scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A Comprehensive Dataset of Surface Water Quality Spanning 1940-2023 for Empirical and ML Adopted Research

Md. Rajaul Karim[1], M. M. Mahbubul Syeed[1,2]✉, Ashifur Rahman[1], Khondkar Ayaz Rabbani[3], Kaniz Fatema[1,2], Razib Hayat Khan[1,2], Md Shakhawat Hossain[1,2] & Mohammad Faisal Uddin[1,2]

Assessment and monitoring of surface water quality are essential for food security, public health, and ecosystem protection. Although water quality monitoring is a known phenomenon, little effort has been made to offer a comprehensive and harmonized dataset for surface water at the global scale. This study presents a comprehensive surface water quality dataset that preserves spatio-temporal variability, integrity, consistency, and depth of the data to facilitate empirical and data-driven evaluation, prediction, and forecasting. The dataset is assembled from a range of sources, including regional and global water quality databases, water management organizations, and individual research projects from five prominent countries in the world, e.g., the USA, Canada, Ireland, England, and China. The resulting dataset consists of 2.82 million measurements of eight water quality parameters that span 1940 - 2023. This dataset can support meta-analysis of water quality models and can facilitate Machine Learning (ML) based data and model-driven investigation of the spatial and temporal drivers and patterns of surface water quality at a cross-regional to global scale.

Background & Summary

Surface water is the body of liquid water found on the Earth's surface in the form of rivers, streams, lakes, wetlands, reservoirs, creeks, sea, etc.[1]. Surface water is an essential natural resource for ecological sustainability and inhibition of various forms of life on Earth[1,2]. However, for several decades, numerous environmental stressors (both anthropogenic and natural factors) have caused significant deterioration in surface water quality. This includes, but is not limited to, the rapid proliferation of urbanization, industrialization, agricultural growth, and natural phenomena that cause constant discharge of effluents, inorganic substances, and contaminants into surface water sources[2–4]. This imbalance in the quality and quantity of surface water can have a significant impact on environmental, sociological, and economic efforts in many parts of the world. As such, the sustainable management of water resources has become a critical challenge[2,4].

Many countries, both developed and developing, have adopted a variety of effective guidelines and policies for the management of water and its associated ecosystems. For example, the European Union (EU) adopted the Water Framework Directive (WFD) in 2000 as an instrument to achieve at least a *good environmental status* of all water bodies[2,5]. The WFD and other similar frameworks rely on the assessment of water quality that provides an evidential means to plan the sustainable use of water resources[1,5].

Consequently, water quality assessment is done periodically through the evaluation of the biological, chemical, and physical properties of the water with reference to a set of standards, natural quality, and effects on human health and ecosystem[6,7]. For this, several Water Quality Index (WQI) models are developed that transform a large number of complex water quality parameter measurements into a unitless, easy-to-comprehend number, which is then assigned to a water quality class, for example, good, fair, moderate, poor, etc.[1]. These models require comprehensive data on the measurement of water quality parameters for a legitimate prediction

[1]RIoT Research Center, Independent University, Dhaka, 1229, Bangladesh. [2]Department of Computer Science and Engineering, Independent University, Dhaka, 1229, Bangladesh. [3]Department of Environmental Science and Management, Independent University, Dhaka, 1229, Bangladesh. ✉e-mail: mahbubul.syeed@iub.edu.bd

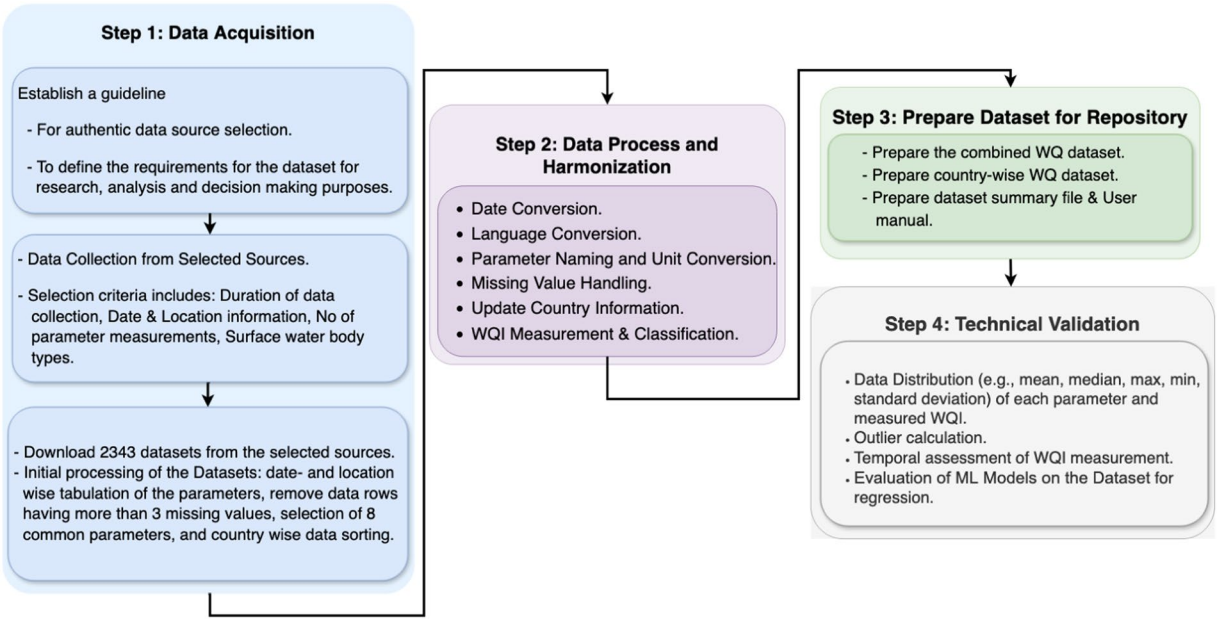


Fig. 1 Step-by-Step process for Data Acquisition, Harmonization, Validation, and Dataset Preparation.

of water quality by minimizing uncertainties, for example, eclipsing and ambiguity issues[1,2,8]. Furthermore, data on water quality parameters are critical for a holistic understanding of the spatial and temporal drivers of surface water quality, and for the development of water management strategies on a global scale[3,9]. Therefore, reliable and standardized data collection and sharing have been communicated as an important step to achieve associated water quality goals as part of the Sustainable Development Goals (SDGs)[10].

Although there are few studies sharing surface water quality data[11–24], hardly any of them extend to the global scale. Furthermore, the number of parameters and their measurement units varies between datasets and is limited in spatio-temporal resolution[12]. To our knowledge, Global River Chemistry (GLORICH)[25] is a publicly available dataset of water quality consisting of 1.27 million data rows, where each row records the reading of 47 water quality parameters. Although this data set has had a significant impact on water quality research, it is not free from limitations. The dataset records only river water quality data. Many of the parameters have less than 100 readings and others have missing values. For example, only 64 readings are available for the parameter *Particulate Organic Nitrogen Concentration* and only 98 are available for the parameter *Particulate Sulfur Concentration*. This imbalance in the dataset should introduce biases and can affect reliable analyses for data-intensive research.

Considering the need to address the limitations of existing datasets, this study attempts to provide a comprehensive and harmonized global dataset on water quality that researchers, policymakers, and professionals can use in a meaningful way. Data are collected from a number of sources, including local, regional, and global water quality databases, governmental organizations, water management commissions, water development boards, and individual research projects from five prominent countries in the world, e.g. the United States of America (USA), Canada, Ireland, England, and China. The data collected are processed and harmonized using a standard data processing method (as summarized in Fig. 1) and an open-access repository is created to make it available under an open-access license agreement. Furthermore, a three-fold validation of the dataset is performed to demonstrate its applicability. This includes (a) standard statistical evaluations in the data set, (b) implementation of a WQI model to measure water quality to complement the dataset, and finally (c) application of machine learning (ML) algorithms for data-driven prediction of water quality[26]. The accumulated dataset consists of 2.82 million water quality measurements in which each data row represents the measurement of eight parameters for a given source at a given location for a day. The data collection period spans between 1940 and 2023. The dataset records measurements of eight water quality parameters, namely, Ammonia (NH_3), Five-day Biochemical Oxygen Demand (BOD_5), Dissolved Oxygen (DO), Orthophosphate (PO_4^{-3}), Potential of Hydrogen (pH), Temperature, Nitrogen (N), and Nitrate (NO_3^-) for several surface water sources, including rivers, lakes, reservoirs, creeks, and coastal area.

The dataset can be used for data and model-driven studies, including (a) the development, calibration and validation of WQI models, (b) the investigation of spatial and temporal drivers and patterns of surface water quality on global and cross-regional scales, (c) the development/application of ML methods for accurate prediction and forecasting of WQI, and (d) to assess the long-term impact of water quality on ecosystem health and human society. This in turn will help develop strategies and policies for the management and preservation of water resources.

Data Acquisition Summary	Selected Data Sources	Selected Datasets	Datasets Sources Link
Google Dataset Search: *datasetsearch.research.google.com* Search Terms: *"Water Quality Dataset" or "Surface Water Quality Dataset"*	Environmental Protection Agency (EPA) Catchments, Ireland (catchments.ie)	982	Total Selected Datasets: *2343* Selected Datasets Sources Access Repository[27] File Name: *DataCollectionSources.xlsx*
	Department for Environment Food & Rural Affairs, United Kingdom (UK) (environment.data.gov.uk)	474	
	Open Government, Canada (open.canada.ca)	281	
	National Water Quality Monitoring Council, United States (US) (waterqualitydata.us)	269	
	Hong Kong Govt. Data (data.gov.hk)	134	
	Find Open Data, UK (data.gov.uk)	50	
	AmeriGEO, Americas (data.amerigeoss.org)	40	
	US Data Catalog & Applications (catalog.data.gov)	39	
	European Data Portal, EU (data.europa.eu)	26	
	NASA Earthdata, US (earthdata.nasa.gov)	6	
	Bright Stripe, UK (brightstripe.co.uk)	6	
	California Open Data (data.ca.gov)	5	
	Ireland Open Data (data.gov.ie)	5	
	Ontario Data Catalogue (data.ontario.ca)	4	
	Other Data Sources	22	

Table 1. Selection of Data Sources and Initial Datasets with Access Links.

Methods

The collection, processing, and preparation of the dataset is performed through the adoption of a well-articulated process. Figure. 1 presents the schematic flowchart of this process, a detailed discussion of which is documented below. Additionally, Table 1 summarizes the sources for data collection, Fig. 2 shows the statistical overview of the dataset according to geographic locations, and Table 3 outlines the dataset variables. Lastly, all scripts written for data processing, harmonization, and technical validation are accessible through the links outlined in Section *Code Availability*.

Data Acquisition. Multiple selection criteria and validation assessments are carried out to select credible data sources for the acquisition of water quality data. To begin with, a thorough study of related research articles and analytical reports on water quality assessment is carried out. This step leads to the establishment of a step-by-step guideline for the selection of authentic data sources for data acquisition, as shown in Step 1 of Fig. 1.

To acquire initial datasets, Google Dataset Search is used with search terms *"Water Quality Dataset" or "Surface Water Quality Dataset"* and the search filter is set to *all year* and *tabular data format* as summarized in column 1 of Table 1. This search leads to a listing of relevant global and regional water quality dataset sources. The authors then performed a manual selection to include the most relevant dataset sources based on the following criteria: (a) number of water quality parameters reported, (b) duration of data collection (i.e. number of years), (c) date and location information, (d) types of surface water body covered, (e) completeness of the dataset, and (f) file format (CSV/XLSX/Tabular). This process leads to a selection of 15 (fifteen) data sources from which 2343 datasets are downloaded for further processing. The detailed documentation of this selection is provided in Table 1 (columns 2 and 3), with access links to all selected datasets in column 4. The *DataCollectionSources.xlsx* file is available in the repository[27] under the *Data Collection Sources* folder, and the folder structure is illustrated in Fig. 3.

The 2343 datasets are then downloaded to the local repository in CSV format. At first, the data in each CSV file are restructured to generate day-/date-wise measurements of all the parameters for a given location. For this, each data row in the CSV files is transposed to columns, resulting in a date- and location-wise tabulation of the parameter measurements. A *Python script* (library used: *Pandas, Numpy*) is written to automate this process. The resulting dataset consists of 450 million data rows that span between 1940 and 2023, occupying a total size of 28 GB. These 450 million rows are then categorized by geo-location (i.e., country), which limits the selection to five countries: the USA, Canada, Ireland, England, and China. Subsequently, common water quality parameters among these countries are identified. This process highlights eight common parameters: Ammonia (NH_3), Five-day Biochemical Oxygen Demand (BOD_5), Dissolved Oxygen (DO), Orthophosphate (PO_4^{-3}), Potential Hydrogen (pH), Temperature, Nitrogen (N), and Nitrate (NO_3^-). Consequently, any data rows with more or fewer than these eight parameters are discarded from the dataset. Additionally, data rows containing over three missing parameter values are omitted, resulting in a final dataset of 2.82 million rows. The selection of these common water quality parameters across the five countries creates a consistent and unified global water quality dataset. Furthermore, reducing missing data improves both the originality and the overall quality and quantity of the dataset for data-driven assessments of water quality. The spatial and temporal distribution of the dataset is shown in Fig. 2.

Data Processing and Harmonization. Data processing and harmonization is a core part of dataset construction to achieve the objectives with this dataset. Step 2 in Fig. 1 illustrates the actions involved in the data processing and harmonization phase. All processing is done using Python libraries (Version: 3.0) and MSExcel

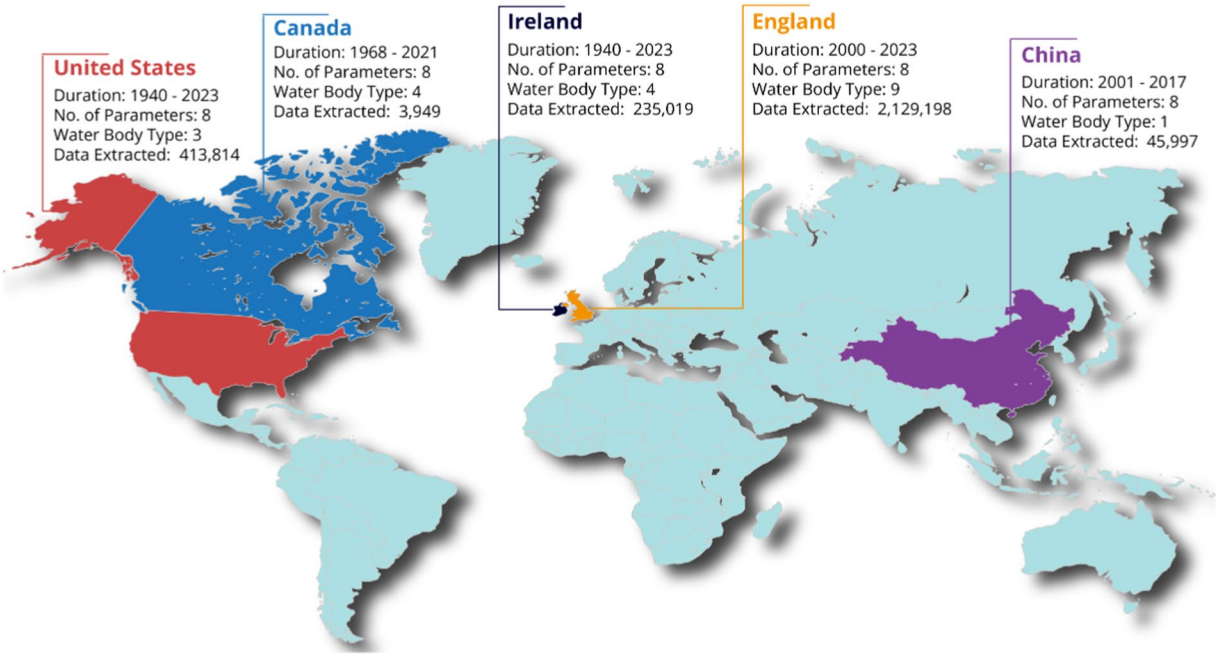


Fig. 2 Data Collection Summary with Geographic Locations.

Detected Terms	Equivalent English Term
氨水(Chines)	Ammonia
Ammoniaque (French)	Ammonia

Table 2. Sample of Language Converstion to English.

macros (Version: 16.83). Since information on the nomenclature of water quality parameters, reported units, language, and date formatting could vary between regional datasets, the harmonization process should standardize all of these cited issues. In addition, missing values are handled and WQIs are calculated using the dataset, which increases its readiness for machine learning / deep learning (DL) adopted data-driven predictive and statistical analysis.

Language Conversion. Multilingual data records are found in the datasets collected for China and Canada. For the earlier, *Chinese* terms are used to define the parameter names and their units, and for the later, *French* language is used for the same. For consistency in data, these terms are converted to their *English* equivalent. To do this, the dataset is searched for the *Chinese* and *French* terms, and a vocabulary is developed for conversion. Then, this vocabulary is translated to their *English* equivalent using *Google Translator*. Finally, a *Python script* (library used: *googletrans, deep-translator*) is written to search for each of the terms and replace them with the *English* equivalent in the dataset. A sample of this language conversion is presented in Table 2 for reference.

Date Conversion. The date format used in the dataset varies significantly between counties. To standradize the format, a *Python script* (library used: *datetime*) is written to convert all the dates in the dataset to the standard *dd-mm-yyyy* format. This conversion allows for convenience in temporal analysis, time series modeling, and statistical assessment using the dataset.

Parameter Naming and Unit Conversion. Since the nomenclature of the water quality parameter and the units reported differ between countries and organizations, the parameter name and its units are converted to harmonize the dataset. The renaming process identified all the variations of a parameter name and then replaced them with the standard name as listed in Table 3. For example, all variations for the parameter *Nitrogen* (e.g., Total identified Nitrogen or Nitrogen Total or Total observed Nitrogen) are renamed to *Nitrogen*(*N*).

Furthermore, the units of each parameter are checked in the dataset and are converted to the standard units as per Table 3. For instance, the unit *Grams/Liter* is converted to *Milligram/Liter*. Also, the unit *Degree Fahrenheit* is converted to *Celcious* (for *Temperature* parameter) using the following standard conversion.

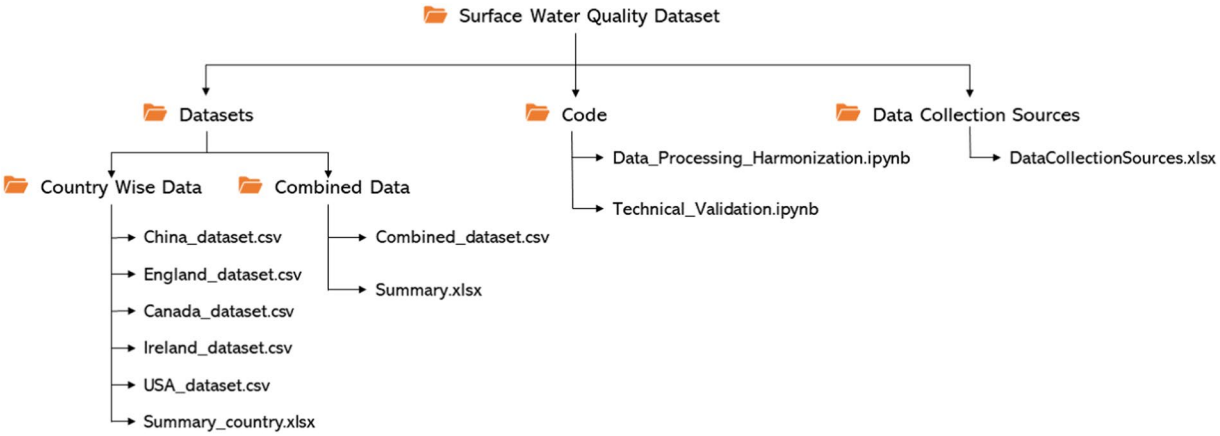


Fig. 3 Folder structure of the Water Quality Dataset Repository[27].

Variable Name	Description	Unit
Country	Name of the Water-body Region	—
Area	Name of the Area in the Region	—
Waterbody Type	Type of the Water-body source	—
Date	Date of the sample collection	dd-mm-yyyy
Ammonia (NH_3)	Concentration of ammonia	mg/l
Five-day Biochemical Oxygen Demand (BOD_5)	Measure of the amount of oxygen required	mg/l
Dissolved Oxygen (DO)	Concentration of dissolved oxygen	mg/l
Orthophosphate (PO_4^{3-})	Concentration of orthophosphates	mg/l
pH	pH of water	pH units
Temperature (Temp.)	Evaluated Temperature	°C (celsius)
Nitrogen (N)	Total nitrogen present	mg/l
Nitrate (NO_3^-)	Concentration of nitrate	mg/l
CCME_Values	The WQI value calculated using CCME WQI Model	Range 0 –100
CCME_WQI	The WQI classification w.r.t CCME_Values	—

Table 3. Comprehensive List of Dataset Variables with Corresponding Descriptions and Units of Measurement.

$$T(C) = (T(F) - 32) \times \frac{9}{5}$$

Where:

- $T(°C)$ is the temperature in degrees Celsius
- $T(°F)$ is the temperature in degrees Fahrenheit

Missing Value Handling. Taking into account the nature of the dataset and its intended use, the missing parameter values are calculated using the *median* imputation technique. This approach also mitigates the influence of data outliers and provides a robust measure of central data tendency. To calculate the country-wise median for missing values of a parameter, the corresponding data column is arranged in ascending order, and then the median is calculated using the following method. Corresponding *Python script* (library used: *numpy, sklearn*) is written to accomplish this task.

For an odd number of elements in a data column:

$$\text{Median} = X_{\frac{n+1}{2}}$$

For an even number of elements in a data column:

$$\text{Median} = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

In both cases, X_i represents the ith element of the sorted dataset X.

There are two main types of missing data patterns: Missing Completely at Random (MCAR) and Missing Not at Random (MNAR). MCAR indicates that the probability of missingness is independent of both observed and unobserved data, while MNAR suggests that missingness is related to unobserved values or specific conditions. For MCAR, median imputation is suitable, as it preserves the central tendency of the dataset without significant skew. However, MNAR can introduce bias if median imputation is used, necessitating more advanced methods such as multiple imputations or model-based approaches. Our dataset is considered MCAR since missing values occur randomly in countries without identifiable patterns. Thus, we used median imputation to fill in the missing values.

Update Country Information. To incorporate country information in the unified dataset, a new column is added that indicates the country of origin for the data records.

WQI Measurement and Classification. WQI measurement and classification are necessary to perform a data-driven evaluation of water quality, e.g., design, development, and performance evaluation of ML or DL models for accurate and reliable water quality prediction and forecasting. These models can also be trained to develop decision support systems by analyzing the trend and patterns of water quality as observed by the time series WQI classification. To better support such studies, the WQI is calculated using day-wise parameter readings and the corresponding classification of the water quality is obtained for the dataset. To document these data in the dataset two additional columns are added (last two variables in Table 3). For WQI computation, Canadian Council of Ministers of the Environment (*CCME*) model is implemented in *Python* (library used: *pandas, numpy, glob, and OS*). CCME calculates WQI using the following mathematical model[1].

F_1: Represents the percentage of parameters that do not meet the specified objectives.

$$F_1 = \left(\frac{\text{number of failed parameters}}{\text{total number of parameters}}\right) \times 100$$

F_2: Indicates the percentage of individual test values not meeting objectives.

$$F_2 = \left(\frac{\text{number of failed tests}}{\text{total number of tests}}\right) \times 100$$

F_3: Measures the deviation of test values from the objectives using an asymptotic function.

$$F_3 = \frac{NSE}{0.01(NSE) + 0.01}$$

Here, the *NSE* is calculated by summing the excursions of individual tests from their objectives and dividing by the total number of tests:

$$NSE = \left(\frac{\sum_{i=1}^{n} \text{excursion}_i}{\text{total number of tests}}\right) - 1$$

The excursion (excursion$_i$) for each test value is determined based on whether it falls below or exceeds the objective value:

$$\text{excursion}_i = \frac{\text{Objective}_j}{\text{failed test value}_i} - 1$$

The CCME WQI is then calculated as:

$$CCME\ WQI = 100 - \frac{\sqrt{F_1^2 + F_2^2 + F_3^2}}{1.732}$$

The divisor of 1.732 in the aggregation equation normalizes the WQI to a range of 0 to 100, where 0 denotes the worst water quality and 100 is the best, considering the maximum values of the individual index factors[2,28]. The specific classification scheme for CCME WQI is generally categorized into five ranges: Excellent (90-100), Good (80-89), Fair (65-79), Marginal (45-64), and Poor (0-44)[1].

Data Records

The dataset and associated contents (e.g. code files and data collection sources) are publicly available in the *figshare* repository[27] with digital object identifier (DOI) as https://doi.org/10.6084/m9.figshare.27800394.v2. The dataset is made available under the license, *CC BY 4.0* that allows reuse without restriction. The folder structure of this repository is presented in Fig. 3 in which the parent directory *Surface Water Quality Dataset* contains 3 folders, namely, *Datasets, Code, Data Collection Sources*.

As shown in Fig. 3, the *Datasets* folder contains two sub-folders, *Combined Dataset* and *Country Wise Dataset*. The *Combined Dataset* folder has two files, (a) the Combined_Dataset.csv file that contains the entire

dataset, and (b) Summary.xlsx file that presents a brief description of the dataset with a summary of the data distribution (e.g., maximum, minimum, mean, standard deviation, etc.). The *Country Wise Dataset* folder contains five CSV files each containing data for one of the five countries and a summary file detailing the country-wise distribution of the dataset.

The *Code* folder contains the *Python* codes (with detailed documentation) that are used for data processing, harmonization, and technical validation. Finally, the *Data Collection Sources* folder contains a file having all the 2343 dataset source links.

Technical Validation

The dataset contains 2.82 million data rows from five countries, each representing readings of eight water quality parameters, the corresponding measurement of WQI and its classification. The following technical validations are performed to assess the quality of the dataset in relation to the study objectives.

Data Distribution. The statistical evaluation (e.g. count, mean, standard deviation (STD), minimum, maximum and quartile value) is carried out in the combined dataset and for each country. Observed results are presented in Table 4. In general, the mean values for all parameters fall within their standard range[2,29,30]. In the combined dataset, the variability of the data around *Mean* for *pH* and Orthophosphate (PO_4^{-3}) are the lowest with STD values of 0.495 and 2.089, while *BOD* and *DO* exhibit relatively high variability with STD values of 16.414 and 1.851, respectively. The *Min* and *Max* values for few parameters are recorded out-of-range, indicating the existence of possible outliers. Therefore, an assessment of data outliers is calculated and reported in the following section. Furthermore, the CCME WQI classification reveals that the overall water quality is good with *Mean* value of 85.047. However, the quality of the water varies between poor (CCME WQI value of 31.304) and excellent (CCME WQI value of 100.00) within the dataset. Country-wise data distribution shows notable distinctions among different regions. For example, China has a relatively low mean value in parameters such as Ammonia (0.101 mg/l), *BOD* (0.912 mg/l), and *DO* (8.314 mg/l), with standard deviations of 0.321, 0.768, and 2.876, respectively, indicating limited variability between these metrics. Consequently, overall water quality in China is reported to be good, with a mean WQI value of 96.508. Similar results can be observed for Canada, Ireland, and the USA, with Ireland and the USA having an overall water quality rating very good (mean WQI values 98.109 and 98.464, respectively). In the case of England, the mean values for most of the water quality parameters are relatively higher than in other countries. For example (ref. Table 4), the higher mean values of nitrogen (6.634 mg/l) and nitrate (5.555 mg / l) indicate higher levels of nutrients in water, which might contribute to a marginally good level of water quality with a mean WQI index of 80.740.

Outlier Detection. In assessing the data variability, *Tukey's outlier detection method*[31] is used. This method effectively handles both symmetric and skewed data, unlike the standard deviation (STD) method (Mean $\pm$ 2 STD, Mean $\pm$ 3 STD), which assumes normality. Tukey's approach is based on the 3rd inter-quartile range (IQR) which makes no distributional assumptions[23]. The IQR represents the spread of the data and calculates the difference between the first quartile (Q1) and the third quartile (Q3):

$$IQR = Q3 - Q1$$

The data points are identified as potential outliers based on inner fences utilizing a 1.5 IQR interval, while the outer fences utilize a 3 IQR interval and identify them as possible outliers[32].

Inner fences are calculated using the below equation.

$$\begin{aligned} \text{Low Potential Outliers} &= Q1 - 1.5IQR \\ \text{High Potential Outliers} &= Q3 + 1.5IQR \end{aligned}$$

Outer fences are calculated using the below equation.

$$\begin{aligned} \text{Low Possible Outliers} &= Q1 - 3IQR \\ \text{High Possible Outliers} &= Q3 + 3IQR \end{aligned}$$

The inner and outer fence calculation for each parameter is presented in Table 5. Comprehending these fences, any value that falls below the *Low Potential Outliers* or above the *High Potential Outliers* is considered the *Potential Outlier*. Likewise, a value that falls below the *Low Possible Outliers* or goes above the *High Possible Outliers* is classified as *Possible Outlier*. The percentile of possible and potential outliers for each water quality parameter can then be calculated using the following equations.

$$\text{possible outliers (\%)} = \frac{(\text{No. of possible outliers} * 100)}{\text{Total number of observations}}$$

$$\text{potential outliers (\%)} = \frac{(\text{No. of potential outliers} * 100)}{\text{Total number of observations}}$$

As can be seen from Table 5, the percentage of outliers detected using Tukey's method for all parameters is relatively low. For example, *Temperature* has the lowest *possible* and *potential* outliers with 0.33% and 2.10%, respectively, and *Orthophosphate* (PO_4^{-3}) has the highest, with *possible* outliers at 13.88% and *potential* outliers at 16.62%. As the percentage of outliers detected is relatively low with respect to the size of the dataset, their impact on models' performance will be insignificant for most applications. Therefore, the outliers are not

Country	Statistics	Ammonia	BOD	DO	Orthophosphate	pH	Temperature	Nitrogen	Nitrate	CCME Values
Combined Dataset	Count	2827977	2827977	2827977	2827977	2827977	2827977	2827977	2827977	2827977
	Mean	1.171	4.887	10.008	0.710	7.736	11.839	5.210	4.767	85.047
	STD	5.669	16.414	1.851	2.089	0.495	5.010	6.200	6.074	17.647
	Min	-0.005	-2.000	0.000	-0.004	-1.000	-5.260	0.000	0.000	31.304
	25%	0.030	1.600	9.860	0.040	7.550	8.990	0.780	1.173	77.153
	50%	0.055	2.700	10.200	0.107	7.780	11.460	4.000	4.500	90.596
	75%	0.317	2.830	11.000	0.227	8.000	14.200	6.320	4.940	100.000
	Max	200.000	255.000	20.000	100.000	30.000	98.000	46.000	155.000	100.000
China	Count	45997	45997	45997	45997	45997	45997	45997	45997	45997
	Mean	0.101	0.912	8.314	0.018	8.017	23.399	0.033	0.136	96.508
	STD	0.321	0.768	2.876	0.036	0.246	4.175	0.049	0.205	4.624
	Min	0.005	0.100	0.000	0.002	2.100	13.000	0.002	0.002	51.076
	25%	0.024	0.500	6.000	0.006	7.900	19.800	0.011	0.026	93.183
	50%	0.046	0.700	7.600	0.011	8.000	24.200	0.019	0.077	100.000
	75%	0.100	1.100	11.684	0.021	8.200	27.000	0.034	0.150	100.000
	Max	10.000	21.000	16.100	1.100	9.300	33.200	1.100	5.900	100.000
England	Count	2129198	2129198	2129198	2129198	2129198	2129198	2129198	2129198	2129198
	Mean	1.534	5.965	10.222	0.907	7.728	11.575	6.634	5.555	80.740
	STD	6.471	18.762	1.559	2.208	0.465	4.128	6.511	5.125	18.134
	Min	0.000	0.000	0.000	0.000	0.000	-5.260	0.000	0.000	31.304
	25%	0.030	1.830	10.100	0.056	7.530	8.800	2.490	2.990	70.224
	50%	0.098	2.700	10.200	0.144	7.780	11.460	5.000	4.500	88.761
	75%	0.500	3.460	10.900	0.442	7.990	14.200	7.980	6.150	93.069
	Max	200.000	255.000	20.000	100.000	14.000	97.500	46.000	153.000	100.000
Canada	Count	3949	3949	3949	3949	3949	3949	3949	3949	3949
	Mean	0.847	3.186	9.897	0.426	7.688	12.421	0.679	10.201	90.217
	STD	3.619	7.393	1.441	2.370	0.634	4.078	1.392	13.711	12.365
	Min	0.000	0.000	0.058	0.000	3.475	0.000	0.000	0.000	37.659
	25%	0.050	2.000	9.300	0.030	7.640	10.172	0.400	2.081	86.597
	50%	0.100	2.133	9.824	0.061	7.790	12.000	0.400	5.837	93.117
	75%	0.196	3.000	10.775	0.150	7.933	14.600	0.400	13.084	100.000
	Max	63.836	231.108	16.930	35.338	19.445	89.000	41.784	150.000	100.000
Ireland	Count	235019	235019	235019	235019	235019	235019	235019	235019	235019
	Mean	0.081	1.529	8.512	0.222	7.684	11.320	1.782	1.511	98.109
	STD	10.940	1.094	3.233	2.585	0.847	4.008	1.511	1.175	6.406
	Min	-0.005	-2.000	0.100	-0.004	3.200	0.000	0.000	0.000	37.192
	25%	0.030	1.400	4.973	0.020	7.400	8.500	0.920	1.300	100.000
	50%	0.038	1.400	9.090	0.025	7.800	11.150	1.422	1.300	100.000
	75%	0.043	1.400	11.684	0.032	8.055	14.000	2.100	1.300	100.000
	Max	134.500	220.000	20.000	100.000	30.000	98.000	46.000	112.452	100.000
USA	Count	413814	413814	413814	413814	413814	413814	413814	413814	413814
	Mean	0.048	1.701	9.946	0.055	7.777	12.201	0.447	3.023	98.464
	STD	0.899	2.089	1.482	0.285	0.361	7.559	0.555	9.974	4.668
	Min	0.000	0.000	0.000	0.000	-1.000	-1.000	0.001	0.000	38.672
	25%	0.021	1.600	9.870	0.040	7.800	9.600	0.400	0.900	100.000
	50%	0.021	1.600	9.870	0.040	7.800	11.100	0.400	0.900	100.000
	75%	0.021	1.600	10.300	0.040	7.800	12.900	0.400	0.900	100.000
	Max	157.000	216.000	20.000	65.600	17.000	96.400	46.000	155.000	100.000

Table 4. Statistical Distribution of the Water Quality parameters and Corresponding WQI Values (Combined and Country-Wise).

removed from the dataset. However, based on the sensitivity level of the research, users can detect and remove outliers using standard methods, e.g., trimming, winsorization, or imputation[33].

Water Quality Trend Analysis. To assess the suitability of the dataset in relation to the development, calibration, and validation of a WQI model and to perform the spatio-temporal trend and pattern analysis, the WQI

Parameters	Q1	Median (Q2)	Q3	IQR	Inner fence		Outer fence		Potential Outliers	Possible Outliers
					Low Potential Outliers	High Potential Outliers	Low Possible Outliers	High Possible Outliers		
Ammonia	0.03	0.055	0.317	0.29	−0.40	0.75	−0.83	1.18	14.31%	11.49%
BOD	1.6	2.7	2.83	1.23	−0.24	4.68	−2.09	6.52	14.60%	10.44%
DO	9.86	10.2	11.0	1.14	8.15	12.71	6.44	14.42	12.47%	6.46%
Orthophosphate	0.04	0.11	0.23	0.19	−0.24	0.51	−0.52	0.79	16.62%	13.88%
pH	7.55	7.78	8.0	0.45	6.88	8.68	6.20	9.35	4.66%	0.96%
Temperature	8.99	11.46	14.2	5.21	1.18	22.01	−6.64	29.83%	2.10%	0.33%
Nitrogen	0.78	4.0	6.32	5.54	−7.53	14.63	−15.84	22.94	7.65%	2.89%
Nitrate	1.17	4.5	4.94	3.77	−4.48	10.59	−10.13	16.24	8.88%	3.99%

Table 5. Summary of water quality parameters with outlier detection ranges (Inner fence and outer fence).

classification measured using the CCME method is plotted on trend charts as shown in Fig. 4. This plot allows to observe the temporal water quality traits for each country to derive data-driven forecasting. According to Fig. 4, CCME WQI in China (Bay water) ranges between 80 (Good) and 100 (Excellent) for 365 days. During the same period, the WQI for England (River water) ranges between 65 (Fair) and 85 (Good), and for Ireland (River water) it remains within the range of 75 (Fair) and 95 (Excellent). However, the WQI in the USA (River water) shows the best quality being in the range of 90 - 100 (excellent), and the WQI in Canada (River water) exhibits sudden fluctuations with a range between 50 (Marginal) and 100 (excellent).

Application of the Dataset for ML Models. One of the core objectives of the dataset is to support the development, training, and testing of cutting-edge ML and DL models for accurate prediction and forecasting of water quality on a global scale considering spatial and temporal drivers and patterns of surface water quality. To demonstrate the preparedness of the dataset in relation to this objective, we have trained and tested four classical ML models, namely, Linear Regression, Decision Tree Regression, Random Forest Regression, and XGBoost Regression, and two deep learning models, namely, artificial neural network (ANN) and long-short-term memory (LSTM) using the dataset. The hyperparameters used to train the models are documented in Table 6. Before training and testing, all outliers for parameters are removed using the IQR test, and normalize the dataset using *the Skearn MinMaxScaler()* function. Moreover, the dataset is split into an 80:20 ratio for training and testing purposes. All experiments are conducted using Google Colab with an allocated Tesla T4 GPU, which provides sufficient computational resources to train the models. The implementation is carried out using Python frameworks, namely TensorFlow, Keras, and Scikit-learn.

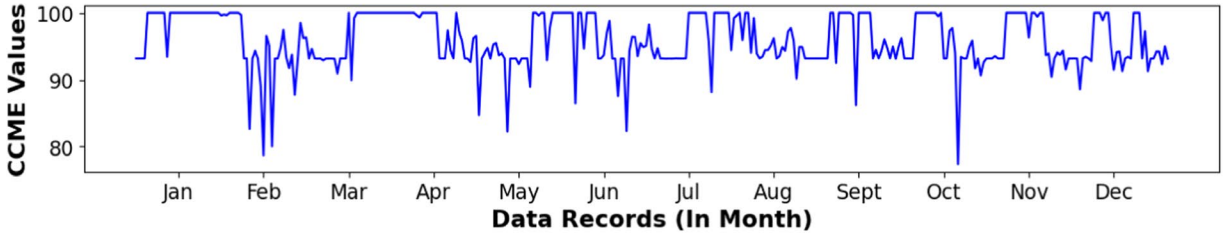
Consequently, the observed results are presented in Table 7 in terms of the mean squared error (MSE), the root mean squared error (RMSE), the mean absolute error (MAE), and the coefficient of determination (R^2), a visualization of which can be found in Fig. 5. Based on the results, the classical ML models, specifically Decision Tree and Random Forest regression, perform better in terms of MSE (0.0002 and 0.0002), RMSE (0.014 and 0.015), MAE (0.0003 and 0.0002), and R^2 (0.99 and 0.99), for the train and test dataset, respectively. These results are reflected in Fig. 5, where the predicted WQI values aligned almost accurately with the perfect prediction line for the Decision Tree and Random Forest regression models.

Usage Notes

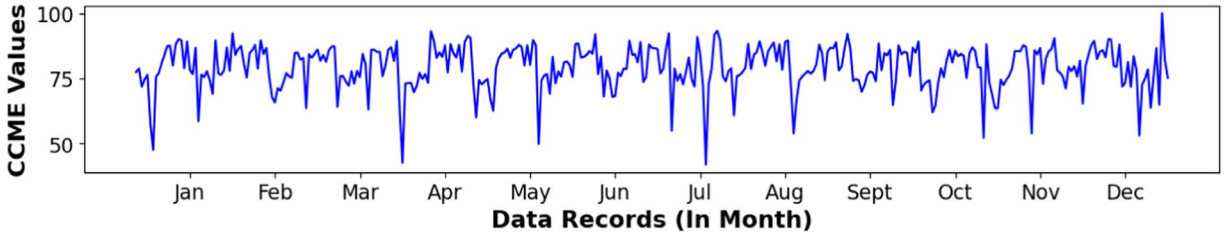
a) The dataset[27] contains 2.82 million records from five different countries, covering eight water quality parameters. Data are gathered from various online sources contributed by governmental agencies, environmental research organizations, and citizen science initiatives. In addition to the combined dataset, country-specific datasets can facilitate regional analysis, allowing for targeted studies and cross-regional comparisons. Available in CSV format, the dataset is compatible with statistical tools such as Python (Pandas, NumPy, Scikit-learn), R, MATLAB, and SQL-based databases.

b) This dataset has broad applications in environmental science, hydrology, and machine learning. It can be used to analyze long-term water quality trends, identify pollution sources, and evaluate human impacts on aquatic ecosystems. Additionally, it supports hydrological modeling, climate change studies, and water resource management. Machine learning and deep learning practitioners can apply it for predictive modeling, anomaly detection, and automated water quality classification, as well as for the development of early warning systems that contribute to public health and environmental sustainability.

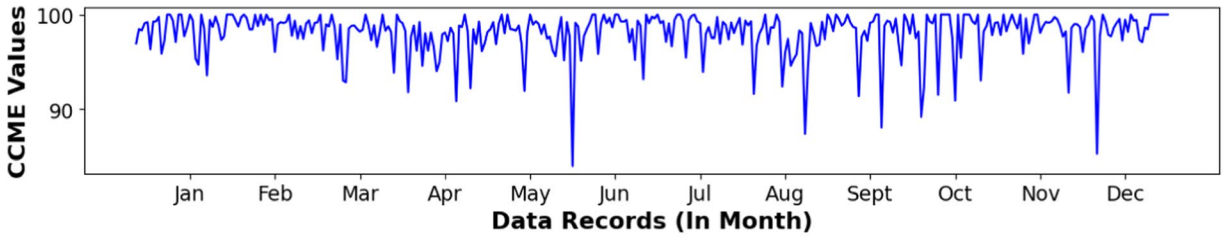
c) To facilitate seamless integration with other water quality datasets, several features have been incorporated, including the standardization of all parameters to align with internationally recognized naming and unit conventions. This harmonization simplifies direct comparisons and enables easy merging to support researchers. Scripts are provided, as outlined in the *Code Availability* section. The *Data_Processing_Harmonization.ipynb* script offers functions for aligning this dataset with external water quality datasets by standardizing parameter names, performing unit conversions, and conducting temporal resampling. Additionally, the *Technical_Validation.ipynb* script provides insights into potential transformations required when integrating with datasets of varying resolutions. By utilizing these scripts, researchers can ensure



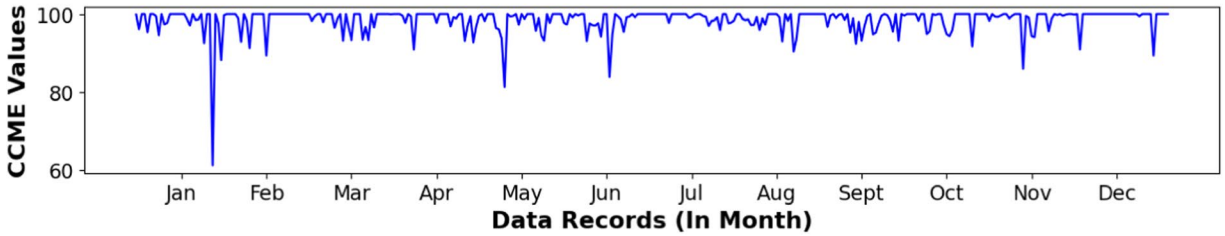
(a) Temporal variation of WQI values in China.



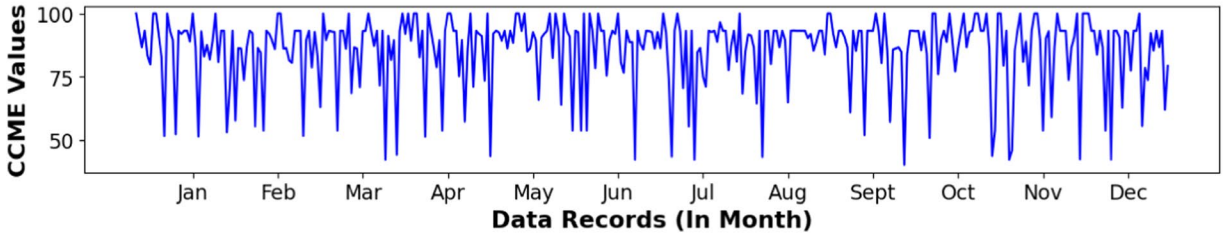
(b) Temporal variation of WQI values in England.



(c) Temporal variation of WQI values in Ireland.



(d) Temporal variation of WQI values in USA.



(e) Temporal variation of WQI values in Canada.

Fig. 4 Temporal Variation of WQI Values for the Five Countries in 365 Days.

consistency in parameter sets, unit conventions, and temporal resolutions when integrating this dataset with others.

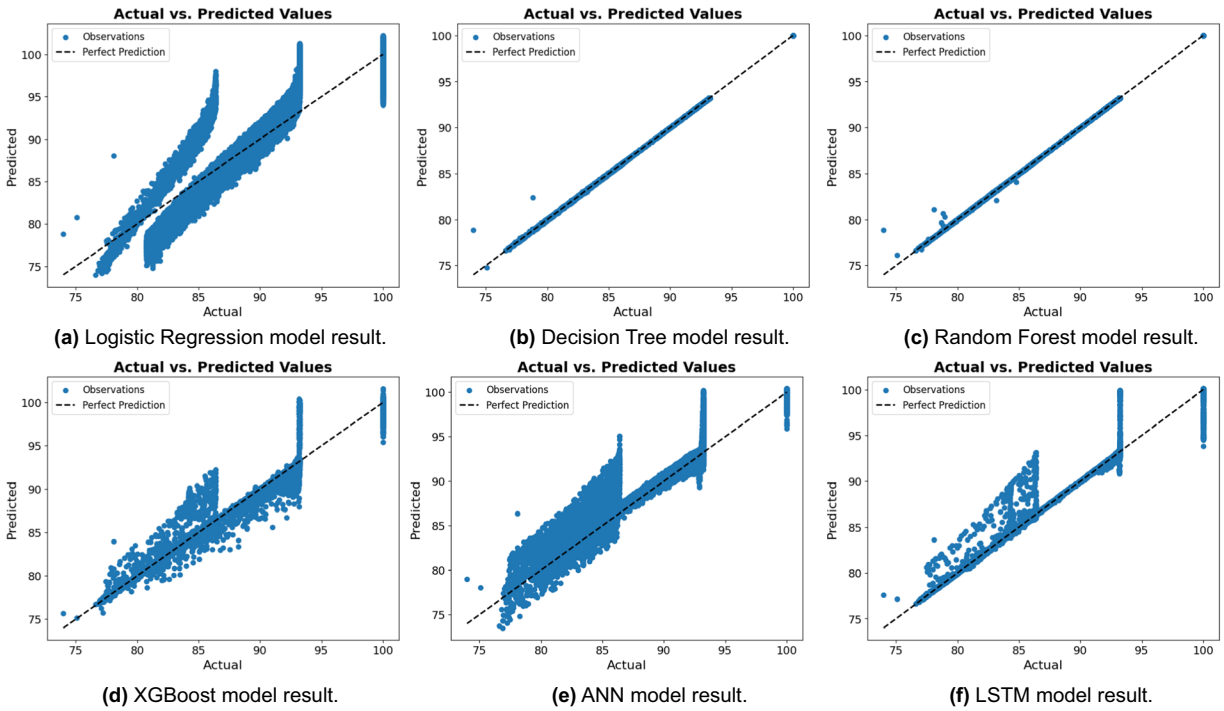
d) While the dataset offers valuable insights, it has some limitations. The records, which cover only five countries, may not represent global water quality conditions comprehensively. Additionally, variations in water quality standards and monitoring methodologies across countries should be considered when making cross-regional comparisons. Despite efforts to ensure consistency, biases may exist in data collection and reporting, potentially influencing certain analyses. Users should recognize these limitations and apply appropriate statistical and ML methods to address potential biases.

Model Name	Hyperparameter
Linear Regression	Default Parameters
Decision Tree Regressor	Default Parameters
Random Forest Regressor	Default Parameters
XGBoost Regressor	Default Parameters
ANN	Optimizer = 'Adam', loss = 'mean_squared_error', Epochs=30, Batch Size = 32, Units = 64, Number of hidden layers = 1
LSTM	Optimizer = 'Adam', loss = 'mean_squared_error', Epochs=30, Batch Size = 32, Units = 50, LSTM layers = 1

Table 6. Hyperparameters used for various regression models.

ML Model	MSE	RMSE	MAE	R^2
Linear Regression	2.63	1.62	1.26	0.91
Decision Tree Regressor	0.0001	0.012	0.0002	0.99
Random Forest Regressor	0.0001	0.012	0.0002	0.99
XGBoost Regressor	0.065	0.255	0.037	0.99
ANN	0.352	0.593	0.347	0.99
LSTM	0.040	0.200	0.027	0.99

Table 7. Performance Assessment of ML models for Prediction of WQI using the dataset.



(a) Logistic Regression model result. **(b)** Decision Tree model result. **(c)** Random Forest model result.

(d) XGBoost model result. **(e)** ANN model result. **(f)** LSTM model result.

Fig. 5 Performance Assessment Curves for the Machine Learning Models.

Code availability

Several scripts are written in *Python* using several libraries (as documented in relevant sections) for data processing, harmonization, and technical validation. All these scripts are categorically available in the *code* folder of the public repository[27] and are accessible directly through the public links below.

- The code used for *Data Processing and Harmonization* (e.g., Language Conversion, Date Conversion, Parameter Naming, and Unit Conversion, Missing Value Handling, WQI Measurement, and Classification) is provided in *Data_Processing_Harmonization.ipynb* file (https://figshare.com/articles/dataset/A_Comprehensive_Surface_Water_Quality_Monitoring_Dataset_1940-2023_2_82Million_Record_Resource_for_Empirical_and_ML-Based_Research/27800394/2?file=50757300).

- The code used for *Technical Validation* (e.g., assessing the Data Distribution, Outlier Detection, Water Quality Trend Analysis, and Verifying the Application of the Dataset for the ML Models) is provided in *Technical_Validation.ipynb* file (https://figshare.com/articles/dataset/A_Comprehensive_Surface_Water_Quality_Monitoring_Dataset_1940-2023_2_82Million_Record_Resource_for_Empirical_and_ML-Based_Research/27800394/2?file=50757303).

Received: 8 May 2024; Accepted: 27 February 2025;
Published online: 06 March 2025

References

1. Syeed, M. M. *et al*. Surface water quality profiling using the water quality index, pollution index and statistical methods: A critical review. *Environ. Sustain. Indic.* **18**, 100247 (2023).
2. Uddin, M. G., Nash, S. & Olbert, A. I. A review of water quality index models and their use for assessing surface water quality. *Ecol. Indic.* **122**, 107218, https://doi.org/10.1016/j.ecolind.2020.107218 (2021).
3. Suresh, K. *et al*. Recent advancement in water quality indicators for eutrophication in global freshwater lakes. *Environ. Res. Lett.* **18**, 063004 (2023).
4. Zaldívar, J.-M. *et al*. Eutrophication in transitional waters: an overview. *Transitional Waters Monogr.* **2**, 1–78 (2008).
5. Uddin, M. G., Nash, S., Rahman, A. & Olbert, A. I. A comprehensive method for improvement of water quality index 270 (wqi) models for coastal water quality assessment. *Water Res.* **219**, 118532 (2022).
6. Ngubane, Z., Bergion, V., Dzwairo, B., Stenström, T. A. & Sokolova, E. Multi-criteria decision analysis framework for engaging stakeholders in river pollution risk management. *Sci. Reports* **14**, 7125 (2024).
7. Kipsang, N. K., Kibet, J. K. & Adongo, J. O. A review of the current status of the water quality in the nile water basin. *Bull. Natl. Res. Centre* **48**, 30 (2024).
8. Akhtar, N. *et al*. Modification of the water quality index (wqi) process for simple calculation using the multi-criteria decision-making (mcdm) method: a review. *Water* **13**, 905 (2021).
9. Deng, S., Li, C., Jiang, X., Zhao, T. & Huang, H. Research on surface water quality assessment and its driving factors: A case study in taizhou city, china. *Water* **15**, 26 (2022).
10. Hegarty, S., Hayes, A., Regan, F., Bishop, I. & Clinton, R. Using citizen science to understand river water quality while filling data gaps to meet united nations sustainable development goal 6 objectives. *Sci. Total. Environ.* **783**, 146953 (2021).
11. Wang, M. *et al*. A triple increase in global river basins with water scarcity due to future pollution. *Nat. Commun.* **15**, 880 (2024).
12. Thorslund, J. & van Vliet, M. T. A global dataset of surface water and groundwater salinity measurements from 1980-2019. *Sci. Data* **7**, 231 (2020).
13. Schraga, T. S. & Cloern, J. E. Water quality measurements in san francisco bay by the us geological survey, 1969-2015. *Sci. Data* **4**, 1–14 (2017).
14. Mohanakavitha, T. *et al*. Dataset on the assessment of water quality of surface water in kalingarayan canal for heavy metal pollution, tamil nadu. *Data brief* **22**, 878–884 (2019).
15. Lehmann, M. K. *et al*. Gloria-a globally representative hyperspectral in situ dataset for optical sensing of water quality. *Sci. data* **10**, 100 (2023).
16. Krasovich, E. *et al*. Harmonized nitrogen and phosphorus concentrations in the mississippi/atchafalaya river basin from 1980 to 2018. *Sci. data* **9**, 524 (2022).
17. Divahar, R., Raj, P. A., Sangeetha, S., Mohanakavitha, T. & Meenambal, T. Dataset on the assessment of water quality of ground water in kalingarayan canal, erode district, tamil nadu, india. *Data brief* **32**, 106112 (2020).
18. Ejoh, A., Unuakpa, B., Ibadin, F. & Edeki, S. Dataset on the assessment of water quality and water quality index of ubogo and egini rivers, udu lga, delta state nigeria. *Data Brief* **19**, 1716–1726 (2018).
19. Lin, J. *et al*. Water quality dataset in china. *Earth Syst. Sci. Data Discuss.* **2023**, 1–14 (2023).
20. Virro, H., Amatulli, G., Kmoch, A., Shen, L. & Uuemaa, E. Grqa: Global river water quality archive. *Earth Syst. Sci. Data Discuss.* **2021**, 1–30 (2021).
21. Wan, W. *et al*. A lake data set for the tibetan plateau from the 1960s, 2005, and 2014. *Sci. Data* **3**, 1–13 (2016).
22. Potapova, M. G., Lee, S. S., Spaulding, S. A. & Schulte, N. O. A harmonized dataset of sediment diatoms from hundreds of lakes in the northeastern united states. *Sci. Data* **9**, 540 (2022).
23. Liu, S. *et al*. A database of water chemistry in eastern siberian rivers. *Sci. Data* **9**, 737 (2022).
24. Mantzouki, E. *et al*. A european multi lake survey dataset of environmental variables, phytoplankton pigments and cyanotoxins. *Sci. data* **5**, 1–13 (2018).
25. Hartmann, J., Lauerwald, R. & Moosdorf, N. A brief overview of the global river chemistry database, glorich. *Procedia Earth Planet. Sci.* **10**, 23–27 (2014).
26. Syeed, M. M. *et al*. An iot intensive ai-integrated system for optimized surface water quality profiling. *In 2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 247-252 (IEEE, 2023).
27. Karim, M. R. *et al*. A Comprehensive Surface Water Quality Monitoring Dataset (1940-2023): 2.82Million Record Resource for Empirical and ML-Based Research. *figshare. Dataset.* https://doi.org/10.6084/m9.figshare.27800394.v2. (2025).
28. Kumar, P., Matta, G. & Kumar, A. Harmonizing water quality: Integrating indices and chemo-metrics for sustainable management in the ramganga river watershed. *Anal. Chem. Lett.* **14**, 29–47 (2024).
29. WHO. Who guidelines for drinking-water quality. https://www.epa.gov/sites/default/files/2014-03/documents/guidelines_for_drinking_water_quality_3v.pdf [Accessed 18-Feb-2025]. (2008).
30. Environmental Protection Agency, U. 2018 edition of the drinking water standards and health advisories tables, epa, usa. https://www.epa.gov/system/files/documents/2022-01/dwtable2018.pdf [Accessed 18-Feb-2025] (2022).
31. Tukey, J. W. *et al*. Exploratory data analysis, vol. 2 (Reading, MA, 1977).
32. Heidarpour, B., Panjalizadeh Marseh, B., Ekramirad, A., Hosseinnezhad, A. & Ghasemian Langroudi, A. Detection of outlier in flood observations: A case study of tamer watershed. *Res. J. Recent Sci.* **2277**, 2502 (2015).
33. Miot, H. A. Anomalous values and missing data in clinical and experimental studies. *Jornal vascular brasileiro* **18**, e20190004 (2019).

Author contributions

Md. Rajaul Karim, Ashifur Rahman, and Khondkar Ayaz Rabbani constructed the dataset, M.M. Mahbubul Syeed and Kaniz Fatema designed the study and wrote the manuscript, Md Shakhawat Hossain and Razib Hayat Khan performed the dataset proofing and assisted manuscript writing, Mohammad Faisal Uddin managed the project and funding.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.M.M.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.