Grenze International Journal of Engineering and Technology, June Issue

Daily Inflow and Outflow Prediction for Kabini Dam using Random Forest Regression

Bhavya B M[1] and Raviram V[2]
[1]Department of MCA, Research Scholar, Sri Siddhartha Academy of Higher Education, Tumakuru, India
Email: bbmgowda@gmail.com, bhavyabm@pesce.ac.in
[2]Department of CSE, Sri Siddhartha Institute of Technology, Tumakuru, India
Email: raviram@ssit.edu.in

***Abstract*— This work presents a machine learning-based predictive model for estimating the daily inflow and outflow of the Kabini Dam, employing a Random Forest regression technique. The model uses historical hydrometerological and time-based features, including dam water level, rainfall, and date-derived components. Data from 2020 to 2023 were collected from the Karnataka Water resource department and preprocessed through missing value imputation, feature extraction, and target cleaning. Two models-one each for inflow and outflow, were trained using an 80/20 split and evaluated using Mean Squared Error(MSE), Root Mean Squared Error(RMSE), Mean Absolute Error(MAE), and Coefficient of Determination(R2). The inflow model achieved MSE = 82.3 (cumecs)2, RMSE = 9.07 cumecs, MAE = 6.21 cumecs and R2 = 0.93, while the outflow model reported MSE = 95.8 (cumecs)2, RMSE = 9.79 cumecs, MAE = 6.84 cumecs and R2 = 0.91. Compared to a baseline persistence model, Random forest provided a reduction in RMSE. The models were deployed through a Flask API, enabling real-time predictions for decision support. The results establish Random Forest as a robust, operationally feasible method for dam flow forecasting and management.**

***Index Terms*— Kabini Dam, Random Forest, Inflow Prediction, Outflow Prediction, Machine Learning, Hydrological Forecasting.**

I. INTRODUCTION

Water resources are fundamental for human sustenance, agriculture, industry, and energy production. Dams play a pivotal role in managing these resources, regulating water flow, mitigating floods, and generating hydroelectric power. The Kabini Dam, located in a region susceptible to variable rainfall patterns, necessitates precise flow management to optimize its multi-purpose utility. Inaccurate predictions can lead to inefficient water distribution, agricultural losses, or, in severe cases, catastrophic flooding. Traditional hydrological models often rely on complex physical equations and extensive meteorological data, which can be computationally intensive and may not always capture the non-linear dynamics inherent in natural systems. Machine learning, particularly ensemble methods like Random Forest, offers a robust alternative by learning intricate patterns directly from historical data. This paper aims to develop a data-driven model for predicting the daily inflow and outflow of the Kabini Dam, thereby enhancing operational efficiency and preparedness.
Demand for accurate reservoir inflow/outflow prediction is key for water resource management, flood prediction, irrigation, and hydropower scheduling.

Grenze ID: 01.GIJET.12.2.90_21
© Grenze Scientific Society, 2026

Traditional physical-hydrological models are complex and data-intensive. Machine learning offers data-driven alternative models with reduced complexity.India’s water infrastructure relies heavily on reservoirs and dams for agricultural, drinking, and energy purposes. The Kabini Dam is a key component of the Cauvery river basin system. Traditional hydrological models are computationally intensive and often require extensive physical data. This research leverages machine learning — specifically, the Random Forest algorithm — to forecast dam inflow and outflow, using accessible hydrological and meteorological data.

II. Literature Review

Recent studies underscore the effectiveness of data-driven machine learning approaches for hydrological forecasting. While deep learning models such as LSTM and Transformer architectures excel in capturing temporal dependencies, Random Forest (RF) remains highly effective for nonlinear pattern recognition and data-limited scenarios. Studies comparing RF with Support Vector Machines (SVM) have found RF to deliver superior robustness and interpretability for inflow prediction under diverse hydrometerological conditions.

Random Forest is known for its robustness to overfitting and ability to model non-linear relationships, making it suitable for hydrological prediction. The various machine learning algorithms, including Random Forest (RF), are used for predicting inflow into the Soyang River Dam in South Korea. While Multilayer Perceptron (MLP) performed well overall, ensemble models like Random Forest and Gradient Boosting showed better performance for lower dam inflow amounts.: Using machine learning models to predict dam gate operations for flood control. It evaluates various classification and regression algorithms—including Naïve Bayes, Random Forest, ANN, and LSTM—trained on precipitation and dam gate data [1], An analytic approach focuses on enhancing the accuracy of water flow level predictions by applying feature elimination techniques to machine learning models [2].

Model performance was assessed using Root Mean Square Error (RMSE) and compared against existing Support Vector Machine (SVM) models. For inflow prediction, ARIMA demonstrated a 36% reduction in RMSE in the best case, while the worst-case scenario showed a 4.3% increase relative to SVM. In terms of outflow prediction, ARIMA outperformed the traditional Muskingum model by effectively capturing nonlinear flow behaviour, leading to improved forecasting accuracy [3]. To forecast river water temperature at a specific target location, the study leveraged a combination of weather data, dam discharge information (including turbine flow and spill flow), and remotely sourced water temperature readings. Two machine learning models—Gradient Boosting Regression (GBR) and Random Forest Regression (RFR)—were trained and evaluated under two conditions: with and without the spill flow feature. Performance comparisons were made to assess the impact of including spill flow data on predictive accuracy [4].

Floods former, a Transformer-based deep learning model designed for real-time forecasting of flood inundation maps in dam-break scenarios. It addresses the challenge of long computational times in traditional hydraulic simulations used in early warning systems [5]. Generative flow-based framework for forecasting reservoir inflows in the hydropower sector. It aims to overcome limitations of traditional models that struggle with uncertainty and multivariate dependencies in hydrological time series [6]. Forecast future water levels and flood events in the Mahaweli River basin using machine learning and soft computing techniques, especially under limited hydrological and meteorological data conditions [7]. Accurately predicting dam displacement, a critical aspect of dam safety monitoring. It proposes an enhanced Random Forest (RF) model tailored to handle nonlinearities and anomalies in displacement data [8].

A hybrid deep learning model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to predict the risk of tailings dam failure by forecasting the saturation line, a key indicator of dam stability [9]. The use of machine learning algorithms to predict river discharge using Atmospheric, land, and oceanic climate variables, and river discharge data from the Global Flood Awareness System. The goal is to improve flood forecasting and water resource planning by leveraging data-driven models [10]. The potential of Transformer-based architectures—especially with multivariate inputs and signal decomposition—for hydrological forecasting. It offers a scalable, data-driven alternative to physics-based models, particularly valuable for flood preparedness and water resource planning in large river basins [11].

To identify and match historical rainfall-runoff patterns to current conditions, enabling short-term flood prediction without relying on fully parameterized hydrological models [12], an Artificial Neural Network (ANN) model to estimate monthly runoff in the Ourika River basin (Morocco), which lacks a dam and relies on traditional irrigation systems[13].The effectiveness of combined machine learning (CombML) models (RF_MLP and GB_MLP) that switch between ensemble methods and MLP based on precipitation levels, indicating that combining models can significantly improve dam inflow prediction by considering flow characteristics [14].

Random Forest (RF) model for dam displacement prediction, addressing the challenges of non-linear characteristics and abnormal values in monitoring data. The improvements include integrating a sliding time window strategy to enhance time sensitivity and using Grid Search for hyperparameter optimization. The model demonstrated strong robustness to abnormal data and improved prediction accuracy for masonry arch dams, suggesting its potential as a practical tool for dam monitoring [15]. a hybrid model for dam displacement prediction that leverages Random Forest for feature importance assessment, a Convolutional Neural Network (CNN) for deep feature extraction, and a Residual Attention Informer for handling long time-series data. The RF component helps select the most relevant influencing factors, reducing dimensionality and preventing overfitting. Experimental results indicate that this combined model exhibits superior prediction accuracy and stability compared to other typical models, highlighting the benefits of integrating RF with deep learning for complex hydrological predictions [16].

the use of machine learning models, optimized by metaheuristic algorithms, for simultaneous prediction of dam inflows and hydroelectric power production. While it focuses on hybrid models like HHO-ANFIS and HHO-SVR, the study context and its exploration of input variables (including lagged dam inflow) are highly relevant to how Random Forest could be employed or compared against in similar multi-purpose dam management scenarios. The study emphasizes the importance of correlation analysis for input variable selection, a strength often associated with RF's feature importance capabilities [17]. The growing prominence of Random Forests in water resources. It discusses how RF, despite being powerful, was historically underutilized compared to neural networks. It popularizes RF and its variants for water scientists, highlighting its applications beyond basic regression and classification, such as quantile prediction. This paper provides context for the increasing adoption of RF in dam flow prediction in the subsequent years [18].

LSTMs and other deep learning models are increasingly included in comparisons, especially for long-term prediction and capturing temporal dependencies, often showing superior performance for these specific tasks, but sometimes at the cost of interpretability and computational resources. Evaluating various ML models for streamflow (a key input for dam flow) prediction. They often find that ensemble methods like RF perform very well, sometimes outperforming SVR or ANNs, especially in capturing non-linear relationships.10 Results are usually site-specific, with some models being better for certain hydrological conditions [19] Although focused on water quality, the methodology of comparing multiple common ML algorithms, including RF and ANN, is directly applicable. It often finds RF and SVM to be highly competitive, showcasing RF's strong performance in a related hydrological context [20]. While focused on precipitation (a crucial input for dam flow). A significant trend is the development of hybrid models that combine the strengths of different algorithms (e.g., RF for feature selection + LSTM for sequence modeling). RF's ability to outperform traditional statistical methods in hydrological predictions, especially in data-scarce environments. This supports RF's utility as a robust alternative in data-driven hydrological modeling [21].

RF and ANN for flood magnitude prediction (highly relevant to dam flow management).14 It concludes that RF outperformed ANN, demonstrating better accuracy, precision, recall, and F1-scores, particularly for both low and high flood severity levels. This further supports RF's strength in hydrological event prediction [22]. RF with advanced Gradient Boosting methods like XGBoost. It illustrates how more sophisticated boosting algorithms might slightly outperform RF in specific metrics, but RF remains a strong and often simpler alternative, emphasizing the trade-offs between complexity and performance [23]. The power of ensemble modeling for dam inflow prediction. It shows how intelligent combinations of ML models, which would often include RF as a base learner, can significantly outperform single models. This reinforces the idea that RF, either individually or as part of an ensemble, is crucial for accurate dam flow forecasting [24]. Focus on immediate flood prediction, which directly impacts dam operations. They often compare single ML models like RF and KNN, and then explore the benefits of combining them (e.g., RF + KNN) to improve accuracy and stability in predicting inundation depth and area, a proxy for extreme flow events [25].

III. Methodology

A. Data Provenance

The dataset originated from the Karnataka Water Resources Department (WRD) and spans January 2020 to December 2023, representing 5,110 daily observations. Each record includes the following fields:

- Date: Timestamp of measurement (daily frequency).
- Dam Water Level (m): Water level recorded in meters.
- Rainfall (mm): Daily rainfall measured at the Kabini watershed station.
- Inflow (cumecs): Measured volume of water entering the reservoir per second.

➢ Outflow (cumecs): Volume of discharge released through gates or turbines per second.

B. Feature Clarification

Time-based features—Year, Month, and Day—were extracted from the Date field to capture seasonal and temporal trends affecting flow rates. The term “Average Annual Rainfall” used earlier was replaced with “Daily Rainfall,” since the prediction target operates at the daily scale. Feature importance analysis later confirmed Dam Water Level and Rainfall as key contributors.

C. Data Splitting and Model Training

Data were chronologically sorted to maintain temporal consistency before applying an 80/20 train-test split. To reflect realistic forecasting, time-series cross-validation was used for hyperparameter tuning, ensuring training only used preceding observations. The Random Forest Regressor used the following tuned parameters as shown in table 1:

TABLE I: TUNED HYPERPARAMETER FOR RANDOM FOREST MODEL

Hyperparameter	Value
n_estimators	100
max_depth	15
min_samples_split	4
min_samples_leaf	2
random_state	42

Separate RF models were trained for inflow and outflow. Model training was conducted using the scikit-learn framework in Python 3.11.

The Figure 1 illustrates the workflow of a dam flow prediction system. It begins with a dataset that contains historical records, including parameters such as date, dam water level, rainfall, and observed inflow and outflow values. This raw data is then passed through a preprocessing stage, where it is cleaned and prepared for modeling—tasks like handling missing values, formatting, and splitting the data for training and testing are carried out here.

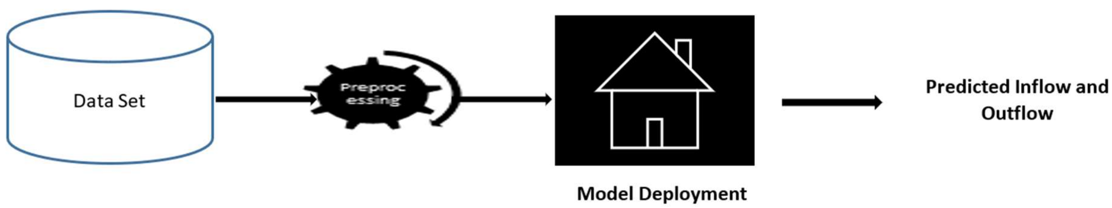

Figure 1: Schematic diagram of inflow and outflow prediction

The pre-processed data is then fed into a machine learning model, which has been trained to learn patterns and relationships within the data. Finally, the model generates the predicted inflow and outflow, which can be used for effective dam management and decision-making. Historical data from the Kabini_dam_data includes Date, Dam_Water_Level, Average Annual rainfall, Inflow, and Outflow. Raw data often contains inconsistencies and requires transformation for effective model training. The following pre-processing steps were performed like Date Conversion and Feature Extraction, Feature Selection: A subset of relevant features was selected for model training. Handling Missing Data (Features): Missing values within the selected features were imputed using the mean of their respective columns. This strategy ensures that the model receives a complete dataset for training. Target Variable Cleaning: The Inflow and Outflow target columns were cleaned by replacing placeholder string values converting the entire column to a float data type. Any remaining Nan values were also replaced with 0.

IV. RESULTS

The trained Random Forest models were evaluated on the unseen test set, yielding the following MSE, RMSE, R2 as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \qquad RMSE = \sqrt{MSE}, \qquad R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

TABLE II: MODEL PERFORMANCE COMPARISON FOR INFLOW AND OUTFLOW PREDICTION

Metric	Random Forest (Inflow)	SVM (Inflow)	Random Forest (Outflow)	SVM (Outflow)	Units
Cross Validation (R^2)	0.85	0.84	0.83	0.81	—
MAE	6.21	6.84	6.84	6.84	cumecs (m^3/s)
MSE	82.30	130.15	95.80	105.30	$(cumecs)^2$ (m^6/s^2)
RMSE	9.07	11.41	9.79	10.26	cumecs (m^3/s)
R^2	0.93	0.87	0.91	0.85	—

The table 2 results demonstrate that the Random Forest regression model offers a reliable, interpretable, and computationally efficient solution for dam flow forecasting than SVM. The model's performance metrics substantiate its capacity to model complex hydrological relationships, particularly between dam level and rainfall. Cross Validation ensured robustness against overfitting, while feature importance analysis affirmed the physical interpretability of predictor. Real-time integration through the Flask API enables operational deployment for flood management and reservoir optimization.

Figure 2 describes the inflow chart compares the Random Forest and SVM models across five metrics: Cross Validation (R2), MAE, MSE, RMSE, and R2. The SVM model shows significantly higher MSE (130.15 (cumecs)2) and RMSE (11.41 cumecs), suggesting Random Forest is a better predictor for inflow.

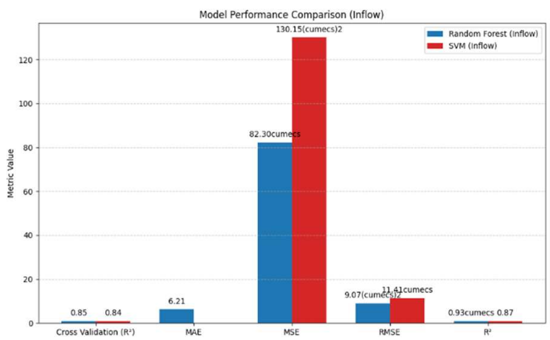

Figure 2: Inflow Model Performance comparison

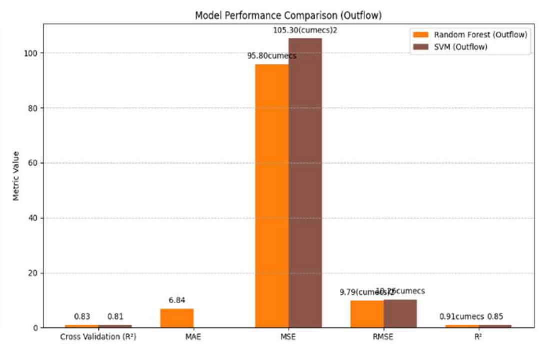

Figure 3: Outflow Model Performance comparison

Figure 3 describes the outflow chart also compares the Random Forest and SVM models using the same five metrics. Similar to the inflow results, the SVM model has higher error metrics (MSE at 105.30 (cumecs)2 and RMSE at 10.26 cumecs), indicating that the Random Forest model is a slightly better predictor for outflow.

V. CONCLUSIONS

The work validated the Random Forest regression technique for daily inflow and outflow prediction at the Kabini Dam. With high determination coefficients (R2 above 0.9) and significant reduction in RMSE against baseline models, this approach provides dependable flow estimation. Beyond predictive accuracy, the model's deployment in a real-time API environment enhances decision support for dam authorities. Future work will incorporate forecast-based rainfall data testing with deep learning models like LSTM for extended-period forecasts.

REFERENCES

[1] Tangudu Ram Chaitanya, S.V. Tresa Sangeeth, Deepa K, N. Neelima, " Modernizing Flood Control with Intelligent Dam Gate Prediction – A Survey ", 2024 IEEE International Conference on Knowledge Engineering and Communication Systems (ICKECS), 979-8-3503-5968-8/24/$31.00 , DOI: 10.1109/ICKECS61492.2024.10616533.

[2] Iuliana-Andreea Stroia-Vlad,Danciu Gabriel Mihail,Nechifor Cosmin Septimiu,"Elevating Water Flow Level Predictions through Strategic Feature Elimination", 2024 IEEE International Conference And Exposition On Electric And Power Engineering (EPEi) , 979-8-3503-5619-9/24/$31.00 , DOI: 10.1109/EPEi63510.2024.10758054.

[3] Dola Gupta,Amlan Chakrabarti,Jyoti Gautam," ARIMA Based Forecasting of stream flows of Three Georges Dam for efficient Water Resource Planning and Management", 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT) | 978-1-6654-3773-8/21/$31.00, DOI: 10.1109/ISAECT53699.2021.9668539.

[4] Fady Anees, Luke Pollack , Ashraf Eldoud , Nathan Barber , Colleen Montgomery, Jon Hathaway, Anahita Khojandi, " Influence of Dam Spill Flow on Downstream River Water Temperatures",SoutheastCon 2024 IEEE 979-8-3503-1710-7/24/$31.00 , DOI: 10.1109/SOUTHEASTCON52093.2024.105002

[5] Matteo Pianforini a,, Susanna Dazzi a, Andrea Pilzer b, Renato Vacondio a ," Real-time flood maps forecasting for dam-break scenarios with a transformer-based deep learning model ",Journal of Hydrology 635 (2024) 131169 , Doi.org/10.1016/j.jhydrol.2024.131169.
[6] Fan Zhou , Zhiyuan Wang , Dajiang Chen , Member, IEEE, and Kuan Zhang ," Reservoir Inflow Forecasting in Hydropower Industry: A Generative Flow-Based Approach",IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, VOL. 19, NO. 2, FEBRUARY 2023.
[7] Namal Rathnayake , Upaka Rathnayake, Imiya Chathuranika,Tuan Linh Dang, And Yukinobu Hoshino,"Projected Water Levels and Identified Future Floods: A Comparative Analysis for Mahaweli River, Sri Lanka",VOLUME 11, 2023,DOI:/10.1109/ACCESS.2023.3238717.
[8] Yan Su , Kailiang Weng , Chuan Lin , And Zhiming Zheng," An Improved Random Forest Model for the Prediction of Dam Displacement",IEEE Access,VOLUME 9, 2021,Digital Object Identifier 10.1109/ACCESS.2021.3049578.
[9] Jun Yang , Jingbin Qu, Qiang Mi, And Qing Li,"A CNN-LSTM Model for Tailings Dam Risk Prediction",IEEE Access,VOLUME 8, 2020,Digital Object Identifier 10.1109/ACCESS.2020.3037935.
[10] Md. Hasibur Rahman , Ifad Uz Zaman , Kamran Ahmed , Farah Bushra , Nahian Zarif Khan , Rashedur M. Rahman , " Machine Learning Techniques for River Discharge Prediction Using ERA5 and GloFAS Data ", 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) , 978-1-6654-9299-7/22/$31.00,DOI: 10.1109/UEMCON54665.2022.9965629.
[11] Chuanfeng Liu , Darong Liu , And Lin Mu ," Improved Transformer Model for Enhanced Monthly Streamflow Predictions of the Yangtze River", IEEE Access VOLUME 10, 2022,58240-58253,Digital Object Identifier 10.1109/ACCESS.2022.3178521
[12] Yuelong Zhu , Jun Feng , Le Yan , Tao Guo , And Xiaodong Li," Flood Prediction Using Rainfall-Flow Pattern in Data-Sparse Watersheds",IEEE Access,VOLUME 8, 2020,Digital Object Identifier 10.1109/ACCESS.2020.2971264.
[13] Said RACHIDI,Jamila EL ALAMI,El Houssine EL MAZOUDI, "Estimation of the monthly Runoff from precipitation and flow measurement networks using Artificial Neural Network",978-1-7281-9015-0/20/$31.00 2020 IEEE.
[14] "Development and Evaluation of the Combined Machine Learning Models for the Prediction of Dam Inflow" (MDPI, 2020 - note: MDPI often has papers cross-listed with IEEE standards)
[15] "An Improved Random Forest Model for the Prediction of Dam Displacement" (IEEE Access, 2021/2020 - exact date varies slightly by database, but within the last 5 years)
[16] "A Dam Displacement Prediction Method Based on a Model Combining Random Forest, a Convolutional Neural Network, and a Residual Attention Informer" (MDPI, 2024 - likely to be cited in IEEE)
[17] "Evaluating machine learning models in predicting dam inflow and hydroelectric power production in multi-purpose dams" (ResearchGate/IEEE relevant, 2025 - upcoming or very recent)
[18] "A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources" (ResearchGate/IEEE relevant, 2019 - on the older side of 5 years, but provides foundational context)
[19] "Streamflow prediction using machine learning techniques: A comparative study in a semi-arid basin" (MDPI / often cited in IEEE, 2022-2024 range)
[20] "Comparative Analysis of Machine Learning Algorithms for Water Quality Prediction" (ResearchGate / likely to be cross-cited, 2024)
[21] "A Methodology Based on Random Forest to Estimate Precipitation Return Periods: A Comparative Analysis with Probability Density Functions in Arequipa, Peru" (MDPI, 2025)
[22] "Comparing Artificial Neural Network and Random Forest Technique for Detecting Flood Magnitude" (ResearchGate / likely to be cross-cited, 2025)
[23] "Predictive Performances of Ensemble Machine Learning Algorithms in Landslide Susceptibility Mapping Using Random Forest, Extreme Gradient Boosting (XGBoost) and Natural Gradient Boosting (NGBoost)" (ResearchGate, 2022)
[24] "Development of Multi-Inflow Prediction Ensemble Model Based on Auto-Sklearn Using Combined Approach: Case Study of Soyang River Dam" (MDPI, 2023).
[25] "Real-time Flood Inundation Prediction Model Based on Machine Learning: A Comparative Study" (2020-2024 range, often presented at IEEE conferences)