows.

Observed daily discharge in the training period was partitioned into five classes using empirical quantile thresholds: very low flow, low flow, normal flow, high flow, and extreme flow. These thresholds, estimated from the training set only, were then applied unchanged to the validation and test periods, and model predictions were assigned to the same flow classes using the identical thresholds. Classification accuracy was computed by comparing predicted and observed class membership on a daily basis, thereby quantifying each model’s ability to detect extreme and near-extreme conditions even when peak magnitudes are underestimated. This regime-based evaluation offers complementary insight into flood detection skills that is not captured by traditional regression metrics such as RMSE or NSE.

Let Q denote the observed daily river discharge. Quantile thresholds are derived from the training dataset and defined as q_{20} , q_{40} , q_{60} , and q_{80} , corresponding to the 20th, 40th, 60th, and 80th percentiles of the discharge distribution, respectively. Based on these thresholds, discharge values are categorized into five flow classes as follows:

$$C(Q) = \begin{cases} 1, & Q \leq q_{20} \text{ (Very low flow)} \\ 2, & q_{20} \leq Q \leq q_{40} \text{ (low flow)} \\ 3, & q_{40} \leq Q \leq q_{60} \text{ (Normal flow)} \\ 4, & q_{60} \leq Q \leq q_{80} \text{ (High flow)} \\ 5, & Q > q_{80} \text{ (Extreme flow)} \end{cases} \quad (11)$$

Using this classification, both observed discharge y_i and predicted discharge $\hat{y}_i$ are mapped to their respective flow classes. The overall classification accuracy is then computed as

$$Accuracy = \frac{1}{n} \sum_{i=1}^n II(C(y_i) = C(\hat{y}_i)) \quad (12)$$

where n is the total number of samples and $II(\cdot)$ is the indicator function, which equals 1 if the condition is true and 0 otherwise.

Table 18 summarizes the overall and extreme-flow classification performance for all models using the five-class quantile scheme. The climatology baseline attains only 19.9% accuracy on all splits and completely fails to detect extreme events (0% extreme-class accuracy), confirming that a long-term mean benchmark cannot represent day-to-day variability in flow regimes. The HBV model shows moderate overall classification accuracy of 49.2% across all data splits, representing clear improvement over the climatology baseline. Its extreme-flow detection accuracy reaches 77.9% on the test set, indicating reasonable capability to identify high-flow conditions despite magnitude underestimation. These results demonstrate that HBV provides useful regime discrimination for operational flood classification at the Hombole station, though deep learning models achieve substantially higher accuracy.

Table 18. Overall and extreme-flow classification performance using five discharge classes.

Model	Accuracy (Training)	Accuracy (Val)	Accuracy (Test)	Extreme-Class Accuracy (Test)
Climatology	19.9%	19.9%	19.9%	0%
HBV	49.2%	49.2%	49.2%	77.9%
CNN	50.8%	57.3%	61.6%	89.9%
LSTM	66.9%	73.1%	73.3%	90.5%
GRU	54.0%	61.3%	66.7%	90.7%
Bidirectional	68.4%	70.5%	76.4%	89.9%
Hybrid CNN–LSTM	58.9%	66.3%	72.8%	89.5%

In contrast, all deep learning models substantially improve regime classification, with test accuracies ranging from 61.6% (CNN) to 76.4% (BiLSTM). Extreme-flow detection skill is consistently high, with extreme-class accuracies above 89% for all architectures on the test set. These results show that, although regression metrics reveal underestimation of peak magnitudes, the deep learning models are highly effective at identifying the occurrence of extreme floods, which is critical for operational flood early-warning and risk management.

4.4. Practical and Policy Implications

Beyond the technical evaluation of deep learning architectures, particularly the Hybrid CNN–LSTM and Bidirectional LSTM models, this study provides essential insights for operational flood forecasting and water resource management in Ethiopia's UAB. Accurate discharge prediction is crucial for early-warning systems, disaster preparedness, and sustainable water allocation, directly supporting flood-prone communities, agricultural production, and infrastructure resilience.

The CNN–LSTM hybrid model, which demonstrated superior generalization and consistency across training, validation, and test datasets, offers substantial potential for real-time flood forecasting. Its ability to capture both short-term rainfall–runoff interactions (via CNN) and long-term temporal dependencies (via LSTM) makes it particularly suited for the basin’s complex hydrological dynamics. Reliable short to medium term forecasts during the July–August rainy season can significantly enhance the capacity of Disaster Risk Reduction (DRR) agencies to issue timely alerts, coordinate evacuations, and mobilize emergency resources. Furthermore, accurate discharge prediction supports optimized reservoir management at key infrastructures such as Koka Dam, balancing flood mitigation with irrigation and hydropower demands.

Beyond immediate disaster management, the integration of high-performing models into operational hydrological services can inform long-term watershed management, infrastructure planning, and climate adaptation strategies. Identifying spatiotemporal trends in flood behavior enables policymakers to prioritize investments in flood control structures such as levees, drainage systems, and retention basins and enhance land-use planning in flood-prone zones.

To maximize these benefits, several policy and implementation measures are recommended:

- Incorporate CNN–LSTM and BiLSTM models into Ethiopia’s hydrological monitoring frameworks for real-time forecasting applications.
- Strengthen hydrometeorological observation networks, enhance data sharing, and ensure continuous quality control of hydro-climatic datasets.
- Develop technical training programs and user-friendly decision-support tools to bridge the gap between model outputs and actionable decision-making.
- Foster partnerships among government agencies, research institutions, and local communities to enhance knowledge transfer and the adoption of AI-based flood forecasting technologies.

Adopting data-driven, AI-enabled approaches for flood forecasting can transform Ethiopia’s disaster and water management sectors. When supported by policy, infrastructure, and institutional capacity, such models can substantially reduce flood-related risks, strengthen early-warning systems, and promote climate-resilient development in vulnerable regions like the UAB.

To further contextualize these findings, Table 19 presents a comparative summary of related studies employing deep learning and machine learning techniques for flood forecasting. This comparative perspective reinforces the robustness of the present study and situates its contributions within the broader hydrological modeling literature.

Table 19. Comparative analysis of flood forecasting studies using DL and ML approaches.

Ref	Objective	Models Tested	Key Findings
[26]	Predict stream stage heights using multi-modal hydrometeorological data	ConvLSTM	Achieved ~26% improvement in model error over state-of-the-art models, effectively capturing spatiotemporal dynamics in flash flood-prone catchments.
[38]	Develop an interpretable hybrid model for flood forecasting	Transformer, LSTM, AGRS	The AGRS–LSTM–Transformer model enhanced interpretability and forecasting accuracy, particularly for extreme events.
[39]	Real-time flood prediction in the Red River of the North, USA	LSTM, 1D-CNN	LSTM outperformed 1D-CNN in predicting flood events, demonstrating better accuracy in capturing temporal dependencies.

Table 19. Cont.

Ref	Objective	Models Tested	Key Findings
[40]	Improve flood forecasting by incorporating vector direction into LSTM	LSTM with vector direction (VD)	The VD-LSTM model improved prediction accuracy by considering the directionality of flood processes.
[41]	Enhance flood prediction using climate parameters	LSTM	Incorporating climate parameters into LSTM models improved the prediction of extreme flood events.
[27]	Operational flood forecasting using ML models	Various ML models	ML models demonstrated potential in operational settings, with some outperforming traditional hydrological models in certain scenarios.
[42]	Utilize satellite data and ML for flood forecasting in the Ou'ém'e basin	ML models with satellite data	Combining satellite data with ML models improved flood prediction accuracy in data-scarce regions.
[43]	Predict flood stages using deep neural networks	LSTM, Dense Neural Networks, CNN	LSTM models provided accurate near real-time flood stage predictions, outperforming other deep learning models.

5. Conclusions and Future Work

This study compared the performance of deep learning models for daily discharge forecasting at the Hombole station in Ethiopia's UAB using ~40 years of rainfall and stream-flow data (1981–2020). Five architectures were tested (CNN, LSTM, GRU, BiLSTM, and a Hybrid CNN–LSTM) and benchmarked against climatology and HBV baseline models. All deep learning models demonstrated substantially superior performance compared to the baseline models in capturing flow variability and extreme events. Among the tested architectures, the Hybrid CNN–LSTM provided the most balanced and stable performance across MAE, RMSE, NSE, KGE, and PBIAS, reflecting consistent generalization over training, validation, and test periods. LSTM and BiLSTM also showed strong daily forecasting skill, with BiLSTM improving flow timing through bidirectional sequence learning. GRU achieved comparable accuracy with fewer parameters, while the standalone CNN captured short-term variability but systematically underestimated peak flows because of its limited temporal memory.

Extreme-event analysis confirmed that all models underpredicted high-magnitude floods, reflecting the known limitation of standard MSE loss functions, which prioritize average performance rather than rare extremes. Despite this, quartile-based flow-regime classification showed high accuracy in detecting extreme-flow occurrences, indicating that the models remain effective for early-warning applications where timely event detection is more critical than precise peak-magnitude estimation. However, the absence of explicit uncertainty quantification limits the direct use of these deterministic forecasts in operational flood management, where decision thresholds, risk tolerance, and warning confidence critically depend on reliable uncertainty information.

Future work should address these limitations by incorporating multi-step forecasting, attention mechanisms, or Transformer-based architectures to better capture rare peak events and improve interpretability. Additionally, the exploration of peak-weighted or extreme-focused loss functions, ensemble methods, and probabilistic approaches would enhance extreme-event representation. Validation across additional Ethiopian basins, such as the Blue Nile and Dire Dawa Dechatu catchment, is recommended to assess spatial transferability and support broader flood risk management and planning.

Author Contributions: Conceptualization, G.M.M., A.G.S., G.T.D. and N.E.B.; data curation, G.M.M. and A.G.S.; formal analysis, G.M.M., A.G.S., G.T.D. and N.E.B.; investigation, G.M.M. and A.G.S.; methodology, G.M.M., A.G.S., G.T.D., N.E.B. and Y.M.; validation, G.M.M., A.G.S., G.T.D. and N.E.B.; resources, G.M.M. and A.G.S.; writing—original draft preparation, G.M.M.; writing—review and editing, G.M.M., A.G.S., G.T.D., N.E.B., E.O.G. and Y.M.; visualization, G.M.M., A.G.S. and N.E.B.; supervision, A.G.S., G.T.D. and E.O.G.; project administration, A.G.S. and E.O.G.; funding acquisition, E.O.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by WACCA-Ethiopia Phase 2: A programme to strengthen the development of climate and water information services in Ethiopia Sida Contribution No. 15585 and the Swedish Meteorological and Hydrological Institute (SMHI) Research Collaboration and Academic Support program. A.G.S. would like to acknowledge funding from NSF through the Learning the Earth with Artificial intelligence and Physics (LEAP) Science and Technology Center (STC) (Award #2019625).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: This work was supported by the Swedish Meteorological and Hydrological Institute (SMHI) Research Collaboration and Academic Support program. Addisu G. Semie would like to acknowledge funding from NSF through the Learning the Earth with Artificial intelligence and Physics (LEAP) Science and Technology Center (STC) (Award #2019625). GTD is currently affiliated with Environment and Climate Change Canada. The authors gratefully acknowledge the Ethiopian Ministry of Water and Energy (MoWE) for providing the discharge data from the Hombole gauge station and the Ethiopian Meteorological Institute (EMI) for supplying the daily rainfall observations used in this study.

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

References

1. WHO. Floods. 2024. Available online: https://www.who.int/health-topics/floods/#tab=tab_1 (accessed on 23 May 2025).
2. Kundzewicz, Z.W.; Kanae, S.; Seneviratne, S.I.; Handmer, J.; Nicholls, N.; Peduzzi, P.; Mechler, R.; Bouwer, L.M.; Arnell, N.; Mach, K.; et al. Flood risk and climate change: Global and regional perspectives. *Hydrol. Sci. J.* **2014**, *59*, 1–28. [CrossRef]
3. OCHA. Ethiopia: Flooding Update. 2024. Available online: <https://www.unocha.org/publications/report/ethiopia/ethiopia-update-flooding-24-may-2024> (accessed on 23 May 2025).
4. WHO. Flooding in Ethiopia: Public Health Situation Analysis (PHSA). 2024. Available online: <https://www.afro.who.int/countries/ethiopia/publication/flooding-ethiopia-public-health-situation-analysis-phsa-24-may-2024> (accessed on 23 May 2025).
5. Emiru, N.C.; Recha, J.W.; Thompson, J.R.; Belay, A.; Aynekulu, E.; Manyevere, A.; Demissie, T.D.; Osano, P.M.; Hussein, J.; Molla, M.B.; et al. Impact of climate change on the hydrology of the upper awash river basin, Ethiopia. *Hydrology* **2021**, *9*, 3. [CrossRef]
6. Sun, M.; Li, Z.; Yao, C.; Liu, Z.; Wang, J.; Hou, A.; Zhang, K.; Huo, W.; Liu, M. Evaluation of flood prediction capability of the WRF-hydro model based on multiple forcing scenarios. *Water* **2020**, *12*, 874. [CrossRef]
7. Semie, A.G.; Diro, G.T.; Demissie, T.; Yigezu, Y.M.; Hailu, B. Towards improved flash flood forecasting over dire Dawa, Ethiopia using WRF-hydro. *Water* **2023**, *15*, 3262. [CrossRef]
8. Dessu, S.B.; Seid, A.H.; Abiy, A.Z.; Melesse, A.M. Flood forecasting and stream flow simulation of the upper awash river basin, Ethiopia using geospatial stream flow model (GeoSFM). In *Landscape Dynamics, Soils and Hydrological Processes in Varied Climates*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 367–384.
9. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
10. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
11. Gon, R. Multivariate Time Series Analysis with Deep Learning. Ph.D. Dissertation, Universidade do Porto, Porto, Portugal, 2021.
12. Chen, C.; Hui, Q.; Xie, W.; Wan, S.; Zhou, Y.; Pei, Q. Convolutional neural networks for forecasting flood process in internet-of-things enabled smart city. *Comput. Netw.* **2021**, *186*, 107744. [CrossRef]
13. Kabir, S.; Patidar, S.; Xia, X.; Liang, Q.; Neal, J.; Pender, G. A deep convolutional neural network model for rapid prediction of fluvial flood inundation. *J. Hydrol.* **2020**, *590*, 125481. [CrossRef]

14. Apaydin, H.; Feizi, H.; Sattari, M.T.; Colak, M.S.; Shamshirband, S.; Chau, K.-W. Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting. *Water* **2020**, *12*, 1500. [\[CrossRef\]](#)
15. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#)
16. Cho, K.; Van Merri, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv* **2014**, arXiv:1409.1259. [\[CrossRef\]](#)
17. Oddo, P.C.; Bolten, J.D.; Kumar, S.V.; Cleary, B. Deep convolutional LSTM for improved flash flood prediction. *Front. Water* **2024**, *6*, 1346104. [\[CrossRef\]](#)
18. Daba, M.H.; Ayele, G.T.; You, S. Long-term homogeneity and trends of hydroclimatic variables in upper awash river basin, Ethiopia. *Adv. Meteorol.* **2020**, *2020*, 8861959. [\[CrossRef\]](#)
19. Gebremichael, Abayneh and Gebremariam, Ephrem and Desta, Hayal, Flood Hazard Area Mapping Using GIS and AHP in Awash River Basin (ARB), Ethiopia. Available online: <https://ssrn.com/abstract=4939877> (accessed on 2 December 2025).
20. Tsige, M.; Malcherek, A.; Seleshi, Y. Estimating the best exponent of the modified universal soil loss equation and regionalizing the modified universal soil loss equation under hydro-climatic condition of Ethiopia. *Preprints* **2022**.
21. Taye, M.T.; Haile, A.T.; Dessalegn, M.; Nigussie, L.; Bekele, T.W.; Nicol, A.; Dyer, E. *Flood Adaptation and Mitigation in the Awash Basin: Responding to New Climate Patterns*; REACH Synthesis Report; University of Oxford: Oxford, UK, 2024.
22. Le, X.-H.; Ho, H.V.; Lee, G.; Jung, S. Application of long short-term memory (LSTM) neural network for flood forecasting. *Water* **2019**, *11*, 1387. [\[CrossRef\]](#)
23. Subramaniyam, C.; Rajapakse, R. Variants of recurrent neural network models for real-time flood forecasting in Kelani River basin, Sri Lanka. In Proceedings of the International Conference on Climate Change, Colombo, Sri Lanka, 9–10 February 2023; Volume 7, pp. 56–74.
24. Lewinson, E. Three Approaches to Encoding Time Information as Features for ML Models. 2022. Available online: <https://developer.nvidia.com/blog/three-approaches-to-encoding-time-information-as-features-for-ml-models/> (accessed on 20 September 2025).
25. Cho, K.; van Merri, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
26. Feng, D.; Shen, C. Deep learning on hydrologic data: Lstm and beyond. *Water Resour. Res.* **2020**, *56*, e2020WR028091.
27. Nevo, S.; Morin, E.; Rosenthal, A.G.; Metzger, A.; Barshai, C.; Weitzner, D.; Voloshin, D.; Kratzert, F.; Elidan, G.; Dror, G.; et al. Flood forecasting with machine learning models in an operational framework. *Hydrol. Earth Syst. Sci.* **2022**, *26*, 4013–4032. [\[CrossRef\]](#)
28. Seneviratne, S.I.; Zhang, X.; Adnan, M.; Badi, W.; Dereczynski, C.; Luca, A.D.; Ghosh, S.; Iskandar, I.; Kossin, J.; Lewis, S.; et al. Weather and climate extreme events in a changing climate. In *Climate Change 2021: The Physical Science Basis*; Cambridge University Press: Cambridge, UK, 2021.
29. Mosavi, A.; Ozturk, P.; Chau, K.-W. Flood prediction using machine learning models: Literature review. *Water* **2018**, *10*, 1536. [\[CrossRef\]](#)
30. Choi, H.S.; Kim, J.H.; Lee, E.H.; Yoon, S.-K. Development of a revised multi-layer perceptron model for dam inflow prediction. *Water* **2022**, *14*, 1878. [\[CrossRef\]](#)
31. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM networks. In Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 31 July–4 August 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 4, pp. 2047–2052.
32. Zhao, X.; Dong, S.; Rao, H.; Ming, W. Water flow forecasting model based on bidirectional long- and short-term memory and attention mechanism. *Water* **2025**, *17*, 2118. [\[CrossRef\]](#)
33. Wegayehu, E.B.; Muluneh, F.B. Multivariate streamflow simulation using hybrid deep learning models. *Comput. Intell. Neurosci.* **2021**, *2021*, 5172658. [\[CrossRef\]](#)
34. Li, X.; Xu, W.; Ren, M.; Jiang, Y.; Fu, G. Hybrid CNN-LSTM models for river flow prediction. *Water Supply* **2022**, *22*, 4902–4919. [\[CrossRef\]](#)
35. Zhang, Y.; Gu, Z.; Th, J.V.G.; Yang, S.X.; Gharabaghi, B. The discharge forecasting of multiple monitoring station for humber river by hybrid LSTM models. *Water* **2022**, *14*, 1794. [\[CrossRef\]](#)
36. Bergstrom, S. *The HBV-Model—Its Structure and Applications*; SMHI Reports RH No. 4; SMHI: Norrköping, Sweden, 1992.
37. Seibert, J.; Vis, M.J. The HBV model—A guide for practitioners. *Hydrol. Process.* **2012**, *26*, 14–26.
38. Li, W.; Liu, C.; Xu, Y.; Niu, C.; Li, R.; Li, M.; Hu, C.; Tian, L. An interpretable hybrid deep learning model for flood forecasting based on Transformer and LSTM. *J. Hydrol. Reg. Stud.* **2024**, *54*, 101873. [\[CrossRef\]](#)
39. Atashi, V.; Kardan, R.; Gorji, H.T.; Lim, Y.H. Comparative study of deep learning LSTM and 1d-cnn models for real-time flood prediction in red river of the north, USA. In Proceedings of the 2023 IEEE International Conference on Electro Information Technology (eIT), Romeville, IL, USA, 18–20 May 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 22–28.

40. Xie, T.; Hu, C.; Liu, C.; Li, W.; Niu, C.; Li, R. Study on long shortterm memory based on vector direction of flood process for flood forecasting. *Sci. Rep.* **2024**, *14*, 21446. [[CrossRef](#)]
41. Atashi, V.; Gorji, H.T. Enhanced flood prediction using LSTM and climate parameters: Multi-station analysis of snowmelt-induced flooding in the red river of the north. *J. Hydroinformatics* **2025**, *27*, 245–260. [[CrossRef](#)]
42. Obada, E.; Biao, E.I.; Zohou, P.J.; Yarou, H.; Hounnondaho, F.Z.; Alamou, E.A. Using machine learning and satellite data to improve flood forecasting: The case of the Ouémé basin at the bétérou outlet. *Hydrol. Res.* **2025**, *56*, 153–166. [[CrossRef](#)]
43. Windheuser, L.; Karanjit, R.; Pally, R.; Samadi, S.; Hubig, N.C. An end-to-end flood stage prediction system using deep neural networks. *Earth Space Sci.* **2023**, *10*, e2022EA002385. [[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.