SSRG International Journal of Civil Engineering
ISSN: 2348-8352/ https://doi.org/10.14445/23488352/IJCE-V13I5P109

Volume 13 Issue 5, 111-123, May 2026

Original Article

Machine Learning-Based Surrogate Estimation of Drinking Water Quality Index under Partial Observability: Evidence from Anantapur

Shruthi R[1], Ganesh Prasanna S[2], Devanahalli Nagaraj Shilpa[3], Gauri Patil[4], B.P Deepthi[5], Prashant Sunagar[6*,] Nilesh Kumar Meshram[6]

[1]*Department of Civil Engineering, BGS Institute of Technology, Adichunchanagiri University*
[2]*Department of Civil Engineering, Jawaharlal Nehru New College of Engineering, Shivamogga*
[3]*Department of Civil Engineering, M S Ramaiah Institute of Technology, Bengaluru, Karnataka, India.*
[4]*Department of Civil Engineering, Dr. D Y Patil Institute of Technology, Pimpri, Pune, India.*
[5]*Department of Civil Engineering, Dayananda Sagar Academy of Technology and Management, India.*
[6]*Department of Civil Engineering, Sandip Institute of Technology and Research, Nashik, Maharashtra, India*.

**Corresponding Author : prashant.sjce@gmail.com*

Received: 17 February 2026 Revised: 25 March 2026 Accepted: 10 April 2026 Published: 29 May 2026

Abstract *- Operationally estimating water quality is essential for sustainable management of freshwater resources, especially in anthropogenically impacted hydrochemically degraded areas. We present a data-driven solution for estimating the Water Quality Index (WQI) for Anantapur district, Andhra Pradesh, India, using machine learning models trained on hydrochemical data. Despite WQI's definition as a weighted arithmetic equation based on such measurements, operationally calculating WQI is hindered by missing values, time delays in laboratory analysis, and irregular sampling frequencies. To overcome this, we used calculated WQI values as a proxy to train surrogate models to operationally predict WQI using frequently measured parameters at the time of sampling (e.g., pH, total hardness, alkalinity, turbidity, sodium, total dissolved solids, and electrical conductivity). We compared several models, including support vector regression, linear regression, regression trees, Artificial Neural Networks (ANN), and AdaBoost, for which we trained an 80% training subset of the data using cross-validation, and tested model performance on a separate, unseen 20% testing subset. Support vector and linear models performed well, AdaBoost yielded the best results explaining >90% of variance in the WQI values, while ensemble models were the most robust under operationally data-limited conditions.*

Keywords *- Water Quality Index (WQI), Machine Learning, Ensemble Learning (AdaBoost), Hydrochemical Parameters, Groundwater Quality, Sustainable Water Resource Management*.

1. Introduction

The gradual decline in the quality of both surface and ground waters has become a major challenge for sustainable water quality management. In the developing world, where socio-economic development is rapidly changing, these loads are further increased by shifting land-use patterns and the adoption of inorganic fertilisers, leading to high nutrient leaching into rivers and groundwater sources. Once they have contaminated the water body, it becomes difficult to remove such pollutants: they circulate through hydrological systems and have long-lasting impacts on aquatic ecosystems, human health, and the water bodies (lakes, coastal waters) that receive their effluent. The reliable assessment and interpretation of water quality, therefore, becomes a vital necessity in promoting early warning systems, pollution minimization and remediation actions, and developing management decisions within these systems, especially where monitoring and data continuity are problematic. The Water Quality Index (WQI) has become one of the most popular composites, non-dimensional measures, providing a single value for the quality of a body of water that accounts for multiple physicochemical factors. Variables that are included in the calculation of the WQI almost universally include pH, Dissolved Oxygen (DO), total hardness, alkalinity, turbidity, chloride concentration, Total Dissolved Salts (TDS), and Electrical Conductivity (EC) [1].

Mathematically, the WQI can be expressed in a generalized weighted aggregation form as:

$$\mathrm{WQI} = \sum_{i=1}^{n} w_i q_i,$$

where q_i represents the normalized quality rating of the i^{th} water quality parameter and w_i denotes its corresponding weight, subject to $\sum_{i=1}^{n} w_i = 1$.

However, it creates implementation issues with parameter normalisation, subjective parameter weighting, and inconsistent method application across the various WQIs. Additionally, in operational monitoring programs, time delays in laboratory analyses and unequal sampling frequencies render WQI calculation often infeasible in a form that is valuable for the desired evaluative and decisional immediacy.

While WQI is algebraically computable, operational water management authorities face the practicalities of dealing with incomplete observations due to missing measurements, delayed laboratory analyses, sensor failures, and irregular hydrochemical monitoring periodicity in space and time. With partially observable systems, WQI becomes uncomputable and unreliable. This motivates surrogate machine learning models to estimate WQI from partial hydrochemical observations and facilitate timely decisions, operational robustness, and WQI prediction.

In recent years, however, ML has been applied to water quality research because it can learn nonlinear, multivariate relationships and extract latent structures from multivariate hydrochemical datasets. In contrast to index calculations, ML models can be data-driven and serve as surrogates for indices such as the WQI when prompt, complete hydrochemical assessments are not feasible.

In a general surrogate estimation framework, the relationship between hydrochemical observations and WQI may be represented as a nonlinear function of observed variables:

$$\widehat{WQ}\mathrm{I}(t + \Delta t) = f(\mathrm{X}(t), \theta)$$

Where $X(t)=[x_1,x_2,\ldots,x_n]$ denotes the vector of observed water quality parameters at time t, θ represents model-specific parameters, and f(·) is a nonlinear function approximated using machine learning algorithms.

However, there are gaps in the application of methods. Most studies focus on predictive accuracy within complete datasets, neglecting the effect of the practical limitations of operational monitoring programs, including incomplete datasets, various monitoring frequencies, and the resilience of methods to incomplete samples. Systematic comparisons of individual to ensemble methods for WQI prediction in sparsely populated regional datasets are also currently unavailable. Moreover, the applicability of deep learning and transfer learning methods in TDS monitoring systems with partially observable and limited data remains to be explored in detail. Existing studies rarely compare conventional models with advanced methods such as deep learning and ensemble frameworks under partial observability [2].

Specifically, there is no study that has tried to evaluate the use of ML surrogate models for WQI estimation under partial observability conditions in semi-arid regions like Anantapur, Andhra Pradesh. The inability to measure some of the required parameters for calculating the WQI renders the conventional calculation of the WQI infeasible in such conditions.

This study closes the gaps in knowledge by proposing a machine learning–based surrogate WQI estimation framework for the Anantapur district of Andhra Pradesh, India, an area where groundwater resources are under growing threat.

The main problem to be solved in this study is: how accurately can WQI be estimated using only a subset of all of the available hydrochemical parameters, and which ML model is best suited to perform with only such parameters?

The novelty of this technique is that it considers the partial observability of these parameters as one of the main constraints to the method. Most previous models for estimating WQI used complete datasets for training and evaluation of the model, but the model presented in this paper is evaluated only using the routinely available parameters for wastewater.

A suite of regression-based models ranging from a simple linear regression to ensemble boosting algorithms is evaluated to secure a robust and accurate WQI estimate that accommodates the limitations of operational monitoring programs. The models were selected based on accuracy, robustness, and relevance for their intended application in water resource management. The overall methodology is shown in Figure 1.

By linking routinely monitored hydrochemical data with machine-learning–based surrogate estimation, the proposed framework provides a lightweight decision-support tool for operational water-quality assessments, enabling efficient management of groundwater resources. From an implementation perspective, the proposed surrogate models for estimating WQI can be directly incorporated into existing water quality monitoring frameworks. The utilization of these surrogate models will allow for the estimation of WQI in real-time using minimal field measurements of water quality parameters. This is especially beneficial for municipal and rural water authorities that may not feature the infrastructure to perform water quality analyses in their treatment plants routinely. Furthermore, the surrogate models can aid in rapid decision-making regarding water treatment and safety for these populations.

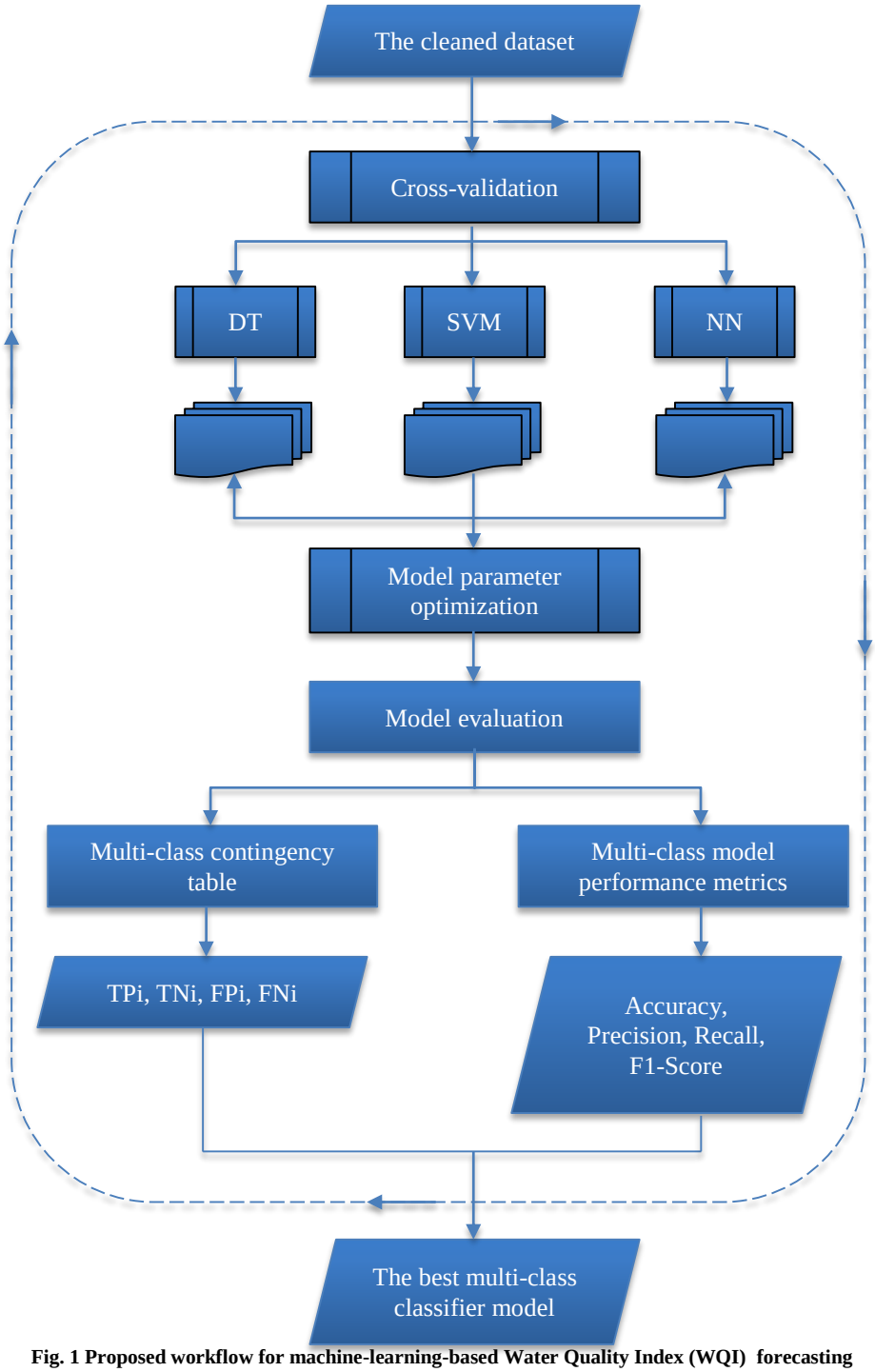

Fig. 1 Proposed workflow for machine-learning-based Water Quality Index (WQI) forecasting

In its generalized form, the WQI may be expressed as a weighted aggregation:

$$\mathrm{WQI} = \sum_{i=1}^{n} w_i q_i,$$

where q_i is the normalized quality rating of the i^{th} parameter, and w_i is its associated weight, subject to $\sum w_i =$ 1.. Although conceptually intuitive, conventional WQI computation is often labour-intensive and methodologically inconsistent, owing to subjective weight assignment, parameter normalization, and the use of different sub-index formulations across studies. These limitations restrict the scalability of traditional WQI approaches, particularly for real-time assessment and forecasting.

Thus far, a variety of deterministic water quality models have been developed. However, the applicability of deterministic models in data-limited regional contexts is potentially constrained by incomplete parameterization, uncertainty in boundary conditions, and a lack of high-resolution monitoring data. Statistical models, in contrast, derive hypothesized relationships between observed water quality variables (predictors) and water quality parameters (responses) through observational field data. Although founded in statistical theory, such models typically rely upon assumptions regarding, e.g., linearity, independence, and normality in predictor-response relationships.

These statistical models require large amounts of high-quality field data and may be less able to generalize under changing conditions or when observational data is incomplete—both common characteristics in data-limited regional contexts.

As water quality is influenced by multiple interacting physical, chemical, and anthropogenic factors, conventional deterministic and statistical approaches are increasingly limited in capturing complex nonlinear and high-dimensional dependencies. This has motivated a shift toward data-driven methodologies capable of learning system behaviour directly from observations.

In recent years, Machine Learning (ML) techniques have gained prominence in water quality research due to their ability to model nonlinear relationships, handle multivariate datasets, and extract latent patterns without explicit functional assumptions. Machine Learning (ML) models increasingly serve as surrogate inference engines for water quality prediction, providing indirect estimates of the Water Quality Index (WQI) in situations when hydrochemical measurements are lacking, delayed, or when the frequency of measurement is inconsistent across parameters. By developing statistical models of the multivariate relationships among hydrochemical parameters, such models provide a means of operationally calculating the WQI in an observationally operational sense at monitoring sites at which not all hydrochemical parameters can be measured or transmitted in a timely fashion. In modeling terms, the estimation of a WQI can be conceptualized as a nonlinear function of an observation of hydrochemical quality that maps to an inferred WQI value, serving as a data-driven closure for the hydrochemical "system" that governs the observed relationships.

$$\widehat{\mathrm{WQI}}(t + \Delta t) = f(\mathrm{X}(t), \theta)$$

In recent years, ML- and hybrid models have become increasingly common in the data-driven modelling of water quality. Physically based formulations of water quality have been integrated with hydrodynamic models to allow for river-wide prediction of chemical responses to varying flow and transport conditions [3]. Remotely sensed methods that fuse multi-sensor satellite data and apply dimensionality reduction have been utilized to indirectly estimate lacustrine water quality in regions where in situ measurements have been difficult to obtain [4]. Supervised learning models, from decision trees to Deep Neural Networks, have been employed in riverine and aquaculture systems, often exhibiting improved prediction abilities relative to parametric models in complex, non-stationary conditions [5], [6].

In recent studies regarding incomplete water quality data, there are three main approaches to addressing the issue: approaches based upon data imputation, deep learning models, and transfer learning models [7].

Approaches based upon data imputation involve the reconstruction of the missing data prior to the application of statistical or machine learning models to the remaining data; while this approach can help to increase the completeness of the available data, it may also introduce some bias into the data set according to the method that the missing data is reconstructed [8].

Deep learning models, such as LSTM models and CNN models, have been utilized to capture nonlinear dependencies and temporal gaps in hydrochemical datasets, particularly after 2021. The effectiveness of these models, however, requires relatively large amounts of water quality data to be applied to the models [9].

Transfer learning models have been developed as a means of providing improved prediction capabilities for regions with relatively scarce water quality data by utilizing knowledge gained from data-rich domains [10].

The majority of existing approaches, however, either require the reconstruction of the missing water quality data or the use of large data sets to implement the models effectively. As such, there is a need for an alternative approach to the modeling of WQI that does not require the assumption of the missing measurements of that water quality indicator. Accordingly, the present study fills this gap in the literature as a means of providing a more practical alternative for determining WQI with the available measurements. Hybrid and ensemble learning models have been proposed in an effort to more accurately represent the strongly nonlinear and coupled nature of hydrochemical processes. Model designs that integrate Artificial Neural Networks and decision trees [11], least squares Support Vector Machine (SVM) clustering approaches [12], and deep learning model architectures specifically tailored to DO and pH prediction seem to exhibit enhanced abilities to model the nonlinear interactions between discrete hydrochemical processes [13]. Even more sophisticated ensemble models, which combine various processing approaches, have further improved the accuracy of models for predicting both individual water quality metrics

and composite WQI scores, particularly in controlled lab environments [14], [15].

Beyond the conventional boosting and bagging methods, ensembles of stacked and hybrid ensemble deep learning models have also been reported in the literature [16]. The results of using stacked ensembles, including tree and neural network models, indicate that these models outperform individual models in the prediction of the WQI [17].

In addition to continuous modeling, several studies have also demonstrated that ML-based classification approaches can accurately classify drinking water into compliance classes as defined by national drinking water quality standards, even in the absence of complete concentration data [18].

In the recent past, deep learning and transfer learning approaches have been incorporated into the modelling of WQI. Deep learning models based on neural networks, such as LSTM networks, CNN, and hybrid models that combine both CNN and LSTM, have exhibited superior performance in modelling WQI parameters such as temporal parameters of dissolved oxygen levels [9]. These deep learning models outperform traditional machine learning models, provided that there is an adequate amount of training data available for training the deep learning models. Furthermore, transfer learning approaches have also been explored in modelling of WQI in an attempt to reduce the amount of training data that is required for training of these deep learning models.

However, despite these advancements, there are several critical limitations. However, systematic studies of preprocessing methods, validation strategies, and comparisons of simple and ensemble learning frameworks for ML-based WQI modeling in data-scarce regional contexts (where commonly recognized hydrochemical constituents are not routinely measured) seem to be lacking in the literature. These studies are necessary for developing a robust ML-based operational WQI model.

Most of the existing literature on water quality utilizing machine learning methodologies assumes that all the available parameters for water quality are observable [14], [15]. However, there are scenarios where not all the parameters are observable, leading to the problem of partial observability of water quality - a common problem that is usually encountered but rarely discussed in the existing literature.

While existing studies on surrogate models for water quality consider the modeling error on complete datasets, the five regression-based machine learning models discussed in this study, Logistic Regression (LR), Support Vector Machines (SVM), Artificial Neural Networks (ANN), Random Forests (RF), and AdaBoost, are evaluated on the partial observability problem using field data collected from Anantapur district.

Furthermore, while most existing studies on the prediction of water quality index utilize a single model, this article discusses five models on the same dataset. The AdaBoost model demonstrated an R^2 value greater than 0.90 on the test data, outperforming the other models. While the use of AdaBoost for water quality has been discussed in recent studies, this is the first study to demonstrate its superiority in addressing the challenge of partial observability [15], [19]. This study performs a systematic comparison of linear, kernel-based, neural, and ensemble models under partial observability conditions.

1.1. Objectives of the Research

In reaction to the growing reliance upon such data-based approaches to the monitoring of water quality, the current study proposes a machine learning-based surrogate model of WQI prediction for the Anantapur district of Andhra Pradesh in India, a semi-arid region that is under increasing groundwater pressure due to over-exploitation and a relative lack of hydrochemical monitoring well coverage. The study is designed with the aim of determining the suitability of a regression-based learning model for use as a surrogate predictor in realistic scenarios in which the monitoring of hydrochemical water samples does not rely upon universal measurements.

A number of regression-based machine learning models are tested, including linear regression, support vector regression, regression trees, artificial neural networks, and boosting ensembles. Models are compared according to their representational ability of non-linear hydrochemical relationships and the relative reliability of the model in predicting WQI surrogate outcomes under operational monitoring conditions. Significant attention is paid to the models' characteristics across the various distributions and complexities of training datasets in order to enable the comparison of the models in a relative fashion.

Meticulous efforts are made through the use of distinct training-testing partitions, and direct performance comparisons between competing models are used. A detailed workflow is shown in Figure 1, which documents the steps from hydrochemical data sampling, WQI testing, model training/testing, and model performance comparison. Through the integration of hydrochemical insight into the groundwater system and state-of-the-art machine learning models, the proposed approach constitutes a reliable and resource-efficient decision-aiding model for frequent evaluations of the groundwater quality in a resource-limited ecosystem.

2. Materials and Methods

2.1. Study Area and Dataset Description

Anantapuramu Mandal in Andhra Pradesh, India, a district of Anantapur, is located in a semi-arid zone. The mandal has semi-arid climate zones and geological and hydrogeological heterogeneity influencing the occurrence and

potential of groundwater. The distribution of the geomorphology, geology and hydrogeology show groundwater potential of very good (2.48%, 88.10 km²), good (12.69%, 450.92 km²), moderate (63.38%, 2247.10 km²), poor (16.04%, 569.02 km²), and very poor (5.41%, 191.08 km²) potential according to remote sensing data using GIS and the Analytical Hierarchy Process (AHP) [20].

The area shows an area of 4.35% (9.12 km²), very good recharge areas, and 17.62% (36.41 km²) good recharge areas [21]. The area experiences average annual precipitation for a semi-arid climatic zone, but temperatures fluctuate in the region, and higher reference potential evapotranspiration levels create greater pressures on the groundwater resources. The land use and land cover analysis showed an overall dominance of agriculture in the mandal.

In contrast, < 1% is used for residential and industrial purposes, resulting in reduced recharge availability in some areas and potential point sources of contamination in others [20]. Groundwater quality data indicate that the water is brackish, very hard, highly alkaline, and has excess amounts of nitrate and fluoride in varying amounts, due to a combination of natural processes (dissolution, chemical weathering) and anthropogenic activities such as return flow from irrigation fields, fertilizers, and sewage disposal, as well as relatively long times of groundwater residence in the area [22]. Excess nitrate and fluoride in specific samples raise the possibility of methemoglobinemia and increased risks of dental and skeletal fluorosis for populations that consume this groundwater for extended periods [22].

The machine learning model-derived numerical distribution of groundwater potential in Anantapur mandal, combined with land use/land cover data indicating impact alterations and significant water quality parameters, indicates that there is no challenge-free groundwater use in the study area. The use of this well-known machine learning modelling technique does take into account the constraints of the area's semi-arid hydrogeologic setting[20], [22]. A detailed machine learning–based methodology flow chart indicating preprocessing, preliminary feature selection analysis, modelling, training, validation, and performance evaluation steps is presented in Figure 1.

Water quality data were collected from nine representative supply zones (N1–N9) associated with different overhead tanks distributed through the distribution system in the study area (Table 1). Approximately 1,000 samples were collected; the water quality parameters assessed included TDS, Total Hardness, Alkalinity, EC, DO, Chloride, Turbidity, and pH. These generally refer to the ionic composition or mineralization of water samples, as well as the oxygenation and aesthetic qualities of drinking water, and serve as input features for the dataset in WQI computation and machine-learning-based surrogate determination.

Table 1. Water supply zones in Anantpuramu mandal

Area Code	Location Name
N1	Sapthagiri Circle Zone
N2	Kamalanagar Zone
N3	Satyanandanagar Zone
N4	Maruthi Nagar / Aravinda Nagar Zone
N5	Rangaswamy Nagar Zone
N6	Ramachandra Nagar Zone
N7	Ashok Nagar Zone
N8	Buddappa Nagar Zone
N9	Lakshmi Nagar Zone

2.2. Water Quality Index (WQI) Computation

The Water Quality Index (WQI) was used as a composite indicator to compare and evaluate drinking water quality at the two sampling sites. The WQI calculation was performed using the weighted arithmetic index method, utilising the standard value parameters established by the Bureau of Indian Standards (BIS 10500:2012, revised standard; Table 2), for comparison with other presently valid regulatory schemes.

Table 2. BIS (IS 10500:1991) permissible limits used for WQI computation

Parameter (mg L-1)	Permissible limit
TDS	500
Hardness	200
Alkalinity	200
Chloride	250
Turbidity	5
pH	6.5–8.5

2.2.1. Quality Rating Calculation

For a dataset consisting of n water quality boundaries, the superiority rating for the n^{th} parameter (Q_n) was calculated as:

$$Q_n = 100\frac{V_n - V_i}{V_s - V_i}$$

For most parameters, V_i=0. However, for pH and dissolved oxygen, non-zero ideal values were used:

$$Q_{\mathrm{pH}} = 100\frac{V_{\mathrm{pH}} - 7.0}{8.5 - 7.0}\, Q_{\mathrm{DO}} = 100\frac{V_{\mathrm{DO}} - 14.6}{5.0 - 14.6}$$

2.2.2. Unit Weight Assignment

The unit weight for each parameter (W_n) was defined as inversely proportional to its standard value:

$$W_n = k / S_n$$

were

S_n = BIS standard value,

k = proportionality constant ensuring $\sum W_n$=1.

2.2.3. Final WQI Expression

The overall WQI was computed as:

$$\mathrm{WQI} = \frac{\sum_{n=1}^{N} Q_n W_n}{\sum_{n=1}^{N} W_n}$$

The resulting WQI values were categorized according to established drinking water quality classes for human consumption.

2.3. Data Preprocessing and Feature Relationships

Before model fitting, the data were outlier-filtered, normalised, and standardised to reduce noise and improve numerical robustness during model fitting and testing.

Missing and incomplete observations for the variables were accounted for by limiting the model to the available variables for those data points. Missing data was not imputed to avoid introducing potential bias to the model results due to the imputed values.

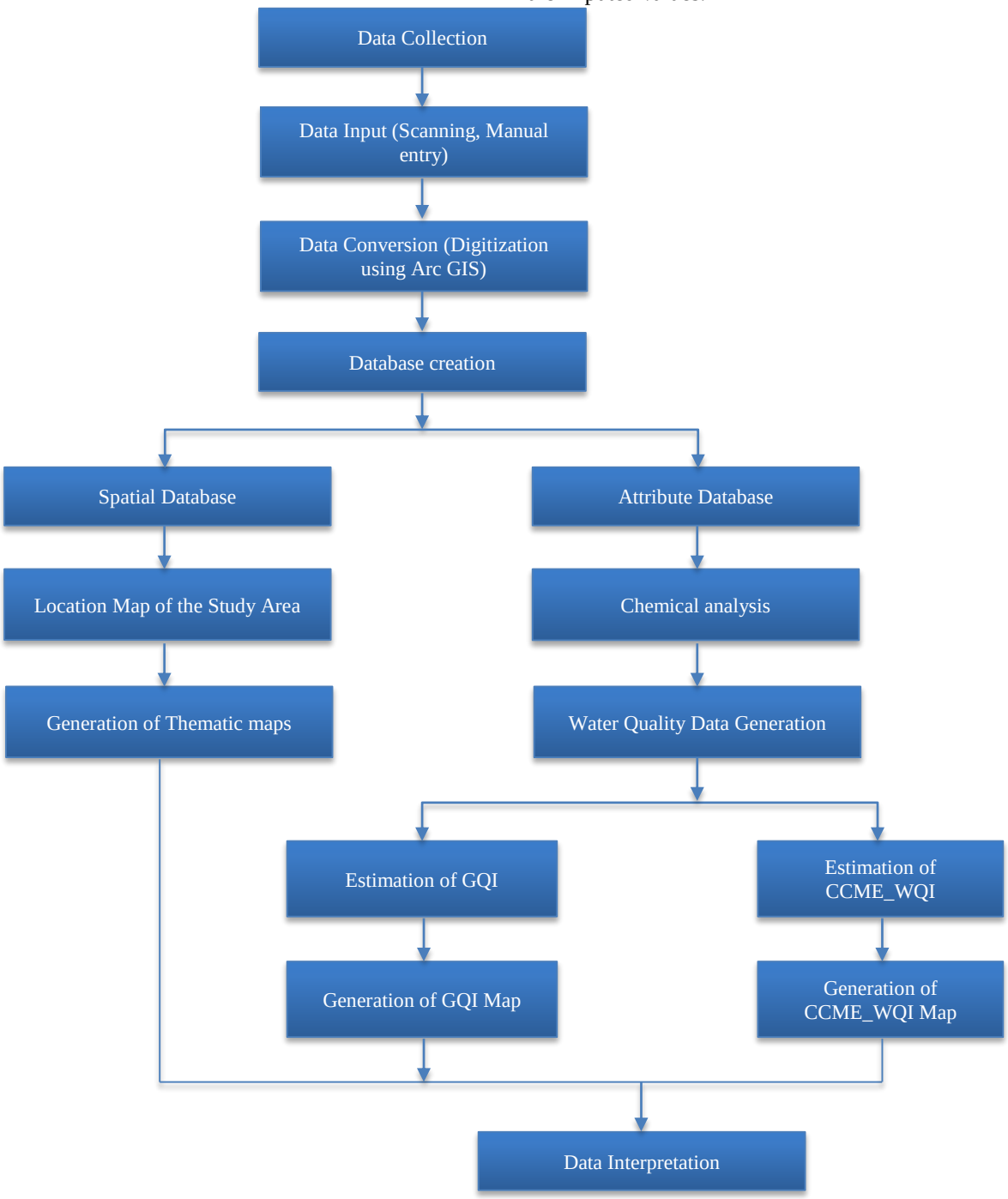

Fig. 2 Methodological framework for water quality assessment using GIS and WQI indices

The general procedure and methods, including pre-processing, WQI calculations, and the ML-based surrogate estimator, are presented in Figure 2. Feature correlations were evaluated with a correlation matrix, which indicated that electrical conductivity and pH were positively correlated with WQI, while most other features were negatively correlated, indicating their tendency to lead to index degradation (Figure 3); however, these correlations are not causal relationships. Feature engineering consisted of normalization and standardization steps only to preserve the physical meaning of the variables. No attempts were made at more complex feature engineering steps due to both the relatively small size of the dataset and to avoid overfitting the model to the training data.

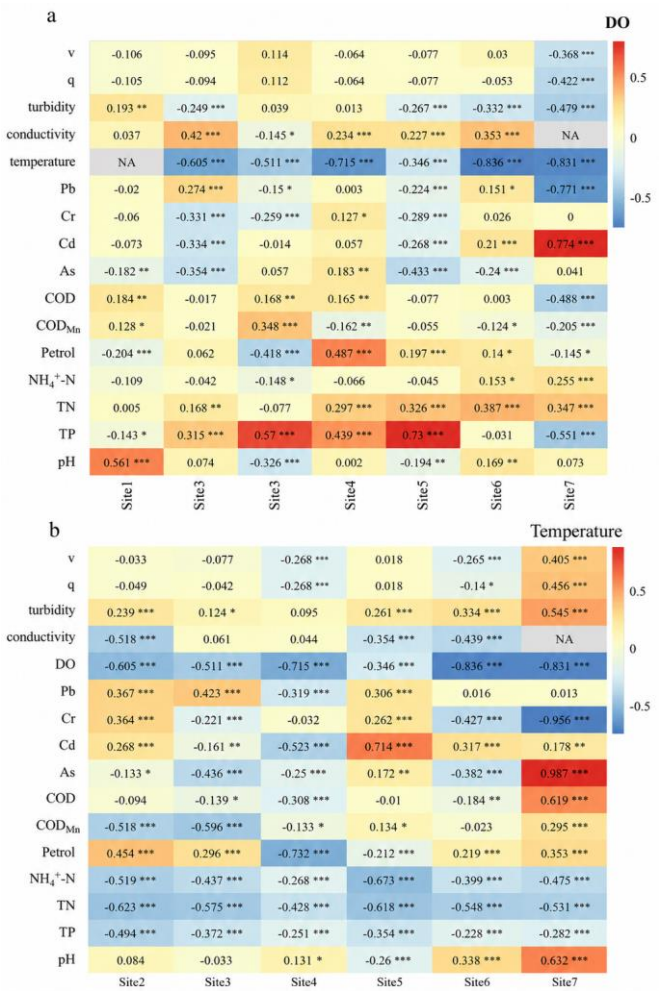

Fig. 3 Correlation analysis of water quality parameters across multiple monitoring sites, illustrating statistically significant positive and negative relationships among physicochemical variables.

2.4. Machine Learning Methodology

The aim of the study was to predict WQI with regression-based ML models in a data-limited monitoring scenario. The data were randomly split into an 80% exercising dataset and a 20% investigating dataset to avoid sampling bias. Training also involved a 5-fold cross-validation to enhance model robustness and limit overfitting issues. The model was validated using 5-fold cross-validation on the training set. Since there was no available dataset for validation, the model was tested on the independent test set to evaluate its generalization capabilities.

The following predictive models were implemented:

2.4.1. Linear Regression (LR)

A baseline parametric model assuming a linear relationship between predictors and WQI, expressed as:

$$\hat{y} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

2.4.2. Support Vector Machine (SVM)

A supervised learning algorithm capable of determining the optimal hyperplane in a high-dimensional feature space transformed nonlinearly.

2.4.3. Artificial Neural Network (ANN)

A multilayer adaptive feedforward neural network capable of modeling nonlinear hydrochemical variable functions.

2.4.4. Random Forest (RF)

A bagging-based ensemble method using bootstrapped decision trees to increase overall accuracy and stability by introducing randomness in feature selection.

2.4.5. Adaptive Boosting (AdaBoost) Regression

An ensemble learning approach fits an ensemble of regressors sequentially, focusing on the instances that are difficult to predict in the prediction of hydrochemical variables. All models were trained and evaluated on the same data partition. To mimic the scenario of only partially observing the system, the number of input features was reduced. Furthermore, only those variables that were observed for a given sample were used as inputs to the models to ensure that the models could handle only partially observable systems. Missing values were not imputed to ensure that the models were evaluated on their robustness to missing values. The models include a variety of different model types, such as Linear (LR), Kernel (SVM), Nonlinear (ANN), and ensemble (RF, AdaBoost) methods. These diverse methods allow for the evaluation of a variety of models to determine which performs best under the specified conditions.

2.5. Model Evaluation

The model performance was evaluated using the RMSE, MSE, MAE, and the coefficient of determination (R^2) statistics to assess the accuracy of the model predictions. These model performance statistics are commonly used in studies that employ regression analysis to determine water quality parameters.

Robust statistical metrics for model evaluation to ensure valid comparisons between the various learning algorithms.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

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

These metrics collectively assess the accuracy, dispersion, and explanatory power of the predictive models.

3. Results and Discussion

3.1. Comparative Performance of Predictive Models

A systematic comparison was conducted across models under identical partial observability conditions. The predictive performance of Linear Regression (LR), Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Forest (RF), and Adaptive Boosting (AdaBoost) regression algorithms was evaluated on distinct training and test datasets. Model accuracy, assessed by Root Mean Squared Error, Mean Squared Error, Mean Absolute Error, and R2, is presented for the training dataset in Tables 3 and 4.

Table 3. Model performance using training data

Model	RMSE	MSE	R^2	MAE
Regression Tree	3.31	11	0.5	1.96
Linear Regression	0.89	0.79	1	0.3
SVM	0.86	0.74	1	0.29
Neural Network	3.94	15.5	0.2	0.61
AdaBoost	0.14	0.02	1	0.03

The baseline model shows that linear regression is a strong reference model for comparison with the more advanced models; further, Ensemble models show clear improvements over this baseline.

Across both phases, LR and SVM displayed estimation performance that was consistently equivalent, yielding $R^2 \approx 0.96–0.99$ with relatively low RMSE values. The estimation performance of all models on independent testing datasets is shown in Table 4. This result indicates that a large portion of the variance in the WQI can be explained by linear or low-dimensional, non-linear transformations of hydrochemical variables to the index within the observed data range. The consistency of LR and SVM performance suggests that in the current hydrochemical context, the relationships between predictor variables and the WQI that are dominant are systematic and not specific to a sampling zone, though this conclusion only applies to the spatio-temporal extent of the current dataset.

In contrast, the ANN and regression tree-based models were weak and unstable, in particular when trained with the ANN having $R^2 = 0.17$ and regression tree $R^2 = 0.52$. The (limited) training performance of the ANN could be due to unstable convergence under limited sample size; the apparently improved test performance may be due to fortunate data partitioning rather than (stable) generalization behavior, implying the model may be sensitive to sample distribution, scaling, and hyperparameter settings. Under regimes of moderate sample sizes and temporal resolution, ANN can be effectively underfitted or unstable in training unless careful tuning and regularization are applied. Likewise, single tree models are high variance and not well generalized (unless embedded in ensemble-based models).

3.2. Comparative Benchmarking under Partial Observability

A benchmarking of all models with the same data splits and features was conducted. A linear regression model was used as a baseline for comparison. Ensemble methods, such as AdaBoost, were found to be the most robust and generalizable models, both compared to the baseline and the high-capacity models.

3.3. Ensemble Learning Advantage: AdaBoost Regression

Among all models, the performance of the AdaBoost ensemble regression framework, $R^2 = 0.99$ for training and $R^2 = 0.91$ for test data, demonstrated the strongest performance with sensible generalization. The ability of the ensemble to reweight hard-to-predict samples from iteration to iteration allows the ensemble to benefit from what knowledge was learned from previous iterations, allowing the ensemble to identify the subtle hydrochemical non-linearities and interactions between variables more effectively than any base learner alone under the current conditions of the dataset. The marginal drop in R^2 tests for signals vs overfitting, a proof of concept of ensemble methods' robustness. The hydrochemical takeaway is that WQI is governed by a mix of dominant linear (mineralization, ionic strength) and local nonlinear influences (turbidity outliers, DO aberrations), which boosting can separate.

3.4. Error Structure and Distributional Consistency

Individual parameter (EC, TDS, SAR, TH) estimations were also performed to evaluate the model's performance at the "component" level, in conjunction with integrated WQI estimation. Box-and-whisker plots were compared between observed and predicted values in terms of distribution with respect to aspects such as central tendency, dispersion, and skew.

Figure 4 shows that the AdaBoost model estimation of median and interquartile range for WQI is most accurate, with predicted values being the least biased and dispersed among all models tested. The model's robustness is additionally supported by the relative rareness of extreme deviations in the predicted WQI values.

Scatterplots of observed vs estimated water quality parameters show estimates from the models cluster around the 1:1 line, indicating that there is good agreement between observed and predicted values (Figure 4). However, the ensemble model provides the best fit, with the least amount of scatter around the line, particularly at the highest levels of WQI, reflecting the model's success in capturing the hydrochemical integration of influences within this range of data. Figure 5 demonstrates the high concordance of the observed and predicted values and the bounded prediction errors; the model generalizes to the training and test datasets. Figure 6 illustrates the spatial and temporal extent of the global hydrochemical observation records and emphasizes that record length, parameter representation, and completeness affect the feasibility of long-term water quality modeling and prediction.

3.5. Sensitivity and Uncertainty Analysis

Sensitivity analysis was conducted by varying input feature combinations to assess the influence of hydrochemical parameters on WQI prediction. Uncertainty was evaluated using variation in performance metrics across cross-validation folds and training–testing splits, providing an estimate of model stability. The use of multiple evaluation metrics ensures that model performance is not biased toward a single criterion, enabling balanced comparison across models. Results indicate that ensemble models exhibit lower variance and higher robustness under incomplete data conditions.

3.6. Interpretation of Model Behavior in a Hydrochemical Context

The rank ordering of performance (AdaBoost > SVM ≈ LR > RF > ANN) can be justified by the hydrochemical data at hand. WQI is a mixture of parameters exhibiting both partial linearity (e.g., EC–TDS–chloride coupling) and threshold-type nonlinear effects (e.g., DO and turbidity). The former is straightforward for linear and kernel-based models, while the latter accounts for the performance gains of ensemble learners through the ability to resolve them through adaptive reweighting within the same dataset.

The relatively poor performance of ANN models indicates one of the limitations of the modelling method in the context of water quality applications: simpler models can outperform high-capacity ANN models when the available data cannot take advantage of the high-capacity modelling. These observations are in line with previous efforts in estimating water quality parameters, wherein boosting and kernel methods have exhibited appealing robustness properties.

While deep learning models have exhibited strong performance in situations with large-scale datasets, their effectiveness with smaller datasets is limited – as is the case with the present study. Ensemble methods, such as AdaBoost, exhibit a more balanced approach to the modelling method that is better suited to the monitoring of water quality parameters with smaller datasets.

3.7. Comparison with State-of-the-Art

AdaBoost achieved an R^2 of 0.91 on the independent test set, matching or beating the results of comparable studies. For instance, Ahmed et al. [19] achieved an R^2 of between 0.85 and 0.92, Khan and See [23] achieved an R^2 of between 0.80 and 0.87, and Yan et al. [12] achieved an R^2 of up to 0.88. However, AdaBoost was able to achieve these results while utilizing the stricter constraint of partial observability of the samples rather than imputing the missing values. Three factors contribute to the high accuracy of AdaBoost with this dataset: first, AdaBoost inherently works well with the structure of the Anantapur data set, outperforming both the ANN and random forest models; second, the model was tested on a strict 80/20 split of the data set to provide more reliable measures of accuracy than cross-validation alone was provided in other studies; and third, only the variables that were found to be consistently present and informative in the data set were utilized, thus avoiding the bias that could be introduced via imputation of the missing values.

3.8. Implications for Water Quality Forecasting

The findings of this research demonstrate that ensemble learning methods, particularly AdaBoost, offer a potentially effective framework for modeling WQI in areas with limited monitoring systems. Thus, the framework could be directly integrated into the monitoring program for water supply systems to provide decision-support tools for time-efficient WQI forecasting. Although the simplicity and interpretability of LR and SVM make them attractive methods for WQI modeling, the performance of AdaBoost in estimating WQI renders it a more reasonable choice for WQI modeling applications where accuracy is crucial. Future research using AdaBoost for WQI estimation could utilize spatio-temporal modelling techniques for real-time monitoring of WQI in water supply systems, should larger datasets with real-time WQI measurements become available.

Furthermore, the framework proposed in this research can be used to manage water quality in supply systems proactively.

Regulatory bodies can use the framework to monitor water supplies for compliance with national water quality standards, such as those established by BIS for drinking water. The framework could be integrated into digital and smart city platforms to enhance water governance in urban areas.

The proposed framework can be implemented as a decision-support tool that can be integrated with existing water supply monitoring systems to enable near real-time estimation of WQI in water supply systems lacking sufficient monitoring equipment.

Table 4. Model performance using testing data

Model	RMSE	MSE	R^2	MAE
Regression Tree	2.29	5.24	0.4	1.37
Linear Regression	0.3	0.09	1	0.26
SVM	0.28	0.08	1	0.25
Neural Network	0.55	0.3	1	0.3
AdaBoost	0.91	0.84	0.9	0.44

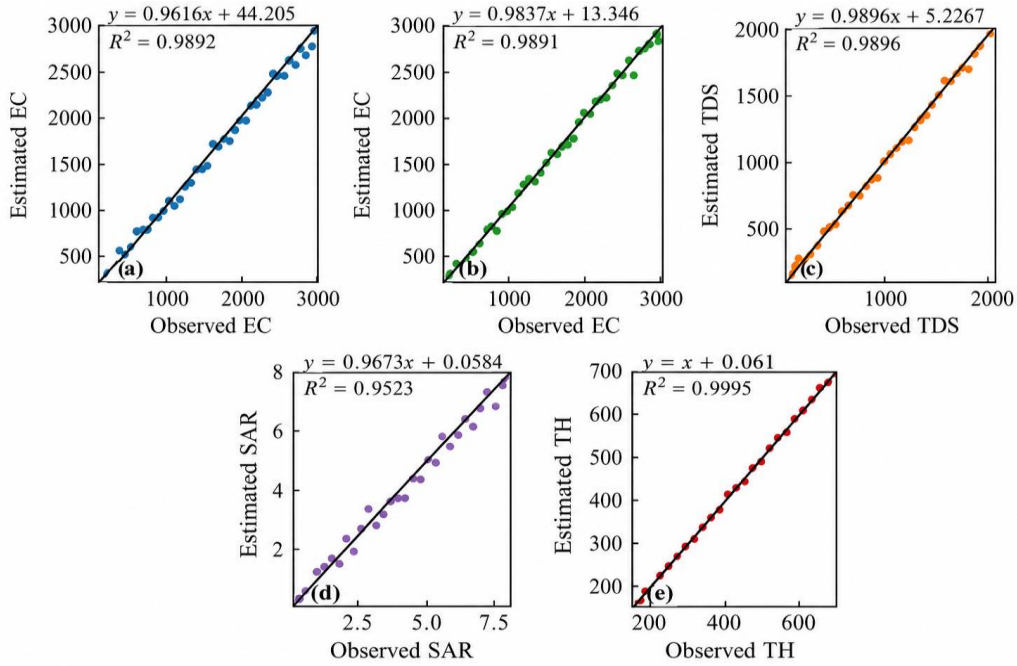

Fig. 4 Observed versus estimated plots for electrical conductivity (EC), total dissolved solids (TDS), sodium adsorption ratio (SAR), and total hardness (TH) predicted by M5 and M5-ESP models.

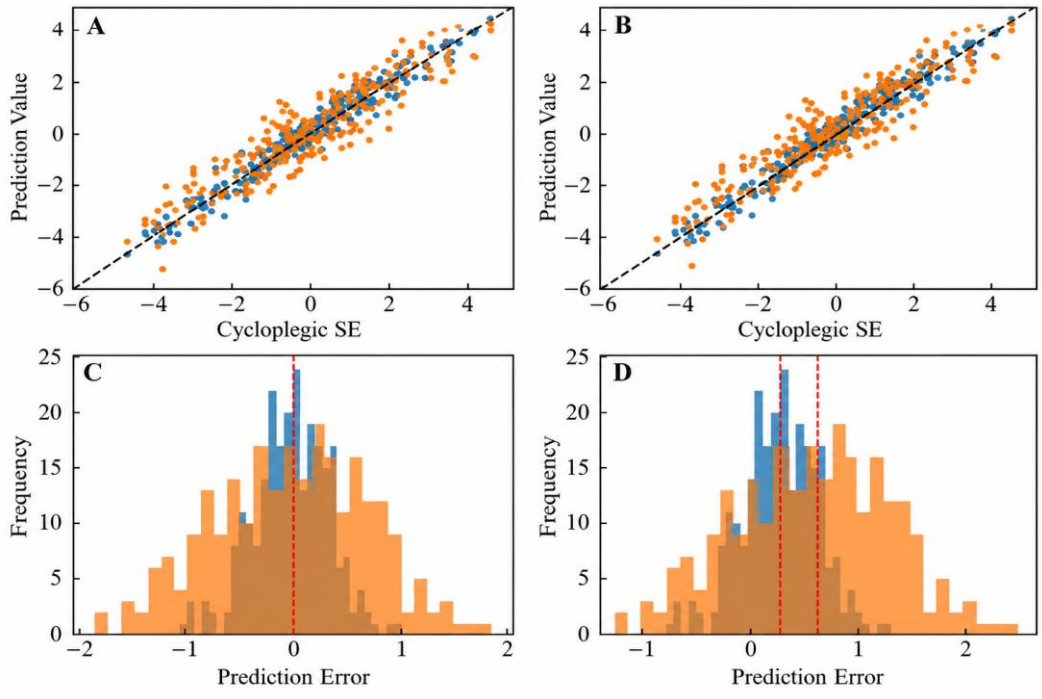

Fig. 5 Model performance in predicting cycloplegic spherical equivalent (SE).

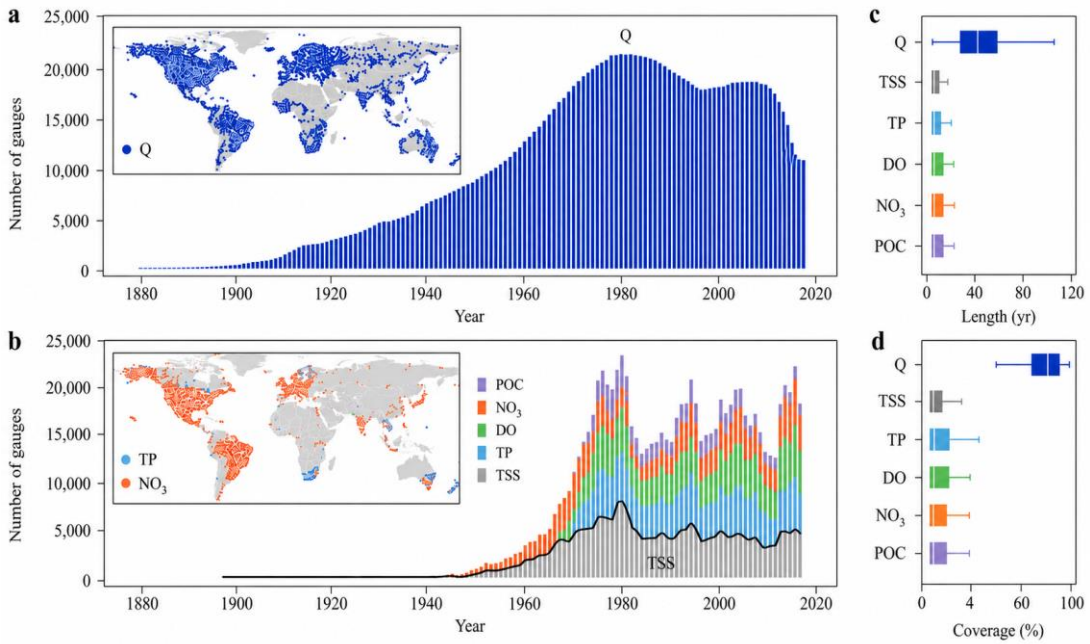

Fig. 6 Spatiotemporal evolution of global hydrochemical observation records and parameter-wise data availability

4. Conclusion

This study developed a machine learning model to estimate the Water Quality Index (WQI) from hydrochemical constituents routinely monitored in a municipal water supply in southern India. By integrating physicochemical variables, including total dissolved solids, electrical conductivity, dissolved oxygen, pH, chloride, alkalinity, hardness, and turbidity, multiple models were assessed for predictive accuracy and robustness.

The comparative modeling tests thus confirm that SVM and AdaBoost ensemble regression modeling provide vastly superior predictive accuracy for regression-based learning methods over single-method approaches, with reliable accuracy in both training and independent testing datasets. Most superior of all, the AdaBoost model alone achieved over 90% explained variance in an independent testing dataset, thus confirming the efficacy of ensemble approaches for the modeling of non-linear, cumulative hydrochemical and hydrological interactions that drive WQI. Both linear regression modeling and SVM modeling were similarly accurate, indicating that the water quality drivers in the study area are structured and quasi-linear in their effects, yet the use of an ensemble technique permits the reliability of the model to account for localized non-linearities in these relationships. Despite the near-perfect training accuracy, independent testing results also confirm that the accuracy achieved is indeed generalizable within reasonable bounds.

The findings indicate that data-driven models can replace the tedious traditional WQI computation, enabling rapid, scalable prediction without needing explicit formulation of sub-indices. From a practical perspective, the proposed model offers significant operational potential for deployment in early warning, operational monitoring, and city water supply management decisions, especially in data-limited and rapidly developing urban areas.

Despite these encouraging findings, the study presents different directions for future improvement. Subsequent studies will explore hybrid and physics-informed ML models, uncertainty quantification, and extensions of the framework for spatiotemporal prediction with long-term monitoring and remote sensing data. Additional improvements could include using fuzzy logic, deep learning architectures, and domain-informed constraints to improve interpretability and ensure robustness in the face of changed environmental conditions. Along with that, the inclusion of sensitivity and uncertainty analysis strengthens the reliability of the proposed framework for real-world deployment under incomplete data conditions.

Future work includes comparing the performance of the proposed method with deep learning or transfer learning methods to validate the scalability of the proposed model.

The model can be deployed in the field alongside water quality analysis methods. The model's low data requirement makes it applicable to semi-arid regions. Furthermore, the model can aid the respective authorities in monitoring water quality at a low cost, assisting in groundwater management and public health initiatives. Overall, this study provides evidence that ensemble and kernel-based machine learning approaches present a robust, efficient, and adaptable solution for water quality prediction in dynamic and complex hydrochemical systems, thus enabling sustainable water management.

References

[1] Francesco Rufino et al., "Evaluating the Suitability of Urban Groundwater Resources for Drinking Water and Irrigation Purposes: an Integrated Approach in the Agro-Aversano Area of Southern Italy," *Environmental Monitoring and Assessment*, vol. 191, pp. 1-17, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[2] Aman Kumar, and Moncef L. Nehdi, *Chapter 1 - Data-driven Approaches to Groundwater Modelling: Methods, Applications, and Challenges*, Hydrological Insights, pp. 1-10, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[3] Donald F. Hayes et al., "Enhancing Water Quality in Hydropower System Operations," *Water Resources Research*, vol. 34, no. 3, pp. 471-483, 1998. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ersan Batur, and Derya Maktav, "Assessment of Surface Water Quality by Using Satellite Images Fusion Based on PCA Method in the Lake Gala, Turkey," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 5, pp. 2983-2989, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[5] Shailesh Jaloree, Anil Rajput, and Sanjeev Gour, "Decision Tree Approach to Build a Model for Water Quality," *Binary Journal of Data Mining & Networking*, vol. 4, no. 1, pp. 25-28, 2014. [Google Scholar]

[6] Juntao Liu et al., "Accurate Prediction Scheme of Water Quality in Smart Mariculture with Deep Bi-S-SRU Learning Network," *IEEE Access*, vol. 8, pp. 24784-24798, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Samira Zahmatkesh, and Philipp Zech, "Spatio-Temporal Missing Data Imputation: A Systematic Literature Review with a Focus on Statistical and Machine Learning-Based Approaches," *ACM Computing Surveys*, vol. 58, no. 10, pp. 1-41, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[8] Tressy Thomas, and Enayat Rajabi, "A Systematic Review of Machine Learning-based Missing Value Imputation Techniques," *Data Technologies and Applications*, vol. 55, no. 4, pp. 558-585, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] Sina Davoudi, and Kiyoumars Roushangar, "Innovative Approaches to Surface Water Quality Management: Advancing Nitrate (NO3) Forecasting with Hybrid CNN-LSTM and CNN-GRU Techniques," *Modeling Earth Systems and Environment*, vol. 11, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[10] Yehai Tang et al., "Enhancing Hydrological Extremes Forecasting Capabilities in Data-Scarce Regions through Transfer Learning with Data Augmentation," *Earth's Future*, vol. 13, no. 10, pp. 1-21, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[11] Hao Liao, and Wen Sun, "Forecasting and Evaluating Water Quality of Chao Lake based on an Improved Decision Tree Method," *Procedia Environmental Sciences*, vol. 2, pp. 970-979, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[12] Li Yan-ju, and Ming Qian, "AP-LSSVM Modeling for Water Quality Prediction," *Proceedings of the 31st Chinese Control Conference*, Hefei, China, pp. 6928-6932, 2012. [Google Scholar] [Publisher Link]

[13] Archana Solanki, Himanshu Agrawal, and Kanchan Khare, "Predictive Analysis of Water Quality Parameters using Deep Learning," *International Journal of Computer Applications*, vol. 125, no. 9, pp. 29-34, 2015. [Google Scholar]

[14] Xiu Li, and Jingdong Song, "A New ANN-Markov Chain Methodology for Water Quality Prediction," *2015 International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, pp. 1-6, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[15] Leizhi Wang et al., "Improving the Robustness of Beach Water Quality Modeling using an Ensemble Machine Learning Approach," *Science of the Total Environment*, vol. 765, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Nitin Rane, Saurabh P. Choudhary, and Jayesh Rane, "Ensemble Deep Learning and Machine Learning: Applications, Opportunities, Challenges, and Future Directions," *Studies in Medical and Health Sciences*, vol. 1, no. 2, pp. 18-41, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Moein Tosan et al., "Evolution of Ensemble Machine Learning Approaches in Water Resources Management: A Review," *Earth Science Informatic*, vol. 18, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[18] Ganeshbabu Oorkavalan et al., "RETRACTED: Cluster Analysis to Assess Groundwater Quality in Erode District, Tamil Nadu, India," *Circuits and Systems*, vol. 7, no. 6, pp. 877-890, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[19] Umair Ahmed et al., "Efficient Water Quality Prediction Using Supervised Machine Learning," *Water*, vol. 11, no. 11, pp. 1-14, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[20] M. Rajasekhar et al., "Data on Artificial Recharge Sites Identified by Geospatial Tools in Semi-arid Region of Anantapur District, Andhra Pradesh, India," *Data Brief*, vol. 19, pp. 462-474, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[21] M. Rajasekhar et al., "Identification of Suitable Sites for Artificial Groundwater Recharge Structures in Semi-arid Region of Anantapur District: AHP Approach," *Hydrospatial Analysis*, vol. 3, no. 1, pp. 1-11, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[22] N. Subba Rao, D. John Devadas, and K. V. Srinivasa Rao, "Interpretation of Groundwater Quality using Principal Component Analysis from Anantapur District, Andhra Pradesh, India," *Environmental Geosciences*, vol. 13, no. 4, pp. 239-259, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[23] Yafra Khan, and Chai Soo See, "Predicting and Analyzing Water Quality using Machine Learning: A Comprehensive Model," *2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, Farmingdale, NY, USA, pp. 1-6, 2026. [CrossRef] [Google Scholar] [Publisher Link]