Modeling Earth Systems and Environment (2025) 11:415
https://doi.org/10.1007/s40808-025-02611-9

ORIGINAL ARTICLE

Comparative insights into independent and hybrid modeling strategies for effective river water level prediction and management

Osama Ragab Ibrahim[1] · Aynaz Vafaei[2] · Sanam Ansari[3] · Erfan Abdi[2] · Maryam Sifaei[4] · Seyed Mahdi Jafari Mohammadi[5]

Received: 25 June 2025 / Accepted: 26 August 2025

Abstract

Accurate river water level prediction is essential for flood risk mitigation, efficient water allocation, and ecosystem protection. It also enables assessment of climate change impacts on hydrological regimes, providing critical data for adaptation strategies. This study applies advanced forecasting models to two major rivers—the Mississippi and the Danube—characterized by distinct climatic and hydrological profiles, over 17 years (2008–2025). A rigorous data preprocessing protocol was implemented, involving outlier removal and imputation of missing values to ensure data integrity. For monthly water level prediction, two standalone models—deep neural network (DNN) and convolution neural network (CNN)—were used alongside two hybrid architectures: CNN with long short-term memory (CNN-LSTM) and CNN with bidirectional LSTM (CNN-BiLSTM). Input features included mean temperature, precipitation, relative humidity, and three lagged water level values. Model performance was evaluated using visual diagnostics and three statistical metrics: root mean square error (RMSE), Pearson's correlation coefficient (r), and Nash–Sutcliffe efficiency (NSE). The CNN-BiLSTM model exhibited superior predictive capability. At the Danube station, it achieved an RMSE of 0.47 m, r of 0.94, and NSE of 0.88. At the Mississippi station, it yielded an RMSE of 0.93 m, r of 0.96, and NSE of 0.95. These results highlight the model's robustness across diverse hydrological settings. The validation results show consistent performance and reliability. This modeling framework provides valuable support for water resources management, enabling informed planning and decision-making in these two basins.

Keywords Flood disaster · CNN-BiLSTM · Rivers water level · Analytical models · Hybrid model

✉ Erfan Abdi
Erabdi78@gmail.com

Osama Ragab Ibrahim
oibrahim@su.edu.om

Aynaz Vafaei
aynazvafaei@gmail.com

Sanam Ansari
sanam.ansari@kapadokya.edu.tr

Maryam Sifaei
maryam.sifaei@connect.qut.edu.au

Seyed Mahdi Jafari Mohammadi
seyedmahdi.jafari@mail.polimi.it

1 Department of Engineering, Sohar University, Sohar, Oman

2 University of Tabriz, Tabriz, Iran

3 Departure of Information Security Technology, Cappadocia University, Nevsehir, Turkey

4 Queensland University of Technology, Brisbane, Australia

5 Politecnico di Milano University, Milan, Italy

Introduction

Rivers are crucial for both ecological and human needs. They support diverse ecosystems, contributing to biodiversity and ecological balance by providing habitats for various species and acting as natural water filters. Rivers are vital freshwater sources for drinking, agriculture, and industry, supporting human settlements and economies (Yamazaki et al. 2012). They also serve as food sources through fishing and support agriculture via irrigation. Historically and currently, rivers are used for transportation and recreational activities, promoting tourism and mental health benefits. Additionally, rivers hold cultural significance, often being revered as sacred sites. Overall, rivers play a pivotal role in shaping landscapes, supporting livelihoods, and fostering economic and cultural activities, making their conservation essential.

Published online: 11 September 2025

Monitoring river water levels helps predict and manage flood risks, providing early warnings to protect people and properties. This is especially important in urban areas where flooding can cause significant damage and disruption. Cities face increasing water risks due to climate change, and maintaining healthy river basins is essential for urban resilience. Enhancing river health helps cities withstand water-related challenges and supports economic stability (Hu et al. 2022). River water levels impact the health of plant and animal species, as many require specific water levels at different times of the year. This ecological balance is vital for maintaining biodiversity in urban environments.

Urban flood resilience depends heavily on precise water level predictions, which enable authorities to activate protective measures and evacuation protocols before crisis events. It also aids in managing water resources sustainably, minimizing risks associated with both floods and droughts. Accurate predictions support urban planning by informing infrastructure projects such as drainage systems and flood defenses, preventing urban flooding, and ensuring smooth drainage (Santos et al. 2024). Additionally, they facilitate emergency response and evacuation efforts during extreme weather events, reducing the impact on communities. Overall, predicting river water levels enhances urban resilience and ensures better management of water-related challenges. Recent advances in hydrological forecasting employ diverse machine-learning approaches. Kim et al. (2022) demonstrated the effectiveness of deep learning architectures (DNNs and LSTMs) for water level prediction in Vietnam's Phan Rang River. Similarly, Atashi et al. (2022) conducted a comparative analysis of SARIMA, Random Forest, and LSTM models using twelve years of flood data (2007–2019), highlighting the strengths of each methodology. Results revealed that the Deep Learning-based LSTM method surpasses SARIMA and RF in accuracy. Le et al. (2021) used GRU-based deep learning models to predict water levels on South Korea's Geum River for short durations (1–9 h). Using only hourly water Level data, the models achieved excellent performance, with NSE exceeding 99% for 1-hour forecasts. Abdi et al. (2024) addressed groundwater surface (GWS) prediction with predictive models (DNN, CNN). Results showed that CNN outperformed DNN in predicting GWS levels.

River water levels are influenced by complex factors like climate variability and spatial-temporal irregularities. Hybrid modeling frameworks address hydrological complexity through complementary architectures - combining, for instance, neural networks for nonlinear pattern recognition with physical models for process-based constraints. Such systems achieve site-specific adaptation through localized parameterization of basin characteristics and flow regimes. This flexibility is essential for managing diverse water resources effectively. More researchers used the hybrid model to improve accuracy and decrease errors in this field. Addressing noise and non-stationarity in lake data, Barzegar et al. (2021) proposed a novel framework pairing wavelet-based signal correction (BC-MODWT) with spatiotemporal deep learning (CNN-LSTM), achieving robust water level forecasts for transboundary water management. Compared to machine learning models like Random Forest (RF) and Support Vector Regression (SVR), the MODWT-based CNN-LSTM model significantly improved forecast accuracy for Lake Michigan and Lake Ontario. Zakaria et al. (2021) evaluated AI-based models—Multilayer Perceptron Neural Networks (MLP-NN) and Adaptive Neuro-Fuzzy Inference System (ANFIS)—for hourly water level forecasting on Malaysia's Muda River. The MLP-NN (RMSE: 0.01299) and ANFIS (RMSE: 0.0130) models demonstrated reliability for forecasts at multiple lead times (1 to 24 h). Dong et al. (2024) conducted a comprehensive evaluation of 48-hour forecasting approaches for the Mekong River's Chiang Saen Station, systematically comparing standalone physical Variable Infiltration Capacity and Hydrology-Hydraulic (VIC, HH) models, pure machine learning Gated Recurrent Unit (GRU), and their hybrid counterparts (VIC-GRU, HH-GRU) to quantify relative predictive performance.

The Mississippi and Danube Rivers are two of the world's most significant waterways, each playing a crucial role in their region's economy, politics, and environment. The Mississippi River, flowing through multiple U.S. states, supports over $400 billion in economic activity and more than 1.3 million jobs across transportation, agriculture, manufacturing, and tourism sectors. It facilitates the shipment of around 175 million tons of freight annually and serves as a vital water source for millions. Its management involves complex coordination among federal and state agencies due to its extensive socio-economic impact (Sarikulov and Bekimbetov 2024). The Danube River, Europe's second-longest river, traverses ten countries, serving as a critical transboundary waterway for commerce, energy production, agriculture, and tourism. It supports rich biodiversity within several protected areas and requires multinational governance to address flood risk, water quality, and sustainable development challenges. Together, these rivers embody distinct hydrological and socio-political contexts, providing a meaningful basis for evaluating the robustness and transferability of river water level forecasting models across diverse and influential fluvial systems.

In this study, we focus on two major rivers—the Mississippi in North America and the Danube in Europe—selected for their contrasting hydrological and climatic characteristics, as well as the availability of long-term, high-quality observational data. The Mississippi, one of the largest river basins in the world, experiences a temperate climate with highly variable seasonal flows, influenced by extensive precipitation and snowmelt patterns. In contrast, the Danube flows through multiple European countries, exhibiting distinct precipitation regimes, flow dynamics, and anthropogenic influences on its hydrology. By examining these two diverse river systems, the study provides a robust framework to evaluate the performance, transferability, and generalizability of independent and hybrid deep learning models, highlighting how model performance may vary across different hydrological and climatic contexts. Also, Prior studies often focus on single river basins or short-term datasets, which limits our understanding of model robustness across diverse hydrological and climatic conditions. Furthermore, limited attention has been given to systematically comparing standalone and hybrid models using comprehensive performance metrics that capture not only accuracy but also operational reliability over extended periods. Addressing critical needs in hydrological monitoring, this study presents an advanced predictive framework for monthly water level forecasting in two major river systems, supporting sustainable water resource management decisions. For this purpose, meteorological data, including average temperature, precipitation, and average relative humidity, for a given station on the Danube River in Europe and the Mississippi River in America, were used. To increase the efficiency and reliability of modeling, in addition to meteorological data, a three-month water level delay was also considered as input. Two independent models, DNN and CNN, and two hybrid models, CNN-LSTM and CNN-BiLSTM, were comparatively evaluated and their performance was evaluated for monthly water level prediction. It is worth noting that the CNN-BiLSTM hybrid model has not been used to model the water level of these two rivers so far.

Materials and methods

Study area and data required

Focusing on two of the world's most hydrologically significant rivers, this research examines water level dynamics in the Danube and Mississippi basins, systems critical for regional water security and ecosystem services. Figure 1 illustrates the locations of the rivers and the stations analyzed. The Mississippi and Danube Rivers were selected for this study due to their significance as two of the world's largest and most hydrologically distinct river systems. The Mississippi River represents a temperate continental climate with substantial seasonal variability and extensive watershed characteristics influenced by varied land use and climatic zones. In contrast, the Danube River flows through multiple climatic zones in Europe, exhibiting diverse hydrological processes and management challenges influenced by

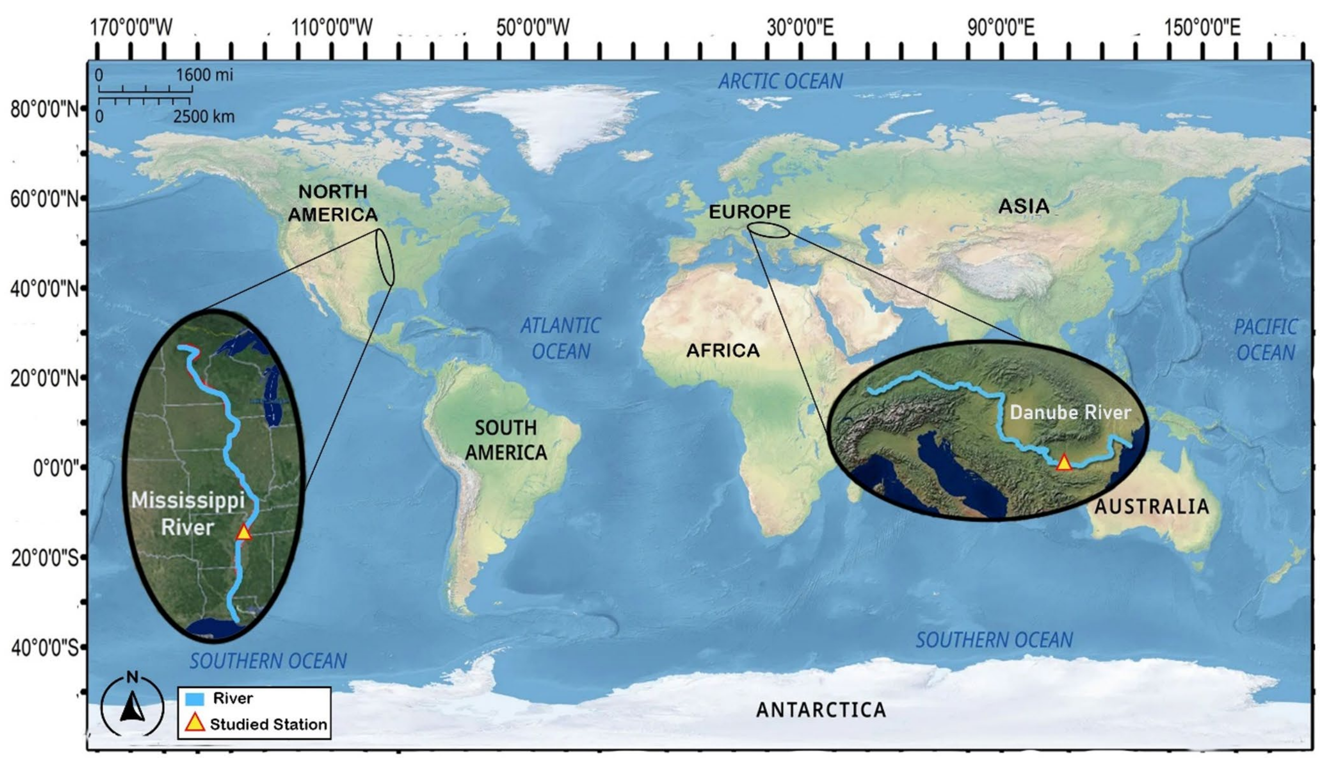

Fig. 1 The location of the two selected rivers and stations

mountainous and lowland regions. By choosing these two rivers, the study captures a wide spectrum of hydrological regimes, enabling the evaluation of model robustness and transferability across diverse climatic, geographical, and hydrological conditions. This selection provides meaningful insights into the applicability of hybrid deep learning models in different fluvial environments, which is critical for developing universally relevant water level forecasting tools.

The Danube River system originates in the Black Forest region of Germany through the confluence of the Brigach and Breg rivers near Donaueschingen. With a 2850 km course traversing 10 countries, it represents Europe's most international river and second-longest after the Volga. It covers about 817,000 km², extending into additional countries beyond those it directly flows through (Mănoiu and Crăciun 2021). The Danube supplies water for drinking, industry, and agriculture. Cities such as Vienna, Budapest, and Belgrade utilize the river for industrial processes and municipal water, supporting manufacturing and urban populations. The river basin is among Europe's most fertile, supporting extensive agriculture. More than 50% of the basin's land is used for farming, with favorable conditions for crops like maize, soybeans, and sunflowers. In Eastern Danube countries, agriculture is a major employer and a key driver of rural economies.

As the core of North America's largest river system, the Mississippi originates at Lake Itasca in northern Minnesota and traverses 3,766 km southward to its deltaic discharge into the Gulf of Mexico, draining portions of 32 United States. The Mississippi River has historically been a critical transportation and trade route, shaping the development of major metropolitan areas along its banks, including Minneapolis–Saint Paul, St. Louis, Memphis, New Orleans, and Baton Rouge (Yin et al. 2023). The Mississippi River is essential to the U.S. agricultural industry by providing a vast transportation network that efficiently moves the majority of the nation's grain exports, especially soybeans and corn, from the fertile farmlands of its expansive basin to global markets via major ports Like New Orleans. Its watershed spans 31 states, supporting over 30 million acres of productive farmland enriched by nutrient-rich sediments and ample irrigation water from the river.

The water level in both the Mississippi and Danube rivers plays a critical role in their ecological health, navigation, flood control, and regional water management. Still, the specifics differ due to their distinct geographic and hydrological contexts. Figure 2 presents the time series of water levels for the Danube and Mississippi rivers at selected stations, along with the models' training and testing series.

The Mississippi's water level dynamics represent a key indicator of watershed health, directly influencing flood risks, sediment transport, and aquatic habitats throughout North America's largest river basin. The water level of the Mississippi River directly impacts many aspects of life in the surrounding areas, including agriculture, transportation, and recreation. A proper water level in the Mississippi River is essential for the irrigation of crops, ensuring a steady water supply for farming communities along the river (Hiatt et al. 2019). Additionally, a consistent water level is necessary for the navigation of commercial vessels and barges that transport goods along the river, contributing significantly to the economy of the region. Lastly, the water level of the Mississippi River is vital for recreational activities such as fishing, boating, and wildlife conservation.

Similarly, the Danube River in Europe is also heavily dependent on water levels for various purposes. The Danube's watershed sustains multiple sectors across Central and Eastern Europe, serving as a vital freshwater resource for crop production, manufacturing processes, and municipal water supplies in riparian countries. Maintaining an optimal water level in the Danube River is essential for ensuring a sustainable supply of water for these purposes. Danube water level fluctuations significantly influence both economic and ecological systems—regulating barge navigation capacity for regional trade while simultaneously maintaining habitat stability for endemic aquatic species (Tadić et al. 2022).

Before diving into the process of modeling, it is crucial to understand the importance of preprocessing data. Data preprocessing encompasses a suite of transformative operations—including cleaning (handling missing values/outliers), normalization, and feature engineering—to convert raw inputs into optimized model-ready datasets. This foundational stage critically determines subsequent analytical validity and predictive performance. By preprocessing the data, the model becomes more efficient in extracting meaningful patterns and insights, which in turn leads to more accurate predictions and decisions (Isik et al. 2012). One of the reasons why preprocessing data is important is to deal with missing values and outliers. Missing data points create spurious patterns that compromise generalization capability, while outliers exert undue leverage on decision boundaries in predictive algorithms. By handling missing values and outliers appropriately, the model becomes more robust and reliable. Standardizing the features ensures that all variables are on the same scale, which improves the convergence of the model and prevents certain features from dominating the others.

To predict and model water levels at two selected stations along the Danube and Mississippi rivers, monthly data were collected from the (https://dahiti.dgfi.tum.de/). The study focused on 17 years, spanning from 2008 to February 2025. Also, considering that multivariate modeling is more

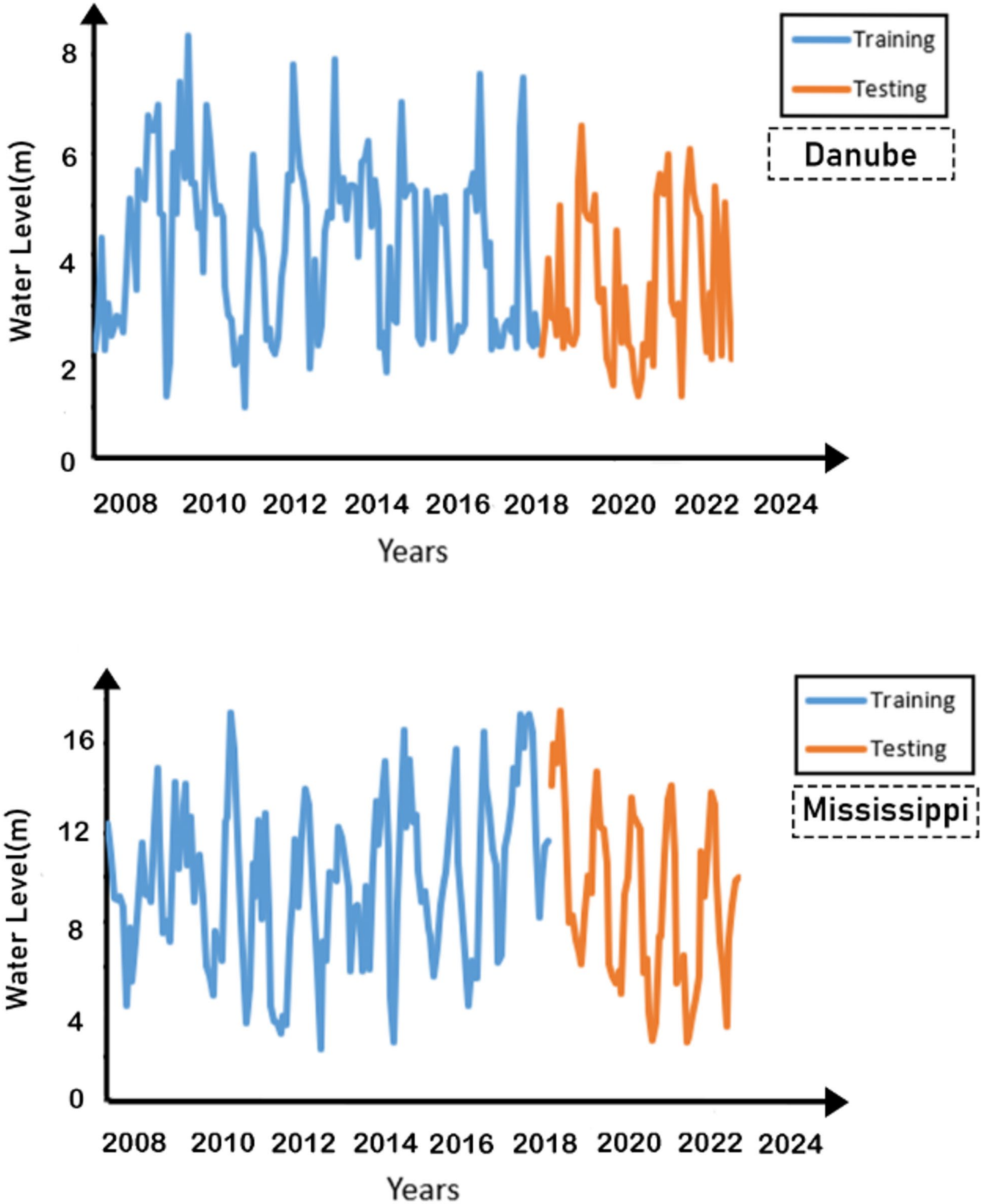

Fig. 2 plot of water level changes along with the division of training and testing for two stations

important than univariate, three parameters of precipitation (PP), average temperature (T_{mean}), average relative humidity (RH_{mean}), and three months of water levels (WL) lag were used. Table 1 shows the statistical characteristics of the parameters used.

Water level measurements for the Danube and Mississippi Rivers were sourced from the Database for Hydrological Time Series of Inland Waters (DAHITI), maintained by the Deutsches Geodätisches Forschungsinstitut at the Technical University of Munich (DGFI-TUM). DAHITI compiles time series data using satellite altimetry from multiple missions, including Jason-1 through Jason-3, Sentinel-6 A, and others. The raw satellite readings are refined through a sophisticated processing algorithm that integrates enhanced outlier detection and Kalman filtering, ensuring the generation of dependable inland water level records. To enhance precision, the methodology incorporates geophysical adjustments such as atmospheric corrections, tidal influences, and range bias compensation. The temporal resolution of the dataset is generally monthly or aligned with satellite pass intervals; for this analysis, monthly data points were utilized.

DAHITI leverages multi-mission satellite altimetry combined through extended outlier rejection and Kalman filtering to retrieve inland water level time series with impressive accuracy. Nevertheless, uncertainties may arise from discrepancies between altimetric and in-situ gauge measurements—reported root-mean-square differences range from approximately 8 cm to 114 cm for rivers, depending on river size and morphology—highlighting potential calibration and measurement biases (Schwatke et al. 2015). While DAHITI applies sophisticated correction algorithms to mitigate these issues, such limitations should be acknowledged as potential sources of residual bias in model training.

Outliers exhibiting substantial deviation from the general data distribution were either excluded or adjusted. To address data gaps, missing entries were imputed using the arithmetic mean, ensuring dataset completeness. Following

Table 1 Statistical characteristics of the parameters used for the two stations

Statistical	PP (cm)	T_{mean} (°C)	RH_{mean}	WL (m)
Danube				
Max	37.39	26.31	0.93	8.35
Mean	4.93	12.31	0.73	3.98
Min	0	−5.58	0.52	1.02
Median	4.29	12.17	0.73	3.96
N Zero	2	0	0	0
Mississippi				
Max	28.67	28.07	0.86	17.40
Mean	7.00	15.38	0.72	9.59
Min	0	−0.39	0.48	2.34
Median	5.415	15.77	0.72	9.51
N Zero	3	0	0	0

preprocessing, all features underwent Z-score normalization, standardizing each variable to a mean of zero and a standard deviation of one. Approximately 2% of the data were modified due to inconsistencies or absence. The dataset was divided into 70% for training and 30% for testing while strictly preserving temporal order. No shuffling was applied to prevent information leakage from future observations, ensuring that model evaluation reflects real-world forecasting conditions where future data are not available during training. To predict monthly water levels using DNN, CNN, CNN-LSTM, and CNN-BiLSTM models, six input parameters were utilized: precipitation, average temperature, and average relative humidity. Additionally, three water level delays were incorporated to enhance the models' reliability. Notably, every stage of this study was carried out using Python version 3.1. Figure 3 illustrates the procedure for conducting monthly forecasts of water levels for both the Danube and Mississippi rivers.

An overview of DNN and CNN

Drawing inspiration from neurobiological architectures, deep neural networks (DNNs) employ hierarchical layers of artificial neurons to progressively extract and transform features. This multi-layered organization enables sophisticated pattern recognition capabilities, excelling in high-dimensional domains like time series and NLP (Wan et al. 2023). DNN architectures employ a hierarchical organization comprising: (1) an input layer for feature representation, (2) multiple hidden layers for nonlinear transformation, and (3) an output layer for prediction generation. Interlayer connections, modulated by trainable weight parameters, enable progressive feature abstraction through forward propagation (Qian et al. 2014).

$$Z = \sum {}_{i=1}^{n} w_i x_i + b \tag{1}$$

where w_i are the weights, x_i are the inputs, and b is the bias term.

DNNs are powerful models capable of learning complex patterns directly from raw data through multiple layers of computation. Their hidden layers enable feature abstraction, allowing for accurate predictions and classifications. The architecture—specifically the depth and width of the network—plays a crucial role in determining its learning capacity, with deeper networks capturing more intricate relationships but requiring greater computational resources and risking overfitting. Similarly, wider networks can enhance learning but may slow training and also face overfitting challenges. A key hyperparameter in training is the learning rate, which influences how quickly the model adapts: a high rate speeds up convergence but may destabilize training,

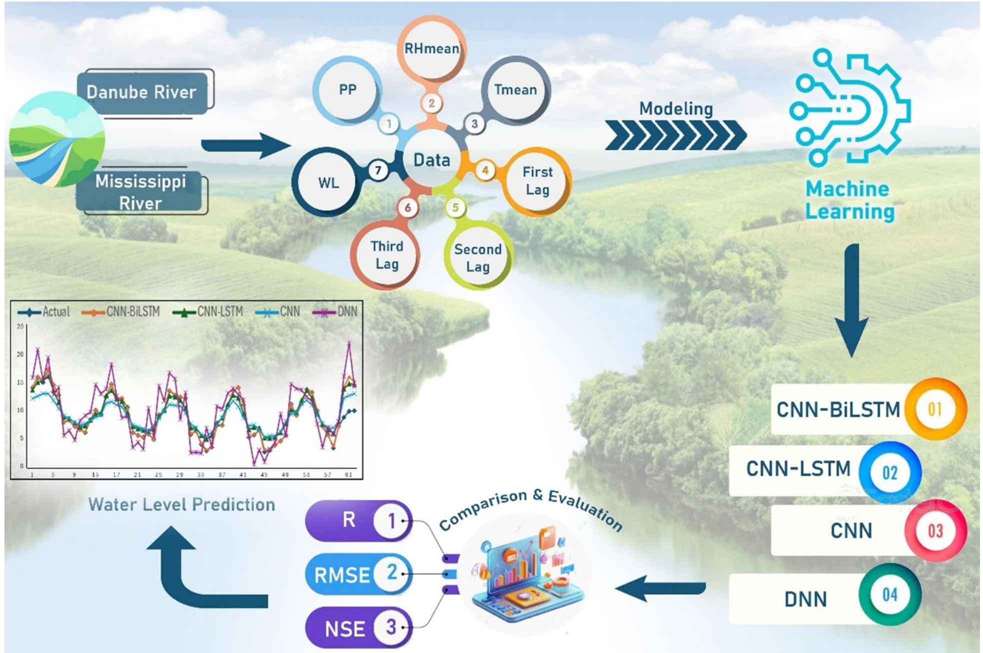

Fig. 3 Flowchart of the steps taken to predict the water level of the selected rivers

while a low rate ensures stability at the cost of slower progress. Effective tuning of these parameters demands a nuanced understanding of both the model and the data to achieve optimal performance.

The model was compiled using the Adam optimizer, which is widely used in deep Learning for its adaptive Learning rate capabilities. In this configuration, the learning rate was set to 2.2, a considerably high value intended to accelerate convergence. However, such a high learning rate may lead to training instability or divergence if not carefully managed, especially in models sensitive to weight updates. The Mean Squared Error (MSE) was selected as the loss function due to its suitability for continuous-valued regression tasks, and the Mean Absolute Error (MAE) was used as a supplementary metric to provide an interpretable measure of average prediction error.

Convolutional Neural Networks (CNNs) are widely used in time series analysis due to their ability to capture temporal patterns through layered processing (Swiderski et al. 2022). The model begins with convolutional filters that slide across the input data, extracting key features via element-wise operations. These features are then passed through pooling layers, which reduce dimensionality and retain essential information while minimizing redundancy. Following this, dense layers perform complex transformations, converting spatial patterns into abstract representations suitable for classification or prediction. Ultimately, the network uses learned weights to map these features to output values, enabling accurate forecasting or pattern recognition in time-dependent data.

One important hyperparameter in the CNN model structure for time series analysis is the kernel size. The kernel size determines the size of the filter that is applied to the input data. A larger kernel size can help capture broader patterns in the data, while a smaller kernel size can capture finer details. The choice of kernel size should be based on the complexity of the time series data and the patterns that need to be extracted. The following formula is used to determine the output size of the convolution layer (Wang et al. 2021).

$$\text{output size} = \left\lfloor \frac{input\ size - filter\ size + 2 \times\ padding}{stride} + 1 \right\rfloor \quad (2)$$

where input size is the length of the input sequence, filter size is the size of the convolution kernel, padding is the number of zeros added to both ends of the input, and stride is the step size of the filter movement. Increasing the filter size or padding generally increases the receptive field, while a larger stride reduces the output length.

Another crucial hyperparameter in the CNN model structure for time series forecasting is the stride. The stride determines how far the filter moves across the input data at each step. A larger stride can reduce the computational cost of the model but may result in information loss, while a smaller stride can capture more details but may increase the computational complexity (Wang et al. 2021). Finding the right balance between stride and kernel size is essential for achieving optimal performance in time series forecasting tasks. In conclusion, selecting the appropriate hyperparameters in the CNN model structure for time series analysis is crucial for achieving accurate and efficient forecasting results. The model structure used to predict the parameter of the water level is shown in Fig. 4.

In this study, a compact deep learning model based on a one-dimensional Convolutional Neural Network (Conv1D) was developed to predict the target variable from univariate time series data. The first layer is a Conv1D layer with 24 filters and a kernel size of 2, using the ReLU activation function to extract local temporal features. To reduce overfitting, a Dropout layer with a dropout rate of 0.5 is applied immediately after convolution. Following the convolution and regularization stages, the output is flattened and passed to a fully connected Dense layer with 10 neurons and ReLU activation. Finally, a Dense layer with a single unit generates the regression output. The model is compiled using the Adam optimizer with an elevated Learning rate of 0.1, which accelerates convergence during training. Mean Squared Error (MSE) is used as the loss function, while Mean Absolute Error (MAE) is reported as an additional performance metric. This architecture offers a lightweight yet effective approach for time series forecasting with minimal input complexity.

CNN-LSTM and CNN-BiLSTM hybrid model

Long Short-Term Memory (LSTM) models represent an advanced type of recurrent neural network (RNN) widely employed for time series prediction. Traditional RNNs often face challenges, such as the vanishing gradient problem, especially when processing lengthy data sequences. LSTMs address this limitation by introducing a memory cell and sophisticated gating mechanisms (Kong et al. 2025). The model is structured around four key elements: the input gate, forget gate, memory cell, and output gate. These components collaboratively manage and update information over time, enabling LSTMs to effectively capture and utilize long-term dependencies in sequential data.

The input gate is responsible for deciding which information is stored in the memory cell and identifies the new details to retain in the current state. The forget gate, on the other hand, determines which data should be removed from the memory cell by assigning a value of either 0 or 1 to the previous state, where 1 signifies retention and 0 represents deletion. Lastly, the output gate regulates the flow of information from the memory cell, selecting specific elements from the current state to be output. The effectiveness of the input gate can be evaluated using the specified mathematical equation (Yan et al. 2023).

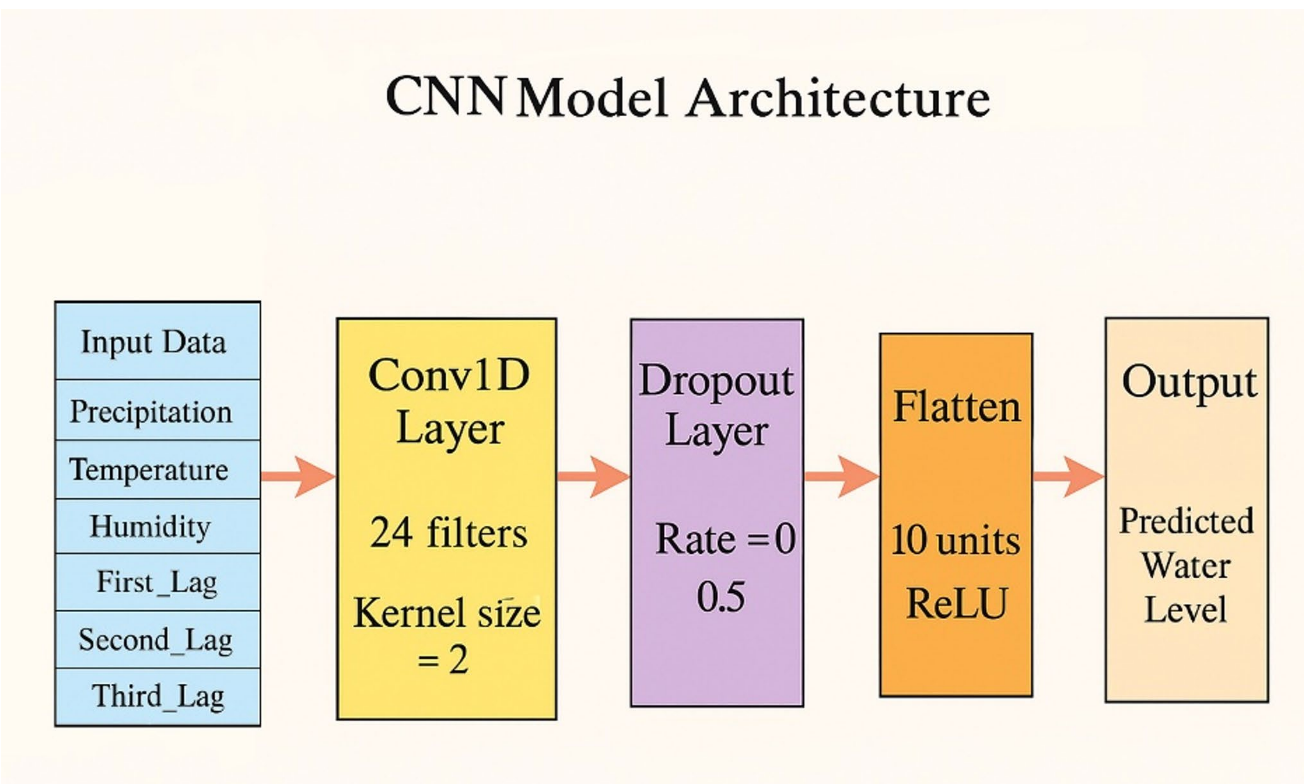

Fig. 4 The structure of the CNN model used in this study

$$q_s = \sigma\left(W_{q[h_s-1,x_s]} + b_s\right) \quad (3)$$

$$v_s = tanh(W_{v[h_s-1,x_s]} + b_v) \quad (4)$$

When considering time s within this process are variables including input gate activation vector q at that moment, which leads to weight matrix W_q impacting previous hidden state h_{s-1} and current input x_s alongside bias vectors bs resulting in a new set of values for cell state stored in v_s followed by two additional variables: weight matrix W_v and bias vector b_v. To find the value of Forget Gate, this formula is used:

$$p_s = \sigma\left(W_{p[h_{s-1},x_s]} + b_p\right) \quad (5)$$

At time s, the prior hidden state is denoted as h_{s-1}, while W_p represents the weight matrix, and the input is denoted as x_s; additionally, the bias vector is b_p, and the forget gate activation vector is referred to as p. The output gate can be mathematically described by the equation below (Yan et al. 2023).

$$f_s = tanh(W_{f[h_s-1,x_s]} + b_f) \quad (6)$$

$$h_s = f_s \times tanh\left(v_s\right) \quad (7)$$

where f_s is the output gate activation and h_s is the new hidden state.

The CNN-LSTM model is a powerful deep learning architecture widely used in time series analysis across domains like finance, marketing, and healthcare. It integrates Convolutional Neural Networks (CNN) for extracting spatial features and Long Short-Term Memory (LSTM) units for modeling temporal dependencies (Widiputra et al. 2021). CNN layers identify multi-scale patterns in the data through convolutional filters, while LSTM layers capture long-term relationships and sequential dynamics. This hybrid approach enables the model to learn both spatial and temporal patterns effectively, making it well-suited for tasks such as forecasting and trend prediction—like estimating river water levels based on historical data (Ali et al. 2025). The model structure used to predict the river water level is shown in Fig. 5.

To model the temporal dependencies and multivariate relationships in the time series data, a sequential deep learning architecture combining Conv1D and LSTM layers was employed. A one-dimensional convolutional layer with 64 filters and a kernel size of 1 is applied as the first layer to extract local temporal patterns and feature-level interactions. This is followed by a MaxPooling1D layer with a pool size of 1, which retains the sequence structure while reducing sensitivity to minor fluctuations. The output of the convolutional layer is then passed to a unidirectional LSTM

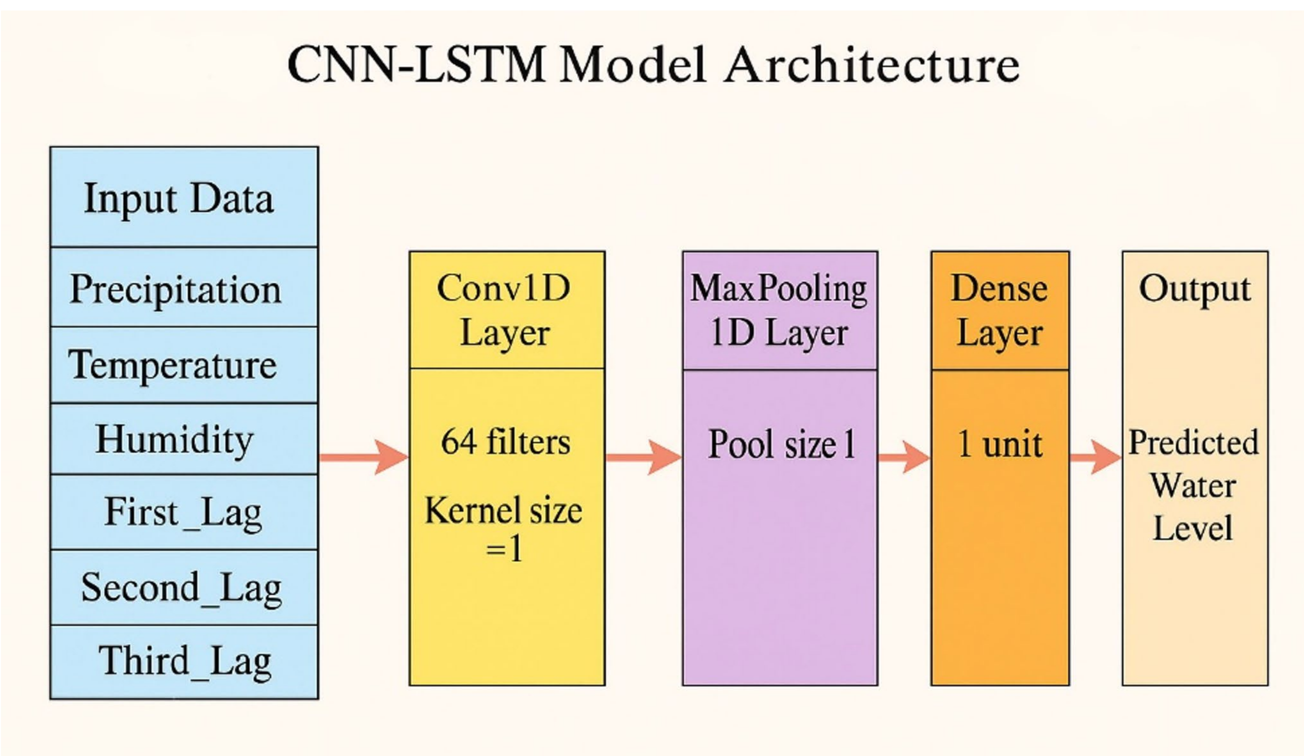

Fig. 5 The structure of the CNN-LSTM model used in this study

layer with 14 units and ReLU activation, enabling the model to capture sequential dependencies and non-linear dynamics over time. A final Dense layer with a single neuron produces the predicted output value. The model is compiled using the Adam optimizer, with a custom loss function tailored to the study's specific requirements. Root Mean Squared Error (RMSE) is used as the evaluation metric, providing an interpretable measure of prediction accuracy. This architecture is designed to balance simplicity and effectiveness in modeling short- to medium-range temporal dependencies.

The Bidirectional Long Short-Term Memory (BiLSTM) architecture employs dual recurrent pathways - forward and backward LSTM chains - that jointly analyze temporal sequences from both chronological directions. This bidirectional processing enables comprehensive context capture, where hidden states integrate information from preceding and subsequent time steps simultaneously, particularly valuable for non-causal prediction tasks. The structure of a BiLSTM model consists of several key components, including input, hidden units, output, and memory cells (Li et al. 2024). The input layer receives the sequential data points, which are then passed through the hidden units where the LSTM cells process and update the information. These memory cells store and update information from previous time steps, allowing the model to maintain context and capture long-term dependencies in the data. The output layer uses the final hidden state to make predictions about future data points in the time series.

In time series forecasting, optimizing a BiLSTM model involves careful tuning of key hyperparameters such as the number of LSTM units, network depth, hidden layer size, dropout rate, and learning rate, all of which influence the model's capacity and generalization (Kim et al. 2023). Techniques like grid search or random search are commonly employed to identify the best configuration by evaluating performance on a validation set while considering computational constraints and the trade-off between complexity and accuracy. A more advanced approach integrates convolutional neural networks (CNNs) with BiLSTM layers, forming a CNN-BiLSTM architecture that captures both local spatial features and long-range temporal dependencies (Staffini 2023). CNN layers extract meaningful patterns from raw time series data, which are then processed bidirectionally by BiLSTM to model contextual relationships across time. This hybrid structure enhances predictive accuracy and robustness, making it highly effective for tasks such as forecasting and classification. The model structure used to predict the river water level is shown in Fig. 6.

Key hyperparameters for a CNN-BiLSTM model in time series tasks include the number and size of convolutional filters, kernel size, and pooling size in the CNN layers; the number of units (hidden size) in the BiLSTM layer; the number of BiLSTM layers; learning rate; batch size; dropout rates; and the number and size of dense (fully connected) layers after BiLSTM. Additional relevant hyperparameters can involve activation functions (commonly ReLU for CNN), optimizer type (such as Adam), sequence length (window size), and the use of regularization techniques like dropout or batch normalization.

A hybrid deep learning model combining a one-dimensional Convolutional Neural Network (Conv1D) and a Bidirectional Long Short-Term Memory (BiLSTM) network was developed to predict the target variable based on multivariate time series inputs. The model begins with an input layer

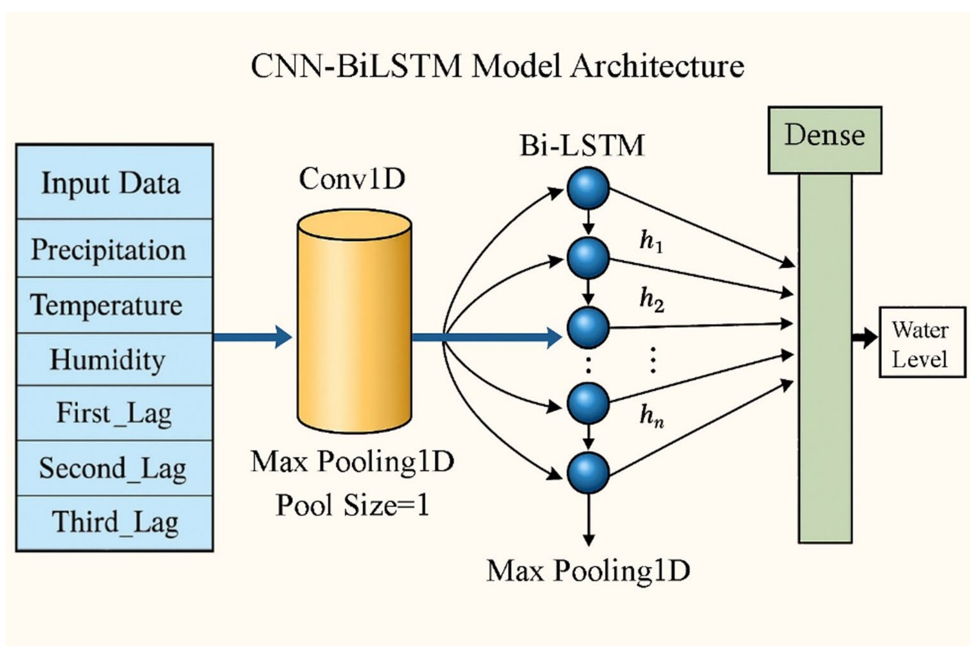

Fig. 6 The structure of the CNN-BiLSTM model used in this study

that accepts data in the shape of (window size, 6), where each time step contains six features. A Conv1D layer with 64 filters and a kernel size of 1 is applied to extract local temporal features, followed by a MaxPooling1D layer with a pool size of 1 to maintain the temporal resolution. The extracted features are then passed to a Bidirectional LSTM layer with 64 units, which captures both past and future dependencies in the sequence. The final output of the LSTM is fed into a Dense layer with a single neuron to generate the predicted value. The model is compiled using the Adam optimizer and trained with the Mean Squared Error (MSE) loss function, making it suitable for continuous variable prediction tasks. This architecture leverages both convolutional feature extraction and recurrent temporal modeling to improve predictive performance. Table 2 shows the hyperparameters used to model DNN, CNN, CNN-LSTM, and CNN-BiLSTM models.

In this study, the selection of hyperparameters was performed using the Random Search optimization method. Unlike Grid Search, which methodically tests every combination of parameters, Random Search randomly samples from the specified parameter space. This approach is more efficient, especially in high-dimensional cases where exhaustive search becomes computationally prohibitive. To ensure the robustness of the findings, the training process was repeated independently 10 times. Model complexity refers to the capacity of a model to capture a wide range of functions. Although complex models such as deep neural networks have strong representational power, they are prone to overfitting, meaning they may capture noise instead of true underlying patterns. By adjusting the model's architecture, size, and number of parameters, this research balanced the trade-off between underfitting and overfitting. The model performance was assessed using 5-fold cross-validation, with mean squared error (MSE) employed as the metric for validation. The relatively high learning rates were determined through a sensitivity analysis and stabilized by batch normalization, feature scaling, and the Adam optimizer, ensuring efficient convergence without gradient instability.

Table 2 Hyperparameters are used to model DNN, CNN, CNN-LSTM, and CNN-BiLSTM models

Type of parameters	Values/Layer
Network Type	Feed-forward propagation
Data Division	Train (70%) Test (30%)
Number of Hidden layers (Neurons)	3–4
Batch Size	32
learning Function	0.1-2.0
Activation Function	Relu
Normalization Function	Batch Normalization
Training function	Adam

Model evaluation metrics

One of the main reasons for employing evaluation criteria in comparative modeling is their ability to create a unified structure for reviewing different methodologies. Without clear reference points, making impartial comparisons and sound decisions becomes challenging. Evaluation criteria provide a formalized approach to examining models based on defined attributes, promoting a more organized, reliable, and open evaluation process.

Root Mean Square Error (RMSE) is a key indicator used across statistics and machine learning to measure how accurately a model predicts outcomes. It reflects the typical difference between predicted results and actual observations in a dataset. As a widely recognized tool for evaluating model performance, RMSE supports unbiased comparisons by highlighting models that demonstrate stronger predictive consistency (Chicco et al. 2021). Generally, a lower RMSE suggests dependable and accurate forecasts, while a higher value may point to diminished precision and the potential need for model adjustments.

In the realm of statistical modeling and data interpretation, the coefficient of determination (R^2) is a pivotal metric for gauging the effectiveness and credibility of regression models. It expresses the extent to which variation in the dependent variable can be explained by the independent variables, thereby serving as an indicator of model adequacy (Saunders et al. 2012). As a standard measure of explanatory capability, a high R^2 value reflects a strong correlation between predicted and actual data points. In contrast, a low R^2 suggests that the model may not sufficiently capture the data's inherent structure, signaling the potential need for improvement or alternative modeling strategies.

The Nash-Sutcliffe efficiency coefficient (NSE) is widely utilized in hydrology and environmental disciplines to evaluate the predictive capability of models. Its appeal lies in its simplicity—by calculating a single efficiency score, researchers can quickly gauge a model's overall accuracy (Gupta et al. 2009). This concise metric enables efficient and practical assessments, making it particularly valuable for applications in climate analysis, water resource planning, and environmental monitoring.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_{pi} - X_{oi}\right)^2} \tag{8}$$

$$R^2 = \frac{\sum_{i=1}^{N}\left(X_{oi} - \bar{X}_o\right)\left(X_{pi} - \bar{X}_p\right)}{\sum_{i=1}^{N}\left(X_{oi} - \bar{X}_o\right)^2 . \sum_{i=1}^{N}\left(X_{pi} - \bar{X}_p\right)^2} \tag{9}$$

$$NSE = 1 - \left[\frac{\sum_{i=1}^{n}(X_{oi} - X_{pi})^2}{\sum_{i=1}^{n}(X_{oi} - \bar{X}_o)^2}\right] \quad (10)$$

where X_{pi} and X_{oi} are the predicted and observed values, $\bar{X}_o$ and $\bar{X}_p$ are the mean observed and predicted values, respectively, and N is the total number of data points.

Experimental results and discussion

Model evaluation is a critical aspect of the data analysis process as it helps ensure the reliability and effectiveness of the predictive models being developed. By evaluating a model, researchers can determine its accuracy and performance in predicting outcomes, which is essential for making informed decisions and drawing meaningful conclusions from the data. Without proper evaluation, there is a risk of developing models that are inaccurate or biased, leading to incorrect interpretations and decisions based on faulty predictions. This research utilized three evaluation metrics—correlation coefficient (r), Nash-Sutcliffe efficiency (NSE), and root mean square error (RMSE)—to analyze and compare the performance of the models. The correlation coefficient is a statistical measure that quantifies the strength and direction of a linear relationship between two variables (Wang et al. 2022a, b). It ranges from −1 to +1, where a value of +1 indicates a perfect positive relationship, −1 indicates a perfect negative relationship, and 0 indicates no relationship. In other words, the correlation coefficient tells us how closely the two variables are related; the closer r is to either −1 or 1, the stronger the relationship between the variables.

Nash-Sutcliffe efficiency is a widely used metric in hydrology and water resources engineering to evaluate the performance of hydrological models. It measures the model's ability to predict observed values of a variable, such as streamflow or groundwater Levels. The NSE value ranges from negative infinity to 1, with 1 indicating a perfect match between predicted and observed values, while a negative value indicates that the model performs worse than simply using the mean of the observed values (Melsen et al. 2025).

Table 3 Evaluation parameters of the studied models

River	Model	RMSE(m)	R	NSE
Danube	CNN-BiLSTM	0.47	0.94	0.88
	CNN-LSTM	1.14	0.86	0.67
	CNN	1.17	0.83	0.65
	DNN	2.82	0.82	0.36
Mississippi	CNN-BiLSTM	0.93	0.96	0.95
	CNN-LSTM	1.58	0.93	0.74
	CNN	1.69	0.92	0.73
	DNN	3.32	0.80	0.61

NSE is a popular metric because it considers both the bias and variability of the model predictions, providing a comprehensive assessment of model performance.

Root means square error is a widely used metric in the field of statistics and machine learning to evaluate the accuracy of a predictive model. It is used to measure the difference between the values predicted by the model and the actual values observed in the data. RMSE provides a single number to represent the overall performance of the model, with lower values indicating better accuracy (Hodson 2022). The formula for calculating RMSE involves taking the square root of the average of the squared differences between the predicted and actual values.

Table 3 outlines the assessment criteria employed for evaluating the four machine-learning models. At the Danube River Station, the DNN model exhibited poor performance, with r=0.82, NSE=0.36, and RMSE=2.82, falling short compared to the other models. The CNN and CNN-LSTM models demonstrated satisfactory performance, achieving correlation coefficients of 0.83 and 0.86, Nash-Sutcliffe coefficients of 0.65 and 0.67, and mean square errors of 1.17 and 1.14, respectively. Meanwhile, the CNN-BiLSTM model stood out with excellent performance and precision, reflected by r=0.94, NSE=0.88, and RMSE=0.47. At the Mississippi River station, the DNN model performed poorly, as did the Danube River (r=0.80, NSE=0.61, and RMSE=3.32). The CNN and CNN-LSTM models were more accurate than the DNN model with correlation coefficients of 0.92 and 0.93, Nash-Sutcliffe coefficients of 0.73 and 0.74, and mean square errors of 1.69 and 1.58, respectively. The CNN-BiLSTM model was more accurate in predicting monthly water levels with r=0.95, NSE=0.96, and RMSE=0.93.

From a hydrological perspective, the predictive errors have different implications for operational use. On the Mississippi River, where major flood stages typically occur at Levels 3–5 m above normal, an RMSE of 0.93 m represents roughly 18–30% of this threshold. While this Level of accuracy supports seasonal planning and medium-range forecasting, it may be insufficient for issuing short-lead flood warnings, where differences of 0.3–0.5 m can alter alert status. Conversely, for the Danube, an RMSE of 0.47 m suggests a stronger operational potential, particularly in managing early warning systems.

Figure 7 presents the time series and scatter plots for the two study stations along the Danube and Mississippi rivers, highlighting the impressive performance of CNN-BiLSTM. The DNN model exhibits significant error at the Danube River station, displaying a time series graph that deviates noticeably from observed values. Additionally, its scatter plot reveals greater dispersion. Both the CNN and CNN-LSTM models produce time series that closely align with

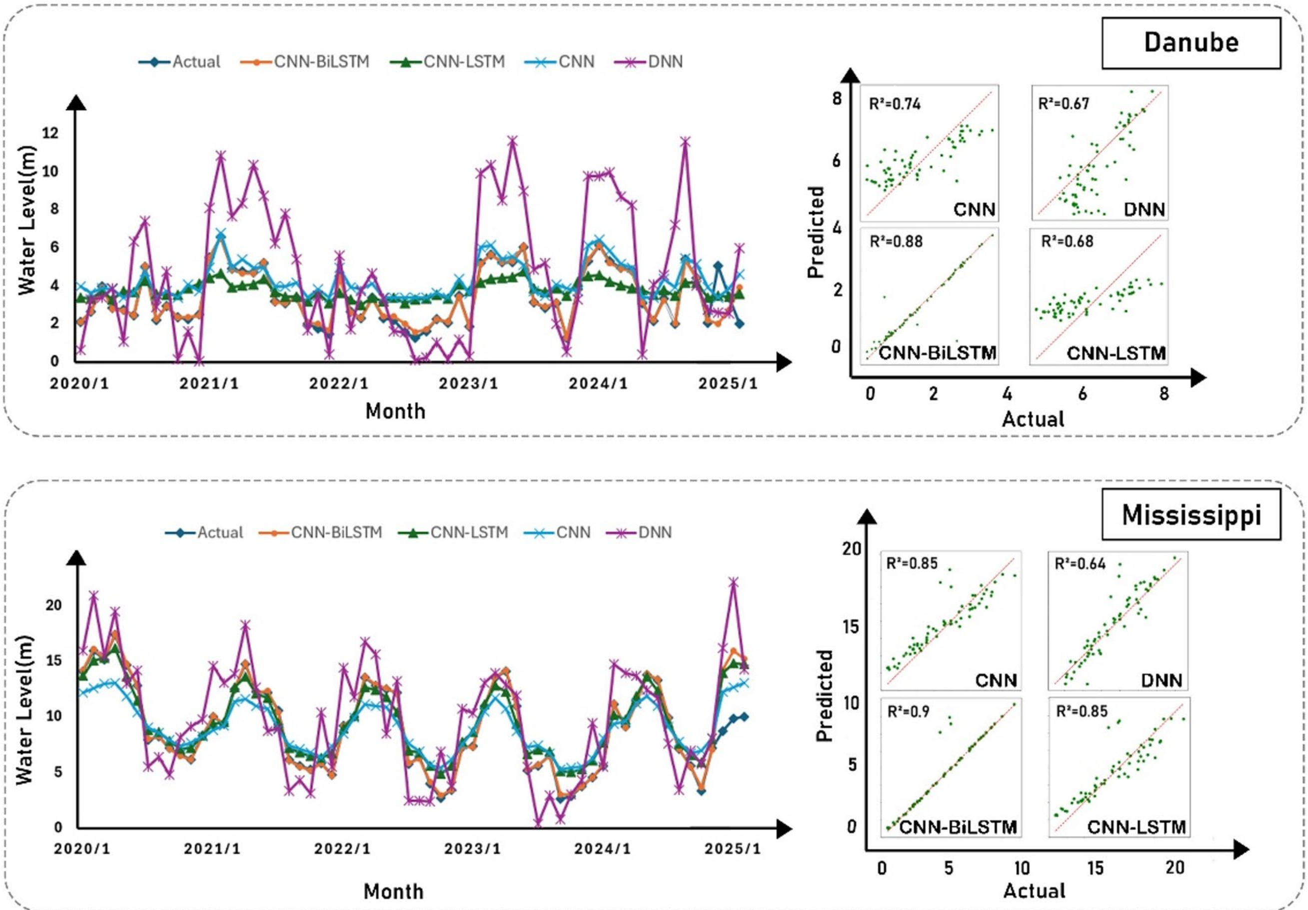

Fig. 7 Time series and scatter plot for two Danube and Mississippi rivers stations

the actual values. However, they encounter notable errors when estimating the maximum and minimum points. On the other hand, the CNN-BiLSTM model with less dispersion has a time series graph similar to the actual values and has less error than the other three models. At the Mississippi station, the scatter plot for the DNN model shows greater dispersion and its time series deviates significantly from the observed values. In comparison, the CNN and CNN-LSTM models exhibit reduced scatter and improved time series plots. Ultimately, the CNN-BiLSTM hybrid model delivers highly accurate and dependable predictions for river water levels.

Although formal statistical significance tests were not conducted, the differences in model performance appear robust and consistent across both river datasets and multiple evaluation metrics. CNN-BiLSTM consistently achieved lower RMSE, higher Pearson correlation, and higher NSE compared to CNN-LSTM, CNN, and DNN throughout the 17-year validation period. Visual inspection of predicted versus observed water levels further highlights its superior ability to capture seasonal peaks and extreme events, whereas other models systematically under- or over-predicted during these periods. The consistency of these patterns across metrics, periods, and hydrological conditions strongly suggests that the improved performance of CNN-BiLSTM is unlikely to result from random variation or model noise.

The CNN-BiLSTM model performs well for both rivers, but differences in performance are observed: errors are slightly higher for the Danube than for the Mississippi. This can be attributed to the Danube's more variable hydrological regime and complex flow patterns influenced by multiple tributaries and seasonal precipitation variability, whereas the Mississippi exhibits more gradual and predictable seasonal flows. Examination of extreme events reveals that while the model captures general trends, predicting flood peaks and very low-flow periods is more challenging, Likely due to their rarity in the 17-year dataset.

The CNN-BiLSTM model achieves higher accuracy compared to the DNN, CNN, and CNN-LSTM models because it effectively combines the strength of convolutional neural networks in extracting local spatial features with the bidirectional LSTM's ability to capture long-range temporal dependencies in both forward and backward directions (Wang 2024). This dual approach enables the model

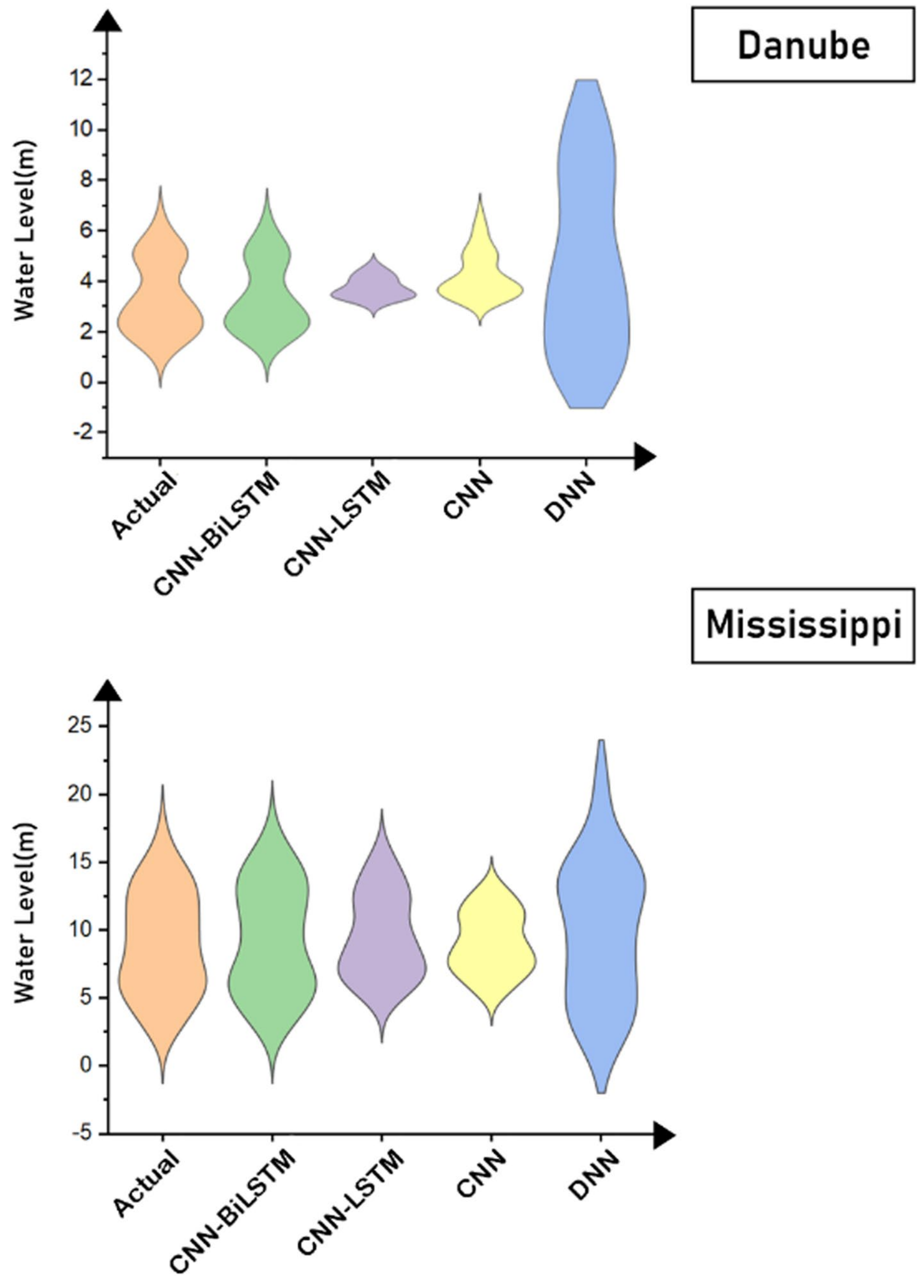

Fig. 8 Violin plot for two Danube and Mississippi rivers stations

to learn richer and more comprehensive feature representations, especially for sequential or spatiotemporal data, where understanding the context from the entire sequence is crucial. In contrast, DNNs lack spatial or temporal feature extraction, CNNs focus mainly on spatial patterns without strong temporal modeling, and CNN-LSTMs model temporal dependencies only in one direction. By integrating CNN with BiLSTM, the model leverages both detailed spatial feature extraction and enhanced bidirectional temporal context, resulting in consistently superior performance and higher accuracy across various tasks such as time series prediction, fault diagnosis, and classification problems (Seoni et al. 2024).

Figure 8 illustrates the violin plots representing the performance of the four models across the two stations, Danube and Mississippi. In the violin plot for the Danube station, the DNN, CNN, and CNN-LSTM models exhibit lower performance, with their violin diagrams showing reduced alignment with the actual values. In contrast, the CNN-BiLSTM model demonstrates reliable accuracy, with its violin plot closely resembling the actual values. At the Mississippi River station, the CNN and CNN-LSTM models outperform the DNN model. Overall, the findings highlight the superior capability of the CNN-BiLSTM model in forecasting monthly water levels for the two studied rivers.

For the Danube River, the CNN-BiLSTM and CNN-LSTM models produce prediction distributions closely aligned with the actual water level data, indicating their superior capacity to capture temporal dependencies and nonlinear hydrological dynamics. The CNN model exhibits a more constrained variance, implying potential underfitting or insufficient model complexity, while the DNN demonstrates a broader distribution with an overestimation bias, suggesting limited generalization for this dataset. In the case of the Mississippi river, a similar trend is observed; hybrid models (CNN-BiLSTM and CNN-LSTM) maintain alignment with observed data distributions, whereas CNN predictions show reduced variance, and the DNN overestimates water levels with high variability.

The predictive performance of the models varied across seasonal cycles, reflecting differences in their ability to capture hydrological seasonality. All models successfully reproduced the general annual trend driven by precipitation and temperature fluctuations; however, CNN-BiLSTM exhibited closer alignment with observed seasonal peaks, likely due to its bidirectional memory structure, which captures both preceding and succeeding temporal dependencies. CNN-LSTM and standalone CNN models, in contrast, tended to underestimate sharp transitions between wet and dry months, indicating limitations in handling long-range dependencies. Although CNN-LSTM showed satisfactory accuracy in general, its effectiveness declined during times of peak flow. This drawback likely results from its unidirectional approach to temporal processing, limiting its capacity to recognize long-term dependencies and anticipate future information. On the other hand, CNN-BiLSTM utilizes bidirectional memory, allowing it to more effectively account for temporal influences from both preceding and succeeding data points, which is essential for accurately identifying peaks.

Model behavior during extreme events such as peak floods or low-flow conditions revealed additional challenges. While CNN-BiLSTM achieved lower error rates than other models during high-flow periods, its performance still degraded when predicting rare, high-magnitude anomalies, suggesting that limited representation of extreme cases in the training dataset constrained learning. Enhancing prediction accuracy under such conditions may require hybrid approaches that integrate physical hydrological knowledge or synthetic data augmentation to better represent extremes. Another critical observation was the presence of prediction lag, particularly in DNN and CNN models, which often responded with a delay to abrupt hydrological changes. This lag likely stems from their reliance on short-term dependencies and limited sequence modeling capability. In contrast, the hybrid architectures—especially CNN-BiLSTM—displayed a reduced lag effect, allowing for more timely adjustments in water level predictions. Despite these improvements, further refinement is necessary to fully eliminate lag under rapid environmental shifts.

The superior accuracy of CNN-BiLSTM models over DNN, CNN, and CNN-LSTM models is consistently demonstrated across diverse applications, such as fault diagnosis, financial prediction, network intrusion detection, and time series. This advantage arises from the hybrid architecture: CNN layers excel at extracting local or spatial features from raw data, while BiLSTM layers capture long-range dependencies and contextual information in both forward and backward directions. In different fields, CNN-BiLSTM models have achieved high accuracy rates as high as, significantly outperforming CNN and LSTM models individually (Rhanoui et al. 2019; Wang 2024; Li et al. 2024).

Some of the studies in which the combination of CNN and BiLSTM models showed high performance in water level prediction include: Nie et al. (2021) presented a novel water level prediction approach tailored for the stochastic and complex nature of hydrological time series data. By integrating Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and an attention mechanism, the proposed model effectively captures both spatial and temporal features while emphasizing key historical patterns. Applied to hourly water level data from the Pinghe Basin in China, the method demonstrated superior accuracy compared to traditional models such as SVM, TCN, and standalone BiLSTM. Hu et al. (2023) introduced a CNN-BiLSTM-Difference Analysis (CNN-BiLSTM-DA) model for accurate water level prediction and real-time flood warning. Applied to Yancheng city in Jiangsu Province, China, the CNN-BiLSTM component achieves over 95% prediction accuracy, while the integrated DA model delivers flood warning accuracy of 85.71% with zero missed alerts. Li et al. (2024) addressed the challenges of accurately predicting sea level height due to the nonlinear and uncertain nature of sea level data by a novel model, SCSSA-CNN-BiLSTM, which combines the sparrow search algorithm (SSA), enhanced by sine–cosine and Cauchy variation strategies (SCSSA), with a convolutional neural network and bidirectional long short-term memory network (CNN-BiLSTM). Another article focused on predicting water quality using a CNN-BiLSTM model. By combining convolutional layers (for spatial feature extraction) and BiLSTM layers (for temporal dependencies), the model achieved a validation accuracy of 92.83%. This outperformed other deep learning models like GRU, LSTM, and CNN-LSTM, highlighting its effectiveness for water quality assessment and management (Pawar et al. 2024). Liao et al. (2024) proposed a high-precision inland water level prediction model that integrates Variational Mode Decomposition (VMD), CNN, and BiLSTM. VMD effectively decomposes complex non-stationary signals and minimizes observational errors, while the CNN-BiLSTM architecture enhances spatiotemporal analysis and captures long-term dependencies. Experimental results show that the combined model excels in noise resistance, prediction accuracy, and detecting abrupt changes.

On the other hand, in addition to the fact that the hybrid model used in water level prediction is highly accurate, it is superior to other models in other hydrological fields. Wang et al. (2022a, b) established a two-dimensional (2D) significant wave height (SWH) prediction model for the South and East China Seas, trained on Wave Watch III (WW3) reanalysis data using CNN-BiLSTM-Attention—a combination of CNN, BiLSTM, and attention mechanism. Under normal conditions, for forecasting up to 24 h, the model achieves high correlation coefficients (0.996 to 0.945) and low root mean square errors (0.063 m to 0.281 m). Similarly, under extreme typhoon conditions, it demonstrates robust performance with correlation coefficients (0.993 to 0.921) and averaged errors (0.159 m to 0.555 m). Li et al. (2023) introduced a CNN-BiLSTM-Attention model for forecasting surface water quality parameters (temperature, pH, dissolved oxygen, ammonia nitrogen, conductivity, turbidity). The model, evaluated against LSTM and CNN-LSTM baselines, showed superior performance in RMSE, MAE, and R^2 metrics. For example, in dissolved oxygen prediction, the hybrid model achieved an R^2 of 0.95801, with significant improvements in error reduction and fit compared to other models. Xiao et al. (2025) presented a novel evaluation and prediction model for water resource carrying capacity (WRCC) and water quality in the Min River Basin. Key water quality factors were identified with Adaptive Lasso Regression, ranked by CatBoost, and predicted using a CNN-BiLSTM-Attention hybrid model. The hybrid model outperformed others, with minimal MAPE (0.0326) and SMAPE (0.0103) values across monitoring points, demonstrating superior accuracy and efficiency for water quality forecasting.

Comparatively, the CNN-BiLSTM model outperforms traditional models such as LSTM and CNN-LSTM in water level prediction tasks. For example, when evaluated against LSTM and CNN-LSTM, the CNN-BiLSTM model demonstrated reductions in key error metrics (MAE, MAPE, MAD, RMSE) by up to 38% compared to LSTM and by about 9% compared to CNN-LSTM, reflecting its superior predictive capabilities (Hu et al. 2023). Additionally, in flood warning applications, the CNN-BiLSTM-based approach achieved an alarm accuracy rate of 85.71% without missing any warning events, further highlighting its practical advantage in real-world scenarios. These results indicate that the CNN-BiLSTM model not only enhances prediction accuracy but also improves the reliability and timeliness of flood warnings compared to other commonly used models (Hu et al. 2023; Ahmed et al. 2024).

Accurate monthly water level predictions have direct implications for water resource management. For instance, improved forecasting supports reservoir operation by enabling more efficient allocation of water for hydropower generation, irrigation, and ecological flow maintenance. Reliable forecasts can also enhance early warning systems for floods, allowing managers to initiate preemptive measures such as controlled releases to reduce flood risk. Similarly, anticipating prolonged low-flow conditions aids in planning for drought mitigation, including optimizing water storage and prioritizing critical uses. These applications underscore the potential of deep learning–based approaches as decision-support tools under increasing climate variability.

Conclusions

Predicting river water levels is crucial for flood prevention, enabling early warnings that protect lives and property, and for effective water resource management to support agriculture, industry, and urban needs. It ensures safe navigation and maintenance of infrastructure like bridges and dams, while also helping preserve aquatic ecosystems by maintaining adequate flows. Additionally, accurate forecasts enhance disaster preparedness and risk mitigation, allowing timely emergency responses and informed long-term planning. Overall, river water level prediction plays a vital role in safeguarding communities, supporting economic activities, and promoting environmental sustainability.

The improved prediction accuracy of CNN-BiLSTM models is crucial for practical applications such as flood forecasting, early warning systems, and water resource management. Accurate river level predictions enable timely alerts, reducing the risk of flood damage and supporting better decision-making for irrigation and ecosystem sustainability. Additionally, the adaptability of CNN-BiLSTM models to different datasets and their potential for optimization make them powerful tools for addressing the challenges posed by changing climatic and environmental conditions in river basin management. This study explores a comparison between two standalone and blended predictive models aimed at forecasting the water levels of two major rivers: the Danube and the Mississippi. And utilized accessible parameters alongside four advanced AI architectures—DNN, CNN, CNN-LSTM, and CNN-BiLSTM. The approach emphasizes combining methodologies to achieve more accurate hydrological predictions. The findings highlight that the hybrid CNN-BiLSTM model outperforms others in forecasting monthly water levels, achieving impressive accuracies—94.47% for the Danube and 95.78% for the Mississippi. This demonstrates the potential of combining Convolutional Neural Networks with Bidirectional LSTMs for accurate hydrological predictions across different geographic locations.

The operational adequacy of model predictions depends on their ability to provide actionable Lead times for flood management. For the Mississippi, an RMSE of 0.93 m indicates substantial improvement over baseline models, but its acceptability depends on the flood stage thresholds, which often range between 2 and 5 m above normal Levels. While such an error margin may not critically hinder seasonal planning, it could pose challenges for real-time flood alerts where early detection of rapid stage increases is essential. In contrast, the 0.47 m RMSE achieved for the Danube suggests higher reliability for operational forecasting. These findings highlight that although deep learning models can significantly enhance predictive accuracy, integrating them with physical models or probabilistic frameworks is necessary for operational flood warning systems.

Predicting river water levels using CNN-BiLSTM models faces several challenges, including the complexity of tuning numerous parameters and the need for extensive, high-quality datasets that capture all relevant influencing factors. These models can struggle with noisy and non-stationary data, often requiring preprocessing to improve feature extraction. Additionally, CNN-BiLSTM architectures may have difficulty accurately forecasting extreme or unprecedented water level events due to their reliance on historical patterns. Ambiguities in model design and insufficient validation can Limit generalizability and practical application across different river systems. Overall, while powerful, CNN-BiLSTM models require careful data preparation, architecture optimization, and thorough testing to overcome these Limitations and achieve reliable predictions. The dataset included roughly 204 monthly observations per station, which is relatively limited for effectively training deep neural networks. Despite using regularization methods such as dropout and early stopping, overfitting cannot be entirely dismissed. Future studies should consider transfer learning techniques utilizing larger hydrological datasets or apply data augmentation to enhance model robustness. Although the proposed framework demonstrated strong performance on two major rivers, its applicability to ungauged or data-scarce basins remains unclear. These basins typically suffer from insufficient climatic and hydrological data, potentially impacting model accuracy. Subsequent research could explore hybrid strategies that integrate physical modeling with data-driven approaches or focus on building models that make use of global datasets to boost transferability.

Future work on river water level prediction using CNN-BiLSTM models should focus on several key areas: although the CNN-BiLSTM model achieved strong predictive results, its output was somewhat influenced by parameter configurations and exhibited sporadic inaccuracies during periods of extreme water level variation. To enhance its ability to manage nonlinear patterns and data noise, incorporating sophisticated decomposition approaches such as Ensemble Empirical Mode Decomposition (EEMD) could be beneficial. Additionally, investigating time-frequency analysis techniques may offer improved feature extraction and greater model resilience. Applying the models within operational contexts—such as flood early warning systems or reservoir management—would enhance their practical value and support real-time water resource decision-making. There is also potential in applying optimization algorithms and their variants to enhance parameter tuning and convergence speed, addressing current limitations in model training efficiency and accuracy. Incorporating additional data sources such as climate indices, rainfall, and upstream alongside water level data can improve the model's ability to capture complex dependencies and predict extreme events.

To enhance broader hydrologic applicability, future research could incorporate stage-discharge rating curves or hydrodynamic modeling frameworks that translate predicted water levels into discharge estimates. This integration would leverage the spatial breadth and consistency of satellite-derived water levels while addressing the site-specific challenges posed by channel morphology. On the other hand, missing values were imputed using arithmetic means and linear interpolation for simplicity and computational efficiency. However, this approach may not preserve important temporal characteristics of hydrological time series, such as seasonality and autocorrelation. Advanced techniques like Expectation–Maximization or Kalman filtering could better account for these dependencies and will be considered in future research to enhance data integrity. Although the 17-year dataset provides extensive temporal coverage for both rivers, it may not fully capture rare extreme events such as exceptional floods or prolonged droughts. Consequently, while the models perform well under typical conditions, their predictive reliability for very rare extremes may be limited. Future work incorporating longer time series or additional extreme event data could improve model robustness under such conditions.

Author contributions O.R.I. Writing – review & editing, Writing—original draft, Visualization, Investigation, Formal analysis, Conceptualization. A.V. Writing—review & editing, Software, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. S.A. Writing—review & editing. E.A. Supervision, Resources, Data curation, Validation, Methodology. M.S. and S.M.J.M. Writing—review & editing. All authors reviewed the manuscript.

Funding The authors did not receive support from any organization

for the submitted work.

Data availability Data and codes used for the case study are available on request.

Declarations

Competing interests The authors declare no competing interests.

Conflict of interest The authors declare no competing interests.

Financial interests The authors declare they have no financial interests.

Ethics approval All authors have read, understood, and have complied as applicable with the statement on "Ethical responsibilities of Authors" as found in the Instructions for Authors.

References

Abdi E, Ali M, Santos CAG, Olusola A, Ghorbani MA (2024) Enhancing groundwater level prediction accuracy using interpolation techniques in deep learning models. Groundwater for Sustainable Development 26:101213

Ahmed AM, Akther S, Nguyen-Huy T, Raj N, Jui SJJ, Farzana SZ (2024) Real-time prediction of the week-ahead flood index using hybrid deep learning algorithms with synoptic climate mode indices. J Hydro-Environ Res 57:12–26

Ali M, Nayahi JV, Abdi E, Ghorbani MA, Mohajeri F, Farooque AA, Alamery S (2025) Improving daily reference evapotranspiration forecasts: designing AI-enabled recurrent neural networks based long short-term memory. Ecol Informatics 85:102995

Atashi V, Gorji HT, Shahabi SM, Kardan R, Lim YH (2022) Water level forecasting using deep learning time-series analysis: a case study of Red River of the North. Water 14(12):1971

Barzegar R, Aalami MT, Adamowski J (2021) Coupling a hybrid CNN-LSTM deep learning model with a boundary corrected maximal overlap discrete wavelet transform for multiscale lake water level forecasting. J Hydrol 598:126196

Chicco D, Warrens MJ, Jurman G (2021) The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci 7:e623. https://doi.org/10.7717/peerj-cs.623

Dong Z, Hu H, Liu H, Baiyin B, Mu X, Wen J, Li X (2024) Superior performance of hybrid model in ungauged basins for real-time hourly water level forecasting–a case study on the Lancang-Mekong mainstream. J Hydrol 633:130941

Duan Y, Wu J (2021) Joint optimization of DNN partition and scheduling for mobile cloud computing. In *Proceedings of the 50th International Conference on Parallel Processing* (pp. 1–10)

Gupta HV, Kling H, Yilmaz KK, Martinez GF (2009) Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J Hydrol 377(1–2):80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hiatt M, Snedden G, Day JW, Rohli RV, Nyman JA, Lane R, Sharp LA (2019) Drivers and impacts of water level fluctuations in the Mississippi River delta: implications for delta restoration. Estuar Coast Shelf Sci 224:117–137

Hodson TO (2022) Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. Geosci Model Dev Discuss 2022:1–10

Hu Y, Li D, Deng J, Yue Y, Zhou J, Chai Y, Li Y (2022) Mechanisms controlling water-level variations in the middle Yangtze river following the operation of the three Gorges dam. Water Resour Res 58(10):e2022WR032338

Hu C, Zhou L, Gong Y, Li Y, Deng S (2023) Research on water level anomaly data alarm based on CNN-BILSTM-DA model. Water 15(9):1659

Isik F, Ozden G, Kuntalp M (2012) Importance of data preprocessing for neural networks modeling: the case of estimating the compaction parameters of soils. Energy Educ Sci Technol Part A: Energy Sci Res 29:463–474

Kim D, Han H, Wang W, Kim HS (2022) Improvement of deep learning models for river water level prediction using complex network method. Water 14(3):466

Kim J, Oh S, Kim H, Choi W (2023) Tutorial on time series prediction using 1D-CNN and bilstm: a case example of peak electricity demand and system marginal price prediction. Eng Appl Artif Intell 126:106817

Kong Y, Wang Z, Nie Y, Zhou T, Zohren S, Liang Y, Wen Q (2025) Unlocking the power of lstm for long term time series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence 39(11):11968–11976

Le XH, Jung S, Yeon M, Lee G (2021) River water level prediction based on deep learning: case study on the Geum River, South Korea. In *Proceedings of the 3rd International Conference on Sustainability in Civil Engineering: ICSCE 2020, 26–27 November, Hanoi, Vietnam* (pp. 319–325). Springer Singapore

Li L, Gu Z, Lu D (2023) Research on Surface Water Quality Prediction based on a CNN-BiLSTM-Attention Combined Deep Learning Approach. In *E3S Web of Conferences* (Vol. 393, p. 01007). EDP Sciences

Li X, Zhou S, Wang F, Fu L (2024) An improved sparrow search algorithm and CNN-BiLSTM neural network for predicting sea level height. Sci Rep 14(1):4560

Liao F, Zhou P, Tang B, Zhou Z (2024), December The Prediction of Water Level Changes in a Coastal Inland River Using VMD and CNN-BiLSTM Model. In *2024 10th International Conference on Systems and Informatics (ICSAI)* (pp. 1–5). IEEE

Mănoiu VM, Crăciun AI (2021) Danube river water quality trends: a qualitative review based on the open access web of science database. Ecohydrol Hydrobiol 21(4):613–628

Melsen LA, Puy A, Torfs PJ, Saltelli A (2025) The rise of the Nash-Sutcliffe efficiency in hydrology. Hydrol Sci J 63:1–12

Nie Q, Wan D, Wang R (2021) CNN-BiLSTM water level prediction method with attention mechanism. In *Journal of Physics: Conference Series* (Vol. 2078, No. 1, p. 012032). IOP Publishing

Pawar P, Singhal N, Sahayaraj K (2024) Water quality prediction using deep learning. In *AIP Conference Proceedings* (Vol. 3075, No. 1). AIP Publishing

Qian Y, Fan Y, Hu W, Soong FK (2014) On the training aspects of deep neural network (DNN) for parametric TTS synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3829–3833). IEEE

Rhanoui M, Mikram M, Yousfi S, Barzali S (2019) A cnn-bilstm model for document-level sentiment analysis. Mach Learn Knowl Extr 1(3):832–847

Santos CAG, Ghorbani MA, Abdi E, Patel U, Sadeddin S, Models CNN (2024) Estimating water levels through smartphone-imaged gauges: a comparative analysis of ANN, DL, and CNN models. Water Resour Manag 63:1–16

Sarikulov MK, Bekimbetov BM (2024) Ecological problems of rivers at the present stage. Int J Sci Trends 3(12):281–287

Saunders LJ, Russell RA, Crabb DP (2012) The coefficient of determination: what determines a useful R2 statistic? Invest Ophthalmol Vis Sci 53(11):6830–6832. https://doi.org/10.1167/iovs.12-10598

Schwatke C, Dettmering D, Bosch W, Seitz F (2015) DAHITI–an innovative approach for estimating water level time series over inland waters using multi-mission satellite altimetry. Hydrol Earth Syst Sci 19(10):4345–4364

Seoni S, Molinari F, Acharya UR, Lih OS, Barua PD, García S, Salvi M (2024) Application of spatial uncertainty predictor in CNN-BiLSTM model using coronary artery disease ECG signals. Inf Sci 665:120383

Staffini A (2023) A CNN–BiLSTM architecture for macroeconomic time series forecasting. Engineering Proceedings 39(1):33

Swiderski B, Osowski S, Gwardys G, Kurek J, Slowinska M, Lugowska I (2022) Random CNN structure: tool to increase generalization ability in deep learning. EURASIP J Image Video Process 2022(1):3

Tadić L, Tamás EA, Mihaljević M, Janjić J (2022) Potential climate impacts of hydrological alterations and discharge variabilities of the Mura, Drava, and Danube rivers on the natural resources of the MDD UNESCO biosphere reserve. Climate 10(10):139

Wan Z, Chang Z, Xu Y, Šavija B (2023) Optimization of vascular structure of self-healing concrete using deep neural network (DNN). Constr Build Mater 364:129955

Wang X (2024) CNN-BiLSTM-Attention Algorithm-Based Stock Prices Prediction During COVID-19. In *2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)* (pp. 441–452). Atlantis Press

Wang Y, Zhang Z, Feng L, Ma Y, Du Q (2021) A new attention-based CNN approach for crop mapping using time series Sentinel-2 images. Comput Electron Agric 184:106090

Wang L, Deng X, Ge P, Dong C, Bethel BJ, Yang L, Xia J (2022a) CNN-BiLSTM-attention model in forecasting wave height over South-east China seas. Comput Mater Continua 73(1):2151–2168

Wang L, Li Y, Hou H, Chen Y, Fan J, Wang P, Hu T (2022b) Analyzing spatial variance of urban waterlogging disaster at multiple scales based on a hydrological and hydrodynamic model. Nat Hazards 114(2):1915–1938

Widiputra H, Mailangkay A, Gautama E (2021) Multivariate CNN-LSTM model for multiple parallel financial time-series prediction. Complexity 2021(1):9903518

Xiao Y, Shen H, You L, Zheng Y, Xie H, Xu Y, You T (2025) Research on water resource carrying capacity assessment and water quality forecasting based on feature selection with CNN-BiLSTM-Attention model of the Min River Basin. Water 17(6):824

Yamazaki D, Lee H, Alsdorf DE, Dutra E, Kim H, Kanae S, Oki T (2012) Analysis of the water level dynamics simulated by a global river model: a case study in the Amazon River. Water Resour Res. https://doi.org/10.1029/2012WR011869

Yan J, Zhang Y, Su Q, Li R, Li H, Lu Z, Lu Q (2023) Time series prediction based on LSTM neural network for top tension response of umbilical cables. Mar Struct 91:103448

Yin S, Gao G, Li Y, Xu YJ, Turner RE, Ran L, Fu B (2023) Long-term trends of streamflow, sediment load and nutrient fluxes from the Mississippi River basin: impacts of climate change and human activities. J Hydrol 616:128822

Zakaria MNA, Malek MA, Zolkepli M, Ahmed AN (2021) Application of artificial intelligence algorithms for hourly river level forecast: a case study of Muda river, Malaysia. Alexandria Eng J 60(4):4015–4028