



A Hybrid Feature-Engineered XGBoost Framework for Rainfall-Runoff Based Inflow Prediction of Ukai Dam, India

Bhavini Shah^{1*}, Dr. Kamalendu Pandey²

^{1*}Department of Computer Science, Veer Narmad South Gujarat University, Surat, Gujarat, India. E-mail: bhavishah5483@gmail.com

²Department of M.Sc(IT/ICT), Veer Narmad South Gujarat University, Surat, Gujarat, India. E-mail: kspandey@vnsgu.ac.in

ABSTRACT

Precise inflow prediction is very important for water resource management, flood control and hydropower optimization, especially for strategically important reservoir like Ukai Dam in Gujarat, India. This study proposes a novel hybrid feature-engineered XGBoost model tailored to the hydrological and meteorological attributes of the Ukai catchment. The model integrates an enhanced Soil Conservation Curve Number (SCS-CN) method, incorporating with LULC, geomorphology, slope and elevation data, alongside a slope-based runoff adjustment factor to better represent watershed variability. Advanced feature engineering techniques, including lag features, rolling statistics, seasonal indicators, and reservoir state variables, are employed to capture temporal dynamics. Variance stabilization is achieved through low-power transformation, while peak-flow-aware weighting enhances the prediction of extreme inflow events. The structured modelling workflow ensures reproducibility and scalability. Evaluation of model is done by matrices like RMSE, MAE, R^2 and NSE. The findings of proposed model is compared against traditional models like Decision Tree, Gradient Boosting, Random Forest and Long-Short-Term Memory. The proposed hybrid framework achieved excellent predictive accuracy ($R^2=0.992$, $NSE=0.992$) with low prediction error ($RMSE=7.415$, $MAE=0.977$) outperforming Decision Tree, Gradient Boosting, Random Forest and LSTM models and demonstrated robust capability for reservoir inflow forecasting. The findings make the proposed model as reliable tool for flood forecasting, reservoir operation and sustainable water management.

Keywords: Reservoir inflow forecasting; XGBoost; Rainfall-runoff modelling; Hydrology; Machine learning; Ukai Dam

1. Introduction

Accurate reservoir inflow prediction is a critical component of modern water resource management. Reliable forecasts support reservoir operation, hydropower scheduling, irrigation planning, and drought management and flood mitigation. Increasing climatic variability and changing precipitation patterns have further increased the importance of developing robust forecasting systems capable of handling nonlinear hydrological processes.

Traditional hydrological models rely on physically-based representations of watershed processes. Although these models provide interpretability, they often require extensive calibration and detailed watershed information. Machine learning methods offer an attractive alternative because they can learn complex nonlinear relationships directly from historical observations.

Among contemporary machine learning techniques, XGBoost has emerged as one of the most successful algorithms for environmental prediction due to its ability to model nonlinear interactions, handle heterogeneous datasets and provide high computational efficiency. However, purely data-driven models often fail to fully exploit physical hydrological knowledge.

This study proposes a hybrid feature-engineered XGBoost framework that combines physically-based runoff estimation with advanced machine learning techniques for inflow forecasting at Ukai Dam.

The major contributions of this study are:

1. Integration of enhanced SCS-CN runoff modelling with machine learning.
2. Development of a hybrid feature engineering framework.
3. Introduction of peak-aware learning for extreme event prediction.
4. Comprehensive comparison against classical and deep learning models.
5. Application to a strategically important reservoir system in India.

2. Literature Review

The hydrological effect of Land Use/Land Cover in Geba Catchment of Ethiopia was assessed previously in [1]. They developed a model by integrating parsimonious hydrological modeling, Alteration of Hydrological Modeling (IHA) and Partial Least Square Regression (PLSR) analysis. This study demonstrated how environmental changes affects hydrology in complex terrains. They developed an integrated approach for the same. Stream flow dynamics, runoff, infiltration are being altered by human induced Land Use/Land Cover. Different land cover types are agriculture (39%), bush/shrub land (30%), bare land (11%), grassland

(7%), wooded land (6%), and forest (1.7%) with changes occurring due to population pressure and watershed management interventions leading to partial vegetation recovery since the mid-1990s. As an input, climate data like precipitation and evapotranspiration were used. Spatial dataset included DEM, soil and Land Use/Land Cover. PLSR model identified agricultural and bare land as major factor for stream flow changes while forest, bush land, and woodland promote dry season flows and soil water storage. The study showed changes in LULC last 40 years affects hydrology, land degradation and vegetation loss increased surface runoff and reduced low flows. But watershed management and vegetation recovery reversed some of effects later.

The runoff at Xiaoqing River Basin was predicted using XGBoost interpreted by SHAP and SWAT model in [2]. The input variables used by system are topography, climate, land use and soil type. Then XGBoost model predicts runoff generation which is further interpreted by SHAP. The result showed that XGBoost can simulate the ability of SWAT and by interpreting XGBoost using the SHAP the factors affecting runoff generation can be identified. Meteorological attributes have different effect on runoff and rainfall at high elevation can increase runoff. Runoff generation is affected by effect of Land Use and Slope. This study provided new method for attribute analysis of runoff.

The inflow of Anadong and Inha Dam was predicted using various deep learning models like ANN, DNN, LSTM and GRU in [3]. The dam hydrology data of 20 years were used to estimate inflow of dam and optimal input variables are derived through various indicators. They achieved R^2 values 0.91-0.94 using ANN based scenario which required multiple configurations. They compared different algorithms like Linear Regression (LR), Gradient Boosting Regressor, Artificial Neural Network, Random Forest Extra Trees Regressor, XGBoost and Long Short-Term Memory (LSTM) for forecasting of daily stream flow. The result showed that LSTM got best performance with maximum NSE and Kling Gupta Efficiency (KGE). Then ANN and LR were at the second position.

The inflow was predicted using LSTM and CNN based Encoder-Decoder model in [4]. The model was model compared with traditional statistical methods like SARIMA, VAR and TBATS. The performance of this model is shown in Table 1.

The Peak Oriented Forecasting by Model Switching (PForMS) was developed in [5] which was a hybrid approach to combine the traditional machine learning and Deep learning methods. They used dam data of JAPAN and USA. Their result is shown in Table 1.

The daily inflow of reservoir using two step algorithm in [6]. In the first step, they had used time series Auto Regressive Integrated Moving Average Model to forecast monthly inflow. This inflow is used as an input further for month-wise daily forecasting of inflow that model combine averaging and boosting methods in machine learning. Result of this model is shown in Table 1.

The hybrid Wavelet Artificial Neural Network was developed to predict short-term daily inflow using Tropical Rainfall Measuring Mission (TRMM) data in [7].

The inflow of Upper Indravati Multipurpose Project was predicted in [8]. The authors concluded that ANN was good at prediction of high flow events while ARIMA was good at prediction of low flow events. They framed reservoir policies through Dynamic Programming.

The framework called Meteorological data long short-term memory (M-LSTM) which is used for monthly inflow forecast was developed in [9]. They selected Xiaowan as case study station for their research. They compared the result of different models like M-LSTM, M-BPNN and M-SVR. Their result is shown in Table 1.

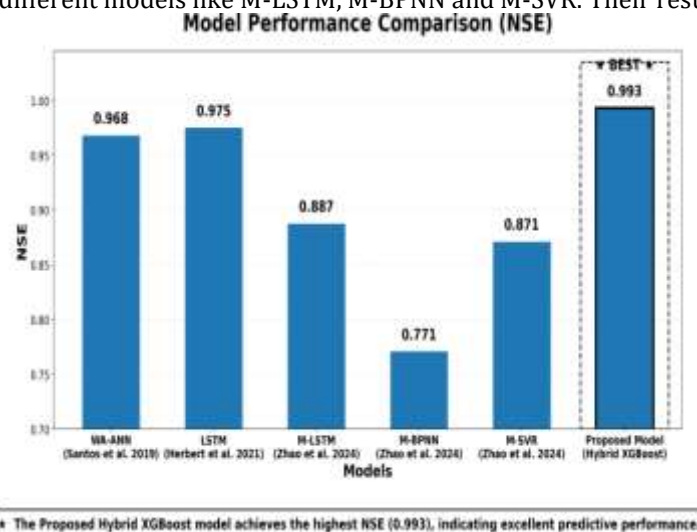


Fig. 1 Model Performance Using NSE matrix.

Table 1 Comparison of proposed model with existing studies

Model	RMSE	MAE	R ²	NSE
Hybrid XGBoost Model (Proposed)	7.415	0.977	0.992	0.992
ANN (Anadong Dam Daily) [3]	–	–	0.94	–

LSTM [4]	–	–	–	0.975
PForMS Model [5]	989.54	–	–	–
Two-Step Model [6]	15.26	9.19	–	–
WA-ANN Model [7]	–	–	0.968	0.968
ANN Model [8]	–	–	0.914–0.949	–
ARIMA Model [8]	–	–	0.824–0.973	–
M-LSTM [9]	247.89	186.29	0.94	0.88
M-BPNN [9]	357.06	267.71	0.93	0.77
M-SVR [9]	266.97	192.43	0.94	0.87

Figure 1 summarizes the performance of various inflow forecasting models reported in previous studies. Deep learning approaches such as LSTM achieved NSE values up to 0.975, while ANN-based methods reported R^2 values between 0.914 and 0.968. Traditional statistical models such as ARIMA exhibited varying predictive performance depending on hydrological conditions. M-BPNN and M-SVR got R^2 values 0.930 and 0.940 respectively. The proposed Hybrid XGBoost model outperforms most reported approaches with RMSE = 7.415, MAE = 0.977, R^2 = 0.992, and NSE = 0.992. The results demonstrate the effectiveness of integrating physically-based runoff estimation, feature engineering, target transformation, and peak-flow weighting within a gradient boosting framework.

3. STUDY AREA

The Ukai Dam is situated on the Tapi River in Gujarat, India. The Tapi River originates in the Satpura Hills of Madhya Pradesh and traverses Maharashtra and Gujarat before discharging into the Arabian Sea. Figure 2 shows study area Ukai Dam within Tapi Basin map.

The reservoir plays a vital role in:

- Flood control,
- Hydropower generation,
- Irrigation supply,
- Municipal water distribution.

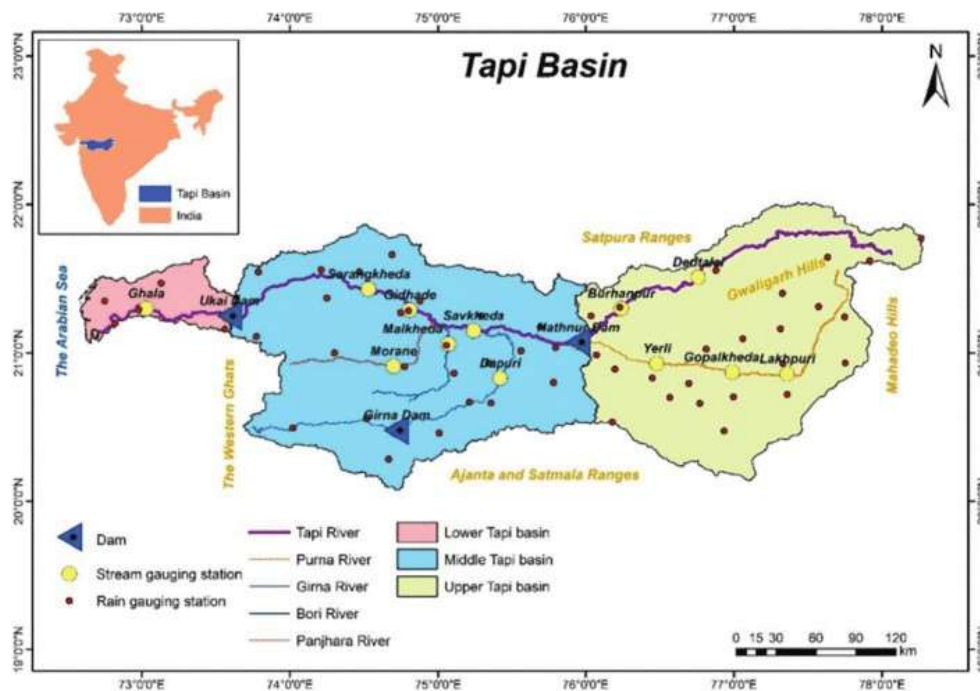


Fig. 2 Location of the Ukai Dam within the Tapi River Basin.

The Ukai Dam catchment receives substantial amount of rainfall water during June to September which results in highly variable inflow patterns. It is characterized by peak and low flow period. The catchment is having different types of Land-Use and Land-Cover (LULC) conditions, different terrain slopes and diverse geographical characteristics, all of them affect the rainfall-runoff processes. To account for watershed variability, runoff was calculated using modified Soil Conservation Service-Curve Number (SCS-CN) method that used land-use, geomorphological, slope and elevation information.

4. Data Collection And Preprocessing

Data were collected from multiple sources representing hydrological, meteorological, and spatial and reservoir characteristics.

We have collected reservoir data from Ukai Dam station. These data contains attributes like Opening Capacity, Average Discharge, Total Volume, Balance Capacity and Inflow of dam. The reservoir operational and hydro meteorological data covering the period 1985-2025 were collected from Ukai Dam and upstream monitoring stations. The raw dataset contained approximately 200,866 reservoir records and 278,899 runoff related records. Following data cleaning, temporal aggregation, pre-processing and integration of station-wise observations, the final modelling dataset consisted of 13,726 daily observations. These processed daily observations were subsequently used for rainfall-runoff estimation and reservoir inflow prediction.

Meteorological data like Minimum Temperature, Maximum Temperature, Dry Bulb Temperature, Wet Bulb Temperature, Pan Water Temperature, Evaporation, Relative Humidity and Wind Speed are collected from 52 Upstream Stations of Ukai Dam. These data are collected from State Water Data Centre, Gandhinagar, Gujarat and Water Resource Department Government, Nashik, Maharashtra. However, some of these 52 upstream stations did not have dedicated hydro meteorological station located in their area, so we have considered nearby hydro meteorological station within a 20 km radius of the original upstream station.

Spatial data like Elevation is collected from Google Earth, Distance is measured using Google Map. Using elevation and distance data slope is calculated and used as one of the spatial parameter.

Watershed characteristics like Land Use/Land Cover and Geomorphology were gathered from bhuvan.nrsc.gov.in

5. Methodology

Figure 3 presents the overall workflow of proposed system.

5.1 Runoff Estimation Using SCS-CN

Runoff is calculated using modified Soil Conservation Service-Curve Number method.

- Curve Number is assigned based on Land Use/Land Cover and Geomorphology attribute.
- Standard Rainfall-runoff relationship is applied using rainfall and retention parameter. Potential maximum retention:

$$S = \frac{25400}{CN} - 254 \quad (1)$$

Where S is maximum retention and CN is Curve number which is calculated using Land type and geomorphological characteristics. Initial abstraction:

$$I_a = 0.2S \quad (2)$$

Where I_a represents initial abstraction Runoff estimation:

$$Q = \begin{cases} 0, & P \leq I_a \\ \frac{(P-I_a)^2}{P-I_a+S}, & P > I_a \end{cases} \quad (3)$$

Where Q is runoff depth and P is rainfall depth.

- A slope-based adjustment factor is introduced:
Higher Slope → Increased Runoff Generation
Lower Slope → Increased Infiltration and Reduced Runoff

Slope correction:

$$Q_{adj} = \begin{cases} 1.2Q, & \text{slope} > 10\% \\ 0.8Q, & \text{slope} < 3\% \\ Q, & \text{otherwise} \end{cases} \quad (4)$$

Above will generate Adjusted Runoff, which will be used as input for inflow prediction.

5.2 Data Integration

- Runoff dataset and hydro meteorological datasets are merged using date column.
- Ensure temporal alignment of all variables.

5.3 Data Preprocessing

The collected dataset for Ukai Dam and 52 upstream stations includes hydrological and meteorological observations recorded at two time intervals each day, i.e., 8:30 AM and 5:30PM. To manage consistency and

obtain a representative daily value, the observations from both time instances were averaged to generate a single daily record for each parameter. Similarly, data from multiple upstream monitoring stations were processed. A total of 52 upstream stations were considered, each providing daily meteorological inputs like:

- Maximum Temperature (MTX)
- Minimum Temperature (MTN)
- Rainfall (MPS)
- Relative humidity (MHS)
- Wet Bulb Temperature (MTW)
- Dry Bulb Temperature (MTD)
- Wind Speed (MWS)

While the attributes related to reservoir are as below:

- Opening Capacity
- Average Discharge
- Total Volume
- Balance Capacity
- Inflow of dam

Since data from both Ukai dam and upstream stations contained missing values, appropriate imputation techniques were applied to ensure data quality. The missing values were handled using Forward-Fill technique, where missing data were replaced with the most recent available previous value. This method preserves the temporal continuity and is suitable for hydrological time-series data. This will help to preserve the overall data distribution without introducing bias. Data from multiple upstream stations were aggregated by calculating daily average values for each meteorological parameter. This generated a single representative observation per day, which was then used for



Fig. 3 Proposed Hybrid Feature Engineered XGBoost Framework For Reservoir Inflow Forecasting subsequent rainfall-runoff and inflow prediction modeling. The date attribute was converted into standard date time format and the dataset was arranged chronologically to preserve the temporal order required for

time-series forecasting. These preprocessing steps ensured:

- Temporal consistency in the dataset
- Reduction of noise and missing data issues
- Effective representation of spatial variability across upstream stations

5.4 Outlier handling

To reduce effects of extreme inflow observation and improve model stability, inflow values are capped at the 99.5 percentile threshold.

$$\text{Inflow}_{\text{capped}} = \min(\text{Inflow}, Q_{0.995}) \quad (5)$$

5.5 Feature Engineering

Feature engineering plays a crucial role in improving the forecasting behavior of any machine learning model by converting raw hydrological data into meaningful input attribute. In this study, both temporal and statistical features are generated to capture the dynamic behavior of reservoir inflow. Lagged inflow variables (1,2,3,5, and 7 days) were included to capture reservoir memory effects and persistence characteristics. To represent short-term hydrological effects and cumulative watershed response seven-day rolling averages of rainfall and runoff were calculated. A month attribute derived from observation date was included to capture the seasonal variability was achieved. The final feature set included rainfall, runoff, seven-day rainfall average, seven-day runoff average, opening capacity, average discharge, balance capacity, total volume of reservoir, month and multiple inflow lag variables. The target variable inflow was transformed using logarithmic power transformation defined as $\log(1 + \text{Inflow}^{0.7})$. It was used to reduce heteroscedasticity and stabilize variance. The Extreme Gradient Boosting (XGBoost) was used to implement the prediction model with 800 decision trees, maximum tree depth of 5, learning rate of 0.025, sub sampling ratio of 0.9, column sampling ratio of 0.9, and minimum child weight of 5 and gamma value of 0.1. To improve prediction accuracy in extreme hydrological events, all the peak-flow observation more than 90th percentile of historical inflows were assigned twice the training weight of normal-flow observation. Gain-based importance metric provided by XGBoost was used to perform the feature importance analysis. According to the analysis, lag variable, runoff estimates, rainfall indicators, and reservoir variables are most effective predictors. To determine future inflow behavior, feature ranking provides insights into the relative contribution of hydrological persistence, watershed response and reservoir operating conditions.

1. Lag features: To capture the temporal dependency and persistence in inflow patterns, lagged values of the inflow variable are included as input parameter. These lag feature represent previous time-step observations. They are defined as below:

$$\text{Lag}_k = \text{Inflow}_{t-k} \quad (6)$$

Where Inflow_{t-k} represents the inflow observed at the previous time step. The inclusion of lag features enables the model to learn autoregressive relationships that are fundamental to hydrological systems. Since current inflow conditions are strongly influenced by antecedent conditions, lag variables help capture temporal dependencies and short-term memory effects, thereby improving forecasting accuracy.

2. Rolling mean: To reduce noise and consider the cumulative hydrological effects, moving average features are calculated over a 7-day window for runoff and rainfall variables.

$$\text{RollingMean}_7 = \frac{1}{7} \sum_{i=0}^6 \text{Inflow}_{t-i} \quad (7)$$

The rolling statistics remove short-term fluctuations and highlight underlying trends in the data. In hydrological system, the rainfall and runoff often exhibit delayed and accumulated effects on reservoir inflow. So, moving average gives a more realistic representation of watershed response.

3. Seasonal features: From the date attribute month feature was extracted which captures seasonal variability.

$$\text{Season}_t = \text{Month}(\text{Date}_t) \quad (8)$$

5.6 Target Transformation

A nonlinear transformation is also applied to stabilize variance, reduce skewness and improve learning behavior for extreme inflow conditions.

$$y_t = \log(1 + \text{Inflow}^{0.7}) \quad (9)$$

This transformation is combined with power transformation (0.7) and Log transformation. With this target transformation skewness can be reduced. With the help of transformation different inflow conditions like low, medium and high can be learned. It also stops large peaks from dominating loss.

Inverse transformation: After prediction of inflow, inverse transformation was applied to recover inflow values in the original scale.

$$Inflow = (e^y - 1)^{1/0.7} \quad (10)$$

5.7 Peak Flow Weighting

$$w_i = \begin{cases} 2, & Inflow_i > P_{90} \\ 1, & otherwise \end{cases} \quad (11)$$

5.8 Time-Series Cross Validation

As the reservoir inflow data has strong temporal dependency, conventional K-fold cross-validation may lead to information leakage by allowing future observations to influence model training. A five-fold TimeSeriesSplit validation strategy was adopted to preserve temporal ordering and avoid information leakage. The complete data set is divided into 5 sequential folds using TimeSeriesSplit method. During each iteration each training set contained historical data available up to specific time period and while the subsequent data was used for validation. This expanding window validation framework mimics real world scenario in which past information is available to predict the future inflows. For each fold XGBoost Hybrid regression model is trained using training data set and evaluated using corresponding validation data set.

5.9 XGBoost Model

The inflow prediction model was implemented using (Extreme Gradient Boosting) algorithm which is represented as following:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (12)$$

Where f_k represents individual boosted trees. The final hyperparameters were selected experimentally based on TimeSeriesSplit validation performance. The model parameters were set as in Table 2.

Table 2 XGBoost Hyperparameter Configuration

Parameter	Value	Description
n_estimators	800	Number of Boosting Trees
max_depth	5	Maximum depth of each tree
learning_rate	0.025	Step size shrinkage
subsample	0.9	Fraction of training samples used
colsample_bytree	0.9	Fraction of features used per tree
min_child_weight	5	Minimum sum of instance weights in a node
gamma	0.1	Minimum loss reduction for split
tree_method	hist	Histogram-based tree construction

To improve model generalization and reduce over fitting, relatively small learning rate is combined with large number of trees. To increase model robustness, sub-sampling and column sampling were included, whereas the minimum child weight and gamma parameters were used to control tree complexity. To improve computational efficiency, the histogram-based tree construction was employed in large-scale hydrological datasets.

To evaluate the effectiveness of proposed hybrid feature-engineered XGBoost framework, its performance was compared with several commonly used machine learning models including Decision Tree (DT), Random Forest(RF), Gradient Boosting Regressor (GBR) and Long Short-Term Memory (LSTM). All models were trained using the same input features and evaluated using the identical TimeSeriesSplit validation framework.

Decision Tree:

- Maximum depth=10

- minimum sample split=5
- minimum sample leaf=2

Random Forest:

- trees=500
- maximum depth=15
- bootstrap sampling enabled

Gradient Boosting:

- estimators=500
- learning rate=0.05
- maximum depth=5

LSTM:

- Input sequence length=7 days
- Two LSTM layers with 64 and 32 units
- dropout rate=0.2
- Adam optimizer ,
- batch size=32
- 100 training epochs.

6. Performance Evaluation

The performance of hybrid Runoff-Rainfall prediction model is evaluated using statistical metrics like Root Means Squared Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2) and Nash-Sutcliffe Efficiency (NSE). The model was validated by 5-fold Time Series Cross Validation to ensure robust temporal evaluation and prevent data leakage. Root Mean Square Error:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (13)$$

Mean Absolute Error:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (14)$$

Coefficient of Determination:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (15)$$

Nash-Sutcliffe Efficiency:

$$NSE = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (16)$$

7. Results And Discussion

The proposed hybrid runoff-rainfall estimation framework was implemented using enhanced SCS-CN based runoff estimation and Extreme Gradient Boosting based Time-series inflow forecasting. The model was evaluated using 5-fold Time-series cross validation to preserve the chronological dependency and to avoid temporal data leakage. The performance of proposed model is evaluated using statistical measures like RMSE, MAE, R^2 and NSE. The result of the proposed model are presented in Table 3 with and without Lag Features. Feature Importance of each feature is shown in Figure 5. Based on the comparison, the model using lag feature significantly outperformed the model without lag feature. The mean RMSE of model with lag feature is approximately 7.415 while the model without lag feature got RMSE 5409.4. The MAE is 0.977 for model with lag feature versus 629.402 without lag feature. Furthermore, the R^2 and NSE scores for the model with lag feature are both high around 0.992 that indicates high fit and predictive power. While in contrast, the model without lag feature exhibits negative R^2 and NSE (-17602.368), which indicating it performs worse than simply predicting the mean. This study highlights the critical importance of lag feature for accurate inflow prediction in this time series model. Table 3 shows the critical role of lag features in inflow forecasting. The inclusion of antecedent inflow information substantially improves predictive performance, reducing prediction errors and increasing both NSE and R^2 values.

Table 3 Impact of lag features on model performance

Model	RMSE	MAE	R ²	NSE
With Lag Features	7.415	0.977	0.992	0.992
Without Lag Features	5409.4	629.4	-17602	-17602

Table 4 summarizes the comparative performance.

Table 4 Performance comparison of forecasting models

Metric	Hybrid	DT	RF	GB	LSTM	RNN
RMSE	7.415	25	39	73	31	32
MAE	0.977	5	13	23	5	6
R ²	0.992	0.40	0.50	0.82	0.08	0.109
NSE	0.992	0.40	0.50	0.82	0.08	0.109
Time(min)	4.05	0.22	67.94	21.26	21.98	6.3

The performance evaluation of different predictive models used for inflow prediction at Ukai Dam is summarized in Table 4. The models considered include the Hybrid XGBoost model, Decision Tree, Random Forest, Gradient Boosting, and LSTM, with their respective evaluation metrics: RMSE, MAE, R² and NSE.

The Hybrid XGBoost model significantly outperformed all models in terms of predictive accuracy. Hybrid XGBoost model achieved RMSE (7.415) and MAE (0.977). It indicates that model achieved very small error in prediction of inflow. Model also got highest coefficient of determination (R²=0.992) and Nash-Sutcliffe Efficiency (NSE=0.992) which shows an excellent ability to capture inflow dynamic.

The decision tree exhibited least execution time (0.22 min) among all algorithms. But its prediction accuracy was poor and got NSE and R² value of 0.40 only. The result indicates that single decision tree is not able to model nonlinear relationship which controls reservoir inflow.

The Random Forest improved upon Decision Tree but required the highest computational time (67.94 minutes). Despite the training cost and complexity, the performance of Random Forest is poor with NSE and R² value of 0.50 only.

Gradient Boosting had error RMSE=73 and MAE=23 with very low NSE and R² with value 0.82 which is better among all except Hybrid model.

While LSTM achieved RMSE=31 and MAE=5 with NSE and R² value 0.08. It shows poor capability to model inflow dynamics.

The last RNN model got RMSE=32 and MAE=6. While NSE and R² got 0.109 which is lowest among all models.

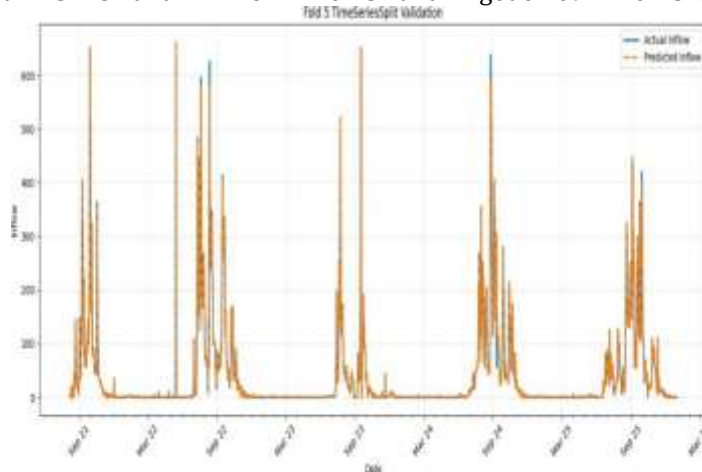
**Fig. 4 Observed and predicted inflow comparison.**

Figure 4 presents a comparison between the observed and predicted inflow values generated by the proposed Hybrid XGBoost framework. The results demonstrate a high degree of agreement between the two series, with model accurately capturing both low-flow and peak-flow conditions. The model efficiency has been improved by including runoff estimation, lag feature, rolling statistics and reservoir state variables. In hydrological systems, during peak flow events some discrepancies are observed attributed to sudden meteorological variation and inherent uncertainties. Overall, the figure confirms the robustness and reliability of the proposed model in reservoir forecasting.

Feature importance analysis revealed that lagged inflow variables were the most influential predictors, confirming the strong temporal persistence of reservoir inflow. Then the runoff related variables ranked second which highlights watershed response

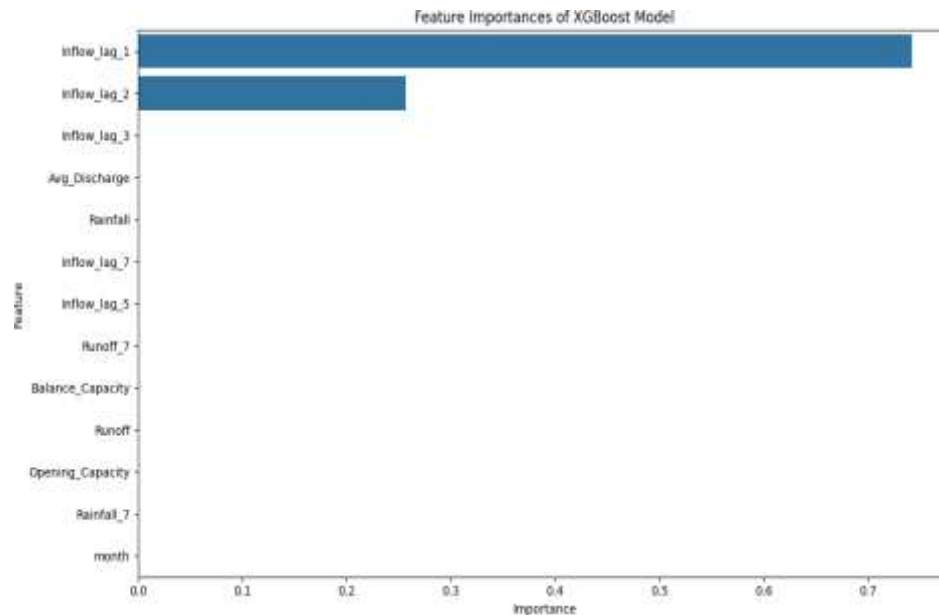


Fig. 5 Feature importance obtained from XGBoost.

characteristics. Further, to improve prediction performance reservoir operation parameters like balance capacity and opening capacity were also included. The seasonal indicators exhibited relatively lower importance. These findings demonstrate that inflow behavior are controlled by hydrological memory and reservoir operating conditions.

8. Implications For Reservoir Management

The proposed framework provides substantial practical value for:

- Real-time flood forecasting.
- Reservoir operation optimization.
- Water allocation planning.
- Hydropower scheduling.
- Climate resilience assessment.

9. Conclusions

This study developed a hybrid feature-engineered XGBoost framework for inflow prediction at Ukai Dam. By integrating physical runoff estimation with advanced machine learning techniques, the proposed framework achieved excellent predictive accuracy. The obtained results (RMSE = 7.415, MAE = 0.977, $R^2 = 0.992$ and NSE = 0.992) demonstrate that hybrid approaches can significantly outperform conventional machine learning and deep learning models. Future research may focus on incorporating radar rainfall products, satellite observations and transformer-based architectures to further enhance forecasting performance.

Declarations

Ethics approval and consent to participate

Not applicable. This study did not involve human participants, human data, human tissue, or animals.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and analyzed during the current study were obtained from the Ukai Dam Authority, State Water Data Centre (SWDC), Gandhinagar, Gujarat, and the Water Resources Department, Nashik, Maharashtra. The processed datasets are available from the corresponding author on reasonable request, subject to the permissions and data-sharing policies of the respective authorities.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable. No external funding was received for this research.

Authors' contributions

Bhavini Shah conceived the study, collected and processed the data, developed the rainfall-runoff model, implemented the machine learning framework, performed the experiments, analyzed the results, and drafted the manuscript. Dr. Kamlendu Pandey supervised the research, contributed to the study design and methodology, reviewed the analysis, and critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors gratefully acknowledge the Ukai Dam Authority for providing reservoir operation data, the State Water Data Centre (SWDC), Gandhinagar, Gujarat, for hydrological data support, and the Water Resources Department, Nashik, Maharashtra, for providing hydro-meteorological datasets used in this study. The authors acknowledge the use of AI-assisted tools for figure drafting and manuscript formatting. Responsibility for all scientific content remains with the authors.

Authors' information

Bhavini Shah is a postgraduate researcher in Computer Science at Veer Narmad South Gujarat University, Surat, India. Dr. Kamlendu Pandey is a faculty member in the Department of M.Sc. (IT/ICT), Veer Narmad South Gujarat University, Surat, India, with research interests in machine learning, data analytics, and intelligent systems.

References

1. Gebremicael, T.G., Mohamed, Y.A., Van Der Zaag, P.: Attributing the hydrological impact of different land use types and their long-term dynamics through combining parsimonious hydrological modelling, alteration analysis and plsr analysis. *Science of The Total Environment* 660, 1155–1167 (2019) <https://doi.org/10.1016/j.scitotenv.2019.01.085>
2. Wang, S., Peng, H., Hu, Q., Jiang, M.: Analysis of runoff generation driving factors based on hydrological model and interpretable machine learning method. *Journal of Hydrology: Regional Studies* 42, 101139 (2022) <https://doi.org/10.1016/j.ejrh.2022.101139>
3. Kim, B.-J., Lee, Y.-T., Kim, B.-H.: A study on the optimal deep learning model for dam inflow prediction. *Water* 14(17), 2766 (2022) <https://doi.org/10.3390/w14172766>
4. Herbert, Z.C., Asghar, Z., Oroza, C.A.: Long-term reservoir inflow forecasts: Enhanced water supply and inflow volume accuracy using deep learning. *Journal of Hydrology* 601, 126676 (2021) <https://doi.org/10.1016/j.jhydrol.2021.126676>
5. Paul, T., Raghavendra, S., Ueno, K., Ni, F., Shin, H., Nishino, K., Shingaki, R.: Forecasting of reservoir inflow by the combination of deep learning and conventional machine learning. In: 2021 International Conference on Data Mining Workshops (ICDMW), pp. 558–565. IEEE, 10.1109/ICDMW53433.2021.00078 (2021).
6. Gupta, A., Kumar, A.: Two-step daily reservoir inflow prediction using arima-machine learning and ensemble models (2020)
7. Santos, C.A.G., Freire, P.K.M.M., Da Silva, R.M., Akrami, S.A.: Hybrid wavelet neural network approach for daily inflow forecasting using tropical rainfall measuring mission data. *Journal of Hydrologic Engineering* 24(2), 04018062 (2019) [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0001725](https://doi.org/10.1061/(ASCE)HE.1943-5584.0001725)
8. Jain, S.K., Das, A., Srivastava, D.K.: Application of ANN for reservoir inflow prediction and operation. *Journal of Water Resources Planning and Management* 125(5), 263–271 (1999) [https://doi.org/10.1061/\(ASCE\)0733-9496\(1999\)125:5\(263\)](https://doi.org/10.1061/(ASCE)0733-9496(1999)125:5(263))
9. Zhao, H., Liao, S., Song, Y., Fang, Z., Ma, X., Zhou, B.: Long-term inflow forecast using meteorological data based on long short-term memory neural networks. *Journal of Hydroinformatics* 26(5), 954–971 (2024) <https://doi.org/10.2166/hydro.2024.196>