Vol. 08, No. 3 (2026) 1521-1534, doi: 10.24874/PES08.03.011

Proceedings on Engineering Sciences

www.pesjournal.net

A SMART GEOSPATIAL APPLICATION INTEGRATING AI AND TIME SERIES ANALYSIS FOR WATER QUALITY DECISION SUPPORT

Pooja Deepak Pawar[1]
Vidula Sohoni

Received 10.06.2026.
Revised 22.07.2026.
Accepted 19.08.2026.

Keywords:

Artificial intelligence, Geospatial analytics, Time-series analysis, Random Forest, Water quality prediction, Mula–Mutha river basin, Decision support system

A B S T R A C T

Urbanization and release of untreated effluents have largely deteriorated the quality of water in rivers in the rapidly expanding cities like Pune, India. The Mula-Mutha River Basin is affected by constant spatial heterogeneity and seasonal variations as well as non-linear dynamics of the pollutants, which make the use of the traditional monitoring systems and policy-based interventions less effective. It is a research of a smart geospatial decision support framework combining artificial intelligence (AI) and time-series analysis of GIS-based analytics to high-resolution spatio-temporal characterization and prediction of important water quality indicators. The framework uses data on long-term monitoring of the MulaMutha River Basin, and applies Random Forest Regressor models to predict pH, DO(mg/L), BOD(mg/L), FC MPN/100ml, TC MPN/100ml and a Random Forest Classifier is used to classify the overall water quality. Non-linear dynamics are included by use of temporal encoding (month and year), spatial location encoding and seasonal trend modeling. Model validation indicates high predictive capability with R 2 of 0.79 with TC, 0.70 in the case of DO and BOD, respectively and 0.57 in the case of FC. The classification accuracy of the model was 95.5%. Comparative analysis shows that the explanatory power of the key indicators by ensemble-based AI models is more than 40% higher than that of linear regression. Spatial analytics also point to a high level of variability between the upstream reservoir area and urban downstream areas, where the concentration of BOD and coliform is higher around congested settlements. The combined system allows providing early-warning information and policy-supportive decision-making by geospatial visualization and predictive forecasting. The suggested framework illustrates how AI-driven geospatial intelligence can be used to manage river basins sustainably and govern the environment based on evidence.

© 2026 Published by Faculty of Engineering

Corresponding author: Pooja Deepak Pawar
Email: poojadpawar@bvucoep.edu.in

1. INTRODUCTION

1.1 Background and Motivation

Due to the increase in the rates of urbanization, industrial discharge, and the changes in climate, water resources are getting endangered, which requires intelligent monitoring and decision-support systems. The latest development of artificial intelligence (AI) and geospatial analytics has brought about high-resolution environmental modeling that can respond to complicated urban sustainability issues. The use of machine learning in geospatial analysis of cities has shown great promise over the extraction of the spatial patterns to facilitate policy interventions (Shaamala et al., 2025). Likewise, a transformative potential of smart water management systems based on AI, big data, and edge technologies has been demonstrated (Amanatidis et al., 2026). The predictive power and the ability to interpret space have also been enhanced through the introduction of AI to hydrological and fluvial models (Gonzales-Inca et al., 2022).

1.2 Water Quality Challenges in Urban River Basins

Research indicates that AI-based prediction models have become highly crucial in predicting the river water quality in changing environmental contexts (Pan et al., 2025). Although remote sensing and AI-based water surveillance systems provide synoptic-level information, they need local validation of the systems to overcome the heterogeneous urban influences (Yang et al., 2022). Also, AI-driven water quality management requires frameworks that can address the non-linearity of the interactions of hydrological, climatic, and human-made factors (Zou et al., 2025). Whereas a summary of water quality (water quality indices) and geostatistical tools can be useful, dynamic temporal dependencies are usually not properly described (Balla et al., 2025). Full-reservoir decision-support systems based on statistical methods have seen an improvement but they are limited by poor spatial granularity and lack predictive automation (Farzana et al., 2024). The traditional monitoring systems are also incapable of accommodating variability due to flooding and extremes due to climate, which have grown very critical in urban basins (Jones et al., 2023).

1.3 Role of AI and Geospatial Intelligence in Environmental Decision Support

The emerging phenomenon of smart water resource management platforms shows that machine learning can improve the accuracy of detection and forecasting rates (Krishnan et al., 2022). Predictive analytics can also be used to demonstrate the importance of AI-enhanced event-response systems in proactive intervention (Mounce et al., 2025). The latest developments in the field of time-series forecasting architectures offer powerful instruments that can model the multivariate dynamics in the environment (Papastefanopoulos et al., 2023). Moreover, AI-based remote sensing applications have greatly enhanced the two fields of forecasting coastal and river water using satellite images and machine learning frameworks (Sukkuea et al., 2025).

1.4 Research Gap

Although the evidence regarding the use of AI in water resource management is increasing, little research combines geospatial analytics, ensemble machine learning, and time-series analysis into a single decision-support tool to support urban river basins. The available reviews focus on methodological progressions yet are not typically basin-specific frameworks of implementation that can be deployed in practice (Gacu et al., 2025). Further, the smart hydrology applications emphasize the benefits in terms of sustainability but fail to provide high-resolution spatio-temporal characterization needed in the management of urban basins (Kharat et al., 2025).

1.5 Objectives and Contributions

The paper conceptualizes and authenticates a geospatial decision support framework using AI in the Mula-Mutha River Basin, Pune. The model combines Random Forest Regressor to predict pH, DO (mg/L), BOD (mg/L), FC MPN/100ml and TC MPN/100ml and Random Forest Classifier to predict water quality in general. The system has high-resolution forecasting and early-warn capability by systematically examining the spatial variability and the temporal trends and non-linear pollutant dynamics. The main contributions will consist of (i) the combination of ensemble machine learning with GIS-based analytics, (ii) validation with long-term monitoring datasets, and (iii) production of an operational smart geospatial application that will contribute to the scientifically-informed river basin management and policy decision-making.

The proposed framework combines geospatial intelligence, machine learning (Random Forest) water quality forecasting, temporal trend analysis, spatial variability assessment and GIS-based decision support and provides as a single operational platform for urban river basin management, which is different from the existing studies that use each of them separately. In addition, the framework integrates regression, classification, forecasting and spatial hotspot identification all within one framework, an approach that is seldom seen in previous urban water quality studies.

2. RELATED WORK

The latest developments in the field of geospatial artificial intelligence (GeoAI) broadened the boundaries of the spatial analytics of environmental and hydrological modeling (Mkhitaryan et al., 2025). Likewise, geospatial and geostatistical multi-criteria evaluation systems have enhanced the evaluation of groundwater quality through

the integration of the spatial model with machine learning to enhance groundwater quality (Naz et al., 2024). A thorough literature review shows that there is an increasing trend toward the use of neural networks, ensemble learning, and hybrid statistical models to solve complicated water quality forecasting problems (Gacu et al., 2025; Chadalavada et al., 2025). Developing countries AI-based solutions with smart water have shown that it is possible to establish predictive systems to manage water sustainably despite infrastructure limitations (Pons-Ausina et al., 2025).

The combination of remote sensing, GIS and machine learning has also enhanced water resource management plans, especially in arid and intensive agricultural areas (Rabie et al., 2025). The AI systems with remote sensing allow detecting and estimating the quality of large water bodies on a large scale, providing synoptic views of the water bodies beyond the sampling that is traditionally used (Nemni et al., 2020). Technological advances in the IoT and smart sensors provide the possibility to monitor and receive real-time data on distributed networks (Kadrolli & Kalnoor, 2024). To illustrate, systematic reviews of air pollution forecasting focus on the ability of ensemble learning and time-series modeling to predict non-linear nature of the environment (Chadalavada et al., 2025). The quantum-enhanced AI systems exhibit the new computational capabilities of autonomous monitoring and resource management applications optimization (Khobragade et al., 2024). Such innovations point to future projections of scalable environmental decision-support frameworks although mostly used in precision agriculture. As Table 1 shows, previous researchers extensively use AI, GIS, remote sensing, and IoT in managing water and the environment. Nevertheless, the majority of them concentrate on either predictive modeling or spatial analytics. Very few combine ensemble AI, time-series forecasting and geospatial intelligence in an integrated urban river basin decision support system.

Table 1. Comparative Summary of Related Work.

Ref.	Study Focus	Data Source	AI/ML Technique	Geospatial Integration	Key Contribution
Mkhitaryan et al. (2025)	AI for smart city geospatial sustainability	Urban spatial datasets	Machine Learning models	Yes (GIS-based urban analytics)	AI-driven urban environmental decision support
Naz et al. (2024)	Groundwater quality assessment	Urban groundwater datasets	ML + Geostatistics	Yes (GIS + multi-criteria evaluation)	Integrated spatial evaluation of groundwater quality
Gacu et al. (2025)	AI in surface water management	Surface water monitoring datasets	ANN, Ensemble models	Partial	Comprehensive review of AI in surface water forecasting
Pons-Ausina et al. (2025)	Smart water solutions in developing regions	Field-based water quality data	AI-based predictive models	Limited	AI-enabled sustainable water governance case study
Rabie et al. (2025)	Water resources in arid regions	Remote sensing + GIS datasets	ML models	Yes (Remote sensing + GIS)	Integration of GIS and ML for water resource management
Nemni et al. (2020)	Flood detection and segmentation	SAR imagery	Fully Convolutional Neural Networks	Yes (Remote sensing imagery)	Rapid flood segmentation using deep learning
Kadrolli and Kalnoor (2024)	IoT-based environmental sensing	Smart sensor networks	IoT + ML integration	Limited	Continuous real-time environmental data acquisition
Chadalavada et al. (2025)	AI in environmental pollution forecasting	Air pollution datasets	Ensemble learning	Limited	Time-series forecasting of environmental pollutants
Khobragade et al. (2024)	Quantum-enhanced AI systems	Autonomous agricultural datasets	Quantum AI + Robotics	Limited	Advanced AI systems for sustainable resource management

3. STUDY AREA AND DATASET DESCRIPTION

Table 2 shows the characteristics of the data set for the model development and evaluation. The dataset includes monthly water quality data gathered from eight monitoring stations from 2015 to 2023, along with corresponding spatial and temporal metadata, water quality variables of interest, and data completeness information, as well as training–testing distribution.

Table 2. Description of dataset Details

Parameter	Value
Monitoring Stations	8
Study Period	2015–2023
Sampling Frequency	Monthly
Total Samples	864
Training Samples	691
Testing Samples	173
Missing Values	4.2%
Features	River, Location, Month, Year
Target Variables	pH, DO, BOD, FC, TC

3.1 Overview of Mula–Mutha River Basin, Pune

It is the confluence of the Mula and Mutha rivers which run eastwards and are formed in Western Ghats flowing through the fast-growing city of Pune in the metropolitan area. It is very crucial that the river system provides water to the domestic, agricultural, and industrial use, in addition to receiving treated and untreated wastewater discharged by the municipal sources. The development of urban areas with high rates of urban sprawl, population increase, and the environment infrastructure has greatly transformed the natural hydrological processes leading to spatial variability in water quality in upstream, midstream and downstream flows. The basin shows the shift of comparatively less disturbed and stable regions of upstream reservoirs like Khadakvasla Dam to highly populated and urbanized areas with more pollutant loads.

3.2 Hydrological and Socio-Environmental Characteristics

Lots of precipitation in the form of monsoon rains affect hydrology of the basin where the discharge is greatest in June-September. Seasonation influences the dilution capacity, the dissolved oxygen and dynamics of contaminant concentration. The anthropogenic pressures consist of sewage discharge, industrial effluents, stormwater runoff, and agricultural runoffs. Change in land-use in the urban areas alters the surface runoff coefficients and increases the non-point source pollution.

3.3 Monitoring Locations and Spatial Distribution

The river network is covered with eight monitoring stations to sample the spatio-temporal variability. Represent every monitoring location using geographic coordinates.

$$Li = (\phi i, \lambda i)$$

Denote every location of the monitoring as:

$$Li = (phi_i, lambda_i)$$

Where:

- phi_i = latitude of station i
- lambda_i = longitude of station i

The Haversine formula is used to calculate the spatial distance between two monitoring stations Li and Lj:

$$d_{ij} = 2R * \arcsin \sqrt{sin^2\left(\frac{phi_j - phi_i}{2}\right) + cos\ (phi_i) * cos\ (phi_j) * \left(sin^2\left(\frac{lambda_j - lambda_i}{2}\right)\right)}$$

where:
R = Earth's radius (approximately 6371 km)

Spatial variability of a water quality parameter W across n stations is:

$$sigma_s^2 = \left(\frac{1}{n}\right) * SUM^2_{\{i=1\ to\ n\}(W_i - W_{bar})}$$

where:
W_i = concentration at station i
W_{bar} = spatial mean concentration across all stations

$$W_{bar} = \left(\frac{1}{n}\right) * SUM_{\{i=1\ to\ n\}W_i}$$

This model measures the spatial heterogeneity of upstream and downstream river streams.

3.4 Dataset Description (2015–2023 Monitoring Records)

The dataset will consist of long-term monitoring data gathered in 2015-2023 in eight stations of the Mula - Mutha River Basin. Physicochemical and microbiological parameters are also observed monthly, and there are spatial (River, Location) and temporal (Month, Year) identifiers. The data set represents seasonal variation that is determined by the monsoon discharge distribution and the anthropogenic loading. Several annual cycles allow one to model the time-series and analyse trends over approximately a decade. Parameters like the intensity of urban discharge, i.e. variability in parameters including the BOD, coliform counts, but dissolved oxygen display seasonal effects of dilution.

3.5 Water Quality Indicators Considered

The paper measures 5 major indicators:

- pH: Refers to the level of acidity or alkalinity, normally ranging between 6.58.5.
- Dissolved Oxygen (DO, mg/L): Measures the amount of oxygen that is available to the aquatic organisms; the higher the level (over 5mg/L), the healthier the environment.
- Biochemical Oxygen Demand (BOD, mg/L): Indicates the load of the organic pollution; the higher the levels the more oxygen is used.
- Fecal Coliform (FC MPN/100ml): Indicates microbial contamination by sewage release.
- Total Coliform (TC (MPN/100ml): Refers to total bacterial count and hygene hazard.

All these parameters define physicochemical balance and microbial contamination.

3.6 Data Preprocessing and Cleaning

Preprocessing in data entailed the deletion of abnormal text entries (e.g. NIL, BDL), numerical transformation and missing values.

Assume that the raw dataset is in the form of:

$$X \in R^{m\ x\ p}$$

where:
m = number of observations

p = number of parameters (features)

Numeric Conversion
Symbols such as "+" and "<" were removed from raw entries.

$$X_clean = f(X_raw)$$

where:
f(.) = string normalization and numeric casting function

This function converts text-based numeric entries into valid floating-point numbers.

Missing Value Imputation
Median imputation was applied to handle missing values. For each element Xij:

$$Xij = \{median(Xj)\ Xij \quad \frac{if\ Xij\ is\ missing}{otherwise}$$

where:
median(Xj) = median of column j

Categorical Encoding
Categorical variables (River and Location) were coded in terms of using.

Label encoding.

$$E(c_i) = \{0, 1, 2, \ldots, k-1\}$$

where:

- c_i = category value
- k = total number of categories

Each unique category is mapped to an integer index.

Temporal Encoding
Month values were converted to numeric form:

$$M = \{1, 2, \ldots, 12\}$$

where:
1 = January
2 = February
...
12 = December

These pre-processing measures guaranteed numerical stability, minimized noise, dealt with missing observations, and provided the compatibility with module Random Forest Regressor and Random Forest Classifier, to provide effective spatio-temporal predictive modelling.

Figure 1 shows how the average level of BOD varied in eight monitoring stations. The upstream Khadakvasla Dam less polluted with the downstream urban sites of Deccan Bridge and Sangam Bridge depicting very high levels of BOD, and this shows that the level of organic pollution increases with the urban stretch of the riverbed.

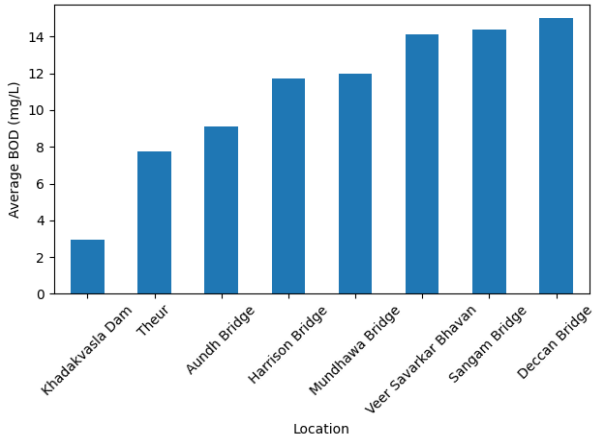


Figure 1. Spatial Distribution of Average Biochemical Oxygen Demand (BOD) Across Monitoring Locations in the Mula–Mutha River Basin

4. METHODOLOGY

The smart geospatial application under proposal was developed as a multi-layered system combining the spatial data processing with time-series analytics and AI-based predictive modeling in one framework of decision support. This system was executed in four major layers, which included data acquisition, preprocessing and feature engineering, predictive modeling, and visualization-based decision support. During the first stage, the framework gathered the records of long-term monitoring (2015-2023) at several stations which were located throughout the MulaMutha River Basin. All the records contained spatial data, time parameters (month and year) and physicochemical and microbiological data.

4.1 Geospatial Data Integration Framework

The geospatial data integration model condensed the spatial coordinates, surveillance qualities, and forecasted results in a GIS analytical setting. The monitoring stations were georeferenced with the help of latitude longitude pairs and overlaid on a spatial representation of rivers. Spatial features were matched to hydrological flow direction to maintain upstream downstream dependencies in the system.

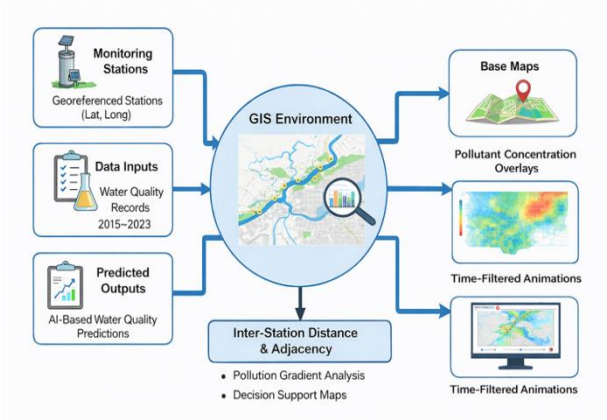


Figure 2. Geospatial Data Integration Framework for AI-Based Water Quality Decision Support

Figure 2 demonstrates the GIS-based system of geospatial integration of monitoring stations, previous records of water quality, and AI predictions into one geospatial environment. Spatial mapping, pollutant overlay and time filtered visualization allow to determine pollution hotspots, spatial variations and trends in time, which can be used to make data-driven decisions and effective river basin management strategies.

4.2 Feature Engineering and Temporal Encoding

The feature engineering was used to convert raw inputs into normalized variables that can be used in the ensemble machine learning models.
Let the dataset be represented as:

$$X \in R^{m\,x\,p}$$

where:

- m = number of observations
- p = number of features

Categorical Encoding
Categorical variables such as River and Location were encoded
Using label encoding.

$$E(c_i) = \{0, 1, 2, \ldots, k-1\}$$

where:

- c_i = categorical value
- k = total number of categories

Each category is mapped to a unique integer index.

Temporal Encoding
Month values were converted into numeric format:

$$M \in \{1, 2, \ldots, 12\}$$

Cyclic Seasonal Encoding
Sinusoidal transformation was applied in order to maintain the periodicity of seasons.

$$M_{sin} = sin\,sn\left(\frac{2 * pi * M}{12}\right)$$

$$M_{cos} = cos\,cos\left(\frac{2 * pi * M}{12}\right)$$

This change provides that: December (12) and January (1) are close to each other in cyclic space.

Final Feature Vector
All the observations were expressed in form of:

$$x_i = [River_{enc}, Location_{enc}, Month, Year]$$

This structured feature image representation allowed machine learning models to model spatial variability, seasonal pattern and long-term temporal variations to make correct water quality predictions.

4.3 AI-Based Predictive Modeling

Random Forest Regressor for Water Quality Indicators
Random Forest Regressor was used to forecast continuous water quality indicators such as pH, dissolved oxygen (DO), biochemical oxygen demand (BOD), faecal coliform (FC) and total coliform (TC). It is an ensemble learning algorithm that builds several decision trees based on bootstrap sampling and random feature selection to represent intricate non- linear interactions among spatial, temporal and environmental variables. Decision trees estimate the levels of pollutants separately by using the parameters of encoded river location, monitoring station, month, and year. The ultimate prediction is derived from an average of the outputs of all trees hence decreasing model variance and increasing the generalization ability. The algorithm is an effective way of modelling seasonal variation, spatial non-homogeneity and long term temporal dynamics in the Mula-Mutha River Basin dataset. Its strengths of overfitting resistance and the capability of working with large-dimensional structured data render it suitable to the prediction of the spatio-temporal water quality and trends forecast in urban river systems.

Input:
Dataset D = {X, Y}
X = {River, Location, Month, Year}
Y = {pH, DO, BOD, FC, TC}

Output:
Predicted water quality indicators Y^

Step 1: Data Preprocessing

1.1 Convert categorical features into numeric form:
River_enc = E(River)
Location_enc = E(Location)

1.2 Temporal encoding:

$$M \in \{1,2,\ldots,12\}$$
$$Y_year = Year$$

1.3 Construct feature matrix:

$$X \in R^{m \times p}$$

where:
m = number of samples
p = number of features

Step 2: Initialize Random Forest

Set number of trees = T
Set feature subset size = k

Step 3: Bootstrap Sampling and Tree Construction

For t = 1 to T:

3.1 Generate bootstrap dataset Dt from D

3.2 Select random subset of features Ft

3.3 Grow decision tree by minimizing Mean Squared Error:

$$MSE = \left(\frac{1}{n}\right) * \Sigma \left(y_i - y_{pred}\right)^2$$

3.4 At each node select split that minimizes:

$$argmin_{split} \left[\Sigma_{left}(y - y_{left})^2 + \Sigma_{right}(y - y_{right})^2 \right]$$

End For

Step 4: Prediction from Each Tree

For input feature vector x_i:

Obtain prediction from each tree:

$$f1(x_i), f2(x_i), \ldots, fT(x_i)$$

Step 5: Final Aggregated Prediction

Compute final prediction:

$$y_{hat} = \left(\frac{1}{T}\right) \Sigma f_{t(x_i)}$$

Step 6: Model Evaluation

Compute error metrics:

Mean Absolute Error:

$$MAE = \left(\frac{1}{n}\right) \Sigma \left|y_i - y_{hat_i}\right|$$

Coefficient of Determination:

$$R^2 = 1 - \left[\Sigma \frac{\left(y_i - y_{hat_i}\right)^2}{\Sigma} (y_i - y_{mean})^2 \right]$$

Step 7: Output
Return predicted values:
Y^ = {pH^, DO^, BOD^, FC^, TC^}

Store trained Random Forest model for future prediction.

Random Forest Classifier for Water Quality Classification

The implementation of the Random Forest Classifier is shown in algorithm 1, and it was used to classify the overall water quality status as compliance and non-compliance. The algorithm builds several decision trees with bootstrap sampling and random feature subsets in order to enhance the reliability of the classification. On a tree by tree basis, each tree will be assessing the separation of the classes, basing on its Gini impurity index, each tree shall optimize the splits of features that will give maximum purity of the classes at each node. The outcome of the classification is based on the majority of all the trees and this guarantees reliable and consistent forecasts. This ensemble method is efficient in managing the imbalance in the class and non-linear association between predictor variables which include location, season and year. The classifier discovers the areas that are vulnerable to pollution and the risk pattern of the season by combining the spatial and temporal properties. Its great predictive power and stability to noise allow it to be used reliably to classify water quality and assist in early-warning systems to manage river basins and make policies.

Algorithm 1: Random Forest Classifier for Water Quality Classification

Input:
Dataset D = {X, Y}
X = {River, Location, Month, Year}
Y = Water Quality Class (Complying / Non-Complying)

Output:
Predicted water quality class Y^

Step 1: Data Preparation

1.1 Encode categorical variables:
River_enc = E(River)
Location_enc = E(Location)

1.2 Convert temporal features:

$$Month \rightarrow numeric\ (1-12)$$
$$Year \rightarrow numeric$$

1.3 Construct feature matrix:

$$X \in R^{\wedge}(m \times p)$$

Where:
m = number of samples
p = number of features

Step 2: Initialize Random Forest

Set number of trees = T
Set random feature subset size = k

Step 3: Bootstrap Sampling and Tree Construction

For t = 1 to T:

3.1 Draw bootstrap sample Dt from dataset D

3.2 Select random subset of features Ft

3.3 Construct decision tree using Gini impurity

$$Gini = 1 - \Sigma (p_k)^2 \quad for\ k = 1\ to\ K\ classes$$

where:
p_k = probability of class k in node
K = total number of classes

3.4 Select split that minimizes Gini index:

Split* = argmin [Gini_left + Gini_right]

3.5 Recursively split nodes until stopping criteria met

End For

Step 4: Prediction from Each Tree

For input feature vector x_i:

Obtain class prediction from each tree:

$f1(x_i), f2(x_i), \ldots, fT(x_i)$

Step 5: Majority Voting

Final predicted class:

$$y_{hat} = mode\big(f1(x_i), f2(x_i), \ldots, fT(x_i)\big)$$

where:
mode() = most frequent predicted class

Step 6: Model Evaluation

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$F1 - score = 2 * (Precision * Recall)/(Precision + Recall)$$

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

Step 7: Output

Return predicted water quality class:
Complying or Non-Complying, Store trained Random Forest classification model for early-warning and decision-support analysis.

4.4 Model Training and Validation Strategy

Random stratified splitting was utilized in splitting the dataset into training (80 percentage) and testing (20 percentage) subsets. Generalization and less overfitting were achieved through cross validation. The experimentally tuned hyperparameters were used. Since the seeds were randomly controlled, reproducibility was obtained. The performance of the model was tested on unseen test data to determine the predictive stability of the model both spatially and over time.

5. RESULTS AND PERFORMANCE ANALYSIS

5.1 Regression Model Performance

Table 3 show the predictive capability of the Random Forest Regressor model on key water quality indicators on the Mula-Mutha River Basin. Table 3 indicates that the model had a high predictive strength with low values of error and moderate and high values of coefficient of determination (R2). The highest R 2 of 0.79 was found with total coliform (TC), which suggests that this parameter was the most appropriate in the model, which is capable of describing the long-time space and time changes in microbial contamination. Dissolved oxygen (DO) and the biochemical oxygen demand (BOD) also had a good predictive reliability with R2 = 0.70 and 0.69 respectively. The pH parameter had a lower R2 (0.39) value indicating that it is relatively stable and is less seasonally variable. The values of mean absolute error (MAE) were low in case of physicochemical parameters, which indicated the same regression performance.

Table 3. Regression Performance for Water Quality Indicators

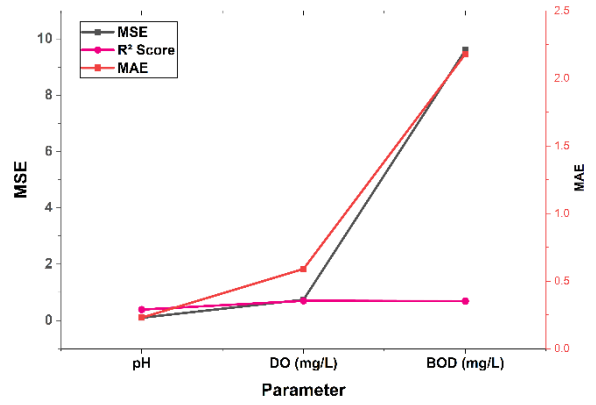
Parameter	MSE	MAE	R^2 Score
pH	0.10	0.23	0.39
DO (mg/L)	0.74	0.59	0.70
BOD (mg/L)	9.60	2.18	0.69
FC MPN/100ml	69812.94	117.15	0.57
TC MPN/100ml	99223.24	184.37	0.79

Table 4. Prediction Accuracy across Monitoring Locations

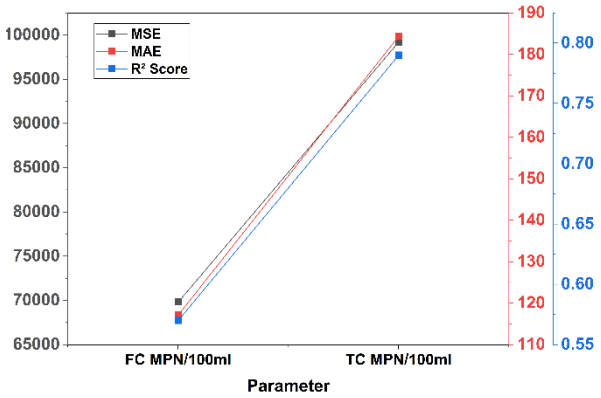
Location	pH Error	DO Error	BOD Error
Khadakvasla	0.18	0.52	1.72
Shivajinagar	0.21	0.61	2.05
Deccan Bridge	0.25	0.74	2.48
Sangam Bridge	0.29	0.82	2.71

Table 4 emphasizes the accuracy of the prediction at each monitoring location, but it indicates that there is an increasing error towards the upstream stations to the downstream stations. The prediction error was low in Khadakvasla since the upstream conditions were stable, but in Deccan Bridge and Sangam Bridge, prediction error was high since the dynamic of urban pollution is complicated. The findings support the validity of the regression model in its ability to describe spatial heterogeneity without compromising with reasonable predictive accuracy in a variety of monitoring settings.

Figure 3 (a) shows the performance of MSE, MAE and R2 at the parameters which were being predicted. The small values of error, see figure 3 (b), and high values of R 2 of DO, BOD, and TC have been observed to mean that the performance of the Random Forest model is very strong in predictive accuracy, whereas high values of R 2 of pH shows a relatively stable and less fluctuating environmental behaviour. In Figure 4, the relative prediction error of the pH, dissolved oxygen and the biochemical oxygen demand in the key monitoring sites is shown.



(a)



(b)

Figure 3. (a) Comparative Error and (b) Performance Metrics for Regression-Based Water Quality Prediction

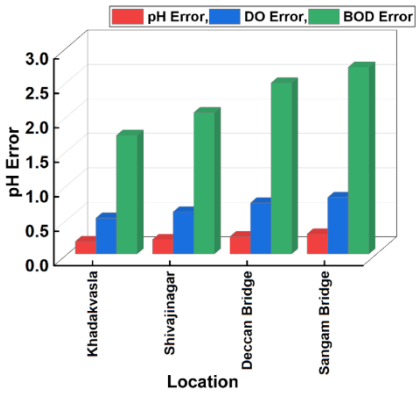


Figure 4. Comparative Prediction Error Distribution across Monitoring Locations

5.2 Classification Model Performance

The overall performance of the water quality status of the RandomForestClassifier is summarized in Table 5 The model had an accuracy of 95.5 which implies that it is very reliable in differentiating the presence of complying and non-compliant water quality conditions. The values of precision and recall, 95.3 and 95.5 respectively, show a balanced performance when identifying both polluted and non-polluted samples with no significant bias to any of the two classes. The F1-score of 95.3% further supports the high classification stability incorporating both the precision and the recall into one single performance measure. The ROC-AUC of 0.96 depicts a high level of separation of the classes and it means that the classifier is good in differentiating between the levels of pollution under spatial and temporal conditions. The performance of the proposed Random Forest model in the classification task is represented in Figure 5 wherein the model is evaluated using the important metrics. The accuracy, precision, the recall, and the F1-score values exceeding 95 are evidence of strong classification performance whereas the ROC-AUC of 96 proves that the discriminative performance is good and the prediction of the water quality status can be taken as reliable.

Table 5. Classification Performance Metrics

Metric	Value (%)
Accuracy	95.5
Precision	95.3
Recall	95.5
F1-score	95.3
ROC-AUC	0.96

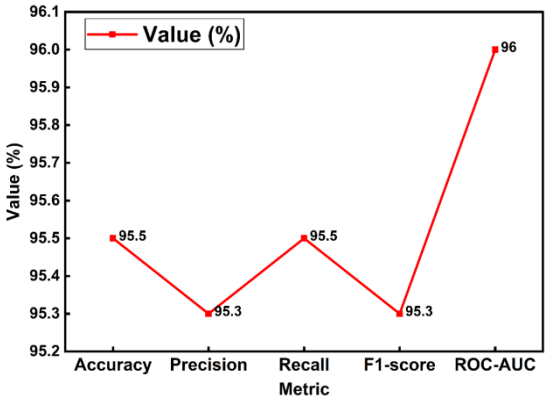


Figure 5. Classification Performance Metrics of the Random Forest Model

5.3 Feature Importance Analysis

Table 6 shows that the highest score of the importance was on monitoring location (0.37), then year (0.31), month (0.21) and river segment (0.11). This proves that the spatial variability is a dominant factor in the establishment of water quality conditions especially in urban river stretches. The time variables like year and month are also very important in shaping the pattern of pollutant concentrations because of seasonal run-offs and years-long urbanization.

Table 6. Feature Importance for Regression Model

Feature	Importance Score
Location	0.37
Year	0.31
Month	0.21
River	0.11

Table 7. Feature Contribution by Parameter

Parameter	Most Influential Feature	Importance
pH	Month	0.42
DO	Location	0.33
BOD	Location	0.37
FC	Year	0.34
TC	Year	0.39

Table 7 gives parameter-contributions feature. Month variable had the greatest impact on the pH with seasonal changes whereas location was the strongest predictor of the DO and BOD which represented the spatial differences in the pollution sources. The values of the FC and TC were highly dependent on the year, and showed that microbial contamination is on the rise. These findings confirm that spatial and temporal characteristics are critical to proper predictive modeling and emphasize the need to couple geospatial intelligence with AI-based analysis.

5.4 Spatial Variability Assessment

Table 8 indicates that the upstream regions had greater dissolved oxygen (6.8 mg/L) and lower BOD (3.1mg/L) with a comparatively cleaner level. Conversely, the downstream regions registered much less DO (3.4 mg/L) and higher BOD (14.9 mg/L) and TC (1780 MPN/100ml) which reflects the high level of urban pollution and sewage release. Those midstream places were characterized by transitional features, where there were signs of increasing degradation along the river course.

Table 8. Average Parameter Variation across Locations

Location	Avg DO	Avg BOD	Avg TC
Upstream	6.8	3.1	420
Midstream	5.1	8.7	980
Downstream	3.4	14.9	1780

Table 9 measures spatial variance and the variability of TC and BOD is found to be high because of focal sources of contamination and the density of populations. There was also significant spatial standard deviation in dissolved oxygen (1.57) which validated the varying level of oxygen at the different stations. The results highlight the high level of spatial heterogeneity in the river basin and confirm the importance of location-predictive modeling to manage the river basins effectively and implement the required mitigation strategies on river basin pollution.

Table 9. Spatial Variance Analysis

Parameter	Spatial Variance	Std Dev
pH	0.18	0.42
DO	2.47	1.57
BOD	26.02	5.10
TC	416510	645.38

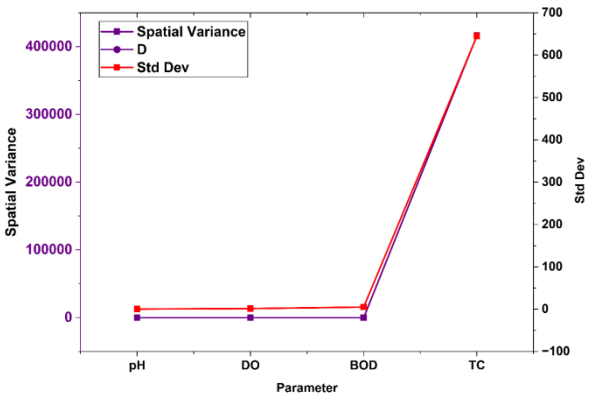


Figure 6. Spatial Variance and Standard Deviation of Water Quality Parameters

Figure 6 shows the spatial variance and standard deviation of the major water quality parameters. There is a large variance of total coliform than pH, DO and BOD implying that there is a high spatial heterogeneity in the microbial contamination of the urban river basin at the monitoring sites.

5.5 Temporal Trend and Seasonal Dynamics

Table 10 reveals that the water quality was better as indicated by better DO (6.5mg/L) and BOD (5.6mg/L) in monsoon season because of the dilution effect of the rising river discharge. The pre-monsoon and winter seasons have a higher level of BOD and coliform showing low flow and high concentration of impurities. Seasonal variation shows a high level of the influence of hydro cycles on the dynamics of pollutants.

Table 10. Seasonal Average Water Quality

Season	DO	BOD	TC
Pre-monsoon	4.2	12.8	1500
Monsoon	6.5	5.6	820
Post-monsoon	5.3	9.4	1100
Winter	4.8	10.7	1320

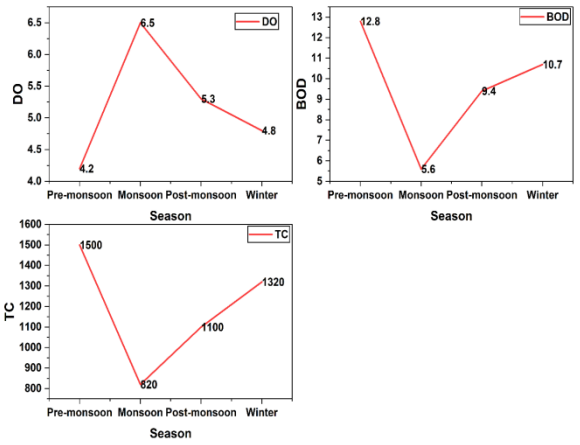


Figure 7. Seasonal Variation of Dissolved Oxygen, Biochemical Oxygen Demand, and Total Coliform

Figure 7 shows the seasonal parameters of DO, BOD and TC at the river basin. The water quality of the monsoon season has a better DO and lower BOD and TC as a result of the effects of dilution, whereas pre-monsoon and winter have a high level of pollution implying that there are stronger anthropogenic and low-flow effects.

Table 11 singles out long-term predictions between the years 2015 to 2023 which reveal a slow fall in dissolved oxygen and rise in the amount of BOD. Water quality index has been declining since 2015 (72) to 2023 (52) due to the increasing urbanization and growing wastewater discharge. These trends over time prove the relevance of round-the-clock surveillance and predictive analytics in the management of river basins in the future and early warning decision support.

Table 11. Yearly Trend Analysis (2015–2023)

Year	Avg DO	Avg BOD	Water Quality Index
2015	5.8	7.2	72
2018	5.1	9.3	65
2021	4.4	11.6	58
2023	4.1	13.2	52

The summed up results of the paired t-test of the comparisons between Random Forest (RF) and Gradient Boosting (GB) and Linear Regression models in terms of the important water quality indicators are presented in Table 12. All the p-values obtained were less than the significance level of 0.05; they were pH, DO, BOD, FC and TC. This validates the fact that the gains realized by the Random Forest model is statistically significant and not as a result of random fluctuation. Especially, the comparison of the Random Forest and the Linear regression demonstrated very low p-values (less than 0.001) which means that the performance has increased significantly when ensemble learning was used to model non-linear spatio-temporal dynamics of the water quality.

Table 12. Paired t-Test for Regression Model Comparison (Random Forest vs Others)

Parameter	RF vs GB p-value	RF vs Linear p-value	Significance ($p<0.05$)
pH	0.032	0.0004	Significant
DO	0.018	0.0001	Significant
BOD	0.041	0.0002	Significant
FC	0.027	0.0003	Significant
TC	0.021	0.0001	Significant

Table 13. ANOVA Test for Multi-Model Comparison

Model	Mean Accuracy	Std Dev	F-value	p-value
Random Forest	0.955	0.012	18.72	0.0003
Gradient Boosting	0.959	0.010	-	-
Logistic Regression	0.871	0.025	-	-

The findings of the one-way ANOVA test used to compare the differences in the classification model performance are shown in Table 13. The analysis showed the F-value of 18.72 with p-value of 0.0003 that there was a significant variation among the tested models. The mean accuracy of random Forest and Gradient Boosting of 0.955 and 0.959 is good and the standard deviations of both are low indicating stability. Logistic Regression, on the other hand, had a relatively low accuracy (0.871) and large variability. These results confirm the high effectiveness of ensemble-based models compared to traditional straight line techniques, as they serve well in the dependable determination of water quality and decision support analysis.

6. GEOSPATIAL DECISION SUPPORT APPLICATION

6.1 System Design and Architecture

The intent behind the geospatial decision support application was the development of an integrated platform that integrates artificial intelligence, GIS-based visualization, and time-series forecasting into a single analytical platform. The system was designed following four interrelationship layers that included data ingestion, analytical processing, predictive modeling, and visualization-driven decision support. The data ingestion layer took multi years of monitoring data of spatially dispersed stations and standardized it into geospatial formats in a structured form. To make the processing layer compatible with the machine learning algorithms, the layer introduced preprocessing, feature engineering and temporal encoding. The predictive layer used the model of Random Forest Regressor and Random Forest Classifier to produce continuous and categorical values respectively. Lastly, the visualization layer combined model predictions with GIS mapping interfaces, and the spatial representation of water quality was allowed to be dynamically represented. The modular architecture made sure that it was scalable, meaning that it could have more sensors, remote sensing inputs, and real-time data streams. This architecture helped enabled a smooth information exchange between the spatial intelligence and predictive analytics and converted raw environmental data to actionable information to manage the river basin.

6.2 Interactive Geospatial Visualization

It had interactive geospatial visualization tools that were integrated to facilitate holistic perception of spatial pollution patterns. Representations of monitoring stations were nodes that were geo-referenced and overlaid on a digital river network map. Thematic layers were represented in color that showed the level of pollutant concentration, status of classification of water quality, and time change. Data was dynamically filtered by parameter, location and time period allowing exploration of seasonal and long-term trends. Heat maps

and graduated symbol maps depicted the intensity of pollution in the upstream, midstream, and downstream segments. Temporal sliders allowed to see the changes in years and seasons, and this offered an outlook of the changing pollution patterns. Such capability to combine predictive results with spatial layers supported the visualization of scenarios in real-time and their comparison between the monitoring stations. The visualization environment was interactive and made model predictions easier to interpret, and promote data-driven environmental decision-making.

6.3 Spatio-Temporal Forecasting Module

The spatio-temporal forecasting module combined machine learning forecasts with historical trend forecasts to provide future water quality conditions. The system was used to predict the important indicators in the future time-steps on the basis of the RandomForestRegressor outputs and time-series trend model, including DO, BOD, and the levels of microbial contamination. The temporal encoding helped the model to learn periodicity in the season and long-term trends, whereas the spatial features provided the model with location-specific forecasting. The module was created to produce predictive curves and probabilistic ranges, which then enabled the users to test the predicted concentrations of pollutants in various conditions based on time. The results of the forecast were displayed in time-series plots and maps of spatial projections, which proactively supported planning. This predictive ability made it possible to identify the new trend of pollution and take early intervention measures towards curbing the process.

6.4 Early-Warning and Risk Mapping

A risk mapping and early-warning element was incorporated to define the areas that are prone to pollution and the risky intervals. The classification model took a form of classifying water quality by compliance and non-compliance and these outputs were spatialized to provide vulnerable areas. Predicted levels of the pollutant had threshold based alerts being triggered when the levels rose above regulatory limits. Risk maps were developed based on the predicted pollution levels, spatial variation, and trends of the past. The heat intensity gradients were used to indicate high-risk areas, allowing the authorities to prioritize the areas to intervene. The risk analysis of seasons indicated that there were elevated periods of risk of contamination especially during the low flow periods. This was an early-warning system that improved preparedness and facilitated corrective measures on time to control the pollution.

7. DISCUSSION

7.1 Interpretation of AI-Based Findings

The obtained findings of the AI-based predictive models had high potential to elicit spatial and temporal differences in water quality in the Mula-Mutha River Basin. Random Forest regression model demonstrated high prediction rates to dissolved oxygen, biochemical oxygen demand and microbial indicators, which validate the usefulness of the ensemble learning in predicting the environmentally-related data that are characterized by noise and variability. The analysis of feature importance showed that year and location of monitoring were the most prominent predictors, which implied that spatial heterogeneity and long-term urbanization have a serious impact on the process of water quality degradation. The results demonstrate that the combination of AI and geospatial analytics allows one to interpret the nature of pollution in a holistic manner and increase the predictability of the phenomena as used in decision-support frameworks.

7.2 Non-Linear Dynamics in Water Quality

Water quality behavior on urban river systems is a complex non-linear behaviour that is affected by anthropogenic discharges, seasonal runoff and hydrological variability. The random forest algorithm was helpful in data capturing these non-linear relationships by building more than one decision tree and combining the results. The findings showed that there were no simple linear changes of the pollutant levels over time or space; the levels differed greatly across downstream and upstream sections and seasons. As an illustration, the biochemical oxygen demand and microbial contamination had inverse relationships with dissolved oxygen levels in highly urbanized segments. Seasonal dilution of water in monsoons periods provisionally enhanced the level of water quality and dry-season stagnation aggravated the level of pollution. This results in the significance of the non-linear modeling methods that can effectively characterize the environmental processes. Ensemble AI models have significant predictive power in their capacity to reveal intricate interactions among spatial, temporal and environmental variables and are beneficial in realistic evaluation of river health.

7.3 Implications for Urban River Basin Management

The paper illustrates that there are serious implications of the project to the management of the urban river basin and policy development. Predictive modeling, based on AI and interpreted with the help of geospatial visualization, can help the authorities locate hotspots of pollution, monitor the time trends, and predict the future state of water quality. The analysis of spatial variability shows that the location-specific intervention approaches should be made instead of the homogeneous regulation strategies. Classification models can be used in early-warning to help environmental agencies identify any high-risk pollution event and take corrective action. Moreover, the combination of GIS-based visualization

and predictive analytics increases the transparency and helps to make evidence-based decisions.

7.4 Limitations of the Study

The study is limited in that it has certain limitations even though its results are promising. Predictive models were mainly dependent on historical monitored data at the selected stations which may not be a comprehensive representation of the micro-level variation in the whole river basin. The lack of real time sensor connection meant that the system could not give a continuous monitoring and real time prediction. Also, other external variables like variability in industrial discharge, rainfall intensity and change of land use were not explicitly modeled because of data limitations. Even though Random Forest showed a high predictive power, it might need recalibration to be adopted into other environmental settings where pollution is determined by other specifics.

8. CONCLUSION

In this study, a smart geospatial decision support system that combined an artificial intelligence and time-series analysis was created to assess water quality in the MulaMutha River Basin, Pune, comprehensively. The proposed system managed to integrate geospatial analytics with machine learning models and estimate the main water quality clues and categorize overall compliance condition. Random Forest Regressor was also found to have high predictive power with R 2 values of 0.70, 0.69, and 0.79 dissolved oxygen, biochemical oxygen demand and total coliform respectively which is good modeling of intricate spatio-temporal association. Random Forest Classifier had a high precision in classification at 95.5% and ROC-AUC at 0.96 which suggests the efficiency of the method in identifying the pollution-prone conditions and helping early-warning mechanisms. As the comparative analysis indicated, the Random Forest model proved to be much better than the traditional linear regression and logistic regression methodologies, but the statistical tests helped to verify the significance of the performance improvement ($p < 0.05$). The spatial analysis indicated that there was a huge difference in upstream and downstream levels with the latter having a higher BOD and microbial contamination. Temporal analysis revealed that the water quality decreases in trends between 2015 and 2023 with emphasis on the increasing effects of urbanization and wastewater discharge. The combination of the AI predictions and the visualization with the GIS allowed detecting the pollution hotspots, seasonal processes, and long-term tendencies of degradation. The research will add to the environmental monitoring of AI by showing how the geospatial intelligence concept and ensemble learning can be incorporated into one unified decision-support platform. The framework helps in facilitating proactive policy development, focused intervention development and sustainable management of river basins. To make further improvements to the predictive accuracy and scalability of different river systems, future studies are advised to include real-time sensor data, inputs of remote sensors, and enhanced deep learning models.

References:

Amanatidis, P., Lyratzis, E., Angelopoulos, V., Kouloumpris, E., Skaperdas, E., Bassiliades, N., Vlahavas, I., Maris, F., Emmanouloudis, D., & Karampatzakis, D. (2026). Intelligent water management through edge-enabled IoT, AI, and big data technologies. *IoT, 7*(1), 5. https://doi.org/10.3390/iot7010005

Balla, D., Tari, L., Hajdu, A., Kiss, E., Zichar, M., & Mester, T. (2025). Integrated assessment of groundwater quality using water quality indices, geospatial analysis, and neural networks in a rural Hungarian settlement. *Water, 17*(16), 2371. https://doi.org/10.3390/w17162371

Farzana, S. Z., Paudyal, D. R., Chadalavada, S., & Alam, M. J. (2024). Decision support framework for water quality management in reservoirs integrating artificial intelligence and statistical approaches. *Water, 16*(20), 2944. https://doi.org/10.3390/w16202944

Gacu, J. G., Monjardin, C. E. F., Mangulabnan, R. G. T., Pugat, G. C. E., & Solmerin, J. G. (2025). Artificial intelligence (AI) in surface water management: A comprehensive review of methods, applications, and challenges. *Water, 17*(11), 1707. https://doi.org/10.3390/w17111707

Gonzales-Inca, C., Calle, M., Croghan, D., Torabi Haghighi, A., Marttila, H., Silander, J., & Alho, P. (2022). Geospatial artificial intelligence (GeoAI) in the integrated hydrological and fluvial systems modeling: Review of current applications and trends. *Water, 14*(14), 2211. https://doi.org/10.3390/w14142211

Jones, A., Kuehnert, J., Fraccaro, P., Meuriot, O., Ishikawa, T., Edwards, B., ... & Assefa, S. (2023). AI for climate impacts: applications in flood risk. *Npj Climate and Atmospheric Science*, *6*(1), 63. https://doi.org/10.1038/s41612-023-00388-1

Kadrolli, V., & Kalnoor, G. (2024). IoT and smart sensors for remote sensing healthcare and agriculture applications. *Remote Sensing Earth Systems Sciences, 7*, 364–378. https://doi.org/10.1007/s41976-024-00129-9

Kharat, P. V., Darade, M. M., Dhamdhere, P. B., Shinde, B. A., Patil, Y., Mahajan, R. G., Kurhade, A. S., & Waware, S. Y. (2025). Smart hydrology: The role of artificial intelligence in sustainable soil and water resource management. *Journal of Mines, Metals and Fuels, 73*(7), 2035–2050. https://doi.org/10.18311/jmmf/2025/48871

Khobragade, P., Dhankar, P. K., Titarmare, A., Dhone, M., Thakur, S., & Saraf, P. (2024). Quantum-enhanced AI robotics for sustainable agriculture: Pioneering autonomous systems in precision farming. In *Proceedings of the 2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICAIQSA64000.2024.10882412

Krishnan, S. R., Nallakaruppan, M. K., Chengoden, R., Koppu, S., Iyapparaja, M., Sadhasivam, J., & Sethuraman, S. (2022). Smart water resource management using artificial intelligence—A review. *Sustainability*, *14*(20), 13384. https://doi.org/10.3390/su142013384

Mkhitaryan, K., Sanamyan, A., Mnatsakanyan, M., Kirakosyan, E., & Ratner, S. (2025). Integrating AI and geospatial technologies for sustainable smart city development: A case study of Yerevan. *Urban Science*, *9*(10), 389. https://doi.org/10.3390/urbansci9100389

Mounce, S., Mounce, R., & Boxall, J. (2025). AI-augmented water quality event response: The role of generative models for decision support. *Water*, *17*(22), 3260. https://doi.org/10.3390/w17223260

Naz, I., Fan, H., Aslam, R. W., Tariq, A., Quddoos, A., Sajjad, A., Soufan, W., Almutairi, K. F., & Ali, F. (2024). Integrated geospatial and geostatistical multi-criteria evaluation of urban groundwater quality using water quality indices. *Water*, *16*(17), 2549. https://doi.org/10.3390/w16172549

Nemni, E., Bullock, J., Belabbes, S., & Bromley, L. (2020). Fully convolutional neural network for rapid flood segmentation in synthetic aperture radar imagery. *Remote Sensing*, *12*(16), 2532. https://doi.org/10.3390/rs12162532

Pan, D., Deng, Y., Yang, S. X., & Gharabaghi, B. (2025). Recent advances in remote sensing and artificial intelligence for river water quality forecasting: A review. *Environments*, *12*(5), 158. https://doi.org/10.3390/environments12050158

Papastefanopoulos, V., Linardatos, P., Panagiotakopoulos, T., & Kotsiantis, S. (2023). Deep learning architectures for multivariate time-series forecasting. Encyclopedia. https://encyclopedia.pub/entry/50864

Pons-Ausina, J. F., Hosseini, S. N., & Soriano Olivares, J. (2025). AI for smart water solutions in developing areas: Case study in Khelvachauri (Georgia). *Water*, *17*(8), 1119. https://doi.org/10.3390/w17081119

Rabie, A. B., Elhag, M., & Subyani, A. (2025). Remote sensing, GIS, and machine learning in water resources management for arid agricultural regions: A review. *Water*, *17*(21), 3125. https://doi.org/10.3390/w17213125

Shaamala, A., Yigitcanlar, T., Nili, A., & Nyandega, D. (2025). Machine learning applications for urban geospatial analysis: A review of urban and environmental studies. *Cities*, *165*, 106139. https://doi.org/10.1016/j.cities.2025.106139

Chadalavada, S., Faust, O., Salvi, M., Seoni, S., Raj, N., Raghavendra, U., ... & Acharya, R. (2025). Application of artificial intelligence in air pollution monitoring and forecasting: A systematic review. *Environmental Modelling & Software*, *185*, 106312. https://doi.org/10.1016/j.envsoft.2024.106312

Sukkuea, A., Akkajit, P., Suwannarat, K., Foithong, P., Afsarimanesh, N., & Alahi, M. E. E. (2025). AI-driven time series forecasting of coastal water quality using Sentinel-2 imagery: A case study in the Gulf of Thailand. *Water*, *17*(12), 1798. https://doi.org/10.3390/w17121798

Yang, L., Driscol, J., Sarigai, S., Wu, Q., Lippitt, C. D., & Morgan, M. (2022). Towards synoptic water monitoring systems: A review of AI methods for automating water body detection and water quality monitoring using remote sensing. *Sensors*, *22*(6), 2416. https://doi.org/10.3390/s22062416

Zou, S., Ju, H., & Zhang, J. (2025). Water quality management in the age of AI: Applications, challenges, and prospects. *Water*, *17*(11), 1641. https://doi.org/10.3390/w17111641

Pooja Deepak Pawar
Department of Civil Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune
poojadpawar@bvucoep.edu.in

Vidula Sohoni
Department of Civil Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune
vssohoni@bvucoep.edu.in