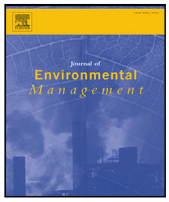
Journal of Environmental Management 352 (2024) 120091

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman



Research article

HDTO-DeepAR: A novel hybrid approach to forecast surface water quality indicators

Rosysmita Bikram Singh [a], Kanhu Charan Patra [a], Biswajeet Pradhan [b,c,*], Avinash Samantra [d]

[a] Department of Civil Engineering, National Institute of Technology, Rourkela, 769008, Odisha, India
[b] Centre for Advanced Modelling and Geospatial Information System, School of Civil and Environmental Engineering, University of Technology Sydney, Sydney, Australia
[c] Institute of Climate Change, Universiti Kebangsaan Malaysia, Bangi, Malaysia
[d] Department of Computer Science & Engineering, National Institute of Technology, Rourkela, 769008, Odisha, India

ARTICLE INFO

Keywords:
Deep learning
HDTO-DeepAR
Time series forecasting
Water quality

ABSTRACT

Water is a vital resource supporting a broad spectrum of ecosystems and human activities. The quality of river water has declined in recent years due to the discharge of hazardous materials and toxins. Deep learning and machine learning have gained significant attention for analysing time-series data. However, these methods often suffer from high complexity and significant forecasting errors, primarily due to non-linear datasets and hyperparameter settings. To address these challenges, we have developed an innovative HDTO-DeepAR approach for predicting water quality indicators. This proposed approach is compared with standalone algorithms, including DeepAR, BiLSTM, GRU and XGBoost, using performance metrics such as MAE, MSE, MAPE, and NSE. The NSE of the hybrid approach ranges between 0.8 to 0.96. Given the value's proximity to 1, the model appears to be efficient. The PICP values (ranging from 95% to 98%) indicate that the model is highly reliable in forecasting water quality indicators. Experimental results reveal a close resemblance between the model's predictions and actual values, providing valuable insights for predicting future trends. The comparative study shows that the suggested model surpasses all existing, well-known models.

1. Introduction

Water quality plays a crucial role for the inhabitants of ecosystems. Significant changes in aquatic environments have occurred during the last several decades due to point and non-point source pollution. As a result, concerns regarding the purity of river water have been raised. Agricultural and industrial activities, municipal wastewater, soil erosion, rising temperatures, and heavy metal pollution (mining) are the anthropogenic factors that possess a negative impact on water quality (Vlad et al., 2012; Uddin et al., 2021; Wątor and Zdechlik, 2021). One of the most challenging issues is identifying both the sources of pollution. Time series forecasting aids in detecting unusual patterns or anomalies in water quality data over time. Sudden increases or decreases in parameter concentration (pH, dissolved oxygen, turbidity), pollutant concentrations, biological contamination, leakages in sewage pipes, or other environmental changes might threaten aquatic ecosystems or human health (Yousefi et al., 2021). In this manuscript, multiple time-series models are employed to forecast water quality indicators based on the temporal pattern of data collected over time. Esterby (1996) reviewed techniques for identifying patterns in water quality. The author discussed parametric and non-parametric techniques that account for seasonality by segmenting the data based on the assumption of a monotonic trend. The analysis is helpful for policymakers and regulators involved in timely water quality management.

In this river basin, numerous short-term studies have been conducted to examine the geographical and temporal variability of physicochemical parameters in connection to anthropogenic influences (Chakrapani and Subramanian, 1990; Subramanian, 1980; Sundaray et al., 2006, 2011). The population residing in the vicinity of the river basin relies entirely on it for irrigation and industrial purposes. Therefore, water quality is a matter of concern for future usage.

The Water Quality Index (WQI) (Horton, 1965) is a valuable tool for analysing and reporting the overall quality of water in a particular region. It is calculated using the weighted arithmetic WQI technique to estimate the suitability of water for human consumption, irrigation, and other applications. The deterioration in water quality in

* Corresponding author at: Centre for Advanced Modelling and Geospatial Information System, School of Civil and Environmental Engineering, University of Technology Sydney, Sydney, Australia.
E-mail addresses: puja00003@gmail.com (R.B. Singh), kcpatra@nitrkl.ac.in (K.C. Patra), Biswajeet.Pradhan@uts.edu.au (B. Pradhan), samantraavinash777@gmail.com (A. Samantra).

https://doi.org/10.1016/j.jenvman.2024.120091
Received 21 September 2023; Received in revised form 23 December 2023; Accepted 8 January 2024
Available online 15 January 2024

rivers, exacerbated by rapid population growth, poses a serious global threat. The WQI model primarily consists of parameter selection, sub-index processing, parameter weighting, and the aggregation function steps (Uddin, 2020). However, subsequent research has revealed concerns about ambiguity with the weighting technique. The rank sum (RS) approach was used by Uddin et al. (2022b). The results indicate that the RS technique is more dependable than other strategies for handling uncertainty in the WQI model. Miyittah et al. (2020) assessed the contamination status of a lagoon system using multivariate statistical tests. Kumar et al. (2021) evaluated pollution levels over 49 years using a Cartesian coordinate system and trend analysis of water quality indices (WQI). In addition, studies were conducted to analyse the influence of urban expansion on river water quality, using a modified NSF WQI to quantify regional changes (Parween et al., 2022). WQI plays a cardinal role in determining the health and quality of river ecosystems, although its values may not accurately represent the true state of water quality. Hence, a new approach using artificial intelligence was undertaken to forecast based on qualitative classes (Georgescu et al., 2023). Long-term physiological characteristics were assessed using the Comprehensive Pollution Index (CPI) and Environmetrics (Dimri et al., 2021), classifying and simplifying large datasets. Machine learning algorithms, including XGBoost and KNN, outperformed other methods in predicting water quality classes. To address uncertainty in WQI models, a Monte Carlo simulation technique was employed, and the Gaussian Process Regression (GPR) approach was used to predict uncertainty at each sample location (Uddin et al., 2023d). An accurate and improved Iris WQI was proposed by Uddin et al. (2023e) for evaluating transitional and coastal water quality, minimizing model uncertainty. The WQM-WQI approach proved to be more successful because of bias-free assessment.

Additionally, a new data-driven technique was established to evaluate trophic levels in intermediate and coastal waters, indicating nutrient levels and potential algal growth. The Assessment Trophic Status Index (ATSI) model, combined with machine learning algorithms, is a viable approach for evaluating trophic conditions (Uddin et al., 2023g). Future research may focus on integrating emerging pollutants and socioeconomic factors into water quality indices. Moreover, the inclusion of socioeconomic factors in indices would offer a more comprehensive assessment. Traditional testing procedures for monitoring water quality characteristics are time-consuming and limited to specific regions due to extensive field sampling and expensive laboratory analysis. These limitations highlight the challenges faced by traditional approaches in estimating water quality at geographical scales (Song and Kim, 2009).

In the age of enormous data collection, the application of computational intelligence (CI) methods in various hydrological settings has been rising with a focus on modelling techniques. Arya and Zhang (Arya and Zhang, 2015) deployed the ARIMA approach for time series analysis to forecast dissolved oxygen and surface water temperature. Antonopoulos et al. (2001) used statistical techniques to assess monthly time series of water quality metrics and discharge. To choose the optimum theoretical distribution for the data, the X^2-test and the Kolmogorov–Smirnov (K–S) test were performed. Arya and Zhang (2017) applied a copula-based Markov process to simulate water quality time series. One of the most crucial indicators for summarizing the state of surface water is the concentration of dissolved oxygen. Csábrági et al. (2017) adopted four linear and non-linear models for forecasting dissolved oxygen concentration. It revealed that non-linear models outperformed the linear models. Deng et al. (2021) implemented and upgraded artificial neural networks (ANN) and support vector machines (SVM) to precisely forecast the algal development and eutrophication in the harbour. The findings showed that ANN was preferable for generating excellent outcomes with immediate response, whereas SVM was appropriate for correctly selecting the optimum model but required a longer training period. To predict the spatial and temporal fluctuations of water quality indicators, machine learning (ML) techniques were developed in conjunction with the Environmental Kuznets Curves (EKC) (Deng et al., 2022). Georgescu et al. (2023) studied non-linear relationships between water quality and flow status parameters. The cascade-forward network (CFN) and radial basis function network models were employed as reference models.

Deep learning models excel in capturing deep correlations in data, making them ideal for modelling the intricacies of water quality dynamics. These models are valuable for predicting water quality metrics, especially when relationships between these indicators are challenging to specify using traditional methods. Hyperparameter tuning in deep learning is an important feature that has a substantial influence on the performance, convergence time, and generalization ability of neural network models. Choosing the best values may result in models with higher accuracy, reduced error rates, and better performance on a variety of tasks. Researchers have developed several metaheuristic techniques to produce accurate results. Takieldeen et al. (2022) developed a novel optimization technique dipper throated optimization (DTO) derived from dipper throated birds. It has an exceptional hunting strategy that involves rapid bowing movements. DTO is scrutinized and contrasted to particle swarm optimization (PSO) (Kennedy and Eberhart, 1995; Poli et al., 2007; Xiang and Jiang, 2009), whale optimization algorithm (WOA) (Ge et al., 2023; Guo et al., 2020), grey wolf optimizer (GWO) (Mirjalili et al., 2014; Faris et al., 2018; Cuong-Le et al., 2022), and genetic algorithm (GA) (Mirjalili and Mirjalili, 2019; Srinivas and Patnaik, 1994; Kuo et al., 2006) to demonstrate its effectiveness. Khullar and Singh (2022) employed the BiLSTM approach to predict river water quality.

Several recent investigations have indicated that the WQI model generates significant ambiguity in its modelling process due to the overestimation of the index by the aggregation function (Abbasi and Abbasi, 2011; Chang et al., 2020; Uddin et al., 2023c). Due to the unreliability of previous WQI methodologies, a few researchers have lately applied the ML methodology to minimize model uncertainty and forecast WQIs accurately (Bui et al., 2020; Hassan et al., 2021; Othman et al., 2020; Kouadri et al., 2021; Khan et al., 2022). Addressing eclipsing and ambiguity issues in water quality forecasting requires careful consideration of model construction and data quality. Numerous studies have been carried out on the prediction of critical water quality indicators. To estimate river water quality, researchers explored multiple prediction models, including independent ML, deep learning (DL), and hybrid models. While dealing with complex or dynamic systems, standalone models may lack resilience. Many standalone models provide point estimates without explicitly quantifying uncertainty (Irwan et al., 2023). Therefore, metaheuristic approaches are intended to effectively search a wide search space. They excel at identifying excellent solutions throughout the whole hyperparameter field, making them beneficial for global optimization (Liu et al., 2021; Morales-Hernández et al., 2023).

However, most research has concentrated on the water quality index and parameter point prediction (Pany et al., 2023; Islam et al., 2022; Zheng et al., 2023). But there are no studies on probabilistic water quality time series forecasting in the Mahanadi River system.

The manuscript determines the pattern of the water quality parameters in the future. To achieve this goal, a novel hybrid hidden dipper throated optimization- deep autoregressive (HDTO-DeepAR) model has been developed. To measure its robustness four standalone methods (DeepAR, BiLSTM, GRU, XGBoost) are employed. The information supplied in the study has numerous ramifications and relevance in the realm of environmental management. BOD is a crucial indicator that determines the degree of organic pollution in water bodies. The variations in pH content, which are reliant on external environmental conditions, emphasize the dynamic aspect of water quality. Elevated sulphate levels can have implications for both ecological and human health. High sodium concentrations may lead to diminished soil permeability, poor aquatic life, and human health difficulties. Elevated chloride levels may harm freshwater ecosystems by harming aquatic plants and animals. Water temperature changes can have a significant

effect on the metabolic rates, reproduction, and dispersal of aquatic species. Temperature changes can also influence nitrogen cycling and overall ecosystem dynamics. The findings of the study give vital insights into environmental management by giving information on changes in major water quality measures and the accuracy of forecasting models. This data may be used to drive decision-making processes aimed at conserving and improving water quality in the study region. As per the author's knowledge, forecasting of water quality parameters in the Mahanadi River basin using the proposed method has not been carried out previously.

The remaining part of the manuscript is organized as follows: Section 2 demonstrates the methodology, whereas Section 3 delves into the results. Section 4 explains the discussion. Section 5 elaborates on the conclusion.

2. Methodology

A detailed breakdown of the study area, data collection, preprocessing methods, and the proposed method is described in this portion.

2.1. Study area

The Mahanadi River basin in India is chosen as the study area. The river's catchment area spans several states including Madhya Pradesh, Odisha, Jharkhand, and Maharashtra, covering a drainage area of $1,41,589$ square kilometres (4.3 per cent of India's geographical area) (Samuel et al., 2017; Kurwadkar et al., 2022; Konhauser et al., 1997). Owing to its rich mineral resources and dependable electricity supply, the basin provides a favourable industrial environment. Prominent industries situated around the basin include aluminium factories in Hirakud and Korba, a paper mill near Cuttack, and an Iron and Steel plant at Bhilai. Additionally, other industries thrive on coal, iron, and manganese mining. An increasing rate of population within a confined area continually stresses the environment and may lead to the degradation of water quality. Major repercussions of urbanization include a significant shift in land use, a rise in built-up areas, solid waste landfilling, and sewage disposal (Ouyang, 2005). The monitoring stations chosen for the analysis are Brajrajnagar D/S, Sambalpur D/S, and Sonepur D/S stations. Increased water use, inadequate sewage infrastructure, and a lack of wastewater treatment facilities have a substantial impact on water resources. As urbanization progresses, agricultural lands and unpaved roads are being paved, resulting in increased surface imperviousness. Surface water bodies suffer severe pollution from untreated sewage and contaminated urban runoff, rendering them unsuitable for supplying fresh water to cities. The majority of the excess dissolved metals are triggered by industrial and urban pollutants (Hussain et al., 2020; Samantray et al., 2009). The alarming pollution not only degrades water quality but also endangers human health, disturbs the balance of the aquatic environment and affects economic growth. Despite the fact that enterprises have taken every precaution to achieve zero effluent status, there is widespread concern about air, land, and water contamination. Therefore, an initiative has been taken for proper water quality forecasting at various stations of the study area for better management in the future. The monitoring stations in the Mahanadi River is shown in Fig. 1.

2.2. Data collection and pre-processing

The data used in this manuscript were obtained from the Central Pollution Control Board (CPCB), India. Water quality data were collected for six indicators from three stations within the Mahanadi River system, specifically Brajrajnagar D/S (S1), Sambalpur D/S (S2) and Sonepur D/S (S3). Fig. 1 shows the details of the monitoring stations in the Mahanadi River. The data is available on a monthly basis from 2001 to 2022. The water quality indicators include sodium, temperature, BOD, pH, chloride, and sulphate. Samples are collected from a well-mixed portion of the river (mainstream) 30 cm below the water's surface using a weighted bottle. The CPCB laboratory is recognized by the National Accreditation Board for Testing and Calibration Laboratories (NABL), a constituent board of the Quality Council of India. Consequently, the water laboratory implements to all quality assurance procedures for diverse analytical tasks. Various inter-laboratory programmes are employed on a regular basis to evaluate laboratory bias. Quality assurance measures primarily include sample control and documentation, standard analytical procedures, analyst qualifications, equipment maintenance, data reduction, validation, and analytical quality control (CPCB, 2020).

Data preprocessing is a pivotal step in deep learning that includes preparing and cleaning the data prior to feeding into a model. The missing values are imputed using the interpolation method. Thereby, the outliers are identified and removed to avoid noise in the data. Later, the features are normalized to a common scale to ensure convergence during training. For performance assessment, 80% of the data is taken for training, 10% for validation, and 10% for testing.

2.3. Hidden dipper throated optimization

The Hidden Dipper Throated Optimization (HDTO) method is based on a new concept that categorizes birds into two groups: swimming birds and flying birds. This approach was motivated by the unique hunting style and quick bow motions. The two forms of birds cooperate with each other to locate food. This presumption is allocated to the exploration and exploitation groups in order to explore the search space for the optimum solution (Abdelhamid et al., 2023; Takieldeen et al., 2022; Abdelhamid et al., 2022). These groups of birds are distinguished by their locations and velocities.

$$s = \begin{bmatrix} s_{1,1} & s_{1,2} & s_{1,3} & \cdots & s_{1,n} \\ s_{2,1} & s_{2,2} & s_{2,3} & \cdots & s_{2,n} \\ s_{3,1} & s_{3,2} & s_{3,3} & \cdots & s_{3,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ s_{l,1} & s_{l,2} & s_{l,3} & \cdots & s_{l,n} \end{bmatrix} \tag{1}$$

$$u = \begin{bmatrix} u_{1,1} & u_{1,2} & u_{1,3} & \cdots & u_{1,n} \\ u_{2,1} & u_{2,2} & u_{2,3} & \cdots & u_{2,n} \\ u_{3,1} & u_{3,2} & u_{3,3} & \cdots & u_{3,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ u_{l,1} & u_{l,2} & u_{l,3} & \cdots & u_{l,n} \end{bmatrix} \tag{2}$$

where, $s_{x,y}$ represents the yth dimension and the x position of the bird, for $x \in [1, 2, 3, \ldots .., l]$ and $y \in [1, 2, 3, \ldots .., n]$ and its velocity in y_{th} dimension is represented by $u_{x,y}$. The fitness functions for birds in the search space are dictated by $f = f_1, f_2, f_3, \ldots, f_l$, which is defined using the matrix below:

$$s = \begin{bmatrix} f_1(s_{1,1},\ s_{1,2},\ s_{1,3},\ \ldots,\ s_{1,n}) \\ f_2(s_{2,1},\ s_{2,2},\ s_{2,3},\ \ldots,\ s_{2,n}) \\ f_3(s_{3,1},\ s_{3,2},\ s_{3,3},\ \ldots,\ s_{3,n}) \\ \cdots \\ f_l(s_{l,1},\ s_{l,2},\ s_{l,3},\ \ldots,\ s_{l,n}) \end{bmatrix} \tag{3}$$

In a fitness assessment that takes into account each bird's success rate in locating food, the fitness score of the mother bird is the greatest conceivable. Numbers are sorted in decreasing order while sorting. The initial HDTO strategy that the optimizer used to monitor the swimming bird is based on the equations shown below to be considered for changes in the position and velocity of the members of the population:

$$s(t+1) = s_{best}(t) - D_1.|D_2.s_{best}(t) - s(t)| \tag{4}$$

Where s_t is a normal bird position, s_{best} is the best bird position and $s(t+1)$ is the updated bird position.

$$D_1 = 2d.p_1 - d \tag{5}$$

$$D_2 = 2p_1 \tag{6}$$

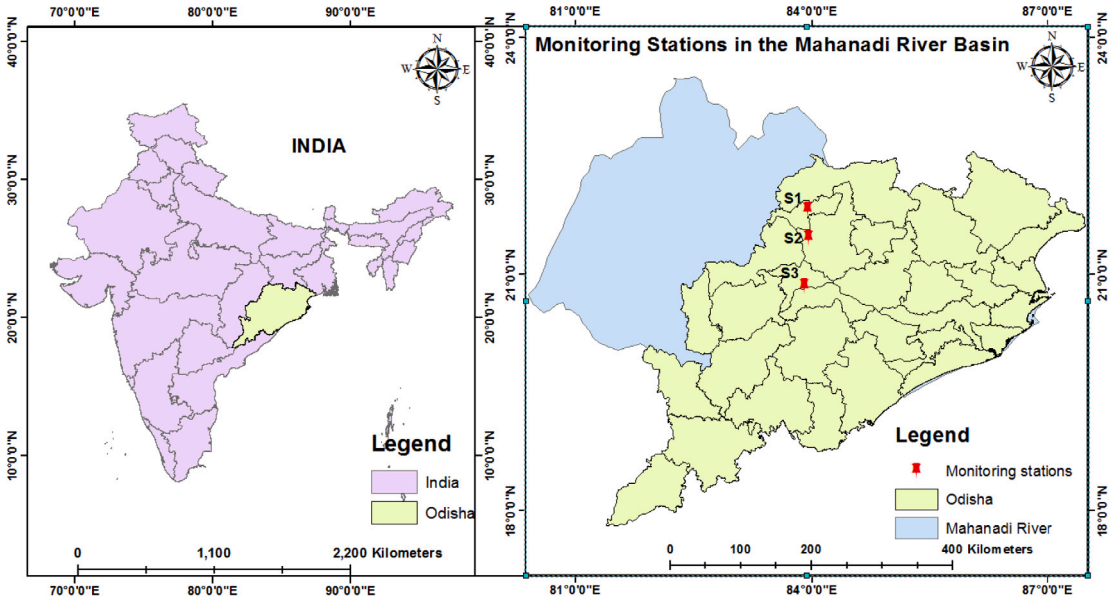


Fig. 1. Water quality monitoring stations in the Mahanadi River.

$$d = 2\left(1 - \left(\frac{t}{T_{max}}\right)^2\right) \tag{7}$$

where, d changes from 2 to 0 exponentially, and T_{max} is the total number of iterations. $u(t+1)$ is the velocities of the bird at iteration $t+1$. The second DTO process is based on the following equations for updating the flying bird's location and velocity. The new position of flying bird is updated as below:

$$s(t+1) = u(t+1) + s(t) \tag{8}$$

where $s(t+1)$ is the new position of a normal bird. The updated velocity of each bird is estimated as:

$$u(t+1) = D_3 u_t + D_4 p_1(s_{best}(t) - s(t)) + D_5 p_2(s_{Gbest} - s(t)) \tag{9}$$

The DTO algorithm can be written as:

$$s(t+1) = \begin{cases} s_{best}(t) - D_1.|K|, \text{if} P < 0.5 \\ u(t+1) + s(t), \qquad otherwise \end{cases} \tag{10}$$

$$K = D_2.s_{best}(t) - s_t \tag{11}$$

Where s_{Gbest} denotes the global best solution. The current iteration's index is marked by t, while the next iteration's index is denoted by $t+1$. The D_1, D_2, and D_3 are weight values. Whereas, D_4, and D_5 are constants. The values of P, p_1, and p_2, are chosen randomly in the range $[0, 1]$.

Algorithm 1 DTO Algorithm()

Require: Optimized solution for the hidden layer of DeepAR
1: **Initialize** population position $s_i(i = 1, 2, 3, \ldots, l)$ with l birds, velocity u_i, objective function f_l, iterations T_{max}, parameters of t = 1, r_1, r_2, R, D_1, D_2, D_3, D_4, D_5
2: **Calculate** f_l for each bird's position s_i
3: **Identify** Best bird position s_{best}
4: **while** $t \leq T_{max}$ **do**
5: **for** $i = 1: i < i+1$ **do**
6: **if** $P < 0.5$ **then**
7: **Update** the position of the swimming bird as:
8: $s(t+1) = s_{best}(t) - D_1.|D_2.s_{best}(t) - s(t)|$
9: **else**
10: **Update** the velocity of the flying bird as:
11: $u(t+1) = D_3 u_t + D_4 r_1(s_{best}(t) - s(t)) + D_5 r_2(s_{Gbest} - s(t))$
12: **Update** the current flying bird position as:
13: $s(t+1) = s(t) + u(t+1)$
14: **end if**
15: **end for**
16: **Calculate** f_l for each bird
17: Set $t = t+1$
18: **Update** P, D_1, D_2
19: **Find** the best position s_{best}
20: Set $s_{Gbest} = s_{best}$
21: **end while**
22: **Return** the best bird s_{Gbest}

2.4. DeepAR

Deep Autoregressive (DeepAR) (Salinas et al., 2020) forecasting method that uses an autoregressive recurrent neural network (AR RNN) to train a global model instead of fitting separate models for each time series like other conventional models. It generates credible probabilistic forecasts and uncertainty estimates based on previous data. Amazon developed the technique, which is renowned for its capacity to expand by leveraging a variety of variables. The model takes a series of past values and predicts the next value in the sequence. It utilizes a combination of LSTM and fully connected layers to capture the underlying patterns and relationships in time series data, and it is trained using a backpropagation and gradient descent technique. The algorithm encodes the input time series during the training phase. Once trained, the model's effectiveness is assessed using assessment measures tailored to the forecasting job. The context window is produced by processing a defined number of historical observations over a given amount of time. To fine-tune the model and increase its accuracy, hyperparameters such as learning rate, layer count, and batch size are modified. The hyperparameter settings of the HDTO-DeepAR is illustrated in the Table 1. The objective of DeepAR is to simulate the conditional distribution as shown in Eq. (12).

$$R(p_{i,t_0:T} | p_{i,1:t_0-1}) \tag{12}$$

where, T is the end of the forecast window, $p_{i,1:t_0-1}$ is the past values, and t_0 is the start of the forecast range. $[1 : t_0-1]$ and $[t_0 : T]$ represents

the conditioning range and prediction range respectively. The DeepAR model forecasts the prediction range as per the conditioning range.

DeepAR assumes that $R(p_{i,t_0:T}|p_{i,1:t_0-1})$ includes likelihood factors. These factors are presented in Eqs. (13) and (14).

$$R(p_{i,t_0:T}|p_{i,1:t_0-1}) = \prod_{t=t_0}^{T} R(p_{i,t}|s_{i,1:t-1}) = \prod_{t=t_0}^{T} R(p_{i,t}|\theta(h_{i,t},\theta)) \tag{13}$$

$$h_{i,t} = h(h_{i,t-1}, p_{i,t-1}, \theta) \tag{14}$$

where, $h_{i,t}$ is the output of the hidden state constructed by LSTM cell parametrized by θ. The standard deviation(σ) and mean(μ) are derived from the hidden state $h_{i,t}$ and become the parameters of the gaussian likelihood function. The model attempts to generate a gaussian distribution that produces predictions that are near the target variable. DeepAR trains (and predicts) a single data point each time, therefore the model is known as autoregressive.

Hence, the Gaussian likelihood (l) is determined in the following Eq. (15).

$$l(p|\mu,\sigma) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-(p-\mu)^2/(2\sigma^2)) \tag{15}$$

2.5. HDTO-DeepAR

Fig. 2 illustrates the overall structure of the proposed optimized hybrid model, featuring densely connected layers. The learning rate, number of hidden layers, dropout rate, batch size, number of cells, and epochs are subjected to optimization using the HDTO algorithm. To optimize the number of instances, the initial training phase begins with an ample number of iterations and early stopping with minimum tolerance intervals. The batch size is the total number of training samples utilized in a particular training cycle. Traditionally, mini-batches are used to train networks faster, and a smaller batch size consumes less memory. Initial training is carried out with a considerable number for the maximum number of epochs to optimize it. The model works reliably at a 0.001 learning rate with 3 hidden layers. A high learning rate may allow the model to converge fast, but it may also exceed the minimum, resulting in oscillations or divergence. A poor learning rate, on the other hand, may cause sluggish convergence, causing the model to get trapped in local minima. The stability of the training process is dependent on hyperparameter optimization. An improper learning rate may cause numerical instability, leading the optimization process to diverge or oscillate. A consistent learning rate leads to a steady and dependable training approach. This technique produced an overfitting-free model and offered an approximate range for the number of instances. Later, The input layer of DeepAR receives a time series of past values. The embedding layers are utilized to process the model's features. The LSTM layers are used to capture temporal dependencies in time series data. These layers include a memory cell that may retain information from earlier timesteps, allowing the model to track long-term relationships in the data. The fully connected layers uncover underlying patterns and correlations in time series data. The output layer generates the predictions for the subsequent time step and takes the form of a probability distribution as per the task.

3. Results

This section summarizes the findings of the innovative methodology for predicting water quality. Table S1 in the supplementary file contains statistical data on water quality indicators.

Table 1
Hyperparameter setting of HDTO-DeepAR.

Hyperparameter	Value
Learning rate	0.001
Number of hidden layers	3
Cell type	LSTM
Dropout rate	0.1
Batch size	64
Number of cells	40
Epochs	70

3.1. Performance metrics

The present research employs several statistical metrics to assess the computational ability to predict water quality. Model performance is evaluated using metrics such as mean squared error (MSE), mean absolute error (MAE), mean absolute percentage error (MAPE), Nash–Sutcliffe efficiency (NSE) and prediction interval coverage probability (PICP). Nash and Sutcliffe (1970), Xiong et al. (2020), Uddin et al. (2023a), Uddin (2020). MSE computes the average squared difference between actual and predicted data. MAE is expressed as the average absolute difference between actual and predicted values. MAPE stands for the mean of the absolute percentage errors (Uddin et al., 2022a, 2023d). The Nash–Sutcliffe efficiency (NSE) is a normalized statistic that indicates the relative amount of the residual variance in comparison to the observed data variance. NSE scores vary from 0 to 1, with 1 signifying a perfect match and values less than 0 indicating that the observed data's mean exceeds the model. The negative NSE result indicates an inefficient model efficiency (Uddin et al., 2023f; Sharif et al., 2022). The statistical indices are calculated using the formula as described below:

$$\text{MSE} = \frac{\sum(y_i - \hat{y}_i)^2}{n} \tag{16}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{17}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| * 100\% \tag{18}$$

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{19}$$

where,
y_i = actual value
$\hat{y}_i$ = predicted value
$\bar{y}_i$ = mean actual value
n = number of observations in the dataset

The lower values of MSE, MAE and MAPE signify reduced prediction bias and higher prediction ability. Naturally, analysts prefer an NSE value near 1 for optimum model performance.

Ensuring the predictability of forecasts is critical for the utility and credibility of any forecasting model. Prediction reliability refers to the consistency and accuracy of projected results in comparison to actual observed events. The prediction interval coverage probability (PICP) is used to assess prediction reliability (Park et al., 2020; Arora et al., 2022). The mean of the decision variable d_i is used to calculate PICP. The expected frequency with which data fall inside the prediction interval range is used to evaluate the prediction values' reliability (Jin et al., 2019).

$$PICP = \left(\frac{1}{n}\sum_{i=1}^{n} d_i\right) \times 100 \tag{20}$$

The prediction values are reliable if the $PICP \geq 95\%$ (Jin et al., 2019).

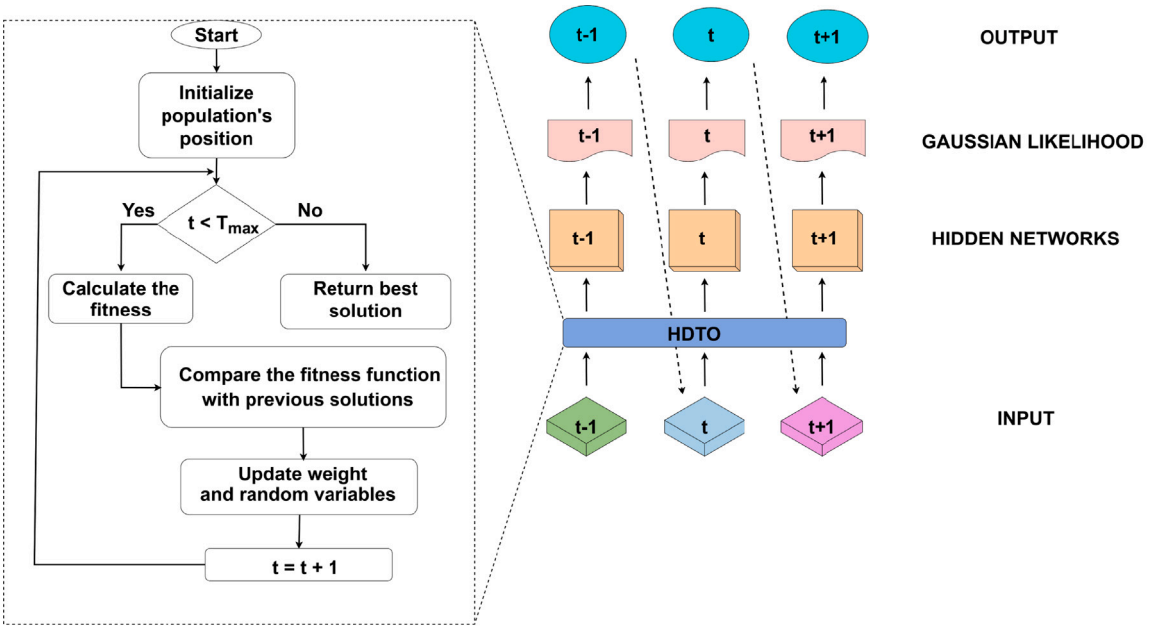


Fig. 2. Proposed architecture for water quality forecasting.

3.2. Forecasting of water quality

The outcome of the proposed methodology is compared to those of the four base models in order to assess its robustness.

Figs. 3(a) to 3(c) shows the probabilistic forecasting of BOD for Brajrajnagar D/S, Sambalpur D/S, and Sonepur D/S station. The value of BOD was found to be decreasing with respect to time. The forecasting result reveals that BOD is lying between 0.5 mg/l to 5 mg/l. It has shown a very good accuracy using HDTO-DeepAR as compared to other standalone deep learning and machine learning methods. Table 4 represents the error estimates of the models in Brajrajnagar D/S, Sambalpur D/S, and Sonepur D/S station. The MAE and MAPE for HDTO-DeepAR in Brajrajnagar D/S are found to be 0.11 and 8.331 respectively and in Sambalpur D/S station, the MAE and MAPE accounted for 0.21 and 10.155. Likewise, the MAE and MAPE for the Sonepur D/S station are found to be 0.11 and 8.851 respectively. The model's NSE is determined to range between 0.89 and 0.96. The proposed model has a closer agreement with the forecasted value. NSE values range from $-\infty$ to 1. A score of 1 shows that the expected and observed values are perfectly matched, whereas lower values suggest a worse fit. A negative NSE indicates that the mean of the observed data is a better predictor than the model. The MSE for Brajarajnagar D/S, Sambalpur D/S and Sonepur D/S are 0.018, 0.113 and 0.018 that can be visualized in Fig. S7 in the supplementary file. The BOD of any aquatic system is an extremely significant indicator for assessing water quality and developing management plans to conserve water resources.

Chloride is an essential water quality characteristic that is often evaluated in surface water bodies owing to its potential consequences for environmental, ecological, and human health. Elevated chloride levels may indicate the existence of pollution sources such as road salt (sodium chloride) runoff, industrial discharges, and wastewater effluents. Monitoring chloride levels may aid in identifying places that need pollution prevention and control measures. The probabilistic forecasting of Chloride is illustrated in Figs. 4(a), 4(b), and 4(c). There is no discernible pattern present in the Chloride dataset. Its forecasted value lies between 3 mg/l to 60 mg/l. The estimated forecast is found to be less than 30 mg/l for Sambalpur D/S and Sonepur D/S stations. Whereas, there is a little spike of about 45 mg/l in the Brajrajnagar D/S station. Table 6 shows the error evaluation of Chloride forecasting. The MAE, MSE MAPE, and NSE of Brajrajnagar D/S using the proposed method are found to be 1.16, 4.134, 8.437, and 0.92. Sambalpur D/S station accounted for 1.26, 3.275, 10.241, and 0.89 of MAE, MSE, MAPE, and NSE. Similarly, the MAE, MSE, MAPE and NSE in the Sonepur D/S station are seen to be 1.11, 2.5398, 10.342, and 0.87 respectively.

pH quantifies the degree of acidity or alkalinity of water on a logarithmic scale ranging from 0 to 14. It is an inherent feature that determines the health, quality, and ecological balance of surface water ecosystems. Extreme pH values may stress or even kill aquatic life, resulting in changes in species composition and probable reductions in biodiversity (Gorde and Jadhav, 2013). Forecasting has been carried out to overcome these severe scenarios and maintain ecological balance in the future. As shown in Figs. 5(a), 5(b) and 5(c), the pH is seen to be maintaining its range between 5.5 to 10 and there exists no trend. Furthermore, the worst-case scenario indicates a value ranging from 4 to 14. Table 5 demonstrates the error evaluation of pH forecasting for the three stations. HDTO-DeepAR has outperformed the other methods in the three cases. The MAE, MSE, MAPE, and NSE are found to lie between 0.67 to 0.76, 0.6 to 0.9, 8.5 to 9.8, and 0.8 to 0.85. The graphical representation of MAPE is shown in Fig. S8 in the supplementary file.

Although sodium is a necessary element, its presence in surface water may have both natural and man-made consequences. Elevated sodium levels in surface water may have a deleterious impact on sensitive freshwater creatures. Therefore, it is necessary to monitor the sodium concentration in water (Khatri and Tyagi, 2015). As shown in Figs. 6(a), 6(b), and 6(c), there is no constant pattern present in three of the stations. The forecasting results indicate the value is approximately lying between 0.5 mg/l to 60 mg/l. Table 2 reflects the error estimation of sodium forecasting. The errors of the proposed HDTO-DeepAR are found to be less as compared to the other methods. The error percentage varies from 8.4 to 9.4. The absolute error is observed to be lying between 0.8 to 1.2 and the squared error ranges from 1.1 to 1.9.

Sulphate is a vital nutrient for aquatic plant and microbial development. It is discharged into bodies of water through natural processes such as weathering of sulphur-rich rocks and minerals. Excessive sulphate levels in runoff may cause eutrophication and oxygen depletion in aquatic bodies. Meanwhile, regular monitoring is required to prevent potentially dangerous disturbances (Bhateria and Jain, 2016).The probabilistic forecasting of Sulphate is shown in Figs. 7(a), 7(b), and 7(c) for Brajrajnagar D/S, Sambalpur D/S and Sonepur D/S respectively. The errors of sulphate forecasting are listed in Table 7.

Table 2
Comparative analysis of HDTO-DeepAR with standalone methods for Sodium forecasting.

WQ Parameter	Station	Model name	MAE	MSE	MAPE	NSE
Sodium (mg/l)	Brajrajnagar D/S	**HDTO-DeepAR**	**1.01**	**1.809**	**9.409**	**0.95**
		DeepAR	1.05	2.245	10.024	0.86
		BiLSTM	1.95	5.852	17.254	0.84
		GRU	2.15	6.261	19.223	0.80
		XGBoost	2.41	7.571	23.941	0.78
	Sambalpur D/S	**HDTO-DeepAR**	**0.96**	**1.219**	**9.203**	**0.95**
		DeepAR	1.21	3.245	11.235	0.89
		BiLSTM	1.62	3.621	15.445	0.85
		GRU	2.09	5.584	20.314	0.8
		XGBoost	2.42	7.398	23.569	0.77
	Sonepur D/S	**HDTO-DeepAR**	**0.83**	**1.154**	**8.453**	**0.93**
		DeepAR	1.01	1.547	9.457	0.87
		BiLSTM	1.478	3.531	14.379	0.82
		GRU	1.94	4.934	20.351	0.79
		XGBoost	2.137	5.996	22.676	0.76

Table 3
Comparative analysis of HDTO-DeepAR with standalone methods for Temperature forecasting.

WQ Parameter	Station	Model name	MAE	MSE	MAPE	NSE
Temperature (°C)	Brajrajnagar D/S	**HDTO-DeepAR**	**2.52**	**7.335**	**9.725**	**0.85**
		DeepAR	3.12	11.245	11.324	0.74
		BiLSTM	3.52	16.323	13.921	0.71
		GRU	3.66	18.024	14.417	0.69
		XGBoost	5.39	39.428	21.194	0.66
	Sambalpur D/S	**HDTO-DeepAR**	**2.64**	**10.714**	**10.832**	**0.87**
		DeepAR	2.79	11.074	10.834	0.86
		BiLSTM	2.84	11.091	10.835	0.83
		GRU	4.97	32.001	18.971	0.79
		XGBoost	5.65	42.341	21.205	0.76
	Sonepur D/S	**HDTO-DeepAR**	**2.52**	**9.664**	**9.191**	**0.82**
		DeepAR	3.42	12.912	11.213	0.8
		BiLSTM	3.57	20.193	12.939	0.79
		GRU	5.76	46.509	21.202	0.75
		XGBoost	6.13	53.265	22.125	0.71

Temperature is a significant variable in aquatic ecosystems impacting a variety of physical, chemical, and biological processes. It has a direct impact on the metabolic rates, growth, and reproductive activities of aquatic species. The metabolic rates of most of the species increase when the temperature rises, resulting in larger energy needs and faster development rates. Warmer water temperatures may reduce dissolved oxygen levels, possibly straining or suppressing aquatic life (Mugwanya et al., 2022; Volkoff and Rønnestad, 2020). As shown in Table 3, the MAE, MSE, MAPE, and NSE using the proposed method are found to be between 2.5 to 2.6, 9.6 to 10.7, 9 to 10.8, and 0.82 to 0.87. As a result, monitoring and comprehending temperature fluctuations is crucial for successful aquatic resource management and conservation. The NSE findings suggest that the model works efficiently. Moreover, the probabilistic forecasting of temperature can be visualized in Fig. 8. The PICP is calculated using Eq. (20) to evaluate the reliability of the forecasting model. Based on the PICP values of BOD, pH, sodium, sulphate, temperature, and chloride forecasting, the expected frequency ranges from a prediction interval of 95% to 98%. Hence, according to a comparative study, it is found that the proposed HDTO-DeepAR method outperformed the other techniques in terms of prediction accuracy and reliability.

4. Discussion

The proposed study contributes by filling the research gap in the methodological approach to forecasting water quality indicators in the Mahanadi River basin, India. It introduces an improvement in the forecasting technique by combining the hidden dipper throated optimization and DeepAR. Training the DeepAR model is computationally challenging and time-consuming. It is difficult to choose the right values of the hyperparameters (Bischl et al., 2023). To overcome this challenge, HDTO method is used in the hidden layer to optimize the hyperparameters.

The study identified the possibility of a decrease in the BOD content in the succeeding years until 2025. A decrease in BOD is often seen as an indication of improved water quality, signifying a reduction in organic pollution and an improvement in the health of aquatic ecosystems (Dutta et al., 2020). Chloride content follows a similar trend as past patterns, as shown in Figs. 4(a), 4(b)and 4(c). The pH level may rise or decrease based on anthropogenic and natural environmental conditions such as hospital trash disposal, acid rain, and other potential influences (Khatri and Tyagi, 2015). Sodium content remains the same as per historical pattern. However, it might rise at a certain stage in Brajrajnagar D/S and Sambalpur D/S stations due to industrial discharge, agricultural runoff and runoff from urban areas. Sulphate content appears to be increasing in the future. The temperature of surface water follows patterns similar to the historical pattern.

There has been essentially no study on water quality time series forecasting in the Mahanadi River. The comparative analysis relies on literature from different river basins. As mentioned in Siami-Namini et al. (2019), Khullar and Singh (2022), BiLSTM operates well due to its backward and forward processing. Due to flaws in regularization approaches, the procedure cannot provide good results. Compared to simpler models, optimizing BiLSTMs might be more difficult because of their increased complexity and several hyperparameters. The optimization techniques can be incorporated to reduce the error rate and improve the forecasting accuracy (Uddin et al., 2022b,a, 2023i; Ding et al., 2023; Uddin et al., 2022c).

Additionally, GRU is chosen for analysis due to its simpler architecture, which can be useful for training the model easily within a short period. Handling information with the single gating mechanism in GRU may seem challenging (Mei et al., 2022).

Furthermore, XGBoost is a popular ensemble high-performance and versatile machine learning method (Uddin et al., 2022b). It has the capability to visualize each decision tree and analyse the feature importance score. Several recent research studies have also indicated that

Table 4
Comparative analysis of HDTO-DeepAR with standalone methods for BOD forecasting.

WQ Parameter	Station	Model name	MAE	MSE	MAPE	NSE
BOD (mg/l)	Brajrajnagar D/S	**HDTO-DeepAR**	**0.11**	**0.018**	**8.331**	**0.95**
		DeepAR	0.19	0.211	11.854	0.9
		BiLSTM	0.35	0.218	24.251	0.64
		GRU	0.41	0.239	31.846	0.59
		XGBoost	0.44	0.303	33.242	0.54
	Sambalpur D/S	**HDTO-DeepAR**	**0.21**	**0.113**	**10.155**	**0.89**
		DeepAR	0.23	0.114	10.098	0.87
		BiLSTM	0.31	0.134	15.939	0.73
		GRU	0.49	0.299	25.207	0.57
		XGBoost	0.57	0.479	29.072	0.55
	Sonepur D/S	**HDTO-DeepAR**	**0.11**	**0.018**	**8.851**	**0.96**
		DeepAR	0.14	0.028	12.312	0.94
		BiLSTM	0.17	0.042	13.116	0.89
		GRU	0.35	0.173	27.759	0.76
		XGBoost	0.41	0.234	29.784	0.67

Table 5
Comparative analysis of HDTO-DeepAR with standalone methods for pH forecasting.

WQ Parameter	Station	Model name	MAE	MSE	MAPE	NSE
pH	Brajrajnagar D/S	**HDTO-DeepAR**	**0.67**	**0.615**	**8.702**	**0.85**
		DeepAR	0.81	1.124	11.234	0.79
		BiLSTM	0.99	1.378	12.823	0.77
		GRU	1.55	3.095	19.793	0.74
		XGBoost	2.38	6.774	29.409	0.68
	Sambalpur D/S	**HDTO-DeepAR**	**0.76**	**0.784**	**9.709**	**0.83**
		DeepAR	0.98	1.459	12.875	0.79
		BiLSTM	1.12	1.668	14.231	0.76
		GRU	1.18	1.689	15.128	0.72
		XGBoost	1.54	3.301	19.674	0.67
	Sonepur D/S	**HDTO-DeepAR**	**0.75**	**0.839**	**9.607**	**0.84**
		DeepAR	1.17	1.124	12.973	0.81
		BiLSTM	1.23	2.005	15.586	0.77
		GRU	1.39	2.709	17.661	0.72
		XGBoost	1.8	4.428	22.916	0.68

Table 6
Comparative analysis of HDTO-DeepAR with standalone methods for Chloride forecasting.

WQ Parameter	Station	Model name	MAE	MSE	MAPE	NSE
Chloride (mg/l)	Brajrajnagar D/S	**HDTO-DeepAR**	**1.16**	**4.134**	**8.437**	**0.92**
		DeepAR	1.7	4.356	9.954	0.91
		BiLSTM	1.8	4.941	16.298	0.85
		GRU	1.97	5.056	17.934	0.82
		XGBoost	2.07	10.856	18.633	0.79
	Sambalpur D/S	**HDTO-DeepAR**	**1.26**	**3.275**	**10.241**	**0.89**
		DeepAR	1.79	4.724	11.571	0.83
		BiLSTM	1.86	5.231	17.828	0.79
		GRU	2.27	7.039	19.953	0.77
		XGBoost	2.48	7.568	23.385	0.75
	SonepurD/S	**HDTO-DeepAR**	**1.11**	**2.539**	**10.342**	**0.87**
		DeepAR	1.38	2.864	11.972	0.85
		BiLSTM	1.45	3.484	14.577	0.83
		GRU	1.74	4.007	17.682	0.81
		XGBoost	2.19	6.495	21.453	0.75

the XGBoost algorithm is good at predicting WQIs (Uddin et al., 2022a, 2023h,g; Khan et al., 2022). Despite the regularization process, the model has faced overfitting (Tong et al., 2023).

In this research, a novel hybrid approach is proposed for water quality forecasting. Additionally, four standalone methods are used to compare the robustness of the hybrid technique. The standalone methods faced challenges of overfitting due to poor hyperparameter tuning with bounded datasets. As a result, an appropriate optimization strategy is employed in HDTO-DeepAR method. Although the proposed approach may include external variables, their investigation and exploitation may be restricted. Integrating external data effectively might be difficult in several cases. It is based on learned data representations and may not gain as much from human feature engineering as certain traditional models. This constraint might be problematic in circumstances when expert knowledge is useful for enhancing predictions.

The findings of this study could assist water resource managers and policymakers in effectively allocating resources. It enables them to foresee changes in water quality and take preventive action to guarantee the supply of clean and secure drinking water. The study generates a probability distribution for forecasting future sequences based on past data. Many times, addressing intermittent and irregular characteristics can be challenging. In such cases, the training procedure in DeepAR can be modified (Jeon and Seong, 2022). Forecasting aids in more effective maintenance activity scheduling, lowering costs and downtime. It can provide early warning systems in situations when water quality unexpectedly declines due to chemical spills or natural catastrophes. Water quality time-series forecasting aims to project the future values of several parameters based on previous data. However, various sources of uncertainty may impair the accuracy of these estimates. Weather factors, seasonal variations, and biological dynamics may all contribute

Table 7
Comparative analysis of HDTO-DeepAR with standalone methods for Sulphate forecasting.

WQ Parameter	Station	Model name	MAE	MSE	MAPE	NSE
Sulphate (mg/l)	Brajrajnagar D/S	**HDTO-DeepAR**	**0.63**	**0.681**	**12.109**	**0.92**
		DeepAR	0.71	1.126	13.754	0.89
		BiLSTM	0.74	1.138	16.534	0.85
		GRU	1.01	1.654	20.614	0.79
		XGBoost	1.32	2.596	27.364	0.73
	Sambalpur D/S	**HDTO-DeepAR**	**0.64**	**1.285**	**10.737**	**0.94**
		DeepAR	0.97	2.269	11.394	0.92
		BiLSTM	1.12	2.509	12.411	0.88
		GRU	1.26	2.534	25.793	0.74
		XGBoost	1.51	3.393	26.481	0.69
	SonepurD/S	**HDTO-DeepAR**	**0.53**	**0.507**	**13.841**	**0.91**
		DeepAR	0.69	0.685	14.246	0.90
		BiLSTM	0.73	0.731	15.743	0.88
		GRU	1.45	2.805	26.751	0.82
		XGBoost	1.58	3.641	31.606	0.78

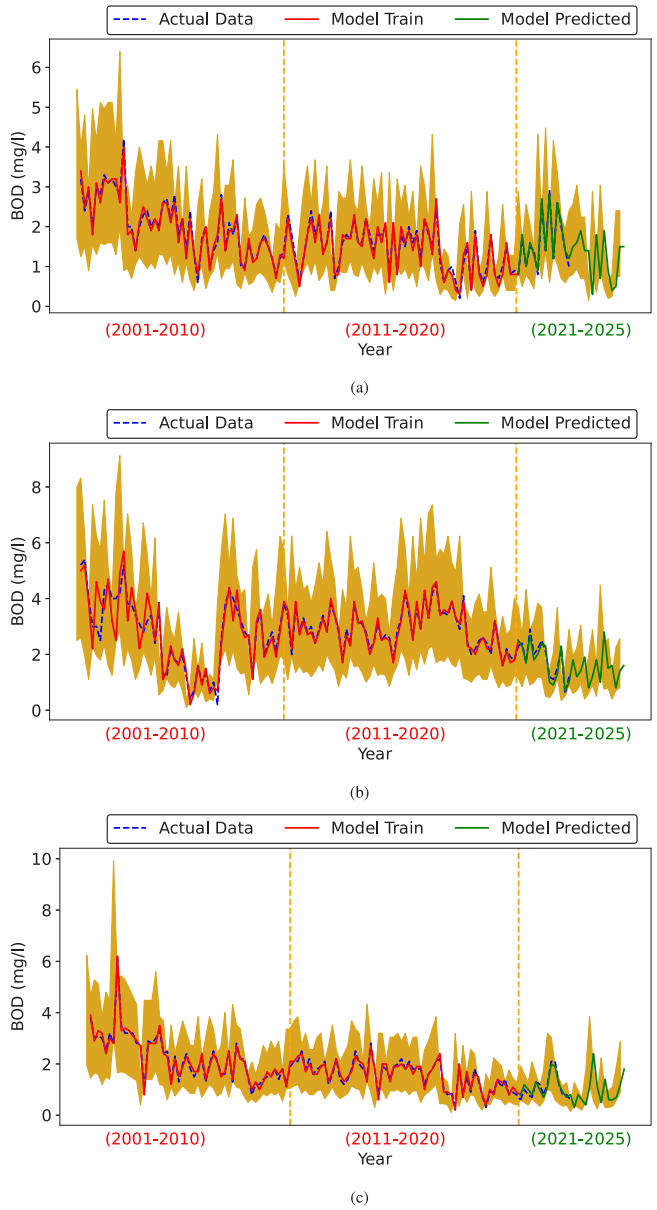


Fig. 3. Probabilistic forecasting of BOD using HDTO-DeepAR for (a) Brajrajnagar D/S. (b) Sambalpur D/S. (c) Sonepur D/S.

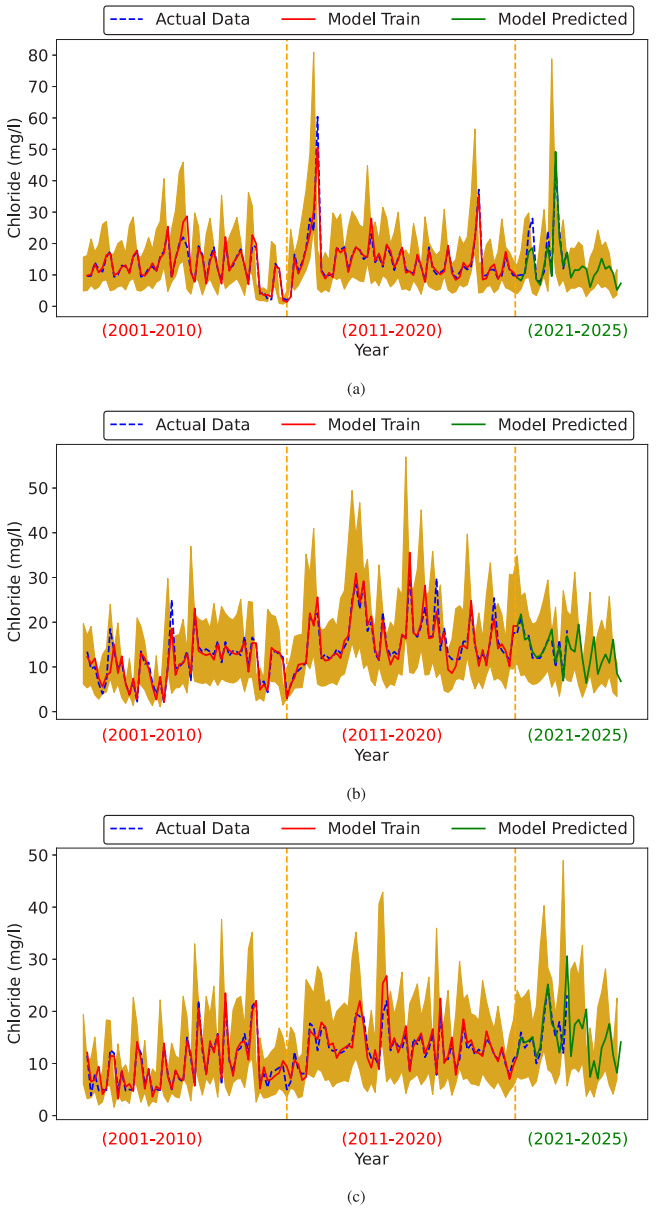


Fig. 4. Probabilistic forecasting of Chloride using HDTO-DeepAR for (a) Brajrajnagar D/S. (b) Sambalpur D/S. (c) Sonepur D/S.

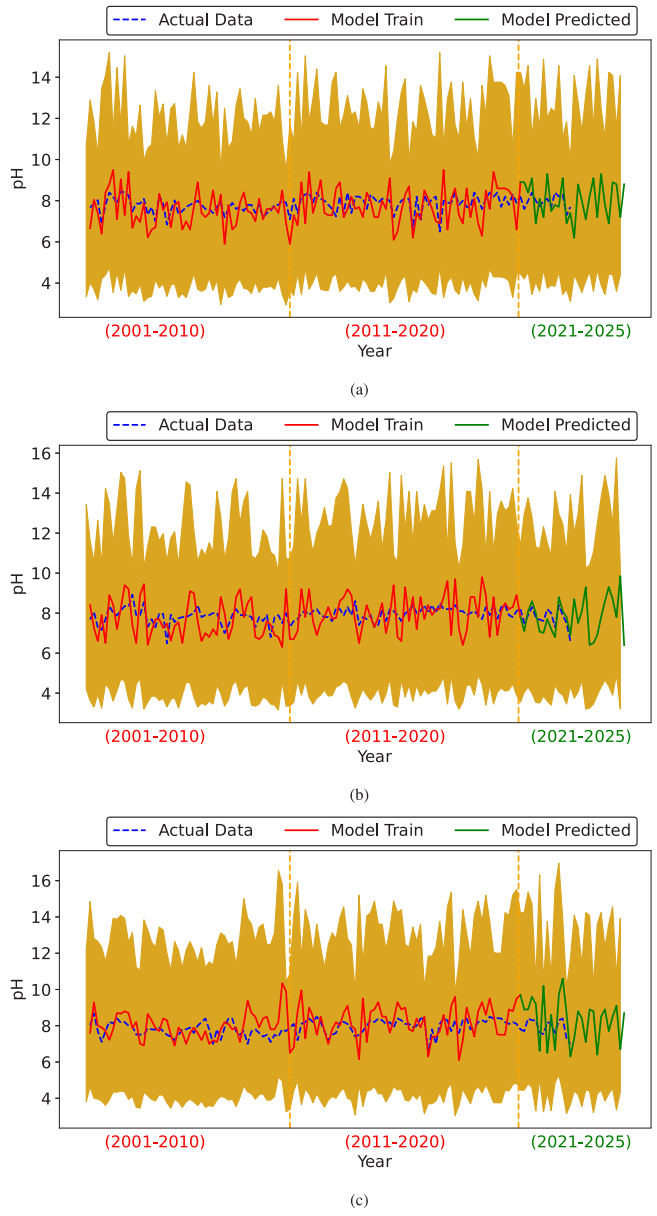


Fig. 5. Probabilistic forecasting of pH using HDTO-DeepAR for (a) Brajrajnagar D/S. (b) Sambalpur D/S. (c) Sonepur D/S.

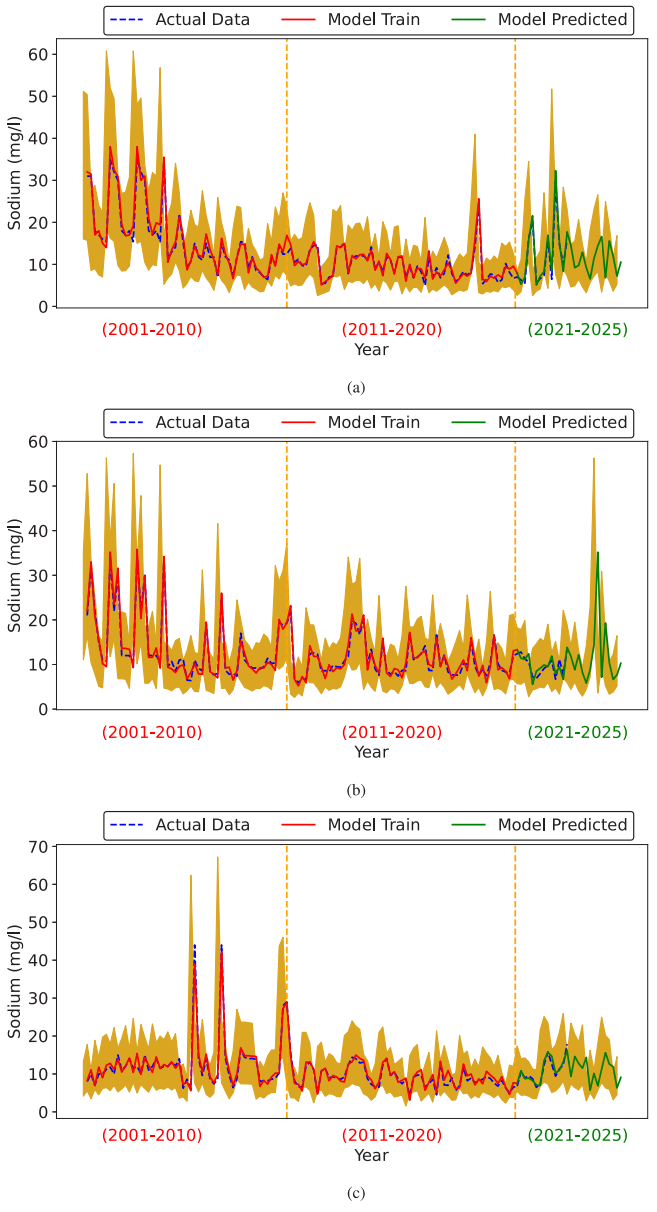


Fig. 6. Probabilistic forecasting of Sodium using HDTO-DeepAR for (a) Brajrajnagar D/S. (b) Sambalpur D/S. (c) Sonepur D/S.

to unpredictable oscillations. Pollutants may enter water bodies as a result of human activities such as industrial discharges, agricultural runoff, and urban expansion. As these activities are dynamic, it is difficult to predict their influence on water quality. Inaccuracies in sampling methodology, equipment calibration, and laboratory analysis may also affect the predictability of historical data. The key component of uncertainty is its capacity to influence decision-making and resource management (Uddin et al., 2023a,b; Gani et al., 2023). However, HDTO-DeepAR operates in uncertain situations by using a probabilistic method. It not only delivers point predictions but also prediction intervals that reflect the degree of uncertainty associated with each prediction.

This research does not consider external environmental factors that affect surface water quality. However, the water quality may also be impacted by external agents and heavy metals from industries. The analysis uses a smaller dataset, while a larger dataset would be more efficient for forecasting. The proposed method does not provide a direct measure of uncertainty in its forecasts. The computing resources needed for training and deploying DeepAR models may be costly, particularly for large-scale forecasting tasks.

5. Conclusion

In this manuscript, six models: HDTO-DeepAR, DeepAR, Bi-LSTM, GRU, and XGBoost are implemented to forecast the water quality indicators in the Mahanadi river basin, India. Parameters used for forecasting include BOD, chloride, pH, sodium, sulphate and temperature. The probabilistic forecasting reveals BOD and chloride concentration might face a decreasing trend over time. On the other hand, there is a possibility that pH might lie between 4–14. Likewise, sodium shows a decreasing trend in Brajrajnagar D/S and Sonepur D/S during 2023 – 2025. whereas it takes a peak in Sambalpur D/S. Sulphate concentration might vary approximately between 2 mg/l–40 mg/l in Sambalpur D/S and Sonepur D/S. However, in Brajrajnagar D/S station the value might lie between 1 mg/l–50mg/l. The temperature can lie between 10 °C–50 °C in all the stations. The PICP value demonstrates that

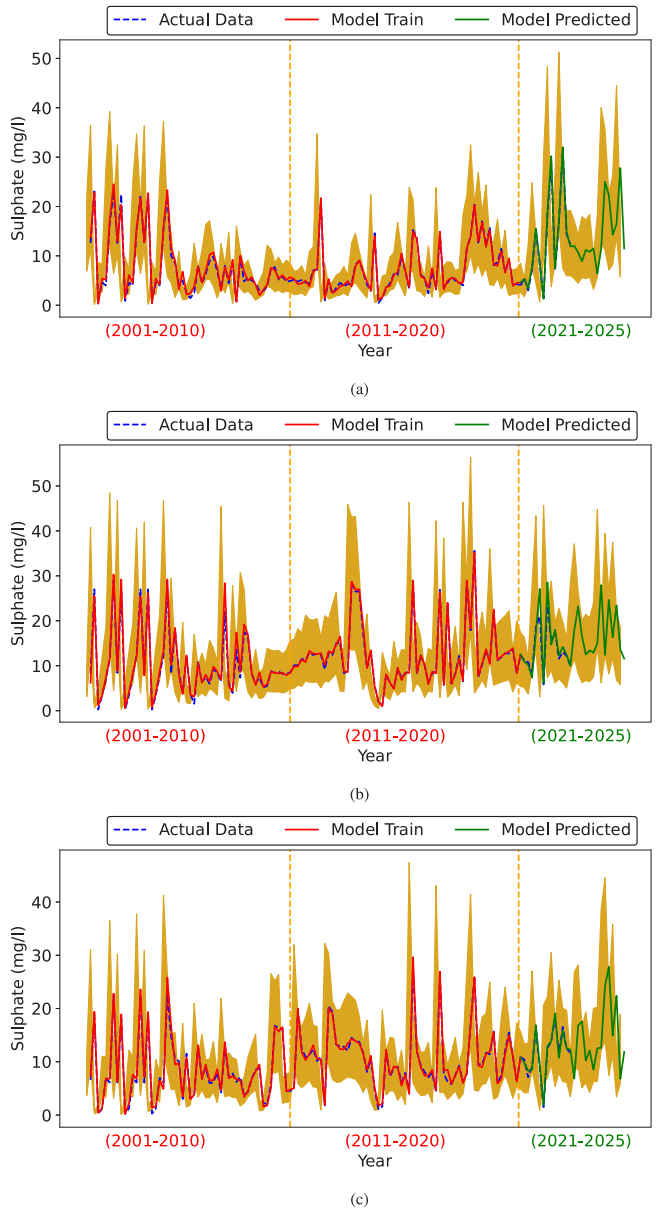


Fig. 7. Probabilistic forecasting of Sulphate using HDTO-DeepAR for (a) Brajrajnagar D/S. (b) Sambalpur D/S. (c) Sonepur D/S.

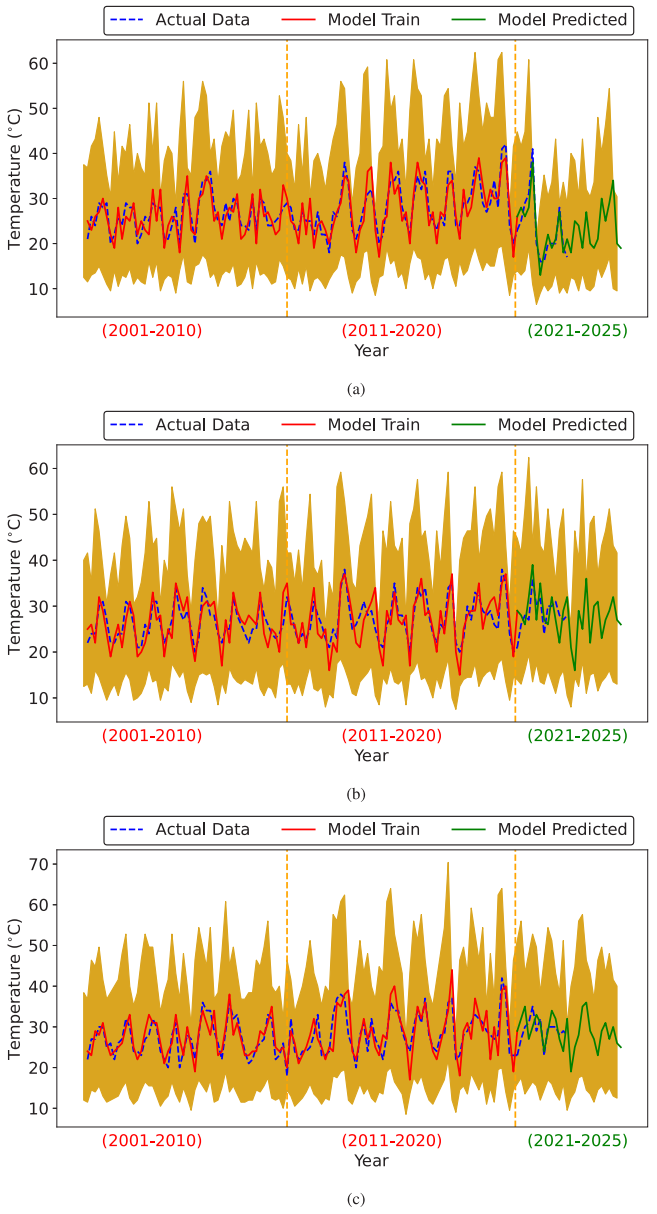


Fig. 8. Probabilistic forecasting of Temperature using HDTO-DeepAR for (a) Brajrajnagar D/S. (b) Sambalpur D/S. (c) Sonepur D/S.

HDTO-DeepAR is the most reliable for projecting water quality. The recommended approach is efficient in producing probabilistic forecasts with high accuracy and can determine complicated patterns like seasonality and uncertainty increase over time. Industrial activities, sewage disposal and agricultural runoff are the most influential factors to affect the water quality. The results of the output show a probable chance of extreme increase and decrease in the water quality concentration in future.

It is evident that water quality is also affected by various other environmental factors. These factors are not considered in this manuscript, which is a limitation. In future, the study can be carried out by incorporating multiple covariates for more precise forecasting.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

CRediT authorship contribution statement

Rosysmita Bikram Singh: Writing – original draft, Investigation, Formal analysis, Data curation. **Kanhu Charan Patra:** Supervision, Conceptualization. **Biswajeet Pradhan:** Writing – review & editing, Resources, Supervision, Project administration. **Avinash Samantra:** Methodology, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jenvman.2024.120091.

References

Abbasi, T., Abbasi, S., 2011. Water quality indices based on bioassessment: The biotic indices. J. Water Health 9 (2), 330–348.

Abdelhamid, A.A., El-Kenawy, E.S.M., Alrowais, F., Ibrahim, A., Khodadadi, N., Lim, W.H., Alruwais, N., Khafaga, D.S., 2022. Deep learning with dipper throated optimization algorithm for energy consumption forecasting in smart households. Energies 15 (23), 9125.

Abdelhamid, A.A., El-kenawy, E.S.M., Ibrahim, A., Eid, M.M., Khafaga, D.S., Alhussan, A.A., Mirjalili, S., Khodadadi, N., Lim, W.H., Shams, M.Y., 2023. Innovative feature selection method based on hybrid Sine cosine and dipper throated optimization algorithms. IEEE Access.

Antonopoulos, V.Z., Papamichail, D.M., Mitsiou, K.A., 2001. Statistical and trend analysis of water quality and quantity data for the Strymon River in Greece. Hydrol. Earth Syst. Sci. 5 (4), 679–692.

Arora, P., Jalali, S.M.J., Ahmadian, S., Panigrahi, B., Suganthan, P., Khosravi, A., 2022. Probabilistic wind power forecasting using optimized deep auto-regressive recurrent neural networks. IEEE Trans. Ind. Inform. 19 (3), 2814–2825.

Arya, F.K., Zhang, L., 2015. Time series analysis of water quality parameters at stillaguamish river using order series method. Stoch. Environ. Res. Risk Assess. 29, 227–239.

Arya, F.K., Zhang, L., 2017. Copula-based Markov process for forecasting and analyzing risk of water quality time series. J. Hydrol. Eng. 22 (6), 04017005.

Bhateria, R., Jain, D., 2016. Water quality assessment of lake water: a review. Sustain. Water Resour. Manag. 2, 161–173.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.L., et al., 2023. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 13 (2), e1484.

Bui, D.T., Khosravi, K., Tiefenbacher, J., Nguyen, H., Kazakis, N., 2020. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Sci. Total Environ. 721, 137612.

Chakrapani, G.J., Subramanian, V., 1990. Preliminary studies on the geochemistry of the Mahanadi river basin, India. Chem. Geol. 81 (3), 241–253.

Chang, N., Luo, L., Wang, X.C., Song, J., Han, J., Ao, D., 2020. A novel index for assessing the water quality of urban landscape lakes based on water transparency. Sci. Total Environ. 735, 139351.

CPCB, 2020. QA/QC measures. URL: https://cpcb.nic.in/index.php.

Csábrági, A., Molnár, S., Tanos, P., Kovács, J., 2017. Application of artificial neural networks to the forecasting of dissolved oxygen content in the Hungarian section of the river Danube. Ecol. Eng. 100, 63–72.

Cuong-Le, T., Minh, H.L., Sang-To, T., Khatir, S., Mirjalili, S., Wahab, M.A., 2022. A novel version of grey wolf optimizer based on a balance function and its application for hyperparameters optimization in deep neural network (DNN) for structural damage identification. Eng. Fail. Anal. 142, 106829.

Deng, T., Chau, K.W., Duan, H.F., 2021. Machine learning-based marine water quality prediction for coastal hydro-environment management. J. Environ. Manag. 284, 112051.

Deng, T., Duan, H.F., Keramat, A., 2022. Spatiotemporal characterization and forecasting of coastal water quality in the semi-enclosed Tolo Harbour based on machine learning and EKC analysis. Eng. Appl. Comput. Fluid Mech. 16 (1), 694–712.

Dimri, D., Daverey, A., Kumar, A., Sharma, A., 2021. Monitoring water quality of River Ganga using multivariate techniques and WQI (Water Quality Index) in Western Himalayan region of Uttarakhand, India. Environ. Nanotechnol. Monit. Manag. 15, 100375.

Ding, F., Zhang, W., Cao, S., Hao, S., Chen, L., Xie, X., Li, W., Jiang, M., 2023. Optimization of water quality index models using machine learning approaches. Water Res. 243, 120337.

Dutta, V., Dubey, D., Kumar, S., 2020. Cleaning the River Ganga: Impact of lockdown on water quality and future implications on river rejuvenation strategies. Sci. Total Environ. 743, 140756.

Esterby, S.R., 1996. Review of methods for the detection and estimation of trends with emphasis on water quality applications. Hydrol. Process. 10 (2), 127–149.

Faris, H., Aljarah, I., Al-Betar, M.A., Mirjalili, S., 2018. Grey wolf optimizer: a review of recent variants and applications. Neural Comput. Appl. 30, 413–435.

Gani, M.A., Sajib, A.M., Siddik, M.A., Moniruzzaman, M., 2023. Assessing the impact of land use and land cover on river water quality using water quality index and remote sensing techniques. Environ. Monit. Assess. 195 (4), 449.

Ge, Z., Feng, S., Ma, C., Dai, X., Wang, Y., Ye, Z., 2023. Urban river ammonia nitrogen prediction model based on improved whale optimization support vector regression mixed synchronous compression wavelet transform. Chemometr. Intell. Lab. Syst. 104930.

Georgescu, P.L., Moldovanu, S., Iticescu, C., Calmuc, M., Calmuc, V., Topa, C., Moraru, L., 2023. Assessing and forecasting water quality in the Danube River by using neural network approaches. Sci. Total Environ. 879, 162998.

Gorde, S., Jadhav, M., 2013. Assessment of water quality parameters: a review. J. Eng. Res. Appl. 3 (6), 2029–2035.

Guo, W., Liu, T., Dai, F., Xu, P., 2020. An improved whale optimization algorithm for forecasting water resources demand. Appl. Soft Comput. 86, 105925.

Hassan, M.M., Hassan, M.M., Akter, L., Rahman, M.M., Zaman, S., Hasib, K.M., Jahan, N., Smrity, R.N., Farhana, J., Raihan, M., et al., 2021. Efficient prediction of water quality index (WQI) using machine learning algorithms. Hum. Cent. Intell. Syst. 1 (3–4), 86–97.

Horton, R.K., 1965. An index number system for rating water quality. J. Water Pollut. Control Fed. 37 (3), 300–306.

Hussain, J., Dubey, A., Hussain, I., Arif, M., Shankar, A., 2020. Surface water quality assessment with reference to trace metals in River Mahanadi and its tributaries, India. Appl. Water Sci. 10 (8), 1–12.

Irwan, D., Ali, M., Ahmed, A.N., Jacky, G., Nurhakim, A., Ping Han, M.C., AlDahoul, N., El-Shafie, A., 2023. Predicting water quality with artificial intelligence: A review of methods and applications. Arch. Comput. Methods Eng. 1–20.

Islam, M.K.B., Newton, M.H., Rahman, J., Trevathan, J., Sattar, A., 2022. Long range multi-step water quality forecasting using iterative ensembling. Eng. Appl. Artif. Intell. 114, 105166.

Jeon, Y., Seong, S., 2022. Robust recurrent network model for intermittent time-series forecasting. Int. J. Forecast. 38 (4), 1415–1425.

Jin, T., Cai, S., Jiang, D., Liu, J., 2019. A data-driven model for real-time water quality prediction and early warning by an integration method. Environ. Sci. Pollut. Res. 26, 30374–30385.

Kennedy, J., Eberhart, R., 1995. Particle swarm optimization. In: Proceedings of ICNN'95-International Conference on Neural Networks, Vol. 4. IEEE, pp. 1942–1948.

Khan, M.S.I., Islam, N., Uddin, J., Islam, S., Nasir, M.K., 2022. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. J. King Saud Univ.-Comput. Inf. Sci. 34 (8), 4773–4781.

Khatri, N., Tyagi, S., 2015. Influences of natural and anthropogenic factors on surface and groundwater quality in rural and urban areas. Front. Life Sci. 8 (1), 23–39.

Khullar, S., Singh, N., 2022. Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation. Environ. Sci. Pollut. Res. 29 (9), 12875–12889.

Konhauser, K., Powell, M., Fyfe, W., Longstaffe, F., Tripathy, S., 1997. Trace element chemistry of major rivers in Orissa State, India. Environ. Geol. 29, 132–141.

Kouadri, S., Elbeltagi, A., Islam, A.R.M.T., Kateb, S., 2021. Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). Appl. Water Sci. 11 (12), 190.

Kumar, A., Taxak, A., Mishra, S., Pandey, R., 2021. Long-term trend analysis and suitability of water quality of River Ganga at Himalayan hills of Uttarakhand, India. Environ. Technol. Innov. 22, 101405.

Kuo, J.T., Wang, Y.Y., Lung, W.S., 2006. A hybrid neural–genetic algorithm for reservoir water quality management. Water Res. 40 (7), 1367–1376.

Kurwadkar, S., Sethi, S.S., Mishra, P., Ambade, B., 2022. Unregulated discharge of wastewater in the Mahanadi River Basin: risk evaluation due to occurrence of polycyclic aromatic hydrocarbon in surface water and sediments. Mar. Pollut. Bull. 179, 113686.

Liu, X., Shi, Q., Liu, Z., Yuan, J., 2021. Using LSTM neural network based on improved PSO and attention mechanism for predicting the effluent COD in a wastewater treatment plant. IEEE Access 9, 146082–146096.

Mei, P., Li, M., Zhang, Q., Li, G., et al., 2022. Prediction model of drinking water source quality with potential industrial-agricultural pollution based on CNN-GRU-Attention. J. Hydrol. 610, 127934.

Mirjalili, S., Mirjalili, S., 2019. Genetic algorithm. Evol. Algorithms Neural Netw. Theory Appl. 43–55.

Mirjalili, S., Mirjalili, S.M., Lewis, A., 2014. Grey wolf optimizer. Adv. Eng. Softw. 69, 46–61.

Miyittah, M.K., Tulashie, S.K., Tsyawo, F.W., Sarfo, J.K., Darko, A.A., 2020. Assessment of surface water quality status of the Aby Lagoon System in the Western Region of Ghana. Heliyon 6 (7).

Morales-Hernández, A., Van Nieuwenhuyse, I., Rojas Gonzalez, S., 2023. A survey on multi-objective hyperparameter optimization algorithms for machine learning. Artif. Intell. Rev. 56 (8), 8043–8093.

Mugwanya, M., Dawood, M.A., Kimera, F., Sewilam, H., 2022. Anthropogenic temperature fluctuations and their effect on aquaculture: A comprehensive review. Aquac. Fish. 7 (3), 223–243.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. J. Hydrol. 10 (3), 282–290.

Othman, F., Alaaeldin, M., Seyam, M., Ahmed, A.N., Teo, F.Y., Ming Fai, C., Afan, H.A., Sherif, M., Sefelnasr, A., El-Shafie, A., 2020. Efficient river water quality index prediction considering minimal number of inputs variables. Eng. Appl. Comput. Fluid Mech. 14 (1), 751–763.

Ouyang, Y., 2005. Evaluation of river water quality monitoring stations by principal component analysis. Water Res. 39 (12), 2621–2635.

Pany, R., Rath, A., Swain, P.C., 2023. Water quality assessment for River Mahanadi of Odisha, India using statistical techniques and Artificial Neural Networks. J. Clean. Prod. 417, 137713.

Park, S., Park, S., Hwang, E., 2020. Normalized residue analysis for deep learning based probabilistic forecasting of photovoltaic generations. In: 2020 IEEE International Conference on Big Data and Smart Computing. BigComp, IEEE, pp. 483–486.

Parween, S., Siddique, N.A., Diganta, M.T.M., Olbert, A.I., Uddin, M.G., 2022. Assessment of urban river water quality using modified NSF water quality index model at Siliguri city, West Bengal, India. Environ. Sustain. Indic. 16, 100202.

Poli, R., Kennedy, J., Blackwell, T., 2007. Particle swarm optimization: An overview. Swarm Intell. 1, 33–57.

Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T., 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. Int. J. Forecast. 36 (3), 1181–1191.

Samantray, P., Mishra, B.K., Panda, C.R., Rout, S.P., 2009. Assessment of water quality index in Mahanadi and Atharabanki Rivers and Taldanda Canal in Paradip area, India. J. Hum. Ecol. 26 (3), 153–161.

Samuel, A., Joy, K., Bhagat, S., 2017. Integrated Water Management of the Mahanadi Basin.

Sharif, O., Hasan, M.Z., Rahman, A., 2022. Determining an effective short term COVID-19 prediction model in asean countries. Sci. Rep. 12 (1), 5083.

Siami-Namini, S., Tavakoli, N., Namin, A.S., 2019. The performance of LSTM and BiLSTM in forecasting time series. In: 2019 IEEE International Conference on Big Data. Big Data, IEEE, pp. 3285–3292.

Song, T., Kim, K., 2009. Development of a water quality loading index based on water quality modeling. J. Environ. Manag. 90 (3), 1534–1543.

Srinivas, M., Patnaik, L.M., 1994. Genetic algorithms: A survey. Computer 27 (6), 17–26.

Subramanian, V., 1980. Mineralogical input of suspended matter by Indian rivers into the adjacent areas of the Indian Ocean. Mar. Geol. 36 (3–4), M29–M34.

Sundaray, S.K., Nayak, B.B., Lin, S., Bhatta, D., 2011. Geochemical speciation and risk assessment of heavy metals in the river estuarine sediments—a case study: Mahanadi basin, India. J. Hazard. Mater. 186 (2–3), 1837–1846.

Sundaray, S.K., Panda, U.C., Nayak, B.B., Bhatta, D., 2006. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of the Mahanadi river–estuarine system (India)–a case study. Environ. Geochem. Health 28, 317–330.

Takieldeen, A.E., El-kenawy, E.S.M., Hadwan, M., Zaki, R.M., 2022. Dipper throated optimization algorithm for unconstrained function and feature selection. Comput. Mater. Contin. 72, 1465–1481.

Tong, S., Li, W., Chen, J., Xia, R., Lin, J., Chen, Y., Xu, C.Y., 2023. A novel framework to improve the consistency of water quality attribution from natural and anthropogenic factors. J. Environ. Manag. 342, 118077.

Uddin, M.G., 2020. Development of a novel water quality index model using data. J. King Saud Univ.-Comput. Inf. Sci. 34, 4773–4781.

Uddin, M.G., Diganta, M.T.M., Sajib, A.M., Rahman, A., Nash, S., Dabrowski, T., Ahmadian, R., Hartnett, M., Olbert, A.I., 2023a. Assessing the impact of COVID-19 lockdown on surface water quality in Ireland using advanced Irish Water Quality Index (IEWQI) Model. Environ. Pollut. 336, 122456.

Uddin, M.G., Jackson, A., Nash, S., Rahman, A., Olbert, A.I., 2023b. Comparison between the WFD approaches and newly developed water quality model for monitoring transitional and coastal water quality in Northern Ireland. Sci. Total Environ. 901, 165960.

Uddin, M.G., Nash, S., Diganta, M.T.M., Rahman, A., Olbert, A.I., 2022a. Robust machine learning algorithms for predicting coastal water quality index. J. Environ. Manag. 321, 115923.

Uddin, M.G., Nash, S., Diganta, M.T.M., Rahman, A., Olbert, A.I., 2023c. A comparison of geocomputational models for validating geospatial distribution of water quality index. In: Computational Statistical Methodologies and Modeling for Artificial Intelligence. CRC Press, pp. 243–276.

Uddin, M.G., Nash, S., Olbert, A.I., 2021. A review of water quality index models and their use for assessing surface water quality. Ecol. Indic. 122, 107218.

Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2022b. A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment. Water Res. 219, 118532.

Uddin, M.G., Nash, S., Rahman, A., Olbert, A., 2022c. Development of a water quality index model-a comparative analysis of various weighting methods. In: Mediterranean Geosciences Union Annual Meeting. MedGU-21. Istanbul, pp. 1–6.

Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2023d. A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches. Water Res. 229, 119422.

Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2023e. A sophisticated model for rating water quality. Sci. Total Environ. 868, 161614.

Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2023f. Assessing optimization techniques for improving water quality model. J. Clean. Prod. 385, 135671.

Uddin, M., Nash, S., Rahman, A., Olbert, A., 2023g. Data-driven modelling for assessing trophic status in marine ecosystems using machine learning approaches. In: Stephen and Rahman, Azizur and Olbert, Agnieszka, Data-Driven Modelling for Assessing Trophic Status in Marine Ecosystems using Machine Learning Approaches.

Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2023h. Performance analysis of the water quality index model for predicting water state using machine learning techniques. Process Saf. Environ. Prot. 169, 808–828.

Uddin, M.G., Rahman, A., Nash, S., Diganta, M.T.M., Sajib, A.M., Moniruzzaman, M., Olbert, A.I., 2023i. Marine waters assessment using improved water quality model incorporating machine learning approaches. J. Environ. Manag. 344, 118368.

Vlad, C., Sbarciog, M., Barbu, M., Caraman, S., Wouwer, A.V., 2012. Indirect control of substrate concentration for a wastewater treatment process by dissolved oxygen tracking. J. Control Eng. Appl. Inform. 14 (1), 38–47.

Volkoff, H., Rønnestad, I., 2020. Effects of temperature on feeding and digestive processes in fish. Temperature 7 (4), 307–320.

Wątor, K., Zdechlik, R., 2021. Application of water quality indices to the assessment of the effect of geothermal water discharge on river water quality–case study from the Podhale region (Southern Poland). Ecol. Indic. 121, 107098.

Xiang, Y., Jiang, L., 2009. Water quality prediction using LS-SVM and particle swarm optimization. In: 2009 Second International Workshop on Knowledge Discovery and Data Mining. IEEE, pp. 900–904.

Xiong, Z., Cui, Y., Liu, Z., Zhao, Y., Hu, M., Hu, J., 2020. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. Comput. Mater. Sci. 171, 109203.

Yousefi, S.R., Alshamsi, H.A., Amiri, O., Salavati-Niasari, M., 2021. Synthesis, characterization and application of Co/Co3o4 nanocomposites as an effective photocatalyst for discoloration of organic dye contaminants in wastewater and antibacterial properties. J. Mol. Liq. 337, 116405.

Zheng, H., Liu, Y., Wan, W., Zhao, J., Xie, G., 2023. Large-scale prediction of stream water quality using an interpretable deep learning approach. J. Environ. Manag. 331, 117309.