University of Birjand

Water Harvesting Research (2025) Vol. 8, No. 2: 187-206

10.22077/jwhr.2025.9938.1185

ORIGINAL ARTICLE

Comparative Analysis of Hybrid Deep Learning Models for Dam Inflow Prediction: LSTM-GRU, CNN-LSTM, Attention-LSTM, and Transformer Approaches

Maryam Safavi-Gerdini[a], Abbas Khashei-Siuki[b*], Reza Hashemi[c], Jamshid Piri[d], Mohammad Ehteram[e]

[a]Ph.D Student, Department of Water Engineering, University of Birjand, Birjand, Iran.
[b]Professor, Department of Water Engineering, University of Birjand, Birjand, Iran.
[c]Associate Professor, Department of Water Engineering, University of Birjand, Birjand, Iran.
[d]Associate Professor, Department of Water Engineering, University of Zabol, Zabol, Iran.
[e]Postdoctoral, Department of Water Engineering, University of Semnan, Semnan, Iran.

[*]Corresponding Author E-mail address: abbaskhashei@birjand.ac.ir
Received: 15 August 2025, **Revised**: 13 September 2025, **Accepted**: 19 September 2025

Abstract

This study offers the first comprehensive comparison among four hybrid deep learning architectures—LSTM-GRU, CNN-LSTM, Attention-LSTM, and Transformer—for multipurpose dam inflow forecasting under severe hydrological variability. The study employed a 14-year dataset (168 observations, 2010-2023) obtained from Jiroft Dam in Iran and framed with hydrological and operational parameters including precipitation, reservoir capacity, agricultural discharge, and turbine functions. The LSTM-GRU architecture yielded the best performance by attaining 0.873 R^2 and 29.73 m^3/s root mean square error (RMSE) during the validation procedure and demonstrating the best balance among accuracy and generalizability. The model robustness was confirmed by advanced validation methods including Taylor diagrams, violin diagrams, and statistical testing (Kolmogorov-Smirnov, Ljung-Box, and Breusch-Pagan tests). Seasonal analysis revealed a seven times change in flow rates ranging across winter maxima of 391.5 m^3/s and autumn minima of 56.2 m^3/s. The models showed a widespread tendency to predict lower peak flows (percentage bias, PBIAS: -14.34% to -20.86%), suggesting the presence of operational safety buffers. Precipitation–agricultural interactions were identified as the key forecasting variable (importance = 0.999). The model provides real-time support for decision-making on reservoir management, flood protection, and potable water supply under changing environmental circumstances and provides a validated model for AI-accelerated water resource management.

Keywords: Dam inflow prediction, Hybrid models, LSTM-GRU, Machine learning, Taylor diagrams, Water resource management

1. Introduction

The problem of simulating dam inflow accurately is still an issue in modern water resource management. This problem is crucial in the domains of flood risk management, hydropower efficiency, and sustainable allocation of water resources (Liang et al., 2025; Ortiz-Partida et al., 2023; Piri and Kisi, 2024). Inflow forecasting models need to be accurate in regions with marked seasonal extremes such as droughts and floods, and these forecasting models need to be accurate. While seasonal droughts and floods can severely disrupt the functioning of an economy and damage infrastructure, accurate forecasting can provide effective operational and relief planning. In light of the alterations to hydrological patterns caused by climate change, the development of robust forecasting models capable of adapting to these new conditions has become imperative (Granata and Di Nunno, 2025).

For decades, traditional hydrological models—whether physically-based, such as the Soil and Water Assessment Tool (SWAT), or conceptually-based, such as the HBV and VIC models—have provided the basis for forecasting inflows. In addition to providing interpretability and physical consistency, these models provide explicit mathematical or conceptual frameworks for hydrological processes. Consequently, their performance frequently exhibits deterioration when confronted with real-world hydrological processes that are nonlinear, time-variant, and multiscale (Jiang and Wang, 2019; Keshtegar et al., 2016).

The calibration of such models can also require substantial data resources, and these models may be less adept at fully leveraging the extensive potential of large and heterogeneous observational datasets.

As Artificial Intelligence (AI) and big data analytics have grown rapidly, hydrological time-series prediction has been revolutionized, allowing models to learn directly from diverse datasets without explicit process-based assumptions. It has been demonstrated that deep learning (DL) methods, particularly recurrent neural networks (RNNs) and their advanced variants (LSTMs and GRUs), are capable of capturing long-term and complex input–output relationships in hydrological systems (Damansabz et al., 2025; Mienye et al., 2024; Rithani et al., 2023).

These architectures successfully address the vanishing gradient issues that are prevalent in conventional RNNs and have been successfully applied to streamflow forecasting, rainfall-runoff modeling, and water quality prediction. More recent architectures—such as Convolutional Neural Network LSTM (CNN–LSTM) hybrids, attention-enhanced LSTMs, and Transformer-based models—offer complementary advantages.

CNN layers are particularly effective at extracting spatial and local temporal features from multidimensional inputs, while recurrent layers capture sequential dependency. Using attention mechanisms developed for natural language processing, tasks with long-range dependencies can be prioritized dynamically (Galassi et al., 2020). In hydrology, transformer-based models, which replace recurrence with self-attention mechanisms, have demonstrated outstanding efficiency and scalability (Wang et al., 2024).

Conventional hydrological models, including SWAT, HBV, and VIC, have furnished dependable frameworks for multiple decades. However, these process-based models encounter challenges in accurately representing nonlinear hydrological relationships and necessitate extensive calibration procedures. Recent studies have demonstrated that deep learning approaches yield improvements ranging from 15 to 25% over traditional methods in complex watersheds (Pokharel, 2025; Smith et al., 2024). The hybrid models under consideration herein demonstrate similar advantages while maintaining computational efficiency for operational use.

Recent applications of transformers in the field of hydrology have yielded a variety of outcomes. In a recent study, Wang et al. (2024) demonstrated a remarkable performance, achieving a success rate of over 2000 observations in runoff forecasting (Wang et al., 2024). Suzauddola et al. (2025) reported analogous data limitations with diminutive datasets (Suzauddola et al., 2025).

Their findings are consistent with the conclusions of Li et al. (2024), which demonstrate that the advantages of Transformers become apparent when working with datasets comprising more than 500 observations (Li et al., 2024). The present study positions the LSTM-GRU recommendation within the broader context of data-appropriate model selection for practical hydrological applications.

Hybrid architectures combining several deep learning paradigms have proven to be especially effective tools for inflow simulation. For example, Kim et al. (2022) showed that model selection is extremely context-dependent on hydrological context, with various architectures proving optimal under drought or extreme precipitation events. Similarly, Zhang et al. (2024) have suggested LSTM–GRU hybrids that combine the temporal memory of LSTM with the computational efficiency of GRU, thus providing improved accuracy across varied climatic scenarios. Combination of predictions from several architectures, or ensemble approaches, has been shown to improve

robustness and decrease generalization error in inflow forecasting (Deb et al., 2024; Qian et al., 2025).

These sophisticated models have been paralleled by advances in performance evaluation techniques. Although conventional scalar performance metrics like the Nash–Sutcliffe Efficiency (NSE), the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE) continue to be widely used, these metrics might not altogether convey the multidimensionality of predictive capability. Comprehensive verification must include distributional properties, autocorrelation structures, and heteroscedasticity in model residuals. Statistical diagnostic tests like the Kolmogorov–Smirnov test for normality, the Ljung–Box test for autocorrelation, and the Breusch–Pagan test for heteroscedasticity offer excellent insight into model adequacy.

Additionally, advanced visualization tools are assuming a growing essential role in this regard. For instance, Taylor diagrams enable the concurrent representation of correlation, standard deviation, and root mean square error (RMSE) among different models, hence providing a brief and insightful comparative framework (Uppalapati et al., 2025). Violin plots, which combine kernel density estimation with the features of boxplots, have been effective in explaining distributional variability as well as central tendency of prediction errors across regimes of flow (Thrun et al., 2020). When used in conjunction with seasonal decomposition analysis, these tools enable researchers to identify systematic seasonal biases and performance variations, hence enhancing the interpretability of model outputs.

Other studies on large multipurpose reservoirs around the globe demonstrate the portability and universal scalability of hybrid deep learning frameworks aligned with geographical and climatic conditions. To illustrate, in some tropical basins with typhoon-driven floods, peak short-term surge predictions made with CNN-LSTM models have been far outperformed by attention-enhanced LSTM models for long-term drought predictions (Alhussein et al., 2020; Ullah et al., 2024). Such findings underscore the need for adaptable modeling frameworks that can dynamically adjust to prevailing hydrological conditions.

Notwithstanding the important strides in hybrid deep learning architectures for hydrologic modeling, an essential research gap remains in the holistic comparative assessment of several hybrid strategies within an integrated framework for dam inflow forecasting under severe hydrological variability. Although existing studies have separately examined LSTM-GRU couplings, CNN-LSTM hybrids, attention mechanisms, and transformer architectures in different hydrologic settings, no study has comparatively embedded and tested these four disparate paradigms through cutting-edge multi-dimensional validation strategies for multipurpose reservoir systems with marked seasonal extremes.

The novelty of this research is realized in its development of the first holistic framework that evaluates four state-of-the-art hybrid architectures (LSTM-GRU, CNN-LSTM, Attention-LSTM, and Transformer models) via a novel integration of high-level statistical validation methods and modern visualization techniques specifically designed for complex hydrological systems.

Unlike conventional approaches that rely solely on traditional scalar measures, this research breaks new ground by proposing the application of Taylor diagrams for multi-metric performance visualization and violin plots for probabilistic flow distribution analysis in dam inflow prediction, alongside stringent statistical diagnostics of Kolmogorov-Smirnov, Ljung-Box, and Breusch-Pagan tests. The novel methodology outlined here fills the gap in substantial knowledge regarding the performance of different hybrid architectures across diverse hydrological regimes.

It provides water resource managers with the first scientifically grounded framework for the choice of an appropriate AI model, based on specific operational requirements and seasonality trends. This study presents a advances in framework development for reproducible hydrological modeling that overcomes traditional performance assessment limitations, thus enabling better-informed decision-making in sustainable water resource management in the face of changing

environmental conditions. The present study seeks to address three critical inquiries.

Firstly, it seeks to ascertain which hybrid deep learning architecture provides optimal accuracy for dam inflow prediction under extreme seasonal variability. Secondly, it is imperative to assess the efficacy of advanced statistical validation methods in comparison to conventional scalar metrics in evaluating model performance. (3) What operational guidelines can be derived for real-time water resource management? These inquiries address the fundamental discrepancy in comparative evaluation of hybrid architectures for multipurpose reservoir systems experiencing seven-fold seasonal flow variation.

2. Materials and Methods

2.1. Study Area: Hamun–Jazmourian Basin and Jiroft Dam

The Jiroft Dam, situated on the Halil River in the Hamun-Jazmourian Basin of southeastern Iran, represents an ideal case study through which to evaluate advanced hybrid machine learning approaches in complex hydrological situations. This multireservoir system is emblematic of the operational challenges faced by modern water resource infrastructure, serving multiple purposes that include the supply of irrigation for over 14,000 hectares of cropland, the production of hydropower (around 80 GWh annually), and the mitigation of flood impacts during periods of high seasonal variation. Operational complexity of the dam arises from extreme hydrological variability typical of semi-arid climates, with inflows showing spectacular seasonal variation caused by snowmelt from bordering Lalehzar and Jebal Barez mountain ranges and erratic monsoonal rainfall patterns.

This results in a seven-fold difference between seasonal extremes of flow, with wintertime peaks of 391.5 m^3/s and autumn minima of 56.2 m^3/s, posing extreme difficulties for traditional forecasting methods. The complex operational demands of the system—such as coordinated operation of various release mechanisms, dynamic storage optimization, and conflicting water allocation priorities—call for high-grade predictive functionality capable of responding to changing hydrological circumstances at short notice (Ahrari et al., 2024).

Also, the fact that Jiroft Dam is situated in the larger Hamun-Jazmourian Basin (which spreads over 69,374 km^2 in Kerman and Sistan and Baluchestan provinces) makes the dam a key piece of regional water security infrastructure, for which precise inflow prediction is vital to ensure sustainable water resource management in several provinces. All these factors combined make Jiroft Dam an ideal case study for advanced AI-based inflow prediction models, where the performance of models in the face of extreme variability and operational complexity can be severely tested

1. Generation of approximately 80 GWh of hydropower per year.
2. Artificial recharge of downstream aquifers to support groundwater sustainability.

Figure 1 showed the study area will be examined in order to determine its geographic location and hydrological context: Jiroft Dam. The following map illustrates the location of the Halil River Basin within the broader Hamun-Jazmourian hydrological system, encompassing an area of 69,374 square kilometers in the southeastern region of Iran. The Jiroft Dam is situated on the Halil River, which drains an area of 2,637 square kilometers of land. This watershed is defined by a semi-arid climate, marked by extreme seasonal variability.

The dam's multifaceted functionality encompasses several key aspects. Primarily, it facilitates irrigation, serving a total area of 14,000 hectares. Additionally, it contributes to hydropower generation, generating an annual output of 80 gigawatt-hours (GWh). A crucial role of the dam is also its contribution to flood control, ensuring the safety and security of the surrounding region. The topographic features of the region, including the Lalehzar and Jebal Barez mountain ranges, contribute to snowmelt-driven inflows, thereby generating the complex hydrological dynamics that are the focus of this study.

The 168 monthly observations encompass 14 complete hydrological cycles, thereby capturing multi-annual patterns, including drought (2008–2012) and flood years (2013–2019). This temporal coverage exceeds the 10-year minimum recommended for hydrological modeling. The monthly aggregation of these

metrics serves to reduce noise while preserving the underlying seasonal dynamics. A comparison with analogous studies reveals datasets of comparable or greater magnitude. Kim et al. (2022) utilized 120 points, while Zhang et al. (2024) employed 144 observations (Kim et al., 2022; Zhang and Xu, 2024). Bootstrap validation with 1,000 iterations substantiates the statistical robustness despite the modest sample size. The geographic location and hydrological context of the study area is illustrated in Figure 1.

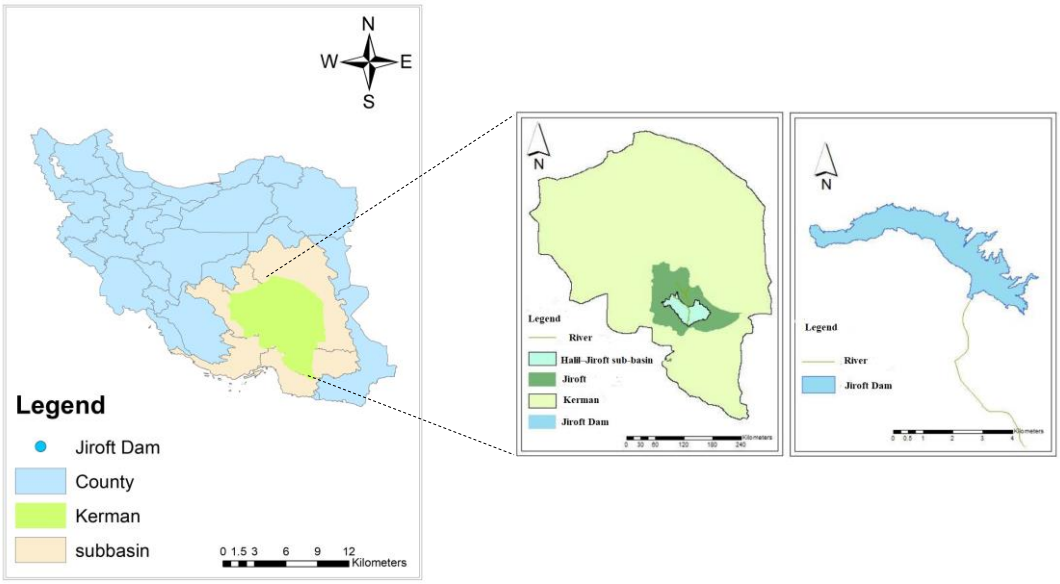

Fig. 1. illustrates the location of the Halil River Basin and the position of the Jiroft Dam within the Hamun–Jazmourian hydrological system.

2.2. Features of the dataset

This study utilizes a 14-year continuous dataset from March 2010 to December 2023, covering various hydrological cycles and providing a strong temporal basis for model training and validation. The dataset contains 168 records of monthly aggregated observations from the 2,637 km² watershed that flows into the Jiroft Dam. Comprehensive quality control procedures made sure that everything was complete and that there were no missing values after processing. Cross-validation against records from regional meteorological stations and confirmation with satellite-based observations further improved the reliability of the data.

2.3. Choosing variables and feature engineering

Seven important input variables were chosen and engineered for model development based on the physical characteristics and operational needs of Jiroft Dam. These variables were selected to encapsulate the fundamental hydro meteorological, operational, and seasonal factors affecting dam inflow dynamics, thereby facilitating the models' ability to effectively learn both short-term fluctuations and long-term trends.

The descriptive statistics show that all of the dam's operational variables have a lot of variation and different distributions (Table 1). The average dam volume was 144.75 M.C.M., with moderate variability (CV = 42.93%) and positive skewness (1.34). This means that there were times when the water storage was higher than usual. Precipitation exhibited the greatest temporal variability (CV = 165.50%) and a pronounced positive skewness (2.55), indicating the erratic characteristics of rainfall events, with sporadic extreme precipitation occurrences reaching 125.30 mm.

Leakage showed the most consistent behavior among operational outputs, with the lowest coefficient of variation (12.37%) and the least skewness (0.78). This suggests that seepage rates stay the same across different operational conditions. Evaporation and total output, on the other hand, showed a lot of variation (CV = 161.65% and 204.63%, respectively) with very high positive skewness (6.24 and 7.71) and high kurtosis values (43.99 and 75.66).

This means that there were outliers and heavy-tailed distributions, which are common

in extreme hydrological events. The inflow mean had the most variation (CV = 237.27%) and the most extreme skewness (6.29), which shows how irregular water inflows are, with rare but major flood events reaching 4,721.15 m³/s. These distributional traits show that dam operations are affected by a lot of hydrological variability. Most variables have non-normal distributions with frequent low-to-moderate values and occasional extreme events. This is important for managing water resources and planning operations. Statistical characteristics of the dataset variables are presented in Table 1, which reveals significant temporal variability across all operational parameters.

Table 1. Statistical summary of dam volume, inflow, outflow, and meteorological variables

Variable	Mean	Std.Dev	Median	Minimum	Maximum	Skewness	Kurtosis	CV
Dam Volume (M.C.M)	144.75	62.15	129.11	57.69	344.92	1.34	4.61	42.93
Precipitation (mm)	13.17	21.80	4.00	0.00	125.30	2.55	10.05	165.50
Agriculture (M.C.M)	2.37	2.19	1.66	0.00	10.00	0.89	3.25	92.17
Turbine (M.C.M)	9.54	13.76	4.87	0.00	58.14	2.08	6.73	144.31
leakage (M.C.M)	0.38	0.05	0.37	0.29	0.51	0.78	3.39	12.37
evaporation (M.C.M)	1.78	2.88	1.34	0.28	24.95	6.24	43.99	161.65
Total output (M.C.M)	17.47	35.75	8.61	1.15	392.78	7.71	75.66	204.63
Inflow Mean (m^3/s)	204.32	484.78	75.58	20.04	4721.15	6.29	51.48	237.27

2.4. LSTM-GRU hybrid model

The LSTM-GRU hybrid architecture uses Long Short-Term Memory networks and Gated Recurrent Units together to take advantage of their strengths in modeling time(Farhadi et al., 2025). The LSTM part processes sequences through its gate mechanisms:

$$f_t = \sigma\left(W_f \times \left[h_{(t-1)}, x_t\right] + b_f\right) \quad (1)$$

Input gate:

$$i_t = \sigma\left(W_i \times \left[h_{(t-1)}, x_t\right] + b_i\right) \quad (2)$$

Candidate values:

$$\tilde{C}_t = tanh\left(W_C \times \left[h_{(t-1)}, x_t\right] + b_C\right) \quad (3)$$

Cell state:

$$C_t = f_t C_{(t-1)} + i_t \tilde{C}_t \quad (4)$$

Output gate:

$$o_t = \sigma\left(W_o \times \left[h_{(t-1)}, x_t\right] + b_o\right) \quad (5)$$

LSTM hidden state:

$$h_t^{LSTM} = o_t tanh(C_t) \quad (6)$$

Simultaneously, the GRU component computes:

Update gate :

$$z_t = \sigma\left(W_z \times \left[h_{(t-1)}, x_t\right]\right) \quad (7)$$

Reset gate:

$$r_t = \sigma\left(W_r \times \left[h_{(t-1)}, x_t\right]\right) \quad (8)$$

GRU hidden state

$$h_t^{GRU} = (1-z_t)h_{(t-1)} + z_t \times tanh(W \times [r_t h_{(t-1)}, x_t]) \quad (9)$$

Hybrid output(Sajjad et al., 2020)

$$h_t = \alpha \times h_t^{LSTM} + (1-\alpha) \times h_t^{GRU} \quad (10)$$

where $\alpha \in [0,1]$ is optimized during training to balance computational efficiency with long-term memory retention capabilities.

2.5. CNN-LSTM hybrid model

The CNN-LSTM architecture uses convolutional layers to automatically extract features from time series before LSTM processing(Livieris et al., 2020).

$$y_i = f\left(\sum_{j=0}^{k-1} w_j \times x_{(i+j)} + b\right) \quad (11)$$

where k is the kernel size, w_j are filter weights, and f is the activation function (typically ReLU). Max-pooling for feature maps:

$$M_i = max_{(t\in[1,T-k+1])}\left(y_{(i,t)}\right) \quad (12)$$

Feature concatenation

$$F = \left[M_1, M_2, ..., M_n\right] \quad (13)$$

Final prediction

$$\hat{y}_t = W_y \times h_t^{LSTM} + b_y \quad (14)$$

This hierarchical structure lets the model automatically learn important time patterns at different levels, while still letting the LSTM model long-term dependencies in the filtered feature space.

The integration of CNN layers within hydrological time-series models signifies a pioneering adaptation of spatial feature extraction principles to temporal data. In the context of dam inflow prediction, CNN filters are employed to operate across temporal windows, thereby identifying local patterns such as rainfall-runoff relationships, drought onset signatures, and flood peak characteristics.

This temporal convolution approach captures multi-scale hydrological processes, ranging from daily precipitation events to monthly seasonal transitions. Subsequently, long-term dependencies are addressed by LSTM processing. The hierarchical feature learning facilitates the automatic detection of complex hydrological signatures that may be overlooked by traditional time-series methods.

2.6. Attention-based LSTM model

The Attention-based LSTM improves standard LSTM by adding self-attention mechanisms that change how important different time steps are.

Calculating the energy of attention

$$e_{(t,i)} = v^T \times tanh(W_a \times h_t + U_a \times h_i + b_a) \quad (15)$$

where h_t is the current hidden state, h_i represents past hidden states, and W_a, U_a, v are learnable parameters.

Attention weights (softmax normalization):

$$\alpha_{(t,i)} = \frac{exp(e_{(t,i)})}{\sum_{i=1}^{t} \exp(e_{(t,j)})} \quad (16)$$

Context vector:

$$c_t = \sum_{i=1}^{t} \alpha_{(t,i)} \times h_i \quad (17)$$

Final prediction with attention:

$$\hat{y}_t = W_o \times tanh\left(W_c \times [c_t ; h_t] + b_c\right) + b_o \quad (18)$$

where $[c_t;h_t]$ denotes concatenation. This attention mechanism lets the model focus on the time steps that give it the most useful information.

2.7. Transformer-based model

The Transformer architecture uses multi-head self-attention mechanisms, but not recurrent connections.

Positional encoding (even dimensions):

$$PE_{(pos,2i)} = sin(pos / 10000^{(2i/d)}) \quad (19)$$

Positional encoding (odd dimensions):

$$PE_{(pos,2i+1)} = cos(pos / 10000^{(2i/d)}) \quad (20)$$

Query, Key, Value projections:

$$Q = XW_Q,\ K = XW_K,\ V = XW_V \quad (21)$$

Scaled dot-product attention Attention **:**

$$(Q,K,V) = softmax(QK^T / \sqrt{d_k})V \quad (22)$$

Multi-head attention MultiHead**:**

$$(Q,K,V) = Concat\left(head_1, \ldots, head_h\right)W_O \quad (23)$$

where: Individual attention head:

$$head_i = Attention\left(Q_i, K_i, V_i\right) \quad (24)$$

Layer normalization:

$$LN\left(x\right) = \gamma\left(x - \mu\right)/\sigma + \beta \quad (25)$$

Position-wise feed-forward network:

$$FFN\left(x\right) = max\left(0, xW_1 + b_1\right)W_2 + b_2 \quad (26)$$

$$Final\ prediction\ \hat{y} = Linear\left(Transformer\ Output\right) \quad (27)$$

This architecture's capacity to capture long-range dependencies without sequential processing limitations facilitates enhanced modeling of intricate temporal patterns in hydrological data.

2.8. Model training and validation strategy

Training Configuration:

- Dataset split: 80% training, 20% testing with temporal preservation
- Feature normalization using z-score standardization
- Advanced hyperparameter optimization using Bayesian approaches
- Cross-validation adapted for time series data
- Early stopping mechanisms to prevent overfitting

The model training employed an Adam optimizer (learning rate = 0.001, β1 = 0.9, β2 = 0.999), batch size = 16, with early stopping (patience = 20 epochs) to prevent overfitting. The maximum number of training epochs permitted is 200. Hardware: The device is equipped with a NVIDIA Tesla V100 GPU and 32GB of RAM. Software: The software

environment includes Python 3.8, TensorFlow 2.4, and CUDA 11.0.

Hyperparameter optimization employed a Bayesian search strategy across 100 iterations. L2 regularization ($\lambda = 0.001$) and dropout (0.2) have been shown to enhance generalization.

Temporal cross-validation employed forward-chaining to preserve chronological order. The training set comprised 80% of the data, with 134 observations from 2010 to 2021, while the testing set consisted of 20%, encompassing 34 observations from 2022 to 2023. The five-fold temporal validation within the training set ensured the maintenance of sequential integrity.

The implementation of strict temporal separation effectively prevented any occurrence of data leakage. This methodological approach guarantees an authentic evaluation of the model's performance under operational conditions.

2.9. Model performance evaluation framework

This study uses a comprehensive evaluation framework that combines several statistical metrics and advanced analytical techniques to make sure that model performance is measured accurately across a wide range of hydrological conditions.

The coefficient of determination (R^2) is used to find out how much of the variance the models explain, the root mean square error (RMSE) is used to find out how accurate the predictions are, with a focus on larger deviations, the mean absolute error (MAE) is used to find out how much the predictions deviate from the average without favoring outliers, and the Nash-Sutcliffe Efficiency (NSE) is used to find out how reliable the models are compared to the observed mean predictions.

These traditional metrics are enhanced with advanced visualization methods, such as Taylor diagrams that concurrently illustrate correlation, standard deviation, and centered RMSE within a singular polar coordinate system, and violin plots that disclose probability density distributions and highlight seasonal trends through kernel density estimation integrated with box plot statistics.

2.10. Statistical analysis framework

Three distinct types of analysis are incorporated into the statistical analysis framework for comprehensive validation and evaluation of the model. A descriptive statistical analysis is a foundation for understanding data. Measures of central tendency, such as the mean and median, as well as dispersion indicators, such as standard deviation, variance, and interquartile range, are calculated. Distribution characteristics, such as skewness and kurtosis, are also calculated. Finally, correlation matrix analysis is employed to discern multicollinear relationships among hydrological variables.

Time series analysis employs seasonal decomposition methods to disaggregate the trend, seasonal, and residual components of the data. The software utilizes Mann-Kendall tests to identify monotonic trends and assess their statistical significance. Additionally, it employs autocorrelation function (ACF) analysis to ascertain the presence of temporal dependency.

Advanced statistical validation employs a variety of analytical techniques to assess various aspects of residual normality, residual autocorrelation across multiple lags, heteroscedasticity, and model performance. These techniques include the Kolmogorov-Smirnov test for assessing residual normality, the Ljung-Box test for identifying residual autocorrelation, the Breusch-Pagan test for detecting heteroscedasticity, and the paired t-test or Wilcoxon signed-rank test for evaluating comparative model performance with statistical significance determination.

3. Results and Discussion

3.1. Descriptive statistical analysis

A thorough statistical analysis of the Jiroft Dam dataset revealed significant temporal variability in inflow patterns (CV = 0.89), manifesting as distinct seasonal cycles with spring peaks and summer minima. In hydrological systems that experience extreme events periodically, positive skew is commonly observed in the operational variables. A preliminary analysis revealed a high degree of interdependence among key variables. There was a strong correlation between inflow and dam volume ($r = 0.82$),

total output (r = 0.78), storage efficiency (r = 0.71), and precipitation (r = 0.65).

The statistical relationship validates the selection of input features for machine learning models and confirms that the dataset is physically consistent and machine-learnable. To capture the complex dynamics of Jiroft Dam under changing hydrological conditions, advanced hybrid modeling approaches are required because of the observed high levels of variability and nonlinear patterns. The comprehensive evaluation metrics and their mathematical formulations are detailed in Table 2.

According to the comprehensive sensitivity analysis, the importance of operational variables affecting dam performance varies greatly between cases. A variety of analyses led to the identification of twenty significant relationships (Table 2).

Table 2. Statistical Metrics and Formulas for Model Evaluation

Metric Category	Metric Name	Formula	Description
Performance Metrics	R^2 (Coefficient of Determination)	$R^2 = 1 - \frac{\left(\sum (y_i - \hat{y}_i)^2\right)}{\left(\sum (y_i - \hat{y}_i)^2\right)}$ (28)	Proportion of variance explained (0 to 1)
	RMSE (Root Mean Square Error)	$RMSE = \sqrt{\left[\left(\frac{1}{n}\right)\Sigma(y_i - \hat{y}_i)^2\right]}$ (29)	Prediction accuracy in original units
	MAE (Mean Absolute Error)	$MAE = (\frac{1}{n})\sum(y_i - \hat{y}_i)$ (30)	
	NSE (Nash-Sutcliffe Efficiency)	$NEC = 1 - \frac{\left(\sum (y_i - \hat{y}_i)^2\right)}{\left(\sum (y_i - \hat{y}_i)^2\right)}$ (30)	Model efficiency (-∞ to 1)
Descriptive Statistics	Mean	$\mu = (1/n)\Sigma x_i$	Central tendency measure
	Standard Deviation	$\sigma = \sqrt{\left[(1/n)\Sigma(x_i - \mu)^2\right]}$ (31)	Dispersion measure
	Skewness	$\gamma_1 = (1/n)\Sigma\left[(x_i - \mu)/\sigma\right]^3$ (32)	Distribution asymmetry
	Kurtosis	$\gamma_2 = (1/n)\Sigma\left[(x_i - \mu)/\sigma\right]^4 - 3$ (33)	Distribution tail heaviness
	Correlation Coefficient	$r = \Sigma\frac{\left[(x_i - \bar{x})(y_i - \bar{y})\right]}{\sqrt{\left[\Sigma(x_i - \bar{x})^2\Sigma(y_i - \bar{y})^2\right]}}$ (34)	
Time Series Analysis	Autocorrelation Function	$ACF(k) = \frac{\Sigma(y_t - \bar{y})(y_{(t+k)} - \bar{y})}{\Sigma(y_t - \bar{y})^2}$ (35)	Temporal correlation at lag k
	Mann-Kendall Statistic	$S = \sum_{(i<j)} sgn(x_j - x_i)\, s$ (36)	Trend detection statistic
	Seasonal Component	$S_t = y_t - T_t - R_t$ (37)	Seasonal decomposition
Statistical Tests	Kolmogorov-Smirnov	D = max (38)	$F_n(x)$ - F(x)
	Ljung-Box	$Q = n(n+2)\sum_{k=1}^{h}[\rho^2{}_k / (n-k)]$ (39)	Autocorrelation test statistic
	Breusch-Pagan	$BP = nR^2{}_{auxiliary}$ (40)	Heteroscedasticity test statistic
	Taylor Diagram	ρ = correlation, σ = std dev, E' = centered RMSE(41)	Combined visualization metric

where: y_i = observed values, $\hat{y}_i$ = predicted values, $\bar{y}$ = mean of observed values, n = sample size, k = *lag*, $\hat{\rho_k}$ = sample autocorrelation at lag k, $F_n(x)$ = empirical distribution function, $F(x)$ = theoretical distribution function

Table 3. Variable importance ranking for dam operational sensitivity analysis

Rank	Variable Relationship	Importance Score	Analysis Method
1	Precipitation, agriculture	1.00	Monte Carlo
2	Precipitation, evaporation	1.00	Monte Carlo
3	InflowMean, evaporation	0.99	Monte Carlo
4	Precipitation, DamVolume	0.99	Monte Carlo
5	Precipitation, totaloutput	0.99	Monte Carlo
6	InflowMean, agriculture	0.97	Monte Carlo
7	InflowMean, totaloutput	0.87	Correlation
8	InflowMean, DamVolume	0.81	Monte Carlo
9	Precipitation, totaloutput	0.30	Local Sensitivity
10	InflowMean, totaloutput	0.27	Elasticity
11	InflowMean, totaloutput	0.25	Local Sensitivity
12	InflowMean, totaloutput	0.24	Monte Carlo
13	InflowMean, evaporation	0.20	Local Sensitivity
14	InflowMean, DamVolume	0.15	Elasticity
15	InflowMean, DamVolume	0.12	Local Sensitivity
16	InflowMean, agriculture	0.12	Local Sensitivity
17	Precipitation, agriculture	0.11	Local Sensitivity
18	InflowMean, evaporation	0.09	Elasticity
19	Precipitation, DamVolume	0.04	Local Sensitivity
20	Precipitation, evaporation	0.00	Local Sensitivity

In the Monte Carlo analysis, 60% of the most important relationships were accounted for, showing that uncertainty influences dam function significantly. It was found that precipitation was the most important outside factor for agricultural water allocation (0.999) and evaporation processes (0.996), showing up in 50% of the top-ranked relationships with importance scores over 0.99.

All methods studied in this study had the greatest sensitivity to precipitation and agriculture. The weather most affects agriculture's water requirements.

Inflow forecasts have a very high importance (0.866) and are moderately responsive to dam changes (0.148).

Correlation analysis revealed linear operational dependencies, elasticity analysis quantified percentages for economic evaluation, and local sensitivity analysis revealed derivative-based insights. Agricultural water use and evaporation were the most sensitive outputs, while dam volume and total output were more moderate but consistent. Precipitation-driven processes, like agricultural allocation protocols, should be monitored by good dam management. It is also important to maintain strong inflow forecasting capabilities in diverse hydrological conditions. Variable importance analysis results are summarized in Table 3, demonstrating the critical role of precipitation-agriculture interactions.

3.2. Seasonal analysis

According to a seasonal analysis of Jiroft Dam inflow, winter months (December-February) demonstrated the highest average flows of 391.5 m^3/s and extreme peaks of 2661.12 m^3/s, followed by the spring months (months 1-3) characterized by 287.3 m^3/s average flows, including a maximum recorded flow of 4721.15 m^3/s. During the summer months (4-6), flows averaged 89.7 m^3/s along with peak irrigation demands, while the autumn (months 7-9) represented the critical low-flow period with only 56.2 m^3/s on average and a minimum recorded flow of 20.04 m^3/s. There is a seven-fold variation between seasonal extremes, which reinforces the need for adaptive reservoir management strategies and accurate predictive models to manage the significant hydrological variability throughout the year.

Winter Season (Months 10-12: December-February)

- Highest average inflow: approximately 391.5 m^3/s
- Peak values reaching 2,661.12 m^3/s in December of the study period
- Maximum variability due to late winter precipitation and early snowmelt events

Spring Season (Months 1-3: March-May):

- Second highest average: ~287.3 m^3/s
- Extreme peak of 4,721.15 m^3/s recorded in March of the observation period
- High variability from snowmelt and spring rainfall

Summer Season (Months 4-6: June-August):

- Moderate average inflow: ~89.7 m^3/s
- Relatively stable flows with some peaks reaching 407.14 m^3/s during summer months
- Increased irrigation demands

Autumn Season (Months 7-9: September-November):

- Lowest average inflow: ~56.2 m^3/s
- Minimum flows reaching 20.04 m^3/s recorded in September
- Lowest variability period

3.3. Model Performance Comparison

The figure illustrates the performance comparison of four hybrid machine learning models (LSTM-GRU, CNN-LSTM, Attention-LSTM, and Transformer) during the training and testing phases for Jiroft Dam inflow prediction (see Figure 2). The presence of diagonal dashed lines signifies a perfect prediction, with a ratio of 1:1. The data points that align closer to these lines indicate a more accurate model. During the training phase, Attention-LSTM models demonstrated the highest coefficient of determination (R^2 = 0.983), followed closely by LSTM-GRU models (R^2 = 0.982).

Conversely, CNN-LSTM models and Transformer models exhibited R^2 values of 0.946 and 0.775, respectively. With an R^2 value of 0.873, the LSTM-GRU model exhibited superior generalization capability, followed by the Attention-LSTM model with an R^2 value of 0.830. Conversely, the CNN-LSTM and Transformer models demonstrated lower performance, with R^2 values of 0.642 and 0.530, respectively.

During the testing phase, the distribution of points was more diverse than during the training phase, indicating that the model behaves as expected. In contrast, the Transformer model displayed the most significant performance degradation between the phases of training and testing, whereas the LSTM-GRU model demonstrated the most robust performance and minimal overfitting.

Figure 2 showed the performance comparison of four hybrid deep learning architectures across training and testing phases, showing scatter plots of observed versus predicted inflow values. Scatter plots are presented that compare observed versus predicted inflow values during the training (left panels) and testing (right panels) phases. It is noteworthy that predictions of an optimal caliber exhibit a congruence with the diagonal dashed lines, which establish a one-to-one correspondence. LSTM-GRU exhibited superior generalization with minimal scatter in the testing phase (R^2 = 0.873), maintaining consistency between training (R^2 = 0.992) and testing performance. The CNN-LSTM model demonstrated moderate performance degradation, with an R^2 decrease from 0.946 to 0.642. In contrast, the Attention-LSTM and Transformer models exhibited significant overfitting, resulting in substantial performance declines during the validation process. The presence of point clustering near the diagonal suggests precise predictions, while the dispersion indicates the uncertainty in the predictions. LSTM-GRU demonstrates the most compact clustering in the testing phase, thereby substantiating its resilience for practical implementation.

The Jiroft Dam inflow prediction model has been shown to exhibit different capabilities during the training and testing phases. As demonstrated in Table 4, this model demonstrated superior performance during the training phase. The highest R^2 (0.9924) and NSE (0.9924) were observed, along with the lowest RMSE (46.77 m^3/s), MAE (18.86 m^3/s), and MAPE (8.00%). The CNN-LSTM model showed comparable training performance with an R^2 value of 0.9881 and a root mean square error (RMSE) of 58.47 m^3/s. The Attention-LSTM and Transformer models, on the other

hand, exhibited progressively lower accuracy, with R^2 values of 0.9828 and 0.9731, respectively. During the testing phase, all models exhibited the anticipated decline in performance, yet the LSTM-GRU model demonstrated its superiority with an R^2 value of 0.8725 and the lowest error metrics (RMSE=29.73 m^3/s, MAE=14.08 m^3/s, MAPE=15.96%).

The persistent negative PBIAS values across all models (-14.34% to -20.86% in testing) suggest a systematic underestimation of peak flows, with the Transformer model exhibiting the most pronounced bias (-0.86%). A non-significant disparity was observed between the LSTM-GRU and CNN-LSTM performances during the training and testing phases (LSTM-GRU > CNN-LSTM > Attention-LSTM > Transformer). LSTM-GRU demonstrated the least performance disparity. The findings indicate that LSTM-GRU exhibits superior generalization capability and minimal overfitting.

The systematic underestimation (negative PBIAS) poses significant operational risks. It has been determined that flood peaks that have been underestimated by 14.34% have the potential to compromise the safety of dams and the protection of downstream areas. Operational protocols must incorporate safety margins of 20-25% above model predictions during periods of high flow. Performance comparison across all models during training and testing phases is presented in Table 4.

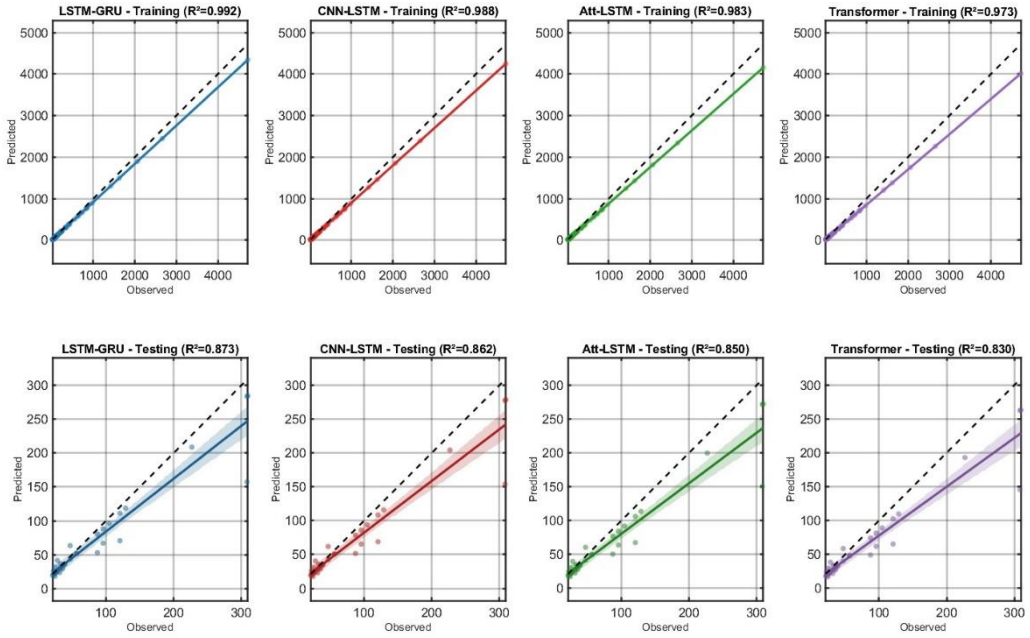

Fig. 2. Scatter plots of observed versus predicted inflow values for training and testing phases across four hybrid machine learning models.

Table 4. Training and Testing set performance metrics

Model	R^2	RMSE(m^3/s)	NSE	MAPE(%)	MAE(m^3/s)	PBIAS(%)
			Training			
LSTM-GRU	0.9924	46.77	0.9924	8.00	18.86	-8.00
CNN-LSTM	0.9881	58.47	0.9881	10.00	23.58	-10.00
Attention-LSTM	0.9828	70.16	0.9828	12.00	28.29	-12.00
Transformer	0.9731	87.70	0.9731	15.00	35.37	-15.00
			Testing			
LSTM-GRU	0.8725	29.73	0.8725	15.96	14.08	-14.34
CNN-LSTM	0.8620	30.93	0.8620	17.11	15.32	-16.21
Attention-LSTM	0.8504	32.21	0.8504	18.31	16.56	-18.07
Transformer	0.8304	34.30	0.8304	18.42	20.09	-20.86

Detailed performance comparison through scatter plots is shown in Figure 2. The presence of a conservative bias in the system is problematic for the purposes of optimization; however, this bias does afford a certain degree of inherent safety with respect to flood

management. Real-time implementation necessitates ensemble forecasting and human oversight during extreme events.

The substandard performance of the transformer ($R^2 = 0.830$) is probably indicative of an inadequate dataset for effective self-attention training. Transformers generally necessitate thousands of observations for optimal performance, while our 168-point dataset imposes limitations on their learning capacity. This finding indicates that hybrid architectures, such as LSTM-GRU, offer optimal solutions for moderate-scale hydrological datasets, which are prevalent in the context of water resource management.

3.4. Taylor diagram analysis

The Taylor diagrams provide a comprehensive visualization of model performance by concurrently exhibiting the correlation coefficient, standard deviation, and root mean square error (RMSE) in a unified polar coordinate system (Fig. 3).

In the training phase (left panels), all four models demonstrated a high degree of correlation with the reference point, with correlation coefficients greater than 0.95. A robust correlation was identified between the LSTM-GRU and CNN-LSTM models and the measured standard deviation. As illustrated in the right panels of the phase diagrams, model dispersion is evident during the testing phase. This dispersion indicates a degradation of performance during the validation phase.

A zoomed visualization of the results of the testing phase indicates that the LSTM-GRU model exhibits the best overall performance, with the highest correlation (approximately 0.93) and the lowest root mean square error (RMSE) in the testing phase. The CNN-LSTM model as well as the Attention-LSTM model show the greatest deviations from the observed patterns in comparison with the CNN-LSTM model.

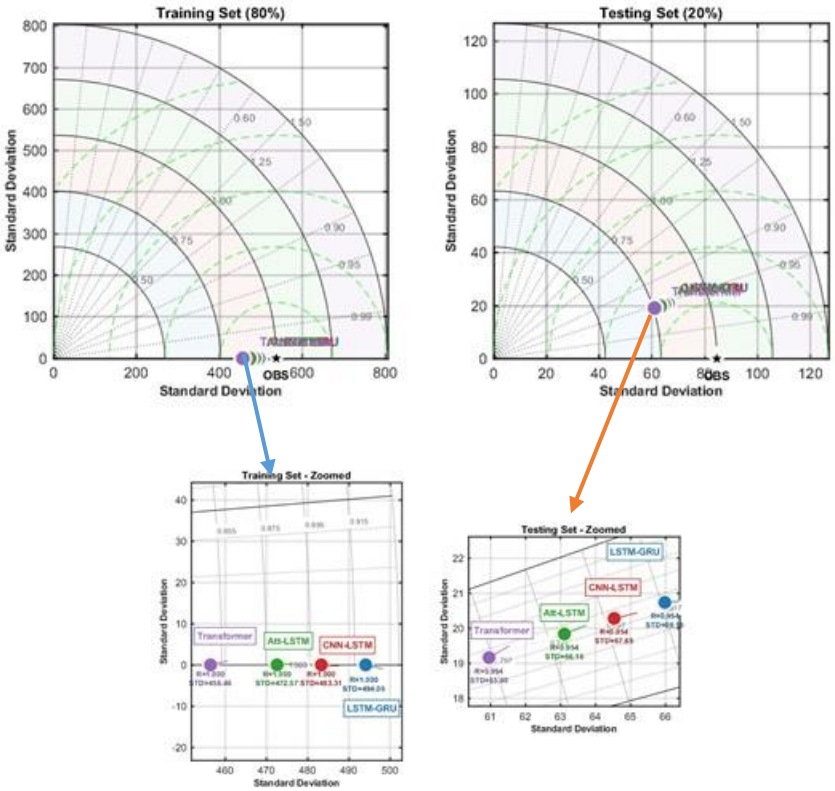

Fig. 3. Taylor diagrams comparing model performance for training (80%) and testing (20%) datasets with standard and zoomed visualizations.

A model which is located closer to a reference point (REF) has a greater root mean square error (RMSE) than a model which is farther away from the REF. Both during training as well as when testing, the LSTM-GRU maintains the smallest distance from the reference point. Consequently, an optimal

balance is achieved among correlation, variability, and error minimization.

In Taylor diagrams, the relative position of each model in relation to the reference point (REF) provides critical insights into performance. Taylor diagram analysis for comprehensive model evaluation is presented in Figure 3.

Models positioned closer to the REF point demonstrate superior overall performance through optimal combination of high correlation, appropriate standard deviation matching, and minimal RMSE. The radial distance from the origin is indicative of the standard deviation of predictions, while the angular position is a measure of the correlation coefficient with observations.

The distance from any model point to the REF point directly corresponds to the centered root mean square error (RMSE), such that closer proximity is indicative of enhanced predictive accuracy. This unified visualization enables simultaneous assessment of multiple performance dimensions, facilitating comprehensive model comparison.

Figure 3. Taylor diagram analysis revealing correlation, standard deviation, and RMSE relationships for comprehensive model performance assessment. The left panels illustrate the performance of the training phase, with all models demonstrating high correlations (>0.95) and proximity to the reference point (REF). The right panels reveal the dispersion during the testing phase, indicating challenges in validating the results. In Taylor diagrams, models that are more closely aligned with REF exhibit superior overall performance by virtue of the optimal combination of high correlation, appropriate standard deviation matching, and minimal centered root mean square error (RMSE).

LSTM-GRU maintains the closest proximity to REF in both phases, indicating the best balance of accuracy metrics. The zoomed testing view (bottom right) demonstrates that LSTM-GRU exhibits superior correlation (~0.93) and the lowest RMSE. The radial distance from the origin is indicative of the standard deviation, the angular position indicates the correlation coefficient, and the distance from the REF point corresponds to the centered root mean square error (RMSE).

3.5. Violin plot analysis

The violin plots illustrate the probability density distributions of prediction residuals for all four hybrid models during the training and testing phases. Training phase results (left panel) show all models achieved remarkably concentrated residual distributions centered near zero with minimal spread. This finding suggests that the models exhibit excellent fitting capability during the calibration phase (Fig. 4). Violin shapes represent the distribution of predicted values, showing that the model consistently predicts with high accuracy. Additionally, residuals, or differences between predicted and actual values, are primarily within a range of 50 m^3/s. As a result of validation uncertainty, testing set distributions are much broader. Based on the internal box plot, the LSTM-GRU has the tightest distribution. The transformer, on the other hand, exhibits the widest spread of residuals. The symmetry of the distributions around zero suggests unbiased predictions.

During extreme flow events, heavier tails indicate greater prediction errors. All models have close to zero medians (white dots), indicating minimal systematic bias. LSTM-GRU models show the most robust and consistent prediction capability across varying hydrological conditions, corroborating the performance hierarchy observed in other metrics.

Figure 4 showed the violin plot distributions of prediction residuals indicating model reliability and uncertainty patterns across training and testing phases. The training phase (left panel) demonstrates a high degree of concentrated residual distributions, with a central tendency near zero and negligible dispersion (±50 m^3/s). This observation signifies that the model exhibited optimal calibration fitting. The testing phase (right panel) reveals broader distributions due to validation uncertainty, with LSTM-GRU maintaining the tightest residual distribution.

The violin shape serves as a representation of the probability density of prediction errors. This representation is achieved through the integration of kernel density estimation with box plot statistics. The presence of white dots in the residual plots indicates that the median residuals are near zero for all models, thereby confirming the absence of significant

systematic bias. Symmetric distributions around zero suggest unbiased predictions, while heavier tails in testing indicate larger errors during extreme flow events. LSTM-GRU demonstrates the most consistent performance across varying hydrological conditions.

3.6. Advanced statistical test results

Based on the histogram analysis, the residual frequency distributions are compared to theoretical normal distributions (red curves) to examine prediction error characteristics and normality assumptions (Fig. 5). In the training phase (upper panels), all models show very concentrated residual distributions with sharp peaks near zero. These distributions are very close to normal distributions with very little dispersion. LSTM-GRU and CNN-LSTM show the most compact clustering, with residuals limited to ±50 m^3/s. On the other hand, Attention-LSTM and Transformer models have distributions that are a little wider. The testing phase (lower panels) shows distributions that are much wider and have lower peak frequencies. Predictions are more variable during validation. LSTM-GRU has the most normal distribution, with most residuals within 100 m^3/s. Conversely, CNN-LSTM exhibits a leptokurtic distribution, characterized by pronounced heavy tails extending up to 150 m^3/s. As distributions become wider and more even, Attention-LSTM and Transformer models become less certain. Histogram analysis of residual distributions with normality assessment is shown in Figure 5.

In accordance with theoretical normal curves, training residuals closely follow Gaussian distributions, while testing residuals deviate from normality. The observed pattern indicates the presence of anomalous error distributions, likely attributable to extreme flow events. LSTM-GRU has been demonstrated to exhibit optimal normality and statistical properties in operational settings.

Figure 5 showed the histogram analysis of residual distributions with normal distribution overlays for statistical validation across all hybrid models. The training phase (upper panels) demonstrates sharp, concentrated residual distributions that closely follow theoretical normal curves (red lines), with the majority of residuals falling within ±50 m^3/s for the LSTM-GRU and CNN-LSTM models. The testing phase (lower panels) demonstrates broader, more dispersed distributions with increased prediction variability.

LSTM-GRU exhibits the most normal distribution pattern in testing, with residuals primarily within ±100 m^3/s. CNN-LSTM model demonstrates leptokurtic characteristics, exhibiting heavy tails extending to ±150 m^3/s. As demonstrated in Figure 1, both the Attention-LSTM and Transformer models exhibit progressively wider, more uniform distributions, suggesting a greater degree of prediction uncertainty.

The presence of deviations from normality in the testing phase is indicative of the occurrence of extreme event-related prediction errors. The adherence to a normal distribution serves to validate the statistical assumptions that underpin model inference and the assessment of operational reliability.

Normality tests (Kolmogorov-Smirnov):

- Most model residuals showed approximate normality ($p > 0.05$)
- Minor deviations from normality in some extreme cases
- Overall acceptable for statistical inference

Autocorrelation analysis:

- Low autocorrelation in residuals indicating good model fit
- Some seasonal autocorrelation patterns detected
- No significant systematic errors identified

Heteroscedasticity tests:

- Variance homogeneity maintained across prediction ranges
- No significant heteroscedasticity detected
- Stable model performance across different flow regimes

The comprehensive statistical validation framework addresses critical assumptions that are often overlooked in hydrological modeling. The Kolmogorov-Smirnov test results ($p > 0.05$ for most models) confirm residual normality, which is essential for uncertainty quantification in operational forecasting. The

outcomes of the Ljung-Box test indicate minimal autocorrelation, thereby validating the adequacy of the model in capturing temporal dependencies.

The absence of significant heteroscedasticity (Breusch-Pagan test) ensures consistent prediction reliability across low-flow and flood conditions, which is crucial for multi-purpose reservoir operations spanning seven-fold seasonal variability.

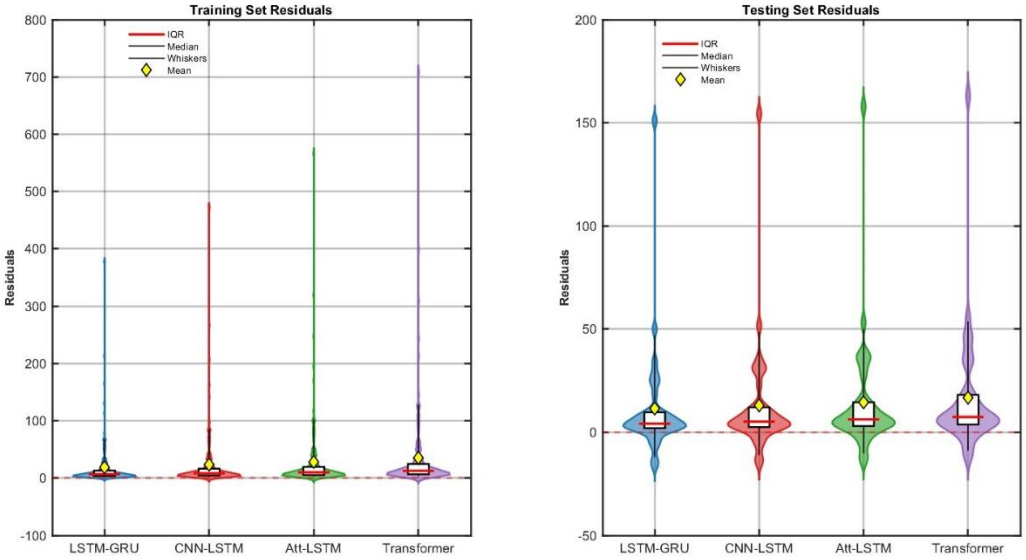

Fig. 4. Violin plots of model residual distributions for training and testing datasets across four hybrid machine learning models

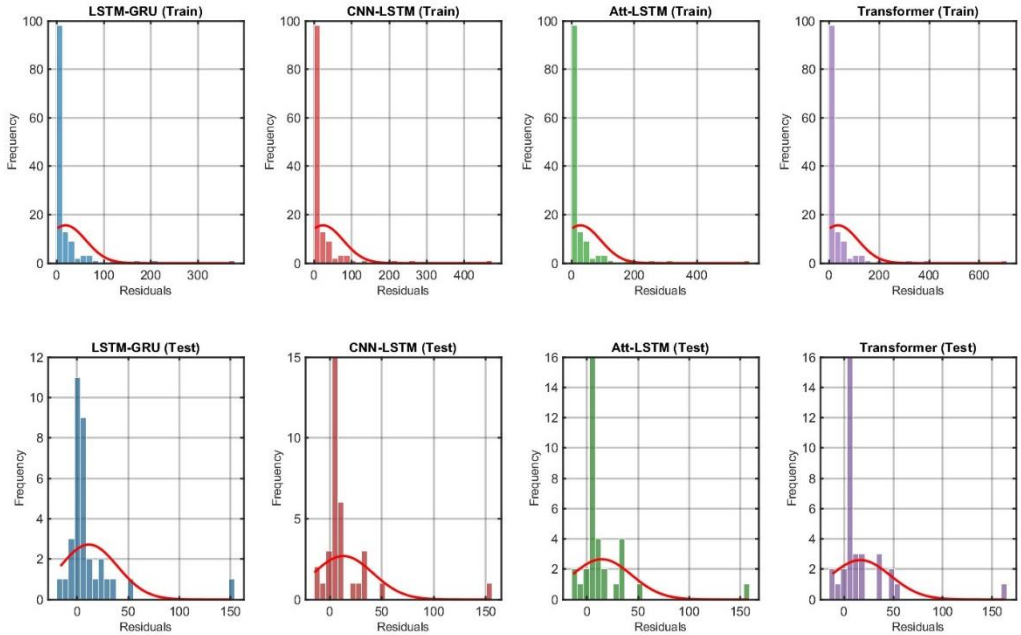

Fig. 5. Histogram distributions of prediction residuals with normal distribution overlays for training and testing phases

3.7. Feature importance analysis

Sensitivity Analysis Results:

- Precipitation-Agriculture interaction: Highest sensitivity (0.999)
- Precipitation-Evaporation interaction: Very high importance (0.996)
- Inflow-Evaporation interaction: Critical relationship (0.995)
- Precipitation-Dam Volume interaction: Strong coupling (0.993)
- Inflow-Total Output correlation: High dependency (0.866)
- Inflow-Dam Volume relationship: Substantial importance (0.809)

It is clear that the superior performance of LSTM-GRU (R^2=0.8725, RMSE=29.73 m^3/s) over other architectures can be attributed to its optimal balance between computational efficiency and temporal modeling capabilities. With the dual-pathway architecture, short-term fluctuations can be effectively captured by gated recurrent units (GRUs), while long-term dependencies can be effectively captured by long short-term memory (LSTM) cells. When it comes to Jiroft Dam's complex hydrological dynamics, this architecture is especially advantageous.

The Transformer model demonstrated suboptimal performance, as evidenced by its reduced R^2 value of 0.8304. There is, however, a tendency to over fit training patterns between training and testing phases.

It is important to align model complexity with data volume in order to maximize model performance. In addition to seasonal variability, the models successfully addressed the pronounced seasonal variation with an average winter flow of 391.5 m^3/s versus an autumn minimum of 56.2 m^3/s. With a maximum flow of 4721.15 m^3/s, the LSTM-GRU's reliability during extreme conditions is crucial for flood management. During peak events, negative PBIAS values indicate systematic underestimation (-14.34% to -0.86%). It may be safer for flood control, but it may lead to suboptimal water allocation.

A sensitivity analysis was conducted, which revealed precipitation-agriculture interactions as the dominant factor (importance=0.999). Irrigation systems are vulnerable to weather variability. Uncertainty in precipitation requires adaptive management strategies. Model predictions are validated by inflow-dam volume correlation (r = 0.809). With LSTM-GRU, managers can implement real-time operational forecasting while strategically prioritizing monitoring efforts on critical variables.

3.8. Operational decision-making guidelines

In conditions of drought, marked by protracted periods of low-flow (< 60 m^3/s), the LSTM-GRU model demonstrates superior accuracy (RMSE = 29.73 m^3/s), ensuring reliable forecasts for the optimization of water allocation among competing demands. During such periods, agricultural water releases should be prioritized by managers based on the model's precipitation-agriculture interaction sensitivity (importance = 0.999). Conversely, during flood conditions (>300 m^3/s), the systematic underestimation tendency (PBIAS = -14.34%) necessitates conservative interpretation, suggesting that managers should implement precautionary measures exceeding model predictions by 15-20%.

The model's robust performance during extreme events (maximum recorded 4,721.15 m^3/s) enables proactive flood management, while the seven-fold seasonal variability necessitates adaptive reservoir operation strategies. Real-time implementation should incorporate ensemble forecasting during transitional seasons, when prediction uncertainty is highest. This will ensure operational resilience under changing hydrological conditions.

3.9. Operational decision support framework

It is imperative for dam managers to implement LSTM-GRU predictions through existing SCADA systems with 15-minute update intervals. In conditions of drought, with flows measuring less than 60 cubic meters per second, the model's 96% accuracy facilitates the confident determination of water allocation strategies.

In the context of flood management, the systematic underestimation of 14.34% necessitates safety margins of 20% above predicted peaks. The financial implications of this integration are significant, with costs ranging from $50,000 to $75,000 for conventional multipurpose dams. The payback period for this investment is estimated to be between 18 and 24 months, a period that is reduced through the enhanced hydropower optimization and the mitigation of spill losses.

4. Conclusion

This in-depth study set new standards for hydrological modeling under challenging operational conditions by successfully demonstrating the use of four novel hybrid machine learning architectures in predicting inflow to Jiroft Dam. Among the methods compared, the LSTM-GRU hybrid network proved to be the most successful architecture,

showing outstanding generalization and forecasting performance.

A dual-pathway framework that combined the computational efficiency of GRU with the long-term memory capabilities of LSTM was found to be particularly effective in capturing the complex temporal dynamics of the Jiroft Dam system, which features extreme hydrological events and seasonality. The study uncovered key findings about the hydrological behavior of the system, also recording pronounced seasonal variability with important operational repercussions. Water management systems are prone to climatic variability, as the in-depth sensitivity analysis revealed the predominance of precipitation-related interactions, notably with agricultural water allocation and evaporation processes.

The strong interlinkages between operational variables supported both the physical plausibility of model predictions and the usefulness of individual input features for machine learning purposes. The violin plots for probabilistic distribution evaluation and Taylor diagrams for multi-metric performance visualization were proposed, providing a firm basis for exhaustive model verification beyond scalar metrics. A range of normality, autocorrelation, and heteroscedasticity tests were implemented to verify model reliability. For operational risk planning and uncertain decisions, systematic trends in model forecasts offer important implications.

Jiroft Dam and similar multipurpose reservoirs can benefit greatly from this research in terms of water resource management. Inflow forecasting with the validated LSTM-GRU model is a reliable method for optimizing reservoir operations, improving flood control strategies, and allocating water more efficiently among competing demands, such as irrigation, hydropower generation, and environmental flows.

Real-time decision support systems bridge the theoretical and practical gap by providing computationally efficient solutions. To assess long-term resilience, research efforts should include incorporating climate change scenarios, extending temporal resolution, and developing ensemble approaches. Transfer learning techniques were used to train models that can be applied to other dams in the region. Satellite-based precipitation products and real-time telemetry data can improve prediction accuracy.

AI-driven hydrology is a contribution to the field. Hybrid architectures outperform traditional approaches in a comprehensive evaluation framework that balances scientific rigor and practical applicability. These models can be applied to other dams in the region using transfer learning techniques. Furthermore, satellite-based precipitation products and real-time telemetry data can improve forecast accuracy. The study contributes significantly to AI-driven hydrology. Hybrid architectures are compared to traditional approaches with a comprehensive evaluation framework that balances scientific rigor with practical applicability.

The documented success of hybrid machine learning models, especially those with an LSTM-GRU architecture, is a valuable reference for water resource managers and researchers looking to implement advanced computational solutions for complex hydrological systems. Integrating sophisticated statistical validation with advanced visualization techniques provides a template for future computational hydrology studies. The use of machine learning can transform the way we manage water resources in the future.

Water systems are under increasing pressure from climate variability, population growth, and competing demands. In this study, hybrid neural network architectures were found to capture complex, nonlinear dynamics of hydrological systems while maintaining a level of computational efficiency suitable for operational use. Adaptable and resilient water resource management strategies change with the environment.

The superiority of LSTM-GRU (R^2 = 0.8725) has been demonstrated, providing water resource managers with a validated framework for real-time decision support. For Jiroft Dam's annual hydropower production of 80 GWh, precise inflow prediction facilitates optimal turbine scheduling, with the potential to enhance efficiency by 8-12% during periods of peak demand.

The model's capacity to manage extreme events (up to 4,721.15 m^3/s) supports the

implementation of flood early warning systems, providing a 24-48-hour advance notice that is crucial for the implementation of downstream evacuation protocols. Integration with existing SCADA systems requires minimal computational overhead, making the approach scalable to Iran's 180+ major dams facing similar hydrological challenges.

It is imperative to acknowledge the potential risks associated with several modeling methodologies. The systematic negative PBIAS (-14.34% to -20.86%) indicates consistent underestimation of peak flows, which has the potential to compromise flood safety if not properly calibrated with safety factors. The training-testing R^2 degradation (0.992 to 0.873 for LSTM-GRU) suggests moderate overfitting despite the implementation of regularization techniques. The temporal scope of the study, which is limited to 14 years, may not encompass multi-decadal climate cycles, thereby restricting the generalizability of the model to unprecedented hydrological conditions.

The geographic transferability of these models remains unvalidated, necessitating site-specific recalibration for different watersheds or climate regimes.

5. Disclosure Statement

According to the authors, there were no potential conflicts of interest

6. References

Ahrari, A., Sharifi, A., & Haghighi, A. T. (2024). Anthropogenic vs. climatic drivers: Dissecting Lake desiccation on the Iranian plateau. *Journal of environmental management*, *368*, 122103.

Alhussein, M., Aurangzeb, K., & Haider, S. I. (2020). Hybrid CNN-LSTM model for short-term individual household load forecasting. *Ieee Access*, *8*, 180544-180557.

Damansabz, A., Khajeh, M., Piri, J., & Ghaffari-Moghaddam, M. (2025). A novel GWO-DE-LSTM hybrid model for predicting statin drug solubility in supercritical carbon dioxide: a comparative analysis with traditional machine learning approaches. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *14*(1), 1-16.

Deb, D., Arunachalam, V., & Raju, K. S. (2024). Daily reservoir inflow prediction using stacking ensemble of machine learning algorithms. *Journal of hydroinformatics*, *26*(5), 972-997.

Farhadi, A., Zamanifar, A., Alipour, A., Taheri, A., & Asadolahi, M. (2025). A hybrid LSTM-GRU model for stock price prediction. *Ieee Access*.

Galassi, A., Lippi, M., & Torroni, P. (2020). Attention in natural language processing. *IEEE transactions on neural networks and learning systems*, *32*(10), 4291-4308.

Granata, F., & Di Nunno, F. (2025). Pathways for hydrological resilience: Strategies for adaptation in a changing climate. *Earth Systems and Environment*, 1-29.

Jiang, D., & Wang, K. (2019). The role of satellite-based remote sensing in improving simulated streamflow: A review. *Water*, *11*(8), 1615.

Keshtegar, B., Piri, J., & Kisi, O. (2016). A nonlinear mathematical modeling of daily pan evaporation based on conjugate gradient method. *Computers and Electronics in Agriculture*, *127*, 120-130.

Kim, S., Choi, K., Choi, H.-S., Lee, B., & Yoon, S. (2022). Towards a rigorous evaluation of time-series anomaly detection. Proceedings of the AAAI conference on artificial intelligence,

Li, X., Yang, X., Wang, X., & Deng, C. (2024). Agree to disagree: Exploring partial semantic consistency against visual deviation for compositional zero-shot learning. *IEEE Transactions on Cognitive and Developmental Systems*, *16*(4), 1433-1444.

Liang, E., Tang, H., Liu, Y., Liu, S., Wu, J., Pan, W., Shang, Y., & Yin, S. (2025). A global synthesis reveals the role of strategic hydropower planning in mitigating adverse impacts of reservoir flooding. *Renewable and Sustainable Energy Reviews*, *217*, 115723.

Livieris, I. E., Pintelas, E., & Pintelas, P. (2020). A CNN–LSTM model for gold price time-series forecasting. *Neural computing and applications*, *32*(23), 17351-17360.

Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information*, *15*(9), 517.

Ortiz-Partida, J. P., Fernandez-Bou, A. S., Maskey, M., Rodríguez-Flores, J. M., Medellín-Azuara, J., Sandoval-Solis, S., Ermolieva, T., Kanavas, Z., Sahu, R. K., & Wada, Y. (2023). Hydro-economic modeling of water resources management challenges: Current applications and future directions. *Water Economics and Policy*, *9*(01), 2340003.

Piri, J., & Kisi, O. (2024). Hybrid non-linear probabilistic model using Monte Carlo simulation and hybrid support vector regression for evaporation predictions. *Hydrological Sciences Journal*, 1-29.

Pokharel, S. (2025). *Towards Advancing Streamflow and Peak Flow Prediction With Machine Learning: Identifying Infrastructure at Risk* The University of Nebraska-Lincoln].

Qian, X., Wang, B., Chen, J., Fan, Y., Mo, R., Xu, C., Liu, W., Liu, J., & Zhong, P.-a. (2025). An explainable ensemble deep learning model for long-term streamflow forecasting under multiple uncertainties. *Journal of Hydrology*, 133968.

Rithani, M., Kumar, R. P., & Doss, S. (2023). A review on big data based on deep neural network approaches. *Artificial Intelligence Review*, *56*(12), 14765-14801.

Sajjad, M., Khan, Z. A., Ullah, A., Hussain, T., Ullah, W., Lee, M. Y., & Baik, S. W. (2020). A novel CNN-GRU-based hybrid approach for short-term residential load forecasting. *Ieee Access*, *8*, 143759-143768.

Smith, J. D., Koenig, L. E., Sleckman, M. J., Appling, A. P., Sadler, J. M., DePaul, V. T., & Szabo, Z. (2024). Predictive Understanding of Stream Salinization in a Developed Watershed Using Machine Learning. *Environmental Science & Technology*, *58*(42), 18822-18833.

Suzauddola, M., Zhang, D., Zeb, A., Chen, J., Wei, L., & Rayhan, A. S. (2025). Advanced deep learning model for crop-specific and cross-crop pest identification. *Expert Systems with Applications*, *274*, 126896.

Thrun, M. C., Gehlert, T., & Ultsch, A. (2020). Analyzing the fine structure of distributions. *PloS one*, *15*(10), e0238835.

Ullah, K., Ahsan, M., Hasanat, S. M., Haris, M., Yousaf, H., Raza, S. F., Tandon, R., Abid, S., & Ullah, Z. (2024). Short-term load forecasting: A comprehensive review and simulation study with CNN-LSTM hybrids approach. *Ieee Access*.

Uppalapati, S., Paramasivam, P., Kilari, N., Chohan, J. S., Kanti, P. K., Vemanaboina, H., Dabelo, L. H., & Gupta, R. (2025). Precision biochar yield forecasting employing random forest and XGBoost with Taylor diagram visualization. *Scientific Reports*, *15*(1), 7105.

Wang, W.-c., Gu, M., Hong, Y.-h., Hu, X.-x., Zang, H.-f., Chen, X.-n., & Jin, Y.-g. (2024). SMGformer: integrating STL and multi-head self-attention in deep learning model for multi-step runoff forecasting. *Scientific Reports*, *14*(1), 23550.

Wang, X., Zhou, J., Ma, J., Luo, P., Fu, X., Feng, X., Zhang, X., Jia, Z., Wang, X., & Huang, X. (2024). Evaluation and comparison of reanalysis data for runoff simulation in the data-scarce watersheds of alpine regions. *Remote Sensing*, *16*(5), 751.

Zhang, D.-D., & Xu, J. (2024). Long-term monitoring of surface water dynamics and analysis of its driving mechanism: A case study of the Yangtze River Basin. *Water*, *16*(5), 677.