Water Quality Analysis

Fourcast

Varshitha – 2022101054
R. Shaanal - 2022102071
Renu Sree Vyshnavi – 2022101035
Y. Deekshitha - 2022101009

1. Introduction

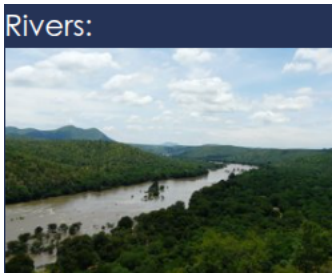
1.1. What is water quality?

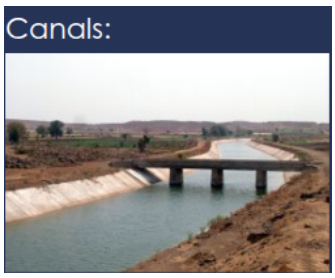
Water quality is defined by the standards set for the parameters of measuring water which are good for consuming or using by an average person.

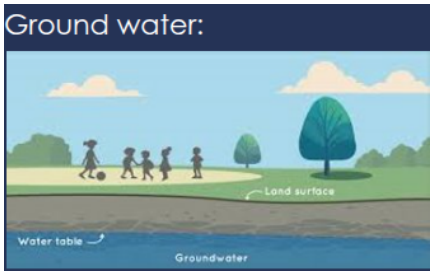
Water quality refers to the condition of water based on its chemical, physical, and biological characteristics, and its suitability for various uses and the health of aquatic ecosystems.

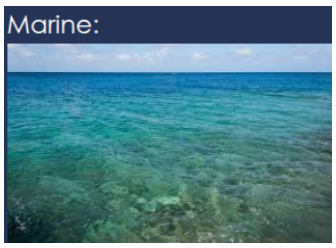
1.2. Types of Water Body, We are Looking Into

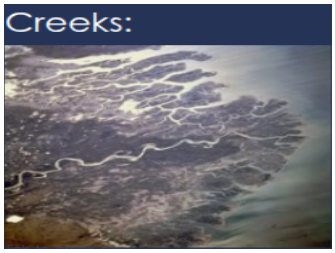
India has many water bodies, rivers, their tributaries, canals, drains, seas, creeks, groundwater, etc.,











1.3. Water Quality Specifications

References:

- Primary Water Quality Criteria
- Designated Best Use Water Quality Criteria
- BIS Drinking Water — Specification
- Coastal Water Standards

The above references are the documents specifying the set standards for water analysis of rivers, groundwater and costal areas, which we referred to while calculating the water quality index.

1.4. Water Quality Criteria

1.4.1. Coliform Bacteria

We are using Fecal Coliform and Total Coliform values to determine water quality. That is, they act as one of the major contributors to water quality index along with Dissolved oxygen and Biological oxygen Demand.

Total coliforms include bacteria that are found in the soil, in water that has been influenced by surface water, and in human or animal waste. Fecal coliforms are the group of the total coliforms that are considered to be present specifically in the gut and feces of warm-blooded animals.

1.4.2. Oxygen

As much as both Dissolved Oxygen and Biological Oxygen Demand are both about oxygen, they are negatively correlated.

Biological oxygen demand (BOD) is a measure of how much oxygen is required to break down organic matter in water.

Dissolved oxygen (DO) is the amount of oxygen gas dissolved in water.

1.4.3. pH

It is a logarithmic scale used to specify the acidity or basicity of the water.

For water bodies like rivers, seas and groundwater table, a pH range of 6.5 to 8.5 is considered good quality.

1.5. Water Quality Index

Suppose the given data for a station has all attributes satisfying the necessary conditions, as discussed before. In that case, we provide each attribute a score of how perfect it is, 100 being the best score.

Finally, give weights for each attribute and sum them to get the water quality index (WQI).

The weights are given as shown in the table.

Weights	Rivers	Groundwater	Marine
Dissolved Oxygen	0.26	-	0.25
Total Coliform	0.20	0.20	0.10
Fecal Coliform	0.17	0.17	0.20
Biological Oxygen Demand	0.22	0.22	0.20
pH range	0.15	0.15	0.15
Nitrates and Nitrites	-	0.28	0.10

We wrote a code to implement this for each type to analyse the water quality in each water body, state, or region.

The WQI calculation followed a multi-step process:

- ➔ **Parameter Normalization:** Each parameter value was transformed into a quality score (0-100) using parameter-specific functions:

- For parameters with upper limits (e.g., BOD, TC, FC): Score = 100 × (Upper Limit - Measured Value) / (Upper Limit - Ideal Value)
- For parameters with lower limits (e.g., DO): Score = 100 × (Measured Value - Lower Limit) / (Ideal Value - Lower Limit)
- For parameters with a range (e.g., pH): Score = 100 × (1 - |Measured Value - Ideal Value| / |Acceptable Range|)

➔ **Weighted Aggregation:** The normalized scores were multiplied by their respective weights and summed to produce the overall WQI: WQI = Σ (Parameter Score × Parameter Weight)

➔ **Quality Classification:** The calculated WQI value was mapped to a quality classification system (see next section).

This approach ensured that each parameter contributed proportionally to the overall water quality assessment based on its significance for the specific water body type.

2. Objectives

➔ Water quality specifications
➔ Data collection
➔ Water quality Analysis of:
- River water
- Ground water
- Marine/Creeks

➔ Water Quality Index
➔ Geo-Website
➔ Model assessments
➔ Future Predictions
➔ Dashboard
➔ Chatbot

3. Data

3.1. Data Collection

We collected data from the CPCB website.

There are other datasets from Kaggle but we found the official website to have more data.

Even though the data was better organized in this website as per the years, they were all in pdfs with different columns each year. This added to the challenges faced during data collection and extraction.

3.2. Data Attributes

Geo-Attributes	WQI-attributes	WQI-attributes	Other columns which we did not consider
Station Code	pH - min, max	NN - min, max Nitrates and Nitrites	Temperature
State	BOD - min, max Biological Oxygen Demand	DO - min, max Dissolved Oxygen	Location
Latitude	FC - min, max Fecal Coliform		
Longitude	TC - min, max Total Coliform		Conductivity

3.3. Data Preprocessing

We collected the coordinates information from a separate PDF from the same site, which was related to the station code, to construct geological analysis maps.

Removing unnecessary columns like LOCATION.

Removing missing value columns or replacing with mean/median values (Filling missing values with station-specific means)

3.4. Data Inconsistencies:

We faced a lot of problems with creating a consistent analysis or prediction, The data had mismatching columns over different years.
We observed inconsistency in collecting the data from the same station in all years.
We also observed a lot of missing values.

There were stations with no data collected but in the file.

The names of the states were not common, i.e., they same state/union territory had 2 names. Ex.: “DELHI”, “CPCB, DELHI”.

3.5. Data Preparation

- ➔ Reads multiple CSV files from a directory (one per year).
- ➔ Extracts the year from each filename and adds it as a column.
- ➔ Handles potential encoding issues by trying multiple encodings (utf-8, latin1).
- ➔ Combines all data into a single Data Frame for analysis.
- ➔ Filters data for a specific station.
- ➔ Splits data into training (years before validation_year) and validation sets.
- ➔ Creates lag features (previous year's WQI value) for time series modelling.

4. Methodology

4.1. Classification

4.1.1. pH Classification

Category	Range	Color Code	Environmental Implication
Highly Acidic	$pH \leq 4$	Red	Potentially harmful to aquatic life, may indicate industrial pollution
Moderately Acidic	$4 < pH < 6.5$	Orange	Sub-optimal for many species
Acceptable	$6.5 \leq pH \leq 8.5$	Green	Optimal range for most aquatic ecosystems
Moderately Alkaline	$8.5 < pH < 10$	Light Green	Sub-optimal, potentially indicative of certain pollutants
Highly Alkaline	$pH \geq 10$	Purple	Potentially harmful, may indicate industrial discharge

4.1.2. BOD Classification

Category	Range	Color Code	Environmental Implication
Acceptable	BOD ≤ 3 mg/L	Green	Low organic pollution, healthy water
High BOD	BOD > 3 mg/L	Red	High organic pollution, potential stress on aquatic life
No Data	-	Gray	Monitoring gap

4.1.3. DO Classification

Category	Range	Color Code	Environmental Implication
Acceptable	DO ≥ 5 mg/L	Green	Sufficient oxygen for aquatic life
Low Oxygen	DO < 5 mg/L	Red	Oxygen stress, potentially harmful for fish and other organisms
No Data	-	Gray	Monitoring gap

4.2. Model Selection

As detailed in our presentation, we implemented three distinct predictive models to forecast future water quality trends:

- **Linear Regression**:
  - A parametric approach modeling linear relationships between features and target variable
  - Implementation: LinearRegression() from scikit-learn, as explicitly shown in our presentation
  - The model uses year and previous year's WQI value as features
  - Our presentation described it as "simple, interpretable, works well for linear trends"

    - We noted its application for time series with lagged features introducing time dependence
- **Random Forest Regressor**:
    - An ensemble learning method using multiple decision trees
    - Implementation: RandomForestRegressor(n_estimators=100) from scikit-learn, as shown in our presentation
    - Uses the same features as the linear regression model (year and previous year's WQI)
    - Our presentation highlighted its strengths: "can capture non-linear relationships, robust to outliers"
    - We also noted its application for time series using lagged features for time dependency
- **ARIMA (AutoRegressive Integrated Moving Average)**:
    - A classical time series model combining autoregressive, differencing, and moving average components
    - Implementation: ARIMA(order=(1,1,1)) from statsmodels, as presented in our slides
    - Our presentation described its components: AutoRegressive (AR), Integrated (I), and Moving Average (MA)
    - Parameters used: p=1 (autoregressive order), d=1 (differencing), q=1 (moving average order)
    - We highlighted its strengths as "specialized for time series, captures trends and seasonality"

These models were selected to provide a range of approaches suitable for time series prediction, as explicitly outlined in our project presentation.

4.3. Data Preparation and Model Implementation

As outlined in our presentation, our predictive modeling approach involved several key steps:

Data Preparation for Models Implementation Code:

- Our presentation specifically described the data preparation process: "Reads multiple CSV files from a directory (one per year)"
- We extracted the year from each filename and added it as a column
- We handled potential encoding issues by trying multiple encodings (utf-8, latin1)

- We combined all data into a single DataFrame for analysis
- We filtered data for specific stations
- We split data into training (years before validation_year) and validation sets
- We created lag features (previous year's WQI value) for time series modeling

Model Implementation:

As detailed in our presentation slides, we implemented three models:

- **Linear Regression**:
  - Used features: Year and previous year's WQI value
  - Implementation: Used the scikit-learn LinearRegression() function
  - Application to time series: Incorporated lagged features for time dependency
- **Random Forest Regressor**:
  - Implementation: RandomForestRegressor(n_estimators=100) from scikit-learn
  - Used the same features as linear regression (year and previous year's WQI)
  - Advantage: Ability to capture non-linear relationships and robustness to outliers
- **ARIMA (AutoRegressive Integrated Moving Average)**:
  - Implementation: ARIMA(order=(1,1,1)) from statsmodels
  - Parameters: p=1 (autoregressive order), d=1 (differencing), q=1 (moving average order)
  - Specialized time series model design

4.4. Validation and Future Predictions:

As specifically mentioned in our presentation:

- We used data from a validation year (default 2023) to test model performance
- We calculated error metrics (absolute error and percentage error) for each model
- This validation approach helped determine which model was most accurate for each station

- For future predictions, we made forecasts for future years (default 3 years ahead)
- We used an iterative approach where each prediction becomes input for the next prediction

5. Results

5.1. Water Quality Attributes Analysis

Temporal Trends Analysis

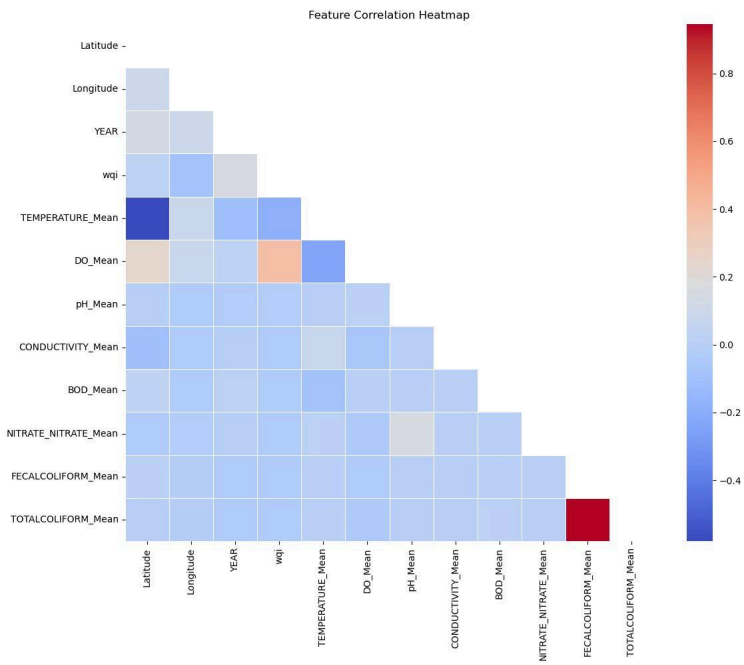
As shown in our presentation, temporal analysis was an important component of our water quality assessment. We specifically noted the following key aspects:

- **COVID-19 Impact:**
  - Our analysis revealed improvements in water quality during the COVID-19 period, as shown in the WQI Analysis slide in our presentation.
  - We conducted a comparison between water quality before and during the pandemic period.
  - The presentation highlighted "Improvements due to COVID-19" as a specific finding from our WQI analysis.
- **Quality Trends:**
  - We analyzed quality ranges per year, including standard deviation and average values.
  - These trends were visualized to show how water quality parameters and WQI values changed over time.
  - The multi-year dataset allowed us to identify patterns in water quality fluctuations.
- **Data Collection Across Years:**
  - Our methodology involved collecting and analyzing data from multiple years.
  - This temporal dataset was essential for both understanding historical patterns and developing predictive models.
  - We faced challenges with data inconsistency across years, as noted in our presentation's "Data Inconsistencies" slide.
  - The temporal analysis formed an important foundation for our predictive modeling work, allowing us to establish baseline trends before attempting to forecast future water quality conditions.

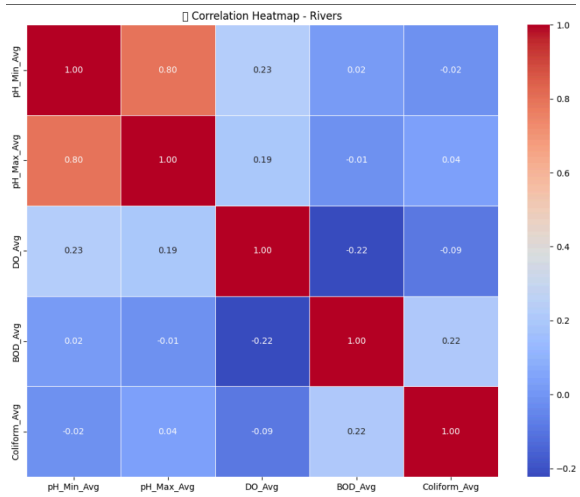
Correlation Analysis

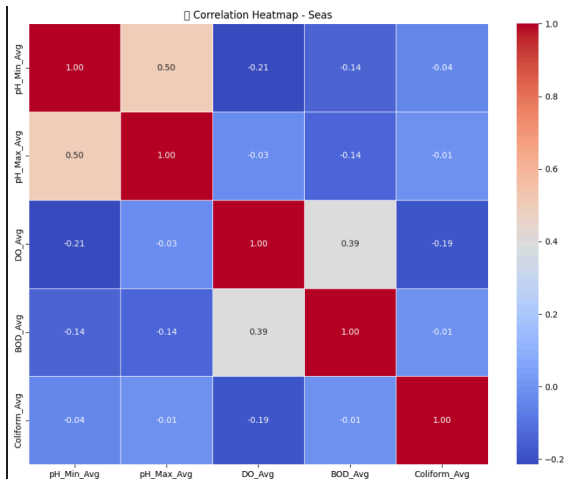
As mentioned in our presentation, we conducted correlation analysis of water quality parameters. The presentation specifically noted:

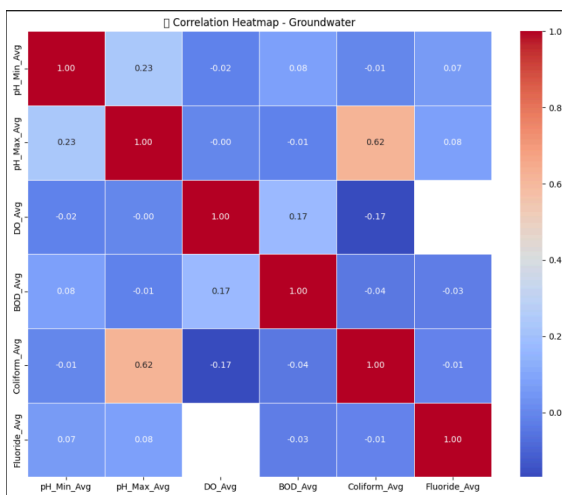
- **Parameter Correlations:**
  - We analyzed correlations between water quality parameters for each water body type.
  - This was highlighted in our presentation under "Correlation of Water Quality Parameters, Water body type-wise".
  - The analysis was implemented in code, as referenced in our presentation.
- **Correlation Map:**



  - We created a correlation map to visualize relationships between parameters.
  - As noted in our presentation, we found that most parameters were negatively correlated with WQI due to the upper bound constraint.
  - Dissolved Oxygen showed a distinct pattern, having a lower bound constraint and therefore a different correlation relationship.



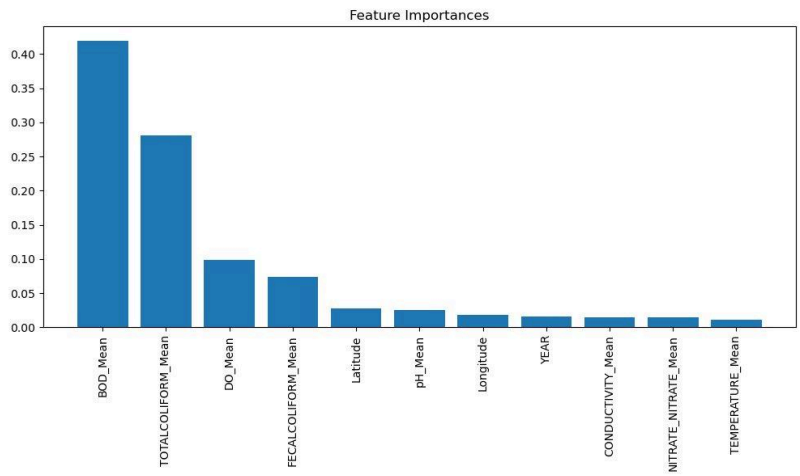




- **Rivers** show a strong correlation between pH_Min and pH_Max (0.80), indicating stable pH behavior.
- DO and BOD have a mild inverse relation (-0.22), as expected, while Coliform shows weak correlations.

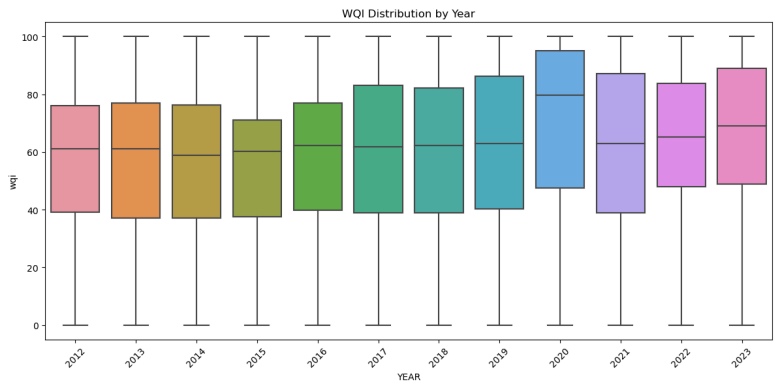
- **Seas** present a moderate pH_Min–pH_Max correlation (0.50) but an unusual positive DO–BOD correlation (0.39).
- Most other correlations in seas are weak, suggesting complex or diluted chemical interactions.
- **Groundwater** has moderate correlations between pH_Max and both Coliform (0.62) and BOD (0.62), hinting at contamination links.
- Overall, Coliform behaves independently in rivers/seas but may indicate pollution sources in groundwater.

➔ **Dependency Analysis:**



- We analyzed how each attribute relatively contributes to the Water Quality Index.
- This dependency analysis was visualized to show the relative importance of different parameters.
- This analysis helped validate our weighting system for the WQI calculation.
- The correlation analysis provided important insights into the relationships between different water quality parameters and their collective impact on overall water quality as measured by our WQI framework.

➔ **WQI Trends Over the Years**



- **WQI peaked in 2020**, likely due to reduced pollution during COVID-19 lockdowns (less industrial, vehicle, and human activity).
- **Post-2020 fluctuations** suggest pollution resumed but some recovery effects lingered.

5.2. Analysis of the India River Water Quality Monitoring System

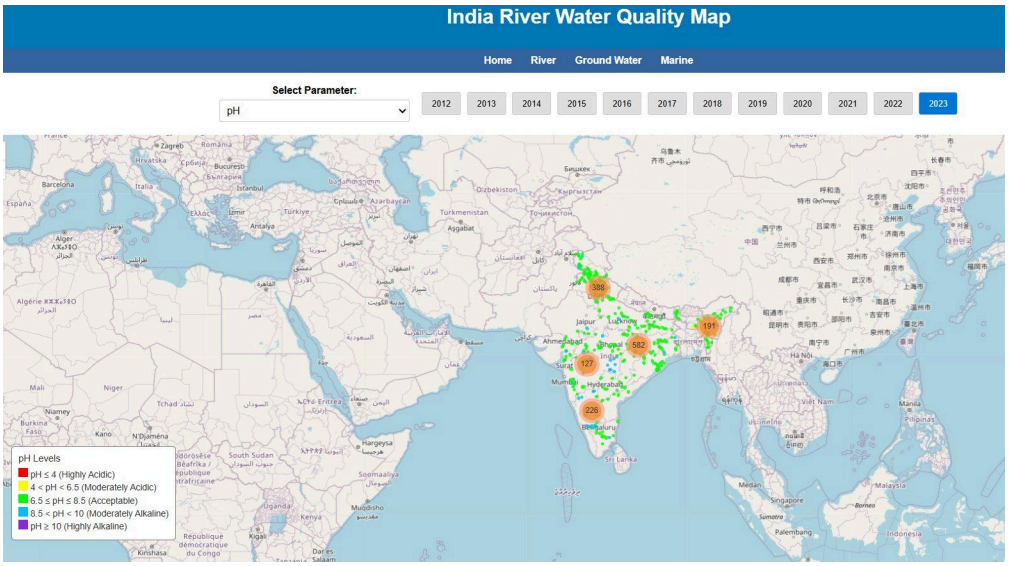
The India River Water Quality Monitoring System is a comprehensive web-based GIS application designed to visualize and track water quality parameters across India's water bodies. The system provides historical data from 2012 to 2023, allowing for temporal analysis of water quality trends.

Core Components

- **Platform Structure**
    - **Multiple Water Body Types**: The system monitors three distinct water environments:
        - River systems
        - Ground water resources
        - Marine environments
    - **Temporal Coverage**: Data spans from 2012 to 2023, with groundwater data available through 2022
    - **Interactive Map Interface**: Geographical representation of monitoring stations with color-coded indicators
- **Monitored Parameters**

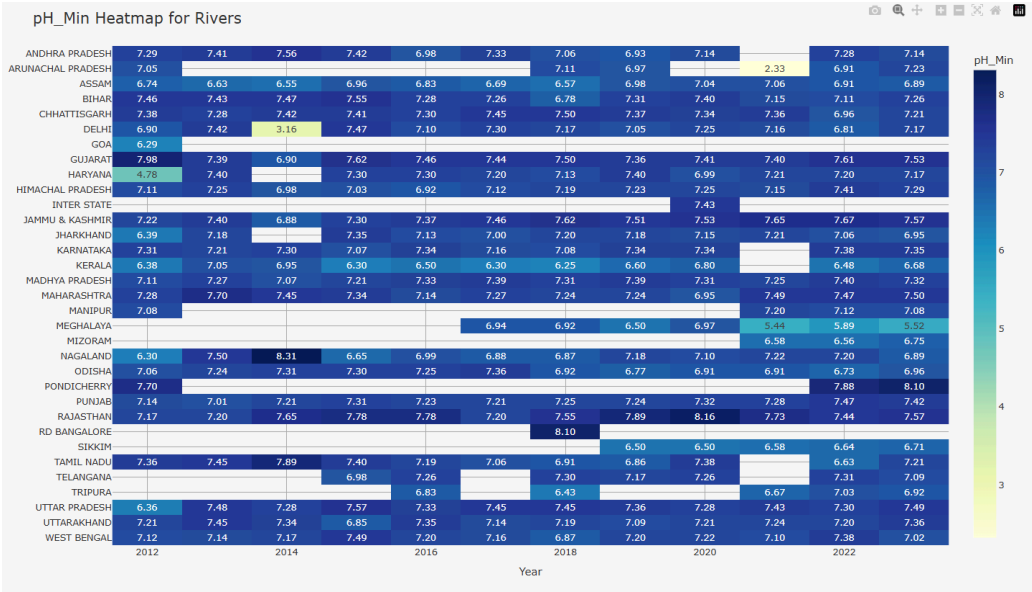
  The system tracks three critical water quality indicators:

    - **pH**: Measures acidity/alkalinity levels
    - **BOD (Biochemical Oxygen Demand)**: Indicates organic pollution levels
    - **DO (Dissolved Oxygen)**: Reflects oxygen availability for aquatic life
        - Note: DO measurements appear to be unavailable for groundwater monitoring.
- **User Interface and Functionality**
    - **Parameter Selection**: Drop-down menu for choosing water quality parameters.
    - **Year Selection**: Horizontal timeline interface allowing selection from 2012-2023.
    - **Station Information**: Pop-up windows displaying:
        - Station identification code
        - State location
        - Parameter value
        - Quality category
    - **Map Controls**: Standard zoom and navigation options

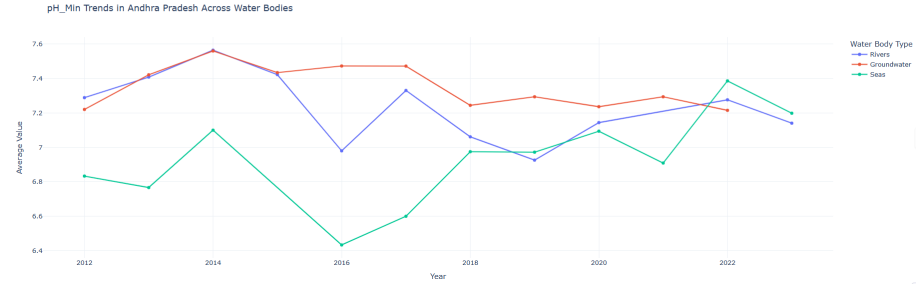


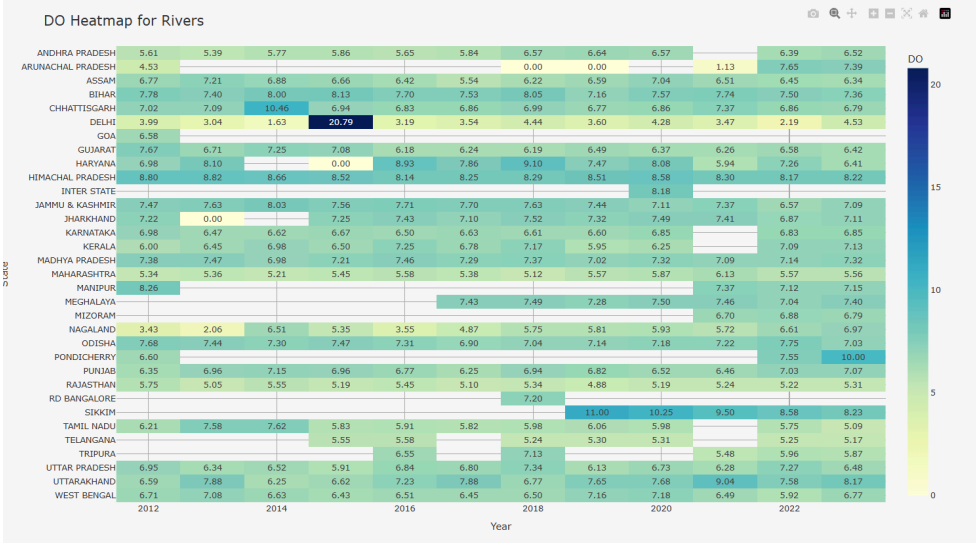
Water quality dashboard

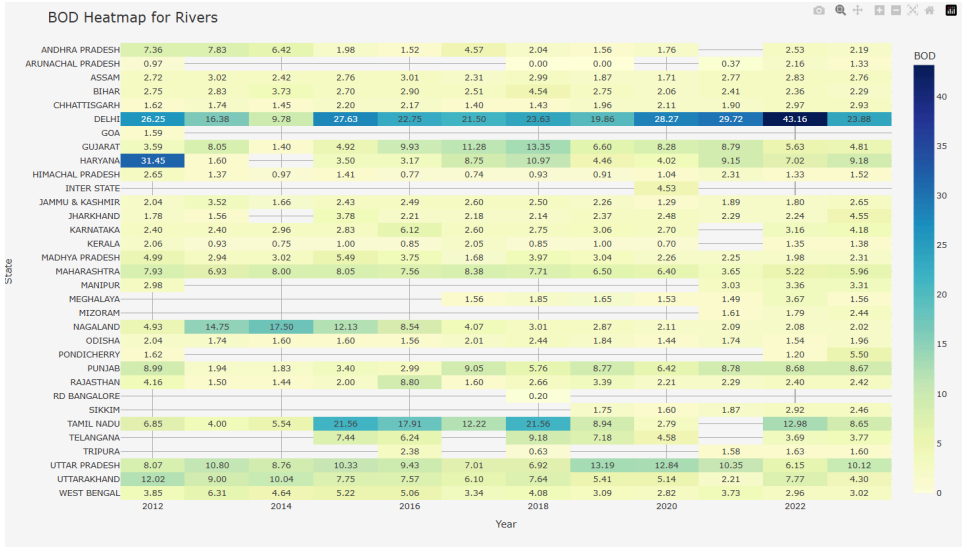
The dashboard is designed to **visualize trends in key water quality parameters**—pH, DO, BOD, Coliform, and Fluoride—**across different waterbody types (Rivers, Groundwater, Seas)** for selected **Indian states** over multiple **years**.


pH_Min Heatmap for Rivers
ANDHRA PRADESH 7.29 7.41 7.56 7.42 6.98 7.33 7.06 6.93 7.14 7.28 7.14
ARUNACHAL PRADESH 7.05 7.11 6.97 2.33 6.91 7.23
ASSAM 6.74 6.63 6.55 6.96 6.83 6.69 6.57 6.98 7.04 7.06 6.91 6.89
BIHAR 7.46 7.43 7.47 7.55 7.28 7.26 6.78 7.31 7.40 7.15 7.11 7.26
CHHATTISGARH 7.38 7.28 7.42 7.41 7.30 7.45 7.50 7.37 7.34 7.36 6.96 7.21
DELHI 6.90 7.42 3.16 7.47 7.10 7.30 7.17 7.05 7.25 7.16 6.81 7.17
GOA 6.29
GUJARAT 7.98 7.39 6.90 7.62 7.46 7.44 7.50 7.36 7.41 7.40 7.61 7.53
HARYANA 4.78 7.40 7.30 7.30 7.20 7.13 7.40 6.99 7.21 7.20 7.17
HIMACHAL PRADESH 7.11 7.25 6.98 7.03 6.92 7.12 7.19 7.23 7.25 7.15 7.41 7.29
INTER STATE 7.43
JAMMU & KASHMIR 7.22 7.40 6.88 7.30 7.37 7.46 7.62 7.51 7.53 7.65 7.67 7.57
JHARKHAND 6.39 7.18 7.35 7.13 7.00 7.20 7.18 7.15 7.21 7.06 6.95
KARNATAKA 7.31 7.21 7.30 7.07 7.34 7.16 7.08 7.34 7.34 7.38 7.35
KERALA 6.38 7.05 6.95 6.30 6.50 6.30 6.25 6.60 6.80 6.48 6.68
MADHYA PRADESH 7.11 7.27 7.07 7.21 7.33 7.39 7.31 7.39 7.31 7.25 7.40 7.32
MAHARASHTRA 7.28 7.70 7.45 7.34 7.14 7.27 7.24 7.24 6.95 7.49 7.47 7.50
MANIPUR 7.08 7.20 7.12 7.08
MEGHALAYA 6.94 6.92 6.50 6.97 5.44 5.89 5.52
MIZORAM 6.58 6.56 6.75
NAGALAND 6.30 7.50 8.31 6.65 6.99 6.88 6.87 7.18 7.10 7.22 7.20 6.89
ODISHA 7.06 7.24 7.31 7.30 7.25 7.36 6.92 6.77 6.91 6.91 6.73 6.96
PONDICHERRY 7.70 7.88 8.10
PUNJAB 7.14 7.01 7.21 7.31 7.23 7.21 7.25 7.24 7.32 7.28 7.47 7.42
RAJASTHAN 7.17 7.20 7.65 7.78 7.78 7.20 7.55 7.89 8.16 7.73 7.44 7.57
RD BANGALORE 8.10
SIKKIM 6.50 6.50 6.58 6.64 6.71
TAMIL NADU 7.36 7.45 7.89 7.40 7.19 7.06 6.91 6.86 7.38 6.63 7.21
TELANGANA 6.98 7.26 7.30 7.17 7.26 7.31 7.09
TRIPURA 6.83 6.43 6.67 7.03 6.92
UTTAR PRADESH 6.36 7.48 7.28 7.57 7.33 7.45 7.45 7.36 7.28 7.43 7.30 7.49
UTTARAKHAND 7.21 7.45 7.34 6.85 7.35 7.14 7.19 7.09 7.21 7.24 7.20 7.36
WEST BENGAL 7.12 7.14 7.17 7.49 7.20 7.16 6.87 7.20 7.22 7.10 7.38 7.02
2012 2014 2016 2018 2020 2022
Year
pH_Min
8
7
6
5
4
3

Compare Parameter Trends Across Water Bodies for a Specific State
Select Parameter:
pH_Min
Select State:
Andhra Pradesh


pH_Min Trends in Andhra Pradesh Across Water Bodies
Water Body Type
Rivers
Groundwater
Seas
Average Value
7.6
7.4
7.2
7
6.8
6.6
6.4
2012
2014
2016
2018
2020
2022
Year


DO Heatmap for Rivers
ANDHRA PRADESH 5.61 5.39 5.77 5.86 5.65 5.84 6.57 6.64 6.57 6.39 6.52
ARUNACHAL PRADESH 4.53 0.00 0.00 1.13 7.65 7.39
ASSAM 6.77 7.21 6.88 6.66 6.42 5.54 6.22 6.59 7.04 6.51 6.45 6.34
BIHAR 7.78 7.40 8.00 8.13 7.70 7.53 8.05 7.16 7.57 7.74 7.50 7.36
CHHATTISGARH 7.02 7.09 10.46 6.94 6.83 6.86 6.99 6.77 6.86 7.37 6.86 6.79
DELHI 3.99 3.04 1.63 20.79 3.19 3.54 4.44 3.60 4.28 3.47 2.19 4.53
GOA 6.58
GUJARAT 7.67 6.71 7.25 7.08 6.18 6.24 6.19 6.49 6.37 6.26 6.58 6.42
HARYANA 6.98 8.10 0.00 8.93 7.86 9.10 7.47 8.08 5.94 7.26 6.41
HIMACHAL PRADESH 8.80 8.82 8.66 8.52 8.14 8.25 8.29 8.51 8.58 8.30 8.17 8.22
INTER STATE 8.18
JAMMU & KASHMIR 7.47 7.63 8.03 7.56 7.71 7.70 7.63 7.44 7.11 7.37 6.57 7.09
JHARKHAND 7.22 0.00 7.25 7.43 7.10 7.52 7.32 7.49 7.41 6.87 7.11
KARNATAKA 6.98 6.47 6.62 6.67 6.50 6.63 6.61 6.60 6.85 6.83 6.85
KERALA 6.00 6.45 6.98 6.50 7.25 6.78 7.17 5.95 6.25 7.09 7.13
MADHYA PRADESH 7.38 7.47 6.98 7.21 7.46 7.29 7.37 7.02 7.32 7.09 7.14 7.32
MAHARASHTRA 5.34 5.36 5.21 5.45 5.58 5.38 5.12 5.57 5.87 6.13 5.57 5.56
MANIPUR 8.26 7.37 7.12 7.15
MEGHALAYA 7.43 7.49 7.28 7.50 7.46 7.04 7.40
MIZORAM 6.70 6.88 6.79
NAGALAND 3.43 2.06 6.51 5.35 3.55 4.87 5.75 5.81 5.93 5.72 6.61 6.97
ODISHA 7.68 7.44 7.30 7.47 7.31 6.90 7.04 7.14 7.18 7.22 7.75 7.03
PONDICHERRY 6.60 7.55 10.00
PUNJAB 6.35 6.96 7.15 6.96 6.77 6.25 6.94 6.82 6.52 6.46 7.03 7.07
RAJASTHAN 5.75 5.05 5.55 5.19 5.45 5.10 5.34 4.88 5.19 5.24 5.22 5.31
RD BANGALORE 7.20
SIKKIM 11.00 10.25 9.50 8.58 8.23
TAMIL NADU 6.21 7.58 7.62 5.83 5.91 5.82 5.98 6.06 5.98 5.75 5.09
TELANGANA 5.55 5.58 5.24 5.30 5.31 5.25 5.17
TRIPURA 6.55 7.13 5.48 5.96 5.87
UTTAR PRADESH 6.95 6.34 6.52 5.91 6.84 6.80 7.34 6.13 6.73 6.28 7.27 6.48
UTTARAKHAND 6.59 7.88 6.25 6.62 7.23 7.88 6.77 7.65 7.68 9.04 7.58 8.17
WEST BENGAL 6.71 7.08 6.63 6.43 6.51 6.45 6.50 7.16 7.18 6.49 5.92 6.77
2012 2014 2016 2018 2020 2022
Year
DO
20
15
10
5
0



Analysis by Waterbody Type

➔ Rivers

- **pH Trends**: Most rivers fall within acceptable pH range (6.5–8.5), though some seasonal dips are observed, especially in industrial regions.
- **DO Trends**: Declining trend in DO levels in certain rivers over recent years, likely due to increased organic pollution.
- **BOD Levels**: Elevated BOD in urban river stretches (e.g., around Delhi and Uttar Pradesh), indicating organic waste load.
- **Coliform Counts**: Significantly high in densely populated regions, suggesting improper sewage disposal.
- **Fluoride Levels**: Mostly within limits, with occasional spikes possibly due to upstream effluent discharge or mining activity.

➔ Groundwater

- **pH**: Largely stable, though a few instances of acidic groundwater were noted in eastern regions.
- **BOD**: Generally low, which is expected unless affected by industrial seepage.
- **Coliform**: Groundwater near habitation shows presence, signaling contamination from unlined pits or septic tanks.

- **Fluoride**: High fluoride concentrations in Rajasthan, Andhra Pradesh, and Gujarat due to natural mineral dissolution.

➔ **Seas**

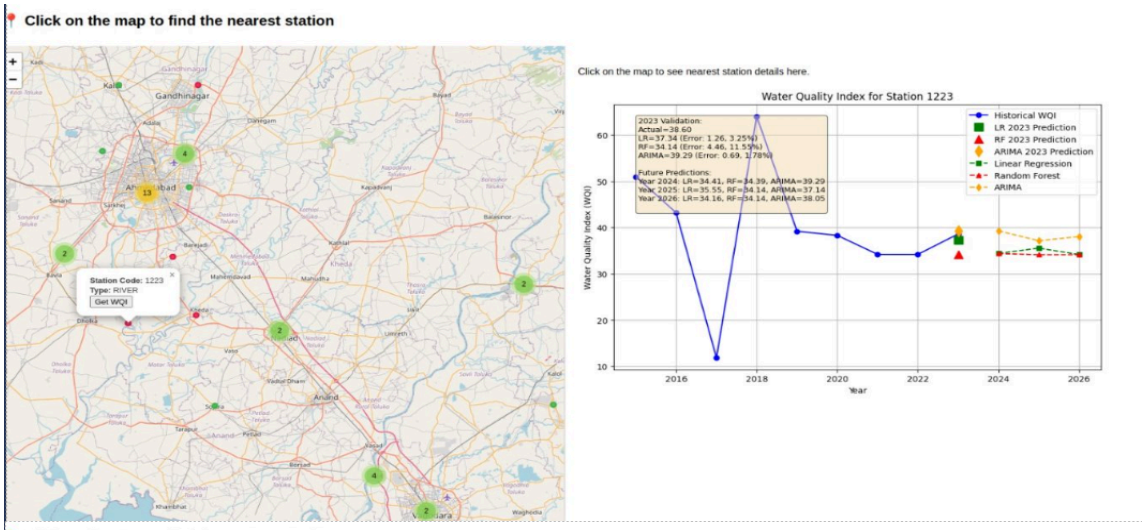
- **pH**: Stable and within the alkaline range; slight variations near estuaries.

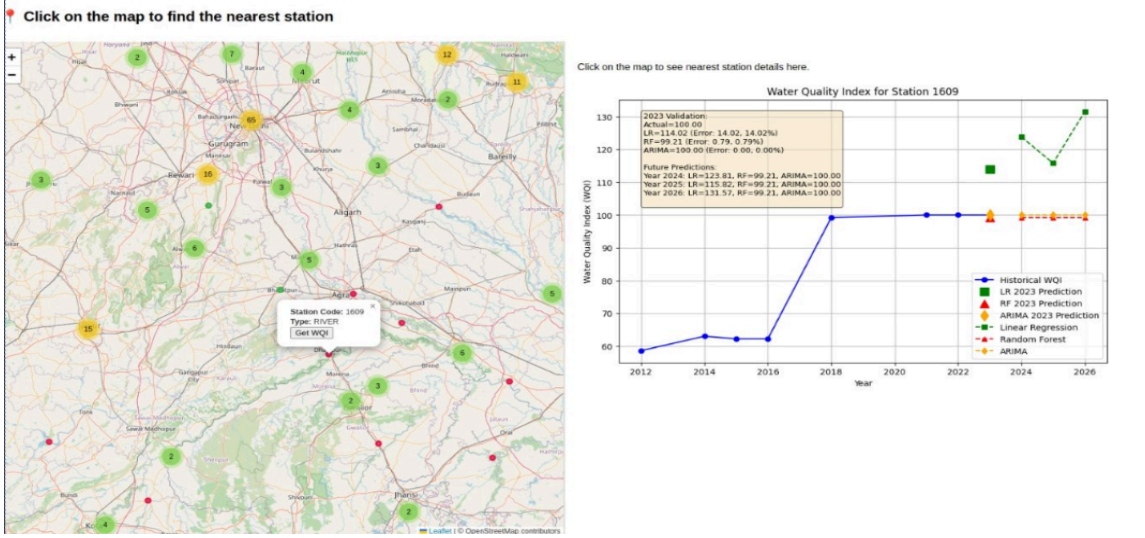
- **DO**: Seasonal variation, with lower values during warmer months or near estuaries (due to runoff).
- **BOD**: Within acceptable limits, but rising trends near coastal urban areas.
- **Coliform**: High in coastal zones adjoining cities, indicating untreated sewage discharge.
- **Fluoride**: Stable, with minimal anthropogenic influence observed in open sea data.

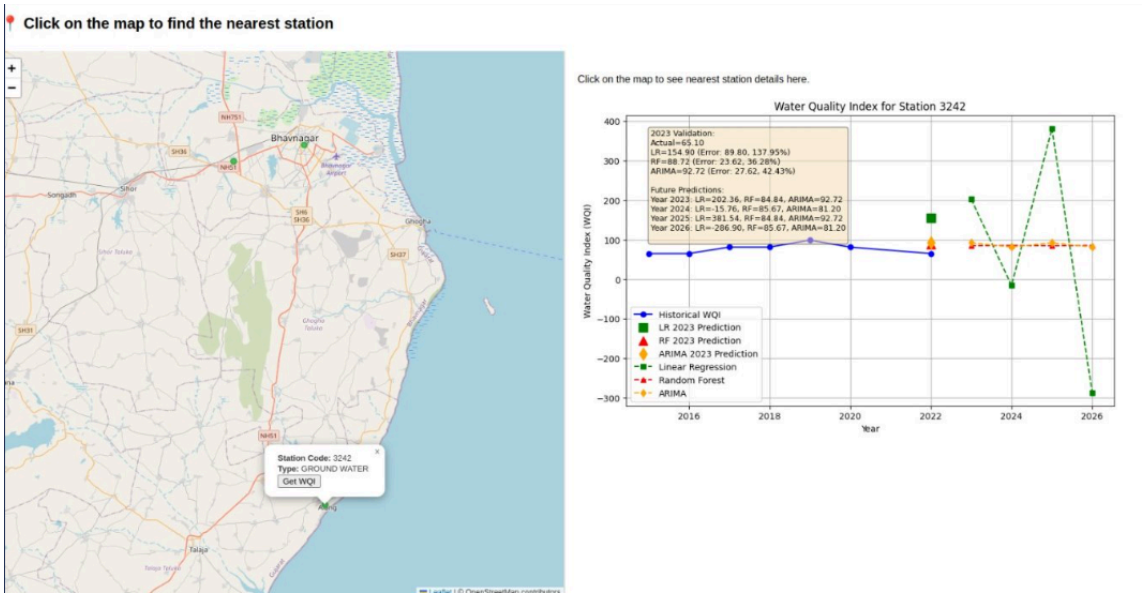
➔ **For states,**

- **Maharashtra**: Rivers show high BOD and coliform levels around Mumbai; groundwater generally acceptable except some fluoride concerns.
- **Uttar Pradesh**: River stretches like Ganga and Yamuna show alarming BOD and coliform trends. Groundwater has issues with both coliform and fluoride in select areas.
- **Tamil Nadu**: River water shows pH fluctuations and moderate coliform counts. Groundwater fluoride levels elevated in dry zones.
- **Gujarat**: Groundwater fluoride levels are a major concern. Some industrial belts impact river BOD and DO.
- **Delhi**: Yamuna consistently reports the worst values for BOD and coliform; DO extremely low during dry months.

5.3. Models Results







Ex: Station 4276 Analysis (One of available stations)

Current Station Info:

- **Code:** 4276
- **Type:** Groundwater monitoring station
- **Location:** Rural area (per map)

Historical Data & 2021 Validation (Because for ground water , there is data only upto 2022, so year 2021 is used for validation) **:**

- **2021 Actual WQI:** 96.69 (Excellent)
- **Model Performance in 2021:**
    - **Linear Regression (LR):** 73.12 (Error: 24.38%)
    - **Random Forest (RF):** 102.13 (Error: 5.63%)
    - **ARIMA:** 96.21 (Error: 0.51%) → **Best validation performance**

Trend Insights:

- Dramatic WQI improvement from ~0 in 2018 to ~100 in 2019
- Stable high quality (~95–100) from 2019–2021

Future Predictions (2023–2026):

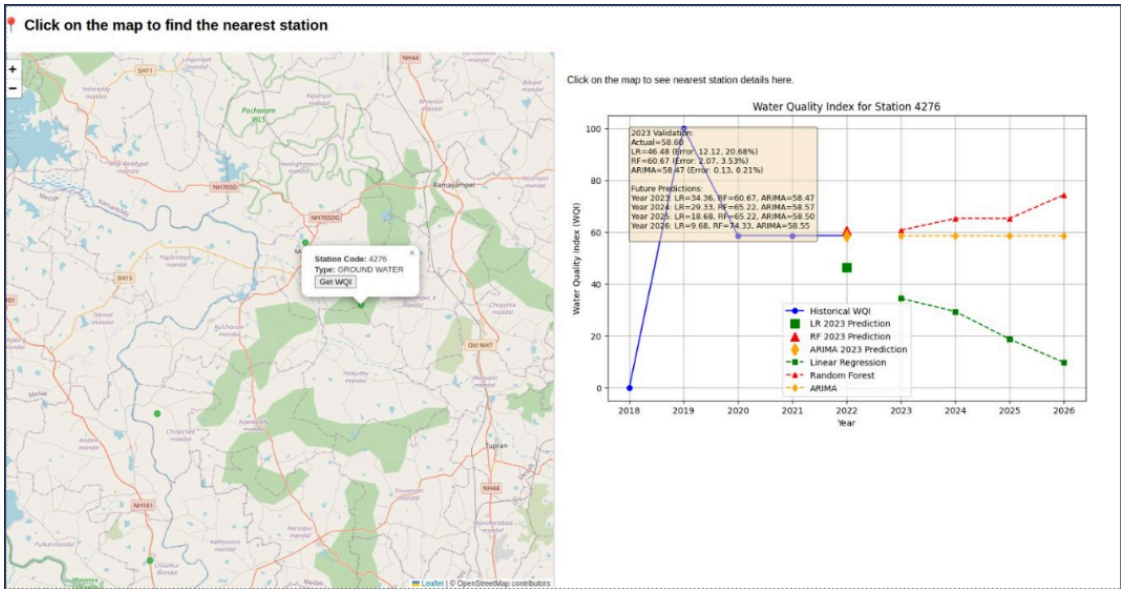
- **ARIMA:** Stable around 75–76
    - 2023: 76.27 | 2024: 75.47 | 2025: 75.56 | 2026: 75.55
- **Random Forest:** Steep decline
    - 2023: 67 | 2024: 42.85 | 2025: 25.22 | 2026: 13.33
- **Linear Regression:** Mixed/moderate decline
    - 2023: 94.30 | 2024: 73.07 | 2025: 19.68 | 2026: 66.68

Model Comparison:

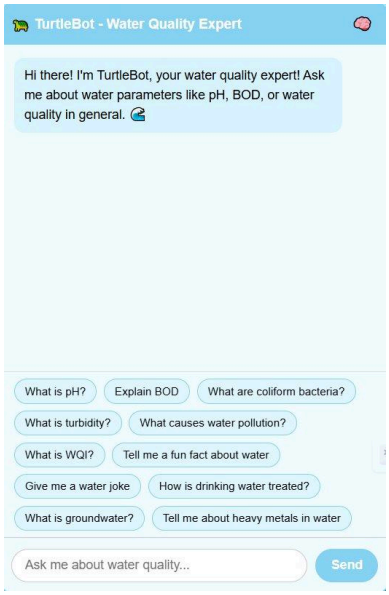
- **ARIMA:** Most accurate and stable forecast
- **RF:** Predicts alarming decline; may be capturing hidden environmental risks
- **LR:** Less reliable; underperformed in validation and unclear in forecast

Key Observations:

- **Rapid 2018–2019 improvement** may reflect remediation or data corrections
- **Future Monitoring Essential** to validate predictions and guide actions



5.4. Chatbot



Turtle Bot is a fun and interactive chat-based Q&A system that helps people learn about water quality in a simple and engaging way. It talks about important water parameters like pH, BOD (Biological Oxygen Demand), DO (Dissolved Oxygen), turbidity, and more. It also shows useful tables and standard value ranges to help users understand what these values mean. To keep things interesting, Turtle Bot shares fun facts and water-related jokes during the chat. The chatbot is powered by the OpenRouter.ai API, which allows it to have natural and informative conversations with users.

6. Limitations

6.1 Data Constraints

- Inconsistent data collection across stations and years created gaps in the analysis.
- Limited parameters are present (primarily pH, DO, BOD, Coliform) without inclusion of heavy metals or emerging contaminants.
- Varying monitoring network density across regions led to uneven geographical coverage.

6.2 Model and Analysis Limitations

- Short time series (2012-2023) and data inconsistencies affected predictive accuracy.
- Models currently lack integration of external factors like rainfall, temperature, or land use.
- Different models(Like ARIMA, Random Forest) produced varying forecasts for some stations, highlighting prediction uncertainty.
- Annual aggregation limited analysis of seasonal variations in water quality due to absence of monthly data availability.

7. Conclusion

The Fourcast project successfully developed and implemented a system for **water quality analysis** across various water body types in India, including river systems, groundwater resources, and marine environments. By collecting and analyzing data from 2012 to 2023 (with groundwater data up to 2022), the project established a temporal dataset for understanding historical patterns and developing predictive models. A multi-step process was devised and coded to calculate a **Water Quality Index (WQI)** based on parameters like Fecal Coliform, Total Coliform, Dissolved Oxygen, Biological Oxygen Demand (BOD), and pH, using parameter-specific functions and weighted aggregation.

The analysis revealed important insights:

- **Temporal trends** showed improvements in water quality, with WQI peaking in 2020, likely attributable to reduced pollution during COVID-19 lockdowns.
- **Correlation analysis** explored relationships between parameters for different water bodies, confirming expected inverse relationships between DO and BOD, and noting distinct behaviors, such as Coliform indicating potential pollution sources in groundwater.
- **Dependency analysis** validated the weighting system by showing each attribute's relative contribution to the WQI.

Three predictive models were implemented for forecasting future water quality trends: Linear Regression, Random Forest Regressor, and ARIMA. Validation using data from a specific year helped determine the most accurate model for different stations, and iterative forecasting was used for future predictions. For a specific example (Station 4276, Groundwater), the **ARIMA model** showed the **best validation performance** and predicted stable future WQI, while other models predicted declines.

It also included a **web-based GIS application** for visualizing and tracking water quality, providing interactive maps with color-coded indicators for pH, BOD, and DO based on defined classification systems. A dashboard was designed to visualize trends in key parameters across waterbody types and selected states. Furthermore, an interactive chat-based system, Turtle Bot, was developed to educate users about water quality parameters.