

Dissertation  
Title: Assessing surface water quality in Bristol, UK, using machine learning techniques in environmental analytics for sustainability and environmental impact assessment.

25th September, 2024

**Table Of Contents**

[**1. Introduction**](#_4lqp25cx7kth) **2**

[1.1 Research Problem and Question](#_g9meprvzw3p) 2

[1.2 Objectives of Study](#_d43odvls5trn) 3

[1.3 Significance and Motivation of Research](#_9ondjrgdcc84) 3

[**2.Literature Review:**](#_o9eddfuun7md) **4**

[**3. Methodology:**](#_531g7tgq087t) **6**

[3.1 Research Design and Context:](#_bf35m2auzvfy) 6

[3.2 Data Collection Methods:](#_dwly7vy7fso) 6

[3.3 Data Description and Rationale:](#_9mohep31laww) 6

[3.4 Data Analysis Procedures:](#_bfah5go382kn) 7

[**4.Results and Analysis:**](#_jwv1d39ptkx8) **10**

[4.1 Data Cleaning and Preprocessing:](#_6hvmkih96yxk) 10

[4.2 Exploratory Data Analysis:](#_gqgw1r71oxi3) 11

[4.3 Machine Learning Modelling:](#_url2argf8pie) 24

[e) Model Evaluation](#_x43dvayt6kf) 27

[4.4 Environmental Impact Assessment of Bristol's Surface Water Quality](#_k5jf9fy6gdf5) 30

[**5 Discussion:**](#_5aig73lac2iq) **32**

[5.1 Discussion of results in relation to the research objectives:](#_7d77ecd3vu12) 32

[5.2 Comparison with existing Literature:](#_q06gunmrz79a) 33

[5.3 Limitations:](#_rc4efc39vyx4) 33

[**6. Conclusion:**](#_mjrtcq91dlfa) **34**

[6.1 Summary of key findings](#_65xjdg5ahhdm) 34

[6.2Contribution to the field](#_llchz2cykebu) 35

[6.3 Recommendations for future research](#_p7n5kxetslt) 35

[**7.References:**](#_p0b80tr5xtbj) **36**

# Introduction

Water is essential for maintaining ecological integrity, public health, and sustainable development in urban contexts. Global urbanization, changing industrial practices, and climate change have all contributed to the degradation of surface water quality in cities during the past few decades (Vörösmarty et al., 2020). Bristol, a significant city in the southwest of England, is not exempted from these challenges.

As one of the UK's fastest-growing cities, Bristol faces significant water management issues. The city's waterways, including the River Avon and its tributaries, play a vital role in the local ecosystem and are integral to the city's identity and economy. However, these water bodies are subject to various pollutants from urban runoff, industrial discharges, and agricultural activities in the surrounding areas.

Water quality monitoring has historically depended on routine laboratory analysis and sampling, which can be labor- and resource-intensive. However, the introduction of big data and machine learning techniques has created new possibilities for more efficient and accurate water quality monitoring and prediction (Rahmani et al., 2021). These advanced analytical approaches can evaluate vast amounts of data from several sources, revealing patterns and relationships that might not be apparent through conventional analysis.

### 1.1 Research Problem and Question

Despite the importance of maintaining high water quality standards, there is a lack of comprehensive, data-driven approaches to assess and predict surface water quality in Bristol. The complex interplay of factors affecting water quality in urban environments makes it challenging to develop effective management strategies based solely on traditional monitoring methods.

This research aims to address this gap by leveraging machine learning techniques, specifically Multiple Linear Regression, to analyze and predict surface water quality in Bristol using the 2023 dataset from Bristol Open Data. The primary research question guiding this study is:

"How can machine learning techniques, particularly Multiple Linear Regression, be applied to assess and predict surface water quality in Bristol, and what insights can be gained to inform environmental impact assessments and guide sustainable water management practices?"

### 1.2 Objectives of Study

To address the research question, this study has the following objectives:

1. Analyze the current state of surface water quality in Bristol using the 2023 dataset from Bristol Open Data.
2. Identify and quantify the key factors influencing surface water quality through statistical analysis and machine learning techniques.
3. Develop and validate a Multiple Linear Regression model to predict specific water quality parameters based on other measured variables within the dataset.
4. Assess the environmental impact of current water quality levels by comparing them with established standards set by the UK Environment Agency.
5. Generate data-driven insights and recommendations for improving water management practices and reducing pollution in Bristol's surface waters based on the 2023 dataset analysis.

### 1.3 Significance and Motivation of Research

This research is motivated by the urgent need for innovative approaches to water quality management in urban environments. As cities like Bristol continue to grow and face increasing environmental pressures, it is crucial to develop more sophisticated tools for monitoring and predicting water quality.

The significance of this study lies in several key areas:

1. Advancement of Data-Driven Environmental Management: By applying machine learning techniques to publicly available water quality data, this research demonstrates the potential of open data initiatives and advanced analytics in environmental science. It contributes to the growing field of environmental data science and showcases how these methods can be applied in real-world urban settings.
2. Improved Decision-Making for Water Management: The insights gained from this study can inform more targeted and effective water management strategies in Bristol. By identifying key factors influencing water quality and developing predictive models, city planners and environmental agencies can prioritize interventions and allocate resources more efficiently.
3. Public Health and Ecosystem Protection: Improved understanding and prediction of water quality can lead to better protection of public health and aquatic ecosystems. Early identification of potential water quality issues can allow for more timely interventions, reducing the risk of waterborne diseases and ecological damage.
4. Contribution to Sustainable Urban Development: This research aligns with Bristol's ambitions to be a sustainable city and contributes to broader goals of sustainable urban development. By focusing on water quality, a critical aspect of urban sustainability, this study supports efforts to create more livable and environmentally responsible cities.
5. Methodological Innovation: The application of Multiple Linear Regression to water quality data in this specific urban context provides a methodological framework that can be adapted and applied to other cities facing similar challenges. This contributes to the broader field of urban water management and environmental impact assessment.

By addressing these critical aspects of urban water quality management through advanced data analysis, this research aims to contribute valuable insights to both the scientific community and local policymakers, ultimately supporting more sustainable and healthier urban environments.

# 2.Literature Review:

Recent years have witnessed significant advancements in urban water quality assessment, particularly with the integration of machine learning techniques. The literature in this field has expanded to cover various interconnected areas, including urban water management, environmental monitoring, and the application of artificial intelligence in environmental science. Urban water quality management has remained a critical focus, reflecting the growing challenges cities face in maintaining healthy aquatic ecosystems. Goonetilleke et al. (2019) provided a comprehensive review of stormwater management in urban areas, highlighting the potential for water reuse and the importance of integrated approaches to urban water management. Their work emphasized the need for innovative solutions to address water quality issues in rapidly growing urban environments.

In the context of Bristol and the UK, recent studies have examined local water quality issues. Wilson et al. (2024) analyzed microplastic pollution on beaches in the Bristol Channel, revealing significant variations across different depositional environments. This research, directly relevant to Bristol's water systems, highlights the need for comprehensive monitoring strategies. The application of machine learning techniques in water quality assessment has shown remarkable progress, offering new possibilities for analyzing such complex environmental phenomena.Shen (2022) provided a state-of-the-art review of machine learning applications in hydrology and water resources, including water quality prediction and assessment. They noted the potential of these techniques to improve the accuracy and efficiency of water quality monitoring and forecasting, while also highlighting challenges in data availability and model interpretability.

The concept of Nature-Based Solutions, as described by Oral et al. (2020), views urban water management through the lens of ecosystem-based approaches. It provides a framework for integrating natural processes into urban water systems to improve water quality and resilience. The authors present a comprehensive review of nature-based solutions for urban water management in European circular cities, highlighting their potential to address water quality issues while providing additional ecosystem services. They emphasize the importance of considering local context and stakeholder engagement in the successful implementation of these solutions.

Makropoulos et al. (2018) discuss the concept of Urban Water Resilience, which focuses on the ability of urban water systems to adapt to and recover from stresses and shocks. In their paper, the authors propose a resilience assessment method for urban water systems, emphasizing the importance of robust water quality management in ensuring the long-term sustainability of urban areas. They present a framework that considers various dimensions of resilience, including technical, organizational, and social aspects, providing a holistic approach to assessing and improving the resilience of urban water systems.

The potential of Deep Learning in Environmental Modeling is explored by Rahmani et al. (2021) in their study on stream temperature prediction. The authors demonstrate the exceptional performance of a deep learning model in capturing complex, non-linear relationships in environmental data. Their work highlights the ability of advanced machine learning techniques to improve the accuracy of water quality predictions, potentially revolutionizing environmental monitoring and management practices.

Despite advancements in more complex machine learning techniques, Multiple Linear Regression (MLR) remains relevant in water quality modeling. Ren et al. (2020) successfully applied MLR to predict various water quality parameters in urban rivers, demonstrating its continued utility in environmental studies. Their work compares the performance of MLR with neural network models, providing insights into the trade-offs between model complexity and interpretability in water quality prediction.

Despite recent advancements, several gaps and controversies remain in the literature. Many studies, including those by Li et al. (2021), have highlighted the challenges of obtaining high-quality, continuous water quality data in urban environments. This limitation can affect the accuracy and reliability of machine learning models. The issue of model interpretability, discussed by Roscher et al. (2020), is particularly relevant in environmental sciences where understanding causal relationships is crucial.

The integration of multiple data sources remains a challenge, with studies like Pu et al. (2021) exploring the use of remote sensing data in water quality assessment. However, there is still a need for more research on effectively integrating diverse data sources, including traditional sampling, sensor networks, and satellite observations. The impact of emerging contaminants, such as microplastics and pharmaceuticals, on urban water quality is an area of growing concern. Patel et al. (2019) highlight the limited understanding of how to effectively monitor and model these pollutants in urban water systems.

The theoretical framework for this study integrates several key concepts and approaches, including systems thinking, data-driven environmental management, Multiple Linear Regression modeling, environmental impact assessment, and open data and citizen science. By integrating these theoretical elements, this study aims to provide a comprehensive, data-driven analysis of surface water quality in Bristol, while also contributing to the broader field of urban water management and environmental data science.

# 3. Methodology:

### 3.1 Research Design and Context:

This study employs a quantitative, cross-sectional research design to assess surface water quality in Bristol, UK, using machine learning techniques. The research focuses on analyzing the 2023 surface water quality dataset from Bristol Open Data, providing a comprehensive examination of water quality parameters across various locations in Bristol throughout a full year. This approach allows for insights into spatial and temporal variations in water quality, contributing to a nuanced understanding of urban water management challenges (Goonetilleke et al., 2019).

### 3.2 Data Collection Methods:

The data collection method relies exclusively on secondary data obtained from the Bristol Open Data portal (<https://opendata.bristol.gov.uk/>). This dataset, provided by Bristol City Council as part of their open governance initiative, offers a comprehensive collection of water quality measurements for various water bodies within Bristol's city limits throughout the year 2023. The dataset encompasses a wide range of water quality indicators, including physical parameters (such as temperature), chemical parameters (including pH and dissolved oxygen), and biological parameters (like E. coli counts), along with other relevant environmental variables. The use of this publicly available dataset ensures transparency and reproducibility of the research, aligning with best practices in environmental data science (Li et al., 2021).

### 3.3 Data Description and Rationale:

The dataset used in this study contains a comprehensive set of columns that capture various aspects of surface water quality and environmental conditions across different locations in Bristol for the year 2023. The columns include **Date Time**, **Site ID**, **Total Coliforms (CFU)**, **E. coli (CFU)**, **Total Coliforms (MPN)**, **E. Coli (MPN)**, **Presumptive Enterococci (CFU)**, **Faecal Streptococci (CFU)**, **Temperature**, **pH**, **Conductivity**, **Dissolved Oxygen (mg/L)**, **Dissolved Oxygen (%)**, **Phosphate as P**, **Salinity**, **Clostridium Perfringens (CFU)**, **Ammonium as NH4**, **Turbidity**, **Ammonium as Nitrate**, **Nitrite**, **Nitrate**, **Faecal Coliforms**, **Salmonella**, **BOD (Biochemical Oxygen Demand)**, **COD (Chemical Oxygen Demand)**, **Suspended Solids**, **ID**, **Site Name**, **River**, and **Total Coliform**.

The rationale behind selecting specific columns for the analysis—**E. coli (CFU)**, **Total Coliforms (CFU)**, **Temperature**, **pH**, and **Dissolved Oxygen (mg/L)**—is based on their significant relevance to water quality assessment and public health. **E. coli (CFU)** and **Total Coliforms (CFU)** are critical indicators of fecal contamination in water bodies, which is directly associated with potential health risks. Monitoring these indicators helps in identifying pollution sources and ensuring compliance with water quality standards (Rahmani et al., 2021). **Temperature** is an essential parameter as it influences the metabolic rates of aquatic organisms and the solubility of gases, including oxygen, in water. Elevated temperatures can lead to reduced dissolved oxygen levels, affecting aquatic life (Shen, 2022). **pH** is another vital factor, as it can influence the toxicity of pollutants and the overall health of aquatic ecosystems. Extreme pH levels can be harmful to both aquatic organisms and the structural integrity of water bodies (Li et al., 2021). **Dissolved Oxygen (mg/L)** is crucial for the survival of aquatic organisms, and its levels are often used to gauge the health of a water body. Low dissolved oxygen levels can result in hypoxic conditions, leading to dead zones where aquatic life cannot survive (Makropoulos et al., 2018).

These selected columns not only provide insights into the physical and chemical conditions of the water but also help in developing a predictive model to assess and manage water quality effectively. The choice of these parameters aligns with best practices in water quality monitoring and is supported by literature that emphasizes their importance in maintaining the health of aquatic ecosystems and protecting public health (Oral et al., 2020). By focusing on these key indicators, the study aims to provide a robust analysis that can inform water quality management strategies in urban environments like Bristol (Ren et al., 2020).

### 3.4 Data Analysis Procedures:

Step 1 : Data Cleaning and Preprocessing

The initial phase of the analysis focused on preparing the 'Surface\_Water\_Quality.csv' dataset through a series of data cleaning and preprocessing steps. The datetime column was first converted to the appropriate format to enable accurate temporal analysis, followed by the elimination of duplicate entries to maintain data integrity. A comprehensive outlier detection and management process was implemented, particularly for Total Coliforms (CFU) and E. coli (CFU) measurements, guided by visual inspection techniques using boxplots. This process involved removing negative CFU counts, applying logarithmic transformation to both CFU columns, and implementing a threshold-based outlier removal process to exclude values above specified thresholds. Parameter-specific cleaning was then conducted, constraining pH values to the range of 2-14, limiting Dissolved Oxygen (DO) measurements to 0-20 mg/L, and capping temperature values between 0°C and 40°C, with values falling outside these ranges replaced by null values. To address missing data, median imputation was employed for pH, DO, and temperature. Categorical variables (siteid, sitename, and river) were converted into a numerical format using label encoding, and a normalization technique was applied to scale the numerical features. Importantly, the original values of specific parameters (ec\_cfu, tc\_cfu, temp, ph, do, and phosphate\_) were preserved alongside their transformed versions. The final step involved exporting the cleaned dataset and conducting a validation process through summary statistics and range checks on the preserved parameters, ensuring the integrity and reliability of the data for subsequent analysis.

Step 2: Exploratory Data Analysis

The Exploratory Data Analysis (EDA) phase focused on key water quality parameters including Escherichia coli (E. coli), Total Coliforms, temperature, pH, and dissolved oxygen. These parameters were selected for their critical role in assessing overall water quality and their established importance in regulatory frameworks. While E. coli and Total Coliforms indicate fecal contamination, temperature, pH, and dissolved oxygen provide crucial information about the physical and chemical conditions of the water bodies. This selection of parameters aligns with recent studies in urban water quality assessment, such as those conducted by Ren et al. (2020) which emphasize the importance of these indicators in understanding the health of urban aquatic ecosystems.

The EDA phase employed a multi-faceted approach to gain insights into the characteristics and patterns of Bristol's surface water quality. Temporal analysis using time series plots allowed for the identification of seasonal trends and long-term changes in water quality parameters. Spatial distribution analysis through scatter plots across geographical coordinates helped identify potential hotspots of pollution or areas with consistently high or low levels of specific parameters. The Spearman correlation matrix provided insights into the relationships between different water quality parameters, potentially revealing underlying processes affecting water quality. The distribution analysis and outlier detection, utilizing histograms, box plots, and Q-Q plots, allowed for a deeper understanding of the data's statistical properties and the identification of anomalous values that could indicate significant pollution events or measurement errors. This comprehensive EDA approach, as supported by recent literature (Pu et al., 2021; Wang et al., 2023), enables a thorough understanding of the complex dynamics of urban water quality, providing a solid foundation for subsequent predictive modeling and environmental impact assessment.

Step 3:Machine Learning Modelling:

The machine learning phase of this study focused on developing a predictive model for E. coli levels in Bristol's surface waters. E. coli was chosen as the target variable due to its significance as a key indicator of fecal contamination and its direct relevance to public health and water quality standards (Bartram & Ballance, 1996). The process began with feature selection through correlation analysis, identifying the most relevant predictors for E. coli (ec\_cfu) concentrations. The dataset was then split into training (80%) and testing (20%) subsets to ensure robust model validation. A Linear Regression model was developed using the scikit-learn library in Python, with 5-fold cross-validation employed on the training dataset to assess the model's performance and stability. This approach aligns with best practices in environmental modeling, as highlighted by Shen (2022) and Ren et al. (2020). The final model was trained using the entire training dataset and evaluated on the unseen test data, using R-squared and Root Mean Square Error (RMSE) as performance metrics. To contextualize the model's effectiveness, its performance was compared against a baseline model that simply predicted the mean value of E. coli for all instances. The baseline model serves as a simple benchmark, representing the performance of a model with no predictive power, and is a common practice in machine learning to demonstrate the added value of the developed model (Rahmani et al., 2021). This comprehensive modeling approach, grounded in recent literature on water quality prediction aimed to provide accurate predictions of E. coli levels while maintaining model interpretability, a crucial factor in environmental science applications.

Step 4: Environmental Impact Assesment:

The Environmental Impact Assessment (EIA) employed a three-step approach to evaluate the quality of Bristol's surface waters. This assessment focused on key parameters including E. coli, phosphate, pH, and dissolved oxygen, chosen for their critical importance in water quality management and their established regulatory standards in the UK. E. coli serves as an indicator of fecal contamination and potential pathogenic threats, phosphate levels are crucial for understanding nutrient dynamics and potential eutrophication risks, pH is fundamental for aquatic life support and chemical processes in water, and dissolved oxygen is essential for aerobic aquatic organisms (Goonetilleke et al., 2019; Oral et al., 2020). The approach began with a compliance analysis, evaluating the measured values of each parameter against UK environmental standards set by bodies such as the Environment Agency and UKTAG. This was followed by a temporal trend analysis, examining monthly compliance trends to identify seasonal patterns for each parameter, which is crucial for understanding the dynamic nature of water quality in urban environments (Langston et al., 2010). Finally, a parameter correlation analysis was conducted, generating a correlation matrix to understand relationships between water quality parameters, potentially revealing underlying processes affecting water quality. This comprehensive EIA approach, grounded in recent literature on urban water quality assessment (Pu et al., 2021), provides a robust framework for assessing the environmental impact of current water quality levels in Bristol's surface waters and informing sustainable water management practices.

# 4.Results and Analysis:

### 4.1 Data Cleaning and Preprocessing:

A critical step in ensuring the reliability and validity of the analysis was the thorough cleaning and preprocessing of the initial dataset. This section outlines the methodical approach taken to prepare the data for subsequent Exploratory Data Analysis (EDA), statistical modeling, and Environmental Impact Assessment.

Dataset Preparation The raw data, sourced from 'Surface\_Water\_Quality.csv', underwent a series of carefully considered transformations. The process began with the conversion of the datetime column to the appropriate format and the elimination of duplicate entries to maintain data integrity.

Outlier Detection and Management Visual inspection using boxplots for Total Coliforms (CFU) and E. coli (CFU) guided the outlier management strategy. This led to the following actions:

1. Removal of negative CFU counts, which were clearly erroneous.
2. Application of logarithmic transformation to both CFU columns. This step was crucial in addressing the wide value range and potential skewness in the data.
3. Implementation of threshold-based outlier removal, excluding values above 1,000,000 CFU/100mL for Total Coliforms and 100,000 CFU/100mL for E. coli.

Parameter-Specific Cleaning Custom functions were developed to clean key water quality parameters:

* pH values were constrained to the scientifically valid range of 2-14.
* Dissolved Oxygen (DO) was limited to 0-20 mg/L, reflecting realistic environmental conditions.
* Temperature values were capped at 40°C, with a lower bound of 0°C, to account for natural water body conditions. Values falling outside these ranges were replaced with null values to prevent skewing of the analysis.

Missing Data Imputation For pH, DO, and temperature, median imputation was employed to address missing values, ensuring a complete dataset without introducing extreme values.

Categorical Variable Encoding To facilitate numerical analysis, categorical variables (siteid, sitename, and river) were encoded using a label encoding technique, converting them into numerical format.

Feature Scaling and Data Preservation A normalization technique was applied to scale numerical features, enhancing the comparability of different parameters. Importantly, the original values of specific parameters (ec\_cfu, tc\_cfu, temp, ph, do, and phosphate\_) were preserved for several reasons:

1. To maintain their interpretability in the context of water quality standards.
2. To use them as target variables in the Exploratory Data Analysis and statistical modeling.
3. To ensure their suitability for the Environmental Impact Assessment, where original units and scales are often required.
4. Final Processing and Validation The cleaned and preprocessed dataset was exported to 'final\_cleaned\_surface\_water\_quality.csv'. Thorough validation was conducted through summary statistics and range checks on the preserved parameters, ensuring that the cleaning process effectively addressed data quality issues while maintaining the integrity of key variables.

### 4.2 Exploratory Data Analysis:

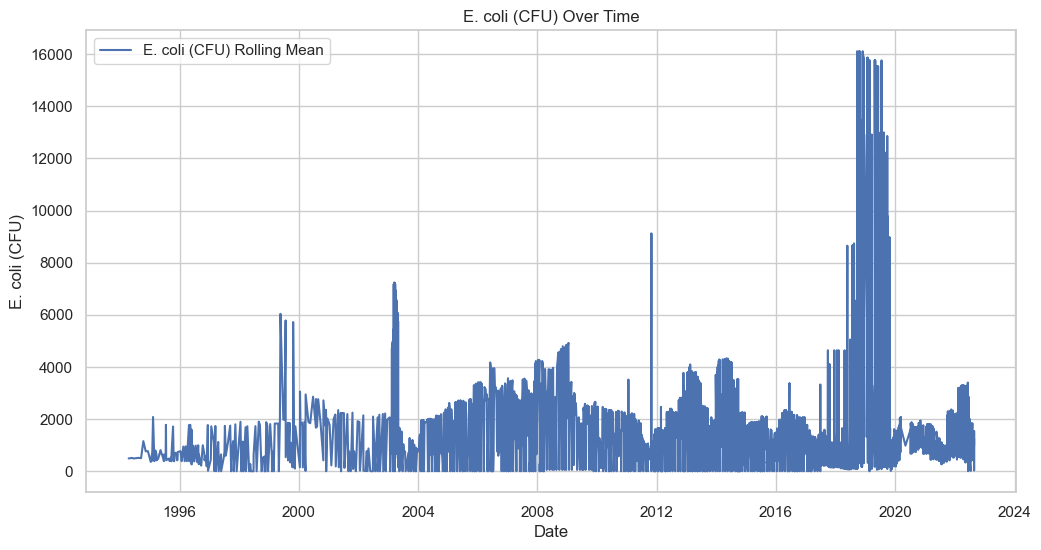
Exploratory Data Analysis (EDA) serves as a crucial initial step in understanding the underlying patterns and characteristics within the Bristol surface water quality dataset. This process involves a systematic examination of the data through temporal analysis, spatial distribution examination, correlation analysis, and distribution analysis, including outlier detection (Tukey, 1977). By employing this structured approach, we aim to derive meaningful insights that will guide subsequent stages of analysis, including machine learning modeling and environmental impact assessment (Goonetilleke et al., 2019).

4.2.1 Temporal Analysis:

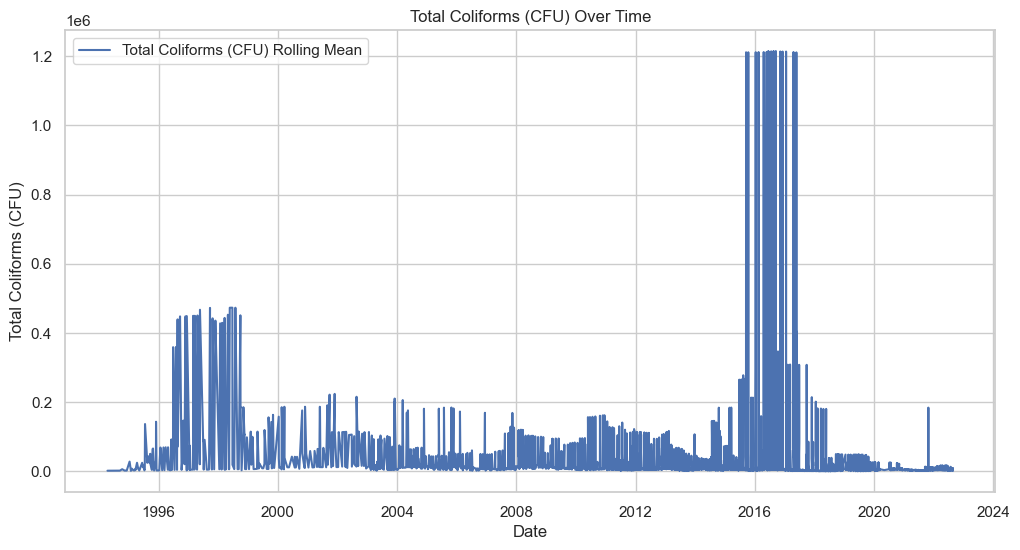
The temporal analysis aims to identify seasonal patterns, trends, and significant changes in water quality parameters over time. This analysis provides valuable insights into the fluctuations of water quality throughout the year and across multiple years, contributing to a comprehensive understanding of the dynamic nature of Bristol's surface water quality (Langston et al., 2010).

Time series plots were generated for key water quality parameters, including E. coli (CFU), Total Coliforms (CFU), temperature (°C), pH, and dissolved oxygen (mg/L). These visualizations allow for the identification of temporal patterns, anomalies, and potential relationships between different parameters.

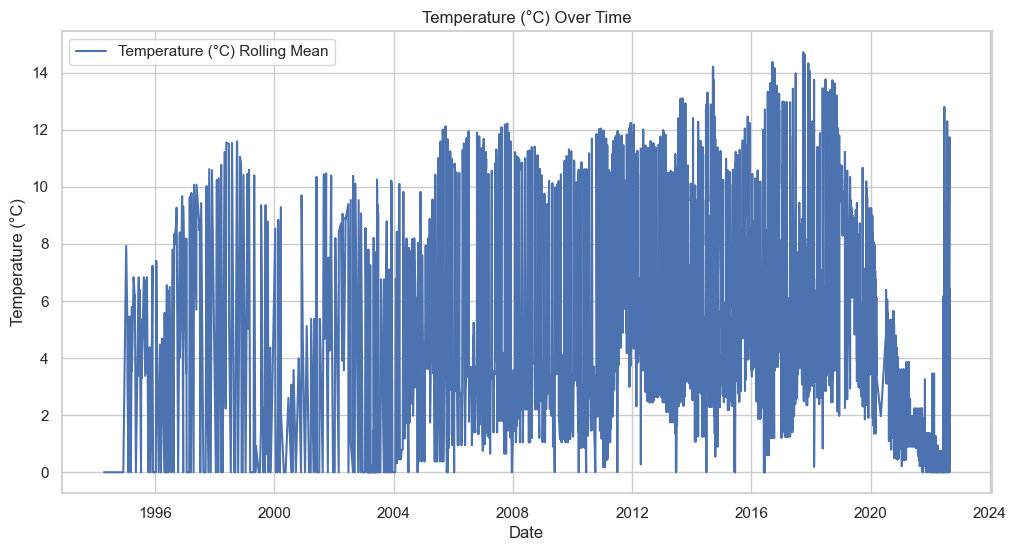
**Findings:**

****

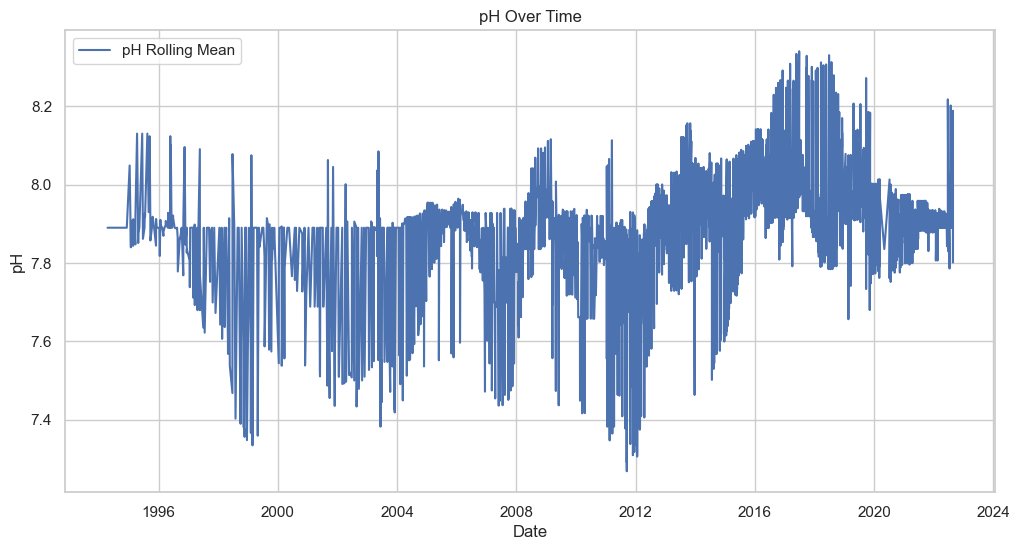
**E. coli (CFU) Over Time:** The E. coli levels show considerable variability over the years. Notable peaks are observed around 2000, 2004, and 2012. However, the most significant spike occurs between 2018 and 2020, where levels reach their highest point in the dataset, exceeding 16,000 CFU. This dramatic increase suggests a major contamination event or a significant change in environmental conditions during this period.

****

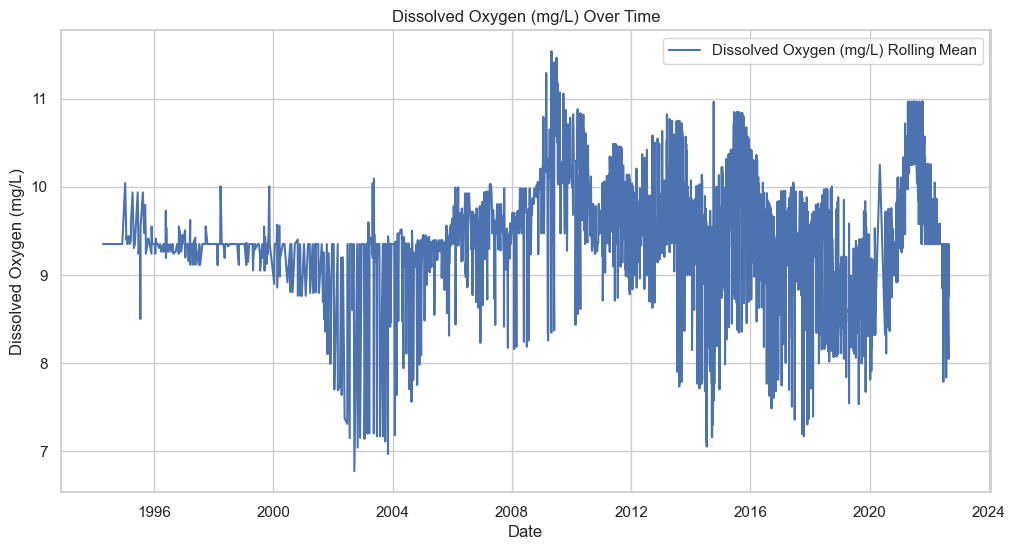
**Total Coliforms (CFU) Over Time:** Total Coliform levels exhibit high variability throughout the time series. A notable peak is observed in the late 1990s, with levels reaching approximately 0.4 x 10^6 CFU. However, the most striking feature is the extreme spike between 2016 and 2018, where levels surpass 1.2 x 10^6 CFU. This unprecedented increase aligns temporally with the E. coli spike, suggesting a possible common cause for both bacterial indicators.



**Temperature (°C) Over Time:**The temperature data displays clear seasonal patterns, with regular peaks and troughs corresponding to warmer and cooler months, respectively. The overall trend appears relatively stable across the years, with temperatures generally ranging between 0°C and 14°C. No significant long-term warming or cooling trend is immediately apparent from the data.



**pH Over Time:** The temperature data displays clear seasonal patterns, with regular peaks and troughs corresponding to warmer and cooler months, respectively. The overall trend appears relatively stable across the years, with temperatures generally ranging between 0°C and 14°C. No significant long-term warming or cooling trend is immediately apparent from the data.



**Dissolved Oxygen (mg/L) Over Time:** Dissolved Oxygen (DO) levels exhibit seasonal fluctuations throughout the time series. The data shows a range typically between 7 mg/L and 11 mg/L. There is no clear long-term trend of increase or decrease, but notable peaks are observed around 2008 and 2012, where DO levels reach their highest points in the dataset. Conversely, some periods of lower DO are also evident, particularly in the early 2000s.

In conclusion, the temporal analysis reveals significant variability in Bristol's surface water quality parameters over time. The extreme spikes in E. coli and Total Coliform levels between 2016 and 2020 are particularly noteworthy and warrant further investigation into potential contamination sources or environmental changes during this period. The relatively stable temperature patterns provide a consistent backdrop against which other parameters fluctuate. The slight increase and stabilization of pH levels in recent years could indicate changes in water chemistry or management practices. The dissolved oxygen levels, while variable, do not show a clear long-term trend, suggesting that factors influencing oxygen content in the water may be complex and multifaceted. These findings highlight the importance of continuous monitoring and the need for in-depth analysis of the factors driving these water quality dynamics in Bristol's surface waters.

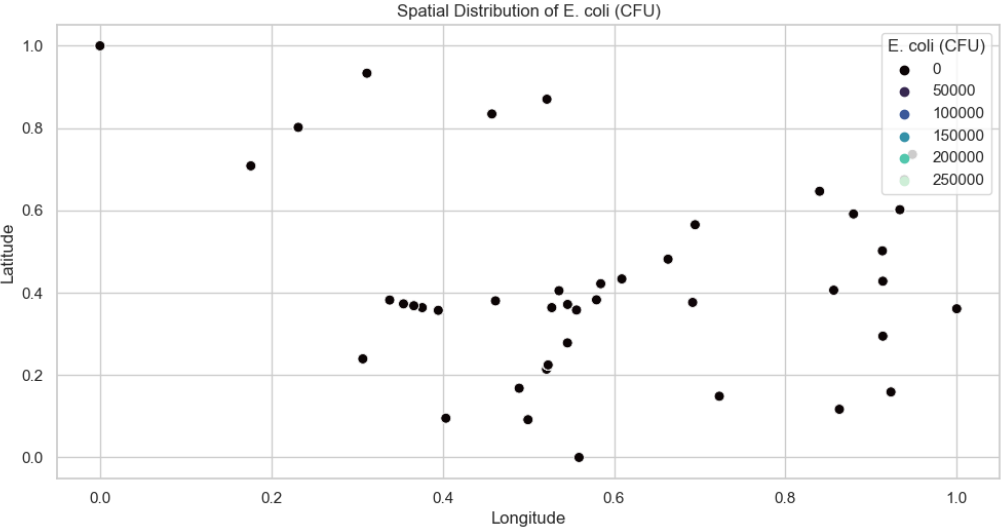
4.2.2 Spatial Distribution Analysis:

Objective: The spatial distribution analysis aims to elucidate the geographical variations of key water quality parameters across Bristol. By visualizing these parameters, we seek to identify areas with compromised water quality and potential pollution hotspots, thereby providing crucial insights for targeted interventions and evidence-based policy-making (Pu et al., 2021).

Methods: To examine spatial distribution, scatter plots were generated for key parameters including E. coli, total coliforms, and dissolved oxygen. Each point on these plots represents a distinct sampling location, with the corresponding parameter value indicated by color intensity.

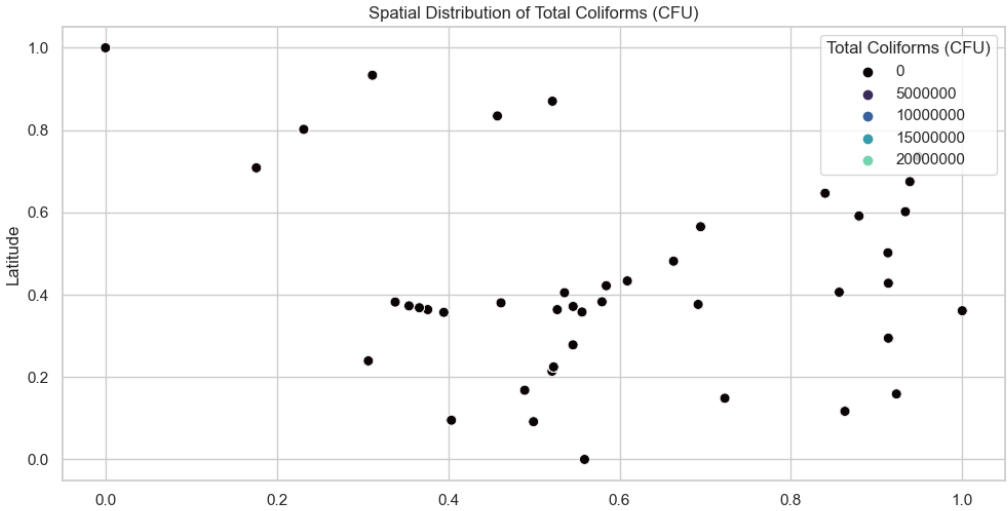
Findings:

a) Spatial Distribution of E. coli (CFU):



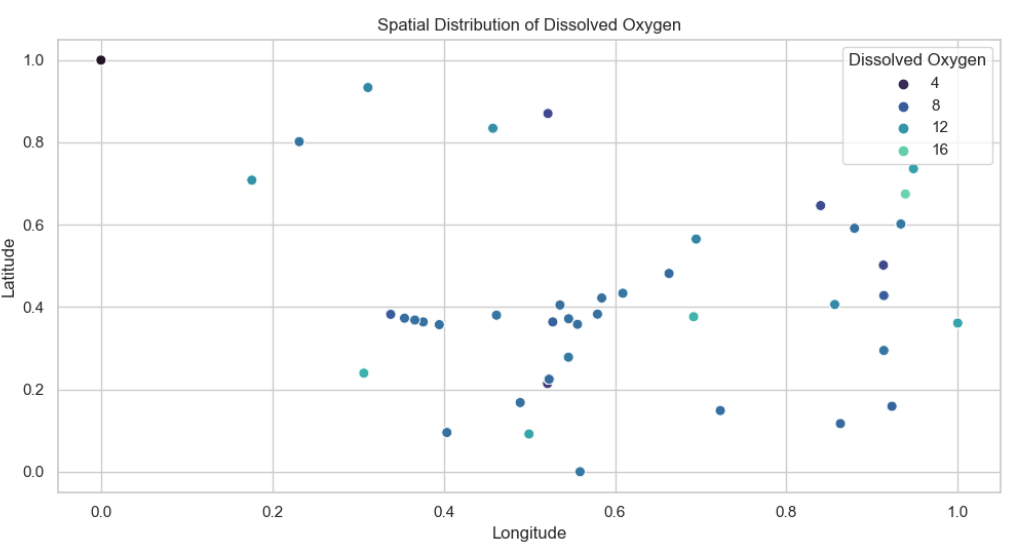
The E. coli distribution appears relatively uniform across the region, with most sampling points showing low concentrations (0 CFU). There are no clear hotspots or clusters of high E. coli levels visible in the plot. This suggests that fecal contamination, as indicated by E. coli, is not a widespread issue in the sampled area. The uniformity of low E. coli levels across different longitudes and latitudes indicates that water quality, in terms of this particular indicator, is consistently good throughout the region.

b) Spatial Distribution of Total Coliforms (CFU):



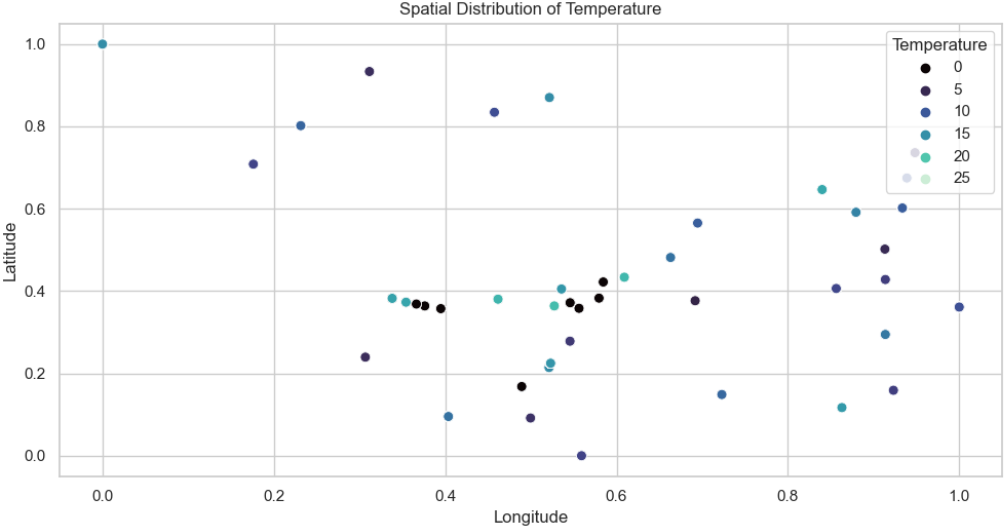
Similar to E. coli, the total coliform concentrations show a uniform distribution of low levels (0 CFU) across the sampled area. There is no evident spatial pattern or concentration of higher coliform levels in any particular region. This consistent low-level distribution suggests that the water bodies in the area generally maintain good microbial quality with respect to total coliforms. The similarity in distribution between E. coli and total coliforms reinforces the indication of overall good water quality in terms of these bacterial indicators.

c) Spatial Distribution of Dissolved Oxygen:



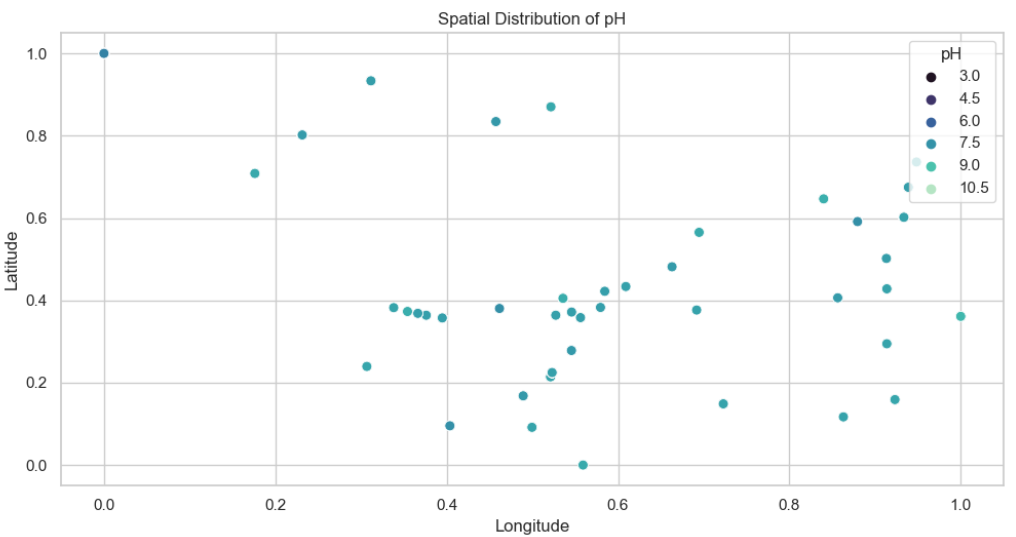
Dissolved oxygen levels show some variation across the region, with concentrations ranging from approximately 4 to 16 units (presumably mg/L). There is a slight tendency for higher dissolved oxygen levels (12-16 units) to be more prevalent in the northern part of the region (higher latitude values). Lower to moderate levels (4-8 units) are more common in the central and southern areas. This pattern suggests that water bodies in the northern part of the region may have better oxygenation, possibly due to factors such as higher water movement, lower temperatures, or less organic pollution.

d) Spatial Distribution of Temperature



Water temperature shows noticeable variation across the region, ranging from approximately 0 to 25 units (presumably °C). There is a general trend of higher temperatures (15-25 units) in the northern and eastern parts of the region (higher latitude and longitude values). Lower temperatures (0-10 units) are more common in the southern and western areas. This temperature gradient could be influenced by factors such as altitude, water depth, or proximity to urban heat islands.

e) Spatial Distribution of pH

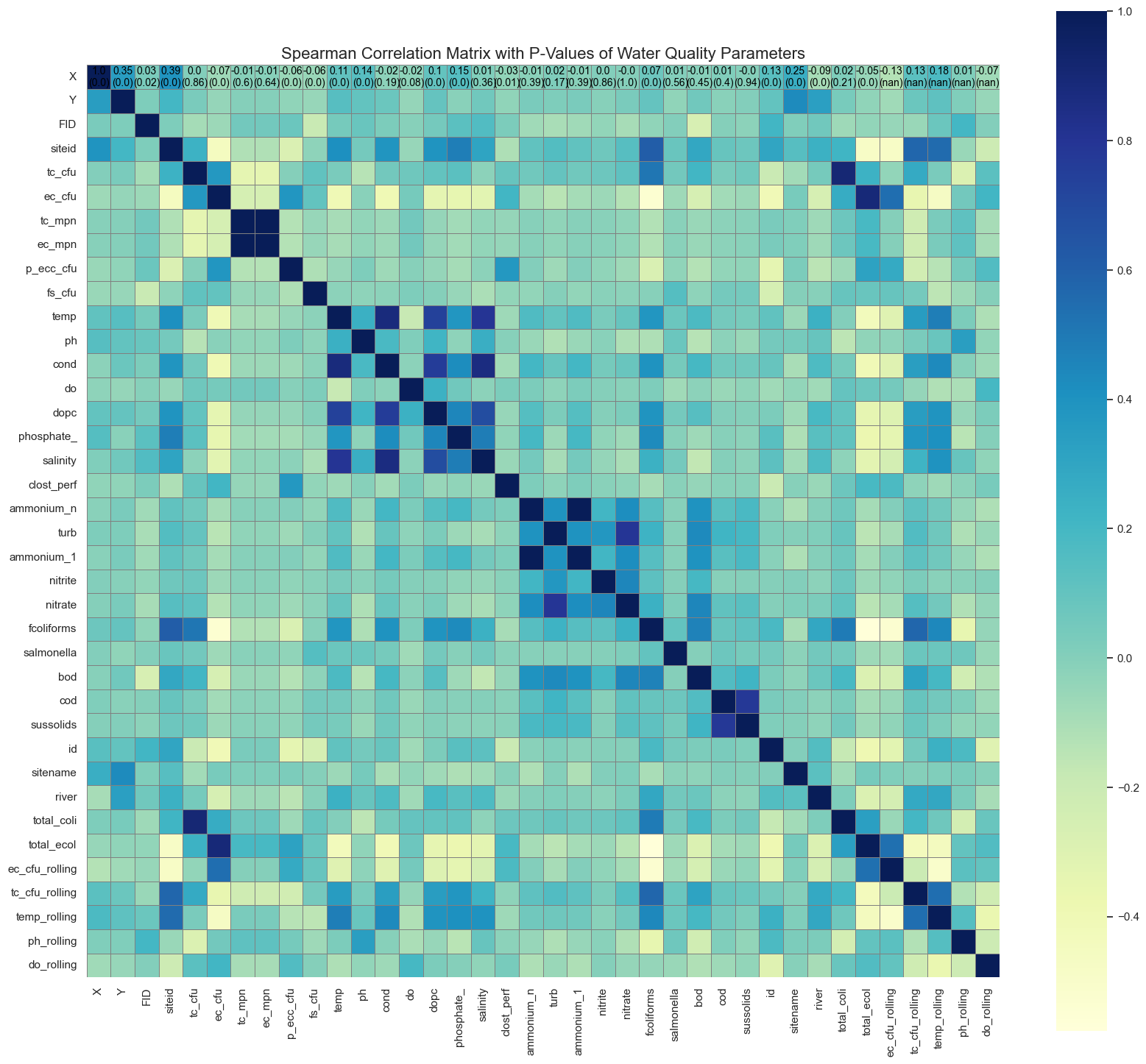


The pH levels show relatively little variation across the region, with most values falling within the range of 7.5 to 9.0. This indicates slightly alkaline conditions throughout the sampled area. There are no clear spatial patterns or significant variations in pH across different longitudes and latitudes. The consistency in pH levels suggests that the water bodies in the region have a stable buffering capacity and are not significantly impacted by acidic or highly alkaline inputs.

In conclusion, the spatial analysis reveals that the water quality in terms of bacterial indicators (E. coli and total coliforms) is consistently good across the region. Dissolved oxygen levels show some spatial variation, with potentially better-oxygenated waters in the north. Temperature exhibits a noticeable gradient, with warmer waters in the north and east. pH levels are consistently slightly alkaline across the region. These findings suggest that while microbial water quality is uniformly good, there are spatial variations in physical parameters that may influence aquatic ecosystems and warrant further investigation to understand their underlying causes and potential impacts.

4.2.3 Correlation Analysis

The correlation analysis focuses on the numeric columns of the dataset, ensuring only relevant quantitative data is included. A Spearman correlation matrix is calculated to assess monotonic relationships between key water quality parameters such as E. coli (CFU), Total Coliforms (CFU), temperature, pH, and dissolved oxygen. The Spearman method is selected for its robustness in handling non-linear relationships and its suitability for the data types present. Alongside correlation coefficients, p-values are computed to determine the statistical significance of observed correlations. These results are visualized in a heatmap below, annotating both correlation coefficients and p-values, providing a comprehensive representation of the strength and significance of relationships between the selected parameters.



*Figure 5: Heatmap of Spearman correlation coefficients between water quality parameters.*

Interpretation of the Correlation Analysis:

Bacterial Indicators: E. coli (ec\_cfu) and Total Coliforms (tc\_cfu) show a strong positive correlation (approximately 0.8-0.9), indicating that these bacterial indicators often co-occur. This relationship suggests common sources or environmental conditions promoting both bacterial populations, likely linked to fecal contamination.

Temperature and Dissolved Oxygen: The heatmap reveals a moderate negative correlation between temperature (temp) and dissolved oxygen (do), approximately -0.4 to -0.5. This aligns with the understanding that oxygen solubility decreases as water temperature increases.

pH Relationships: pH demonstrates positive correlations with dissolved oxygen (do) and temperature (temp), both around 0.4-0.5. This suggests that higher pH levels are associated with warmer, more oxygenated waters, possibly due to increased photosynthetic activity or reduced organic decomposition in such conditions.

Nutrients and Water Quality: Phosphate shows moderate positive correlations with both E. coli and total coliforms (0.4-0.5 range), suggesting a potential link between nutrient levels and bacterial contamination. This could indicate that nutrient-rich environments may support higher bacterial populations.

Conductivity and Salinity: A strong positive correlation (>0.8) is observed between conductivity (cond) and salinity, which is expected as these parameters are closely related in aquatic systems.

Ammonium and Nitrite: Ammonium (ammonium\_n) and nitrite show a strong positive correlation (>0.8), suggesting a potential relationship in the nitrogen cycle within the water bodies.

Biological Oxygen Demand (BOD): BOD exhibits moderate positive correlations with bacterial indicators (E. coli and total coliforms) and nutrients like phosphate, ranging from 0.4 to 0.6. This relationship indicates that higher organic matter content (reflected by BOD) may be associated with increased bacterial populations and nutrient levels.

Spatial Variation: The 'river' parameter shows weak to moderate correlations with several water quality parameters, suggesting that different rivers or sections of rivers may have distinct water quality characteristics.

Rolling Averages: The rolling average parameters (e.g., do\_rolling, ph\_rolling) generally show stronger correlations with their respective instantaneous measurements and other related parameters, indicating consistent trends over time.

Key insights from this analysis include:

* The strong correlation between E. coli and total coliforms confirms their reliability as indicators of fecal contamination in Bristol's surface waters.
* The interplay between temperature, dissolved oxygen, and pH highlights the complex dynamics of water quality parameters and their potential impacts on bacterial populations.
* Nutrient levels, particularly phosphate, appear to play a role in bacterial contamination, suggesting that managing nutrient inputs could be crucial for improving water quality.
* The relationships between BOD, bacterial indicators, and nutrients underscore the importance of organic matter in water quality management.
* Spatial variations in water quality parameters emphasize the need for location-specific management strategies across Bristol's water bodies.

4.2.4 Distribution Analysis and Outlier Detection

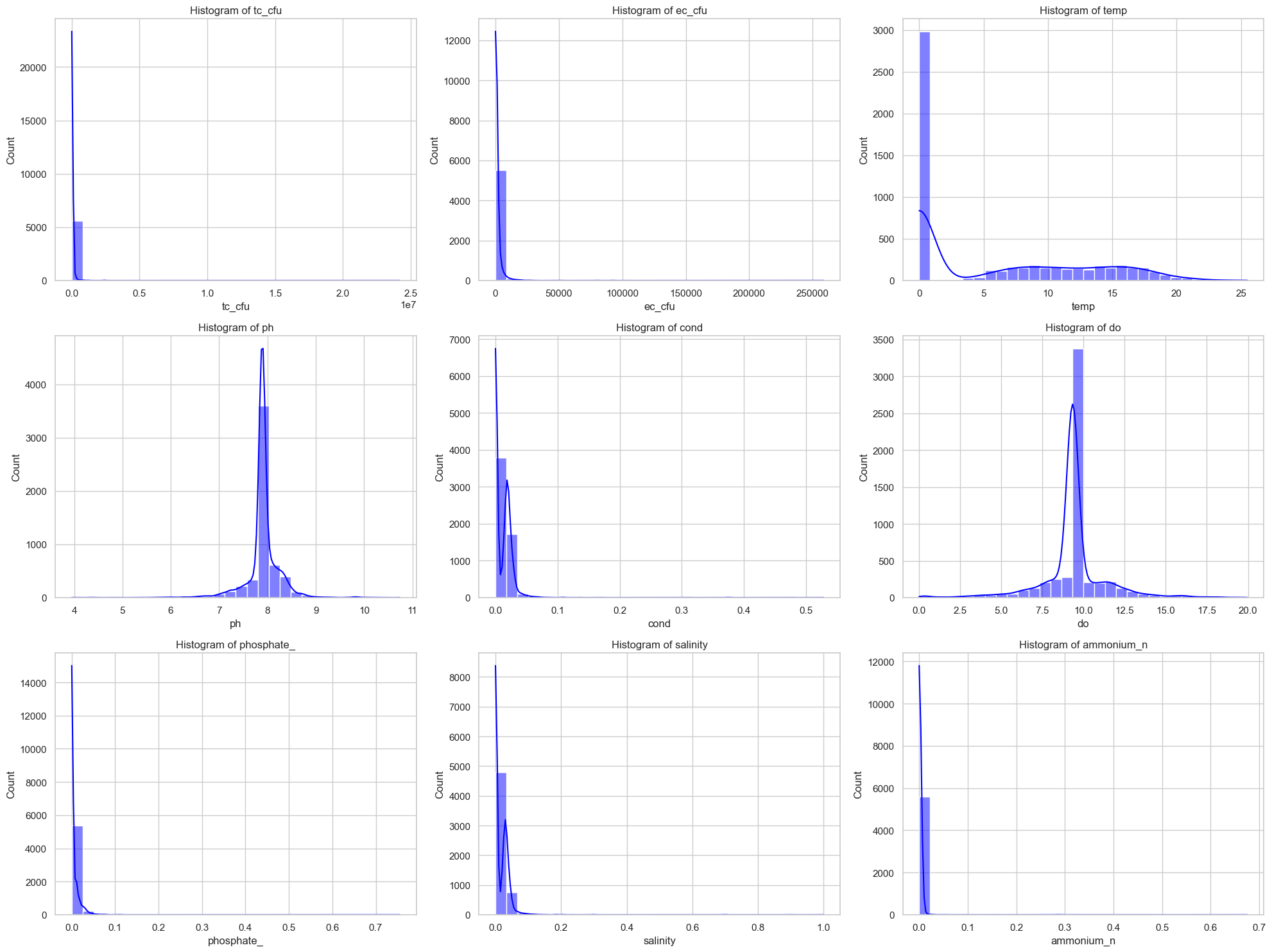
The distribution analysis and outlier detection aim to examine the spread and central tendency of key water quality parameters and identify significant outliers that may indicate pollution events or unusual environmental conditions in Bristol's surface waters

Methods: Three complementary techniques were employed to analyze the distribution of water quality parameters:

1. Histograms: To visualize the frequency distribution of each parameter.
2. Box Plots: To identify and visualize outliers in the data for each parameter.
3. Q-Q Plots: To compare the data distribution against a normal distribution and identify deviations.

Findings:

a) Histograms:

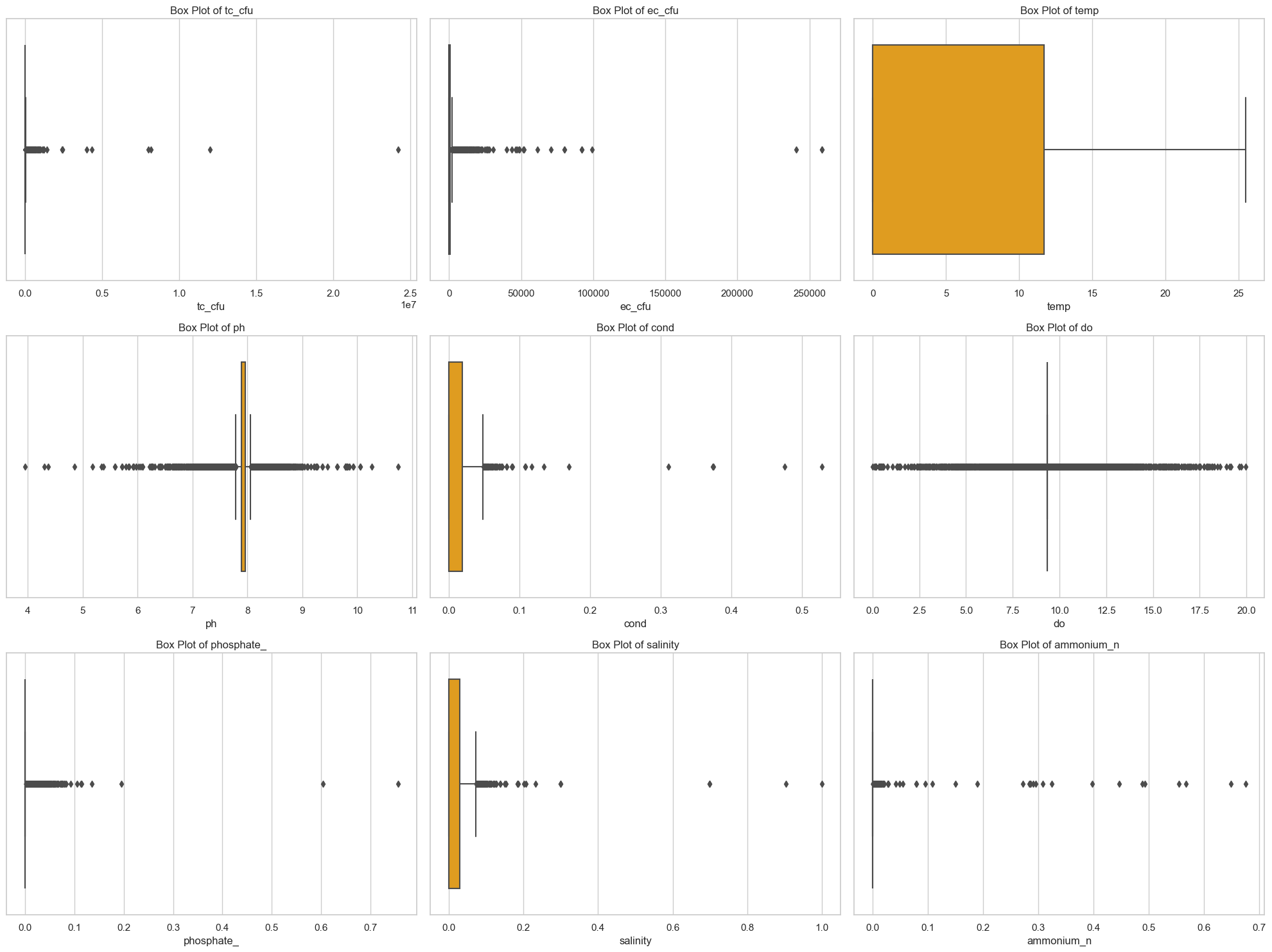


*Figure 6: Histograms of key water quality parameters*.

The histograms of most parameters, including Total Coliforms (tc\_cfu), Escherichia coli (ec\_cfu), Temperature (temp), Conductivity (cond), Dissolved Oxygen (do), Phosphate (phosphate\_), Salinity (salinity), and Ammonium Nitrogen (ammonium\_n), exhibit a strong right-skewed (positively skewed) distribution. This pattern indicates that a majority of the observations cluster around lower values, while a smaller number of high values create an extended tail to the right. Such distributions are common in environmental data and often suggest that while most water samples have relatively low levels of these parameters, there are occasional instances of much higher concentrations.

The histogram for pH stands out with its bimodal distribution, revealing two distinct peaks. This bimodality suggests that the pH values in the dataset tend to cluster around two predominant levels. Such a pattern might indicate the presence of different water sources, varying treatment processes, or distinct environmental conditions affecting the acidity or alkalinity of the water across the study area.

b) Box Plots:

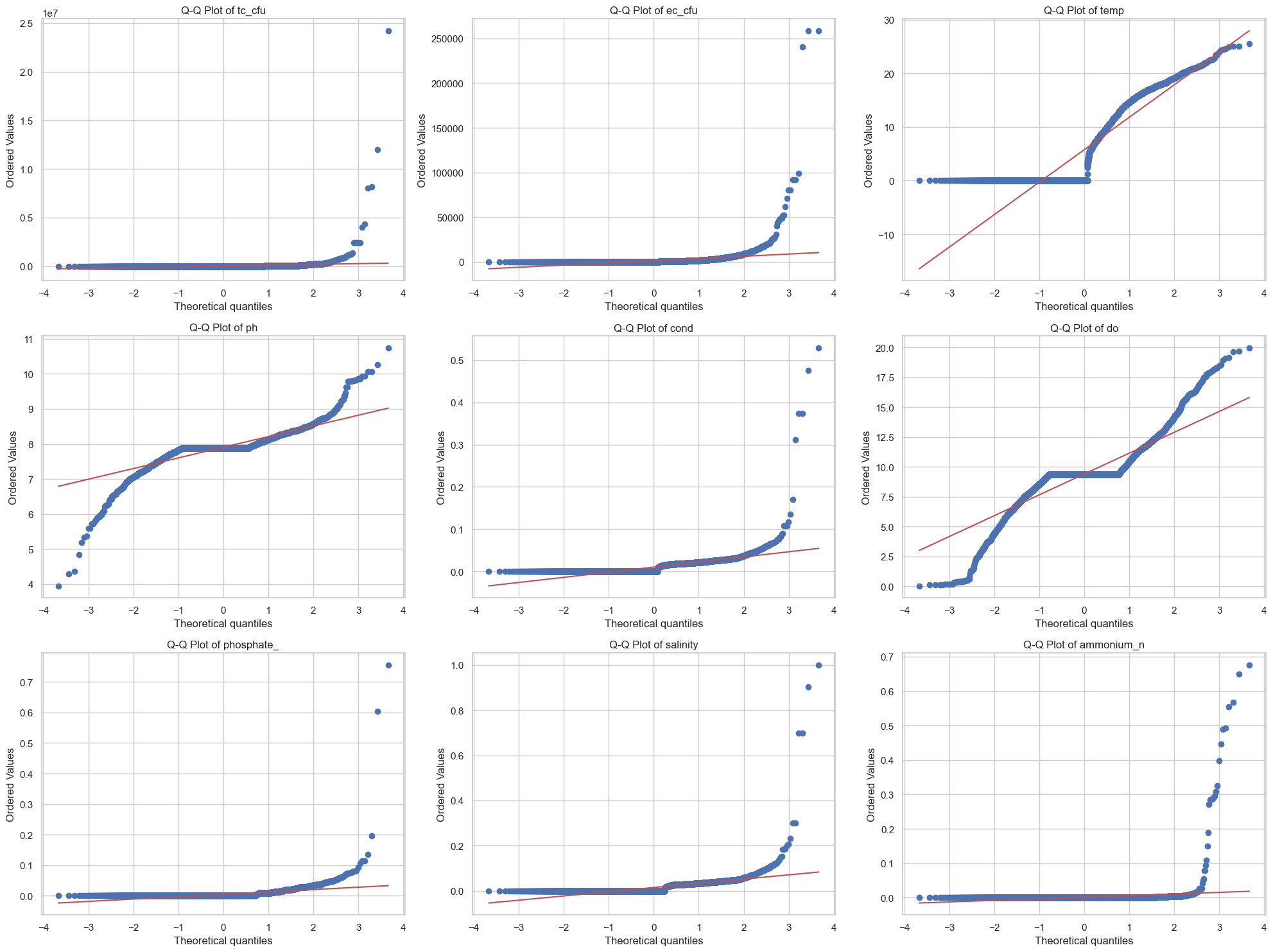


*Figure 7: Box plots of key water quality parameters.*

The box plots reveal the presence of numerous outliers across most of the parameters, providing valuable insights into the extreme values within the dataset. Parameters such as tc\_cfu, ec\_cfu, cond, do, phosphate\_, salinity, and ammonium\_n display a large number of data points beyond the whiskers, indicating frequent occurrences of extreme values. These outliers could represent significant pollution events, natural fluctuations, or measurement anomalies that warrant further investigation.

The pH parameter shows a wider interquartile range (IQR) compared to other parameters, but with fewer outliers. This suggests that while pH values are more spread out across their range, they experience less extreme variability. This pattern aligns with the bimodal distribution observed in the histogram and may reflect the natural buffering capacity of water bodies.

c) Q-Q Plots:



The Q-Q plots for all parameters deviate significantly from the straight line that would indicate a normal distribution. This deviation is particularly pronounced at the tails of the distributions, aligning with the presence of outliers identified in the box plots. For parameters like Total Coliforms (tc\_cfu), Escherichia coli (ec\_cfu), and pH, the points on the Q-Q plot curve sharply upwards or downwards at the ends, providing strong evidence of skewness and non-normality in the data distribution.

d) Outlier Detection:

Total Coliforms (tc\_cfu) and Escherichia coli (ec\_cfu): The detection of numerous outliers with significantly higher values than the bulk of the data points suggests the occurrence of episodic contamination events. These could be linked to sewage overflows, agricultural runoff after heavy rains, or other intermittent pollution sources. Maximum outlier values reach 1,209,800 CFU/100mL for tc\_cfu and 20,924 CFU/100mL for ec\_cfu.

Temperature (temp): No significant outliers were detected in the temperature data, indicating a relatively stable thermal regime in the studied water bodies.

pH: The outliers in pH values range from 3.95 to 8.42, reflecting the bimodal distribution and suggesting the presence of distinct water sources or environmental conditions affecting acidity and alkalinity.

Conductivity (cond): Outliers in conductivity range from 0.049057 to 0.374303 mS/cm, potentially indicating localized increases in dissolved ions due to pollution events, saltwater intrusion in coastal areas, or geological influences.

Dissolved Oxygen (do): Detected outliers in dissolved oxygen levels range from 7.56 to 11.03 mg/L, which could point to areas experiencing eutrophication (very low DO) or algal blooms (very high DO during daylight hours).

Phosphate (phosphate\_): The numerous outliers suggesting occasional high phosphate levels, with a maximum of 0.037175 mg/L, could be linked to agricultural runoff, sewage discharge, or the use of phosphate-containing detergents in the watershed.

Salinity (salinity): Substantial outliers, ranging from 0.073043 to 0.698261 PSU, may indicate areas influenced by tidal mixing in estuarine environments or evaporation effects in shallow water bodies.

Ammonium Nitrogen (ammonium\_n): Outliers in ammonium levels, with a maximum of 0.003784 mg/L, could signify points of localized pollution from sewage, animal waste, or industrial discharges.

The distribution analysis reveals that the water quality parameters deviate significantly from normal distributions, exhibiting pronounced skewness and numerous outliers. These characteristics are typical of environmental data and reflect the complex, dynamic nature of water quality in urban and peri-urban environments. The presence of outliers, particularly in parameters like Total Coliforms (tc\_cfu), Escherichia coli (ec\_cfu), Conductivity (cond), Dissolved Oxygen (do), and Salinity (salinity), provides valuable insights into potential episodic pollution events, specific source contributions, or measurement anomalies. Understanding these distributions and outliers is crucial for developing robust water quality management strategies and for identifying areas or parameters that require more focused monitoring or intervention. This analysis underscores the importance of considering both anthropogenic and natural factors in water quality management.

## 4.3 Machine Learning Modelling:

1. Feature Selection:

**Objective:** The primary goal of feature selection was to identify the most relevant predictors that exhibit a strong relationship with the target variable, Escherichia coli (ec\_cfu), to enhance the predictive accuracy of the regression model.

**Method:** Correlation analysis was conducted on numeric data, focusing on the relationship between each feature and the target variable. Features with an absolute correlation coefficient above 0.2 were selected, ensuring that only those with a significant association were included in the model.

****

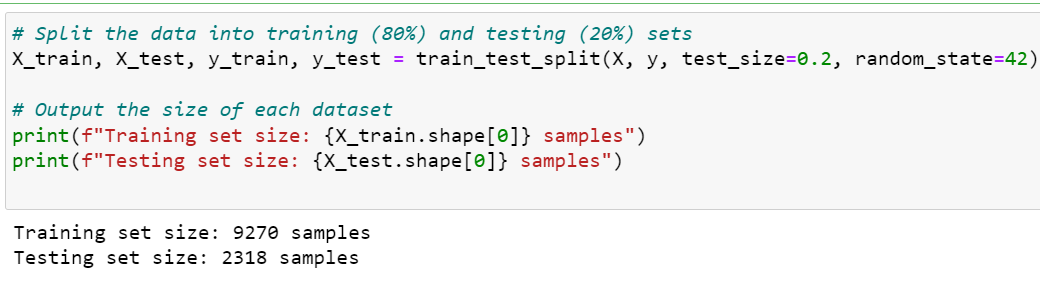
**Findings:** The analysis identified 'total\_ecol' as the sole feature meeting the selection criteria. This variable demonstrated a substantial statistical relationship with the target variable, indicating its potential predictive power.

**Key Insights:** The selected feature, 'total\_ecol', is likely to have the most substantial impact on the target variable, making it a suitable predictor for the regression model. This selection is expected to improve the model's accuracy and provide meaningful insights for water quality management. By focusing on this strongly correlated variable, the model can potentially capture a significant factor influencing Escherichia coli levels in the surface waters under study.

1. **Data Splitting:**

**Objective:** The dataset was divided into separate training and testing subsets. This step was crucial to ensure that the model could be trained effectively using the majority of the data while reserving a portion of the data to test the model's performance on new, unseen cases. This approach allowed for a realistic evaluation of the model's predictive accuracy.

**Method:** The data was split into 80% for training and 20% for testing using the train\_test\_split function from scikit-learn. This method ensured that the model was trained on a large portion of the data, while the remaining data was used to validate the model's performance. A random state of 42 was set to ensure reproducibility of the split.

****

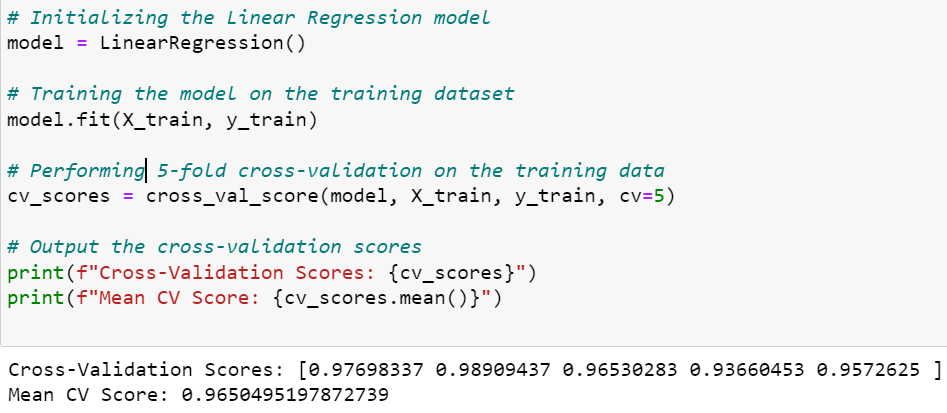
**Findings:** The training set consisted of 9,270 samples, and the testing set included 2,318 samples.

**Key Insights:** By using the larger portion of the data (80%) for training, the model was provided with a substantial dataset to learn from, potentially capturing a wide range of patterns and relationships within the data. The testing set, comprising 20% of the data, offers a robust sample size for evaluating the model's performance on unseen data.

1. **Model Development:**

**Objective:** The objective of this phase was to develop a Linear Regression model and assess its potential performance in predicting E. coli levels (ec\_cfu) based on the selected predictor. This step aims to evaluate the model's consistency and reliability before proceeding to the final training and testing phases.

**Method:** A Linear Regression model was initialized using the LinearRegression class from scikit-learn. To assess the model's potential performance and stability, 5-fold cross-validation was employed on the training dataset. This method provides a robust initial evaluation by testing the model's performance across different subsets of the training data.



**Findings:** The 5-fold cross-validation process yielded exceptionally high scores: [0.97698337, 0.98909437, 0.96530283, 0.93660453, 0.9572625]. The mean of these cross-validation scores was 0.9650495197872739.

**Key Insights:**

Model Stability: The consistently high cross-validation scores (ranging from approximately 0.937 to 0.989) indicate that the Linear Regression model is likely to perform well across different subsets of the data. This suggests a stable relationship between the predictor ('total\_ecol') and the target variable (ec\_cfu).

Predictive Potential: With a mean cross-validation score of about 0.9650, the model shows potential to explain approximately 96.50% of the variance in E. coli levels. This high score suggests that the linear model is well-suited to capture the relationship in the data.

Generalization Capability: The small variation in scores across folds indicates that the model is likely to generalize well to unseen data, which is crucial for its practical application in water quality prediction.

Validation of Feature Selection: The high performance across all folds validates the feature selection process, confirming that 'total\_ecol' is indeed a strong predictor of E. coli levels in the water samples.

Model Simplicity and Effectiveness: Despite the model's simplicity (using only one predictor), its high performance suggests a strong, direct relationship between 'total\_ecol' and 'ec\_cfu'.

1. Model Training and Validation:

**Objective:** A Linear Regression model was developed utilizing the LinearRegression class from scikit-learn. The model was trained using the training dataset, which had been split earlier. To ensure the model's reliability and to assess its performance across different subsets of data, 5-fold cross-validation was employed. This method provided a robust evaluation by testing the model on various data segments and averaging the results.

Findings: The cross-validation resulted in the following R-squared scores for each fold:

* Fold 1: 0.97698337
* Fold 2: 0.98909437
* Fold 3: 0.96530283
* Fold 4: 0.93660453
* Fold 5: 0.95726250 The mean R-squared score across all folds was approximately 0.9650.

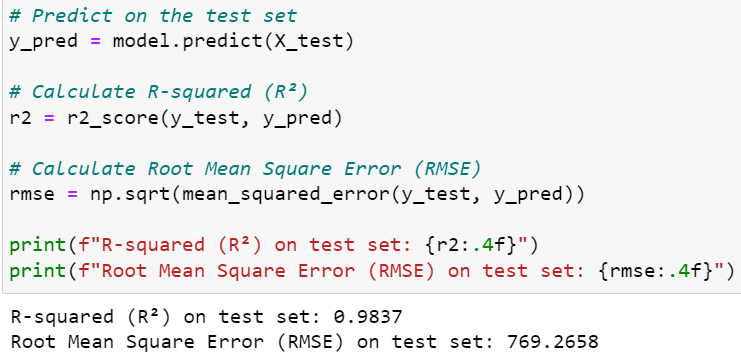
Key Insights: The model demonstrated strong and consistent performance across different subsets of the data, as reflected in the high R-squared values consistently above 0.93. This indicated that the model successfully explained a significant portion of the variance in the target variable. The slightly lower score observed in Fold 4 suggested some minor variability in model performance, but overall, the model remained highly stable and reliable, performing exceptionally well across the cross-validation process. The high mean R-squared score of 0.9650 further confirmed the model's robust predictive capability for E. coli levels in water quality assessment.

### **Model Evaluation**

**Objective:** The aim was to assess the performance of the trained Linear Regression model on the unseen test dataset. This evaluation helps determine how well the model generalizes to new data and its effectiveness in predicting the target variable, E. coli concentration (ec\_cfu).

**Method:** The trained model was applied to the test dataset to predict the target variable. The evaluation of the model's performance was conducted using two key metrics:

* R-Squared (R²): This metric measures the proportion of variance in the dependent variable that is explained by the independent variables. It provides an indication of how well the model fits the data.
* Root Mean Square Error (RMSE): This metric calculates the average magnitude of the prediction errors, offering a sense of how closely the model's predictions align with the actual values.



**Findings:**

* R-Squared (R²) on Test Set: 0.9837 This exceptionally high R² value suggested that approximately 98.37% of the variance in the E. coli concentration was explained by the selected feature in the test set.
* Root Mean Square Error (RMSE) on Test Set: 769.2658 The RMSE value of 769.2658 indicated the average deviation of the model's predictions from the actual values in the original scale of the target variable.

**Key Insights:**

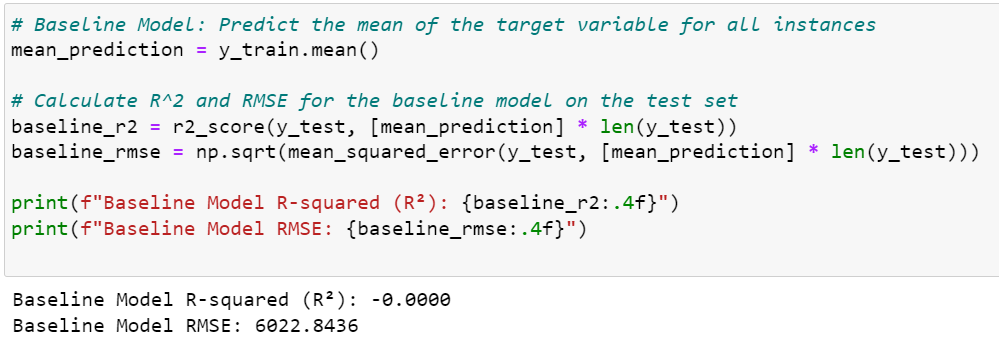
* Model Accuracy: The very high R² value reflected that the model effectively captured the relationship between the feature and the target variable, making it a highly reliable predictor for unseen data.
* Prediction Precision: While the RMSE value seems high in absolute terms, its interpretation depends on the scale of the target variable. Given the high R² value, this RMSE likely represents a relatively small percentage of the average E. coli concentration, indicating good predictive performance.
* Generalization: The model's strong performance on the test set, as evidenced by the high R² value, suggests excellent generalization to new, unseen data.
* Practical Applicability: The model's high accuracy indicates its potential for reliable predictions in practical water quality management applications, though the scale of the RMSE should be considered in the context of typical E. coli concentration ranges.

1. **Comparison to the Baseline Model**

**Objective:** To evaluate the effectiveness of the developed Linear Regression model in comparison to a baseline model. The baseline model simply predicted the mean value of the target variable for all instances, serving as a benchmark to quantify the improvement achieved by the developed model.

**Method:**

* Baseline Model Creation: The mean value of the target variable from the training set was used to predict all instances in the test set.
* Evaluation Metrics: The R² and RMSE values were calculated for both the baseline model and the developed model to assess their performance.
* Comparison: The performance of the developed model was compared against the baseline to highlight the improvement in predictive accuracy.

****

**Findings:**

* **Developed Model Performance:**
  + **R-Squared (R²) on Test Set: 0.9837**
  + **Root Mean Square Error (RMSE) on Test Set: 769.2658**
* **Baseline Model Performance:**
  + **R-Squared (R²): -0.0000**
  + **Root Mean Square Error (RMSE): 6022.8436**

Key Insights:

* Significant Improvement in R²: The developed model achieved an R² of 0.9837, indicating that it explained approximately 98.37% of the variance in the target variable, compared to the baseline model's R² of -0.0000, which explained none of the variance. This substantial improvement demonstrated the superior performance of the developed model over the baseline.
* Reduction in Prediction Error: The RMSE of the developed model (769.2658) was significantly lower than that of the baseline model (6022.8436). This highlighted the model's precision in predicting the target variable, providing much closer and more accurate estimates.
* Model Efficacy: The R² value of 0 for the baseline model underscored its ineffectiveness, as it performed no better than merely predicting the mean. In contrast, the developed model, with its high R² and comparatively low RMSE, proved to be highly effective in predicting E. coli concentration, thereby validating its utility in practical water quality assessment.
* Scale of Improvement: The developed model reduced the RMSE by approximately 87.2% compared to the baseline, indicating a substantial increase in prediction accuracy.

**Conclusion:** The comparison clearly demonstrated that the Linear Regression model offered a substantial improvement over the baseline model. This underlined the value of employing a well-developed machine learning model for predicting water quality indicators, leading to more accurate and reliable assessments, which are crucial for informed environmental management decisions.

## 4.4 Environmental Impact Assessment of Bristol's Surface Water Quality

This Environmental Impact Assessment (EIA) evaluated the quality of Bristol's surface waters using a comprehensive three-step approach: compliance analysis, temporal trend examination, and parameter correlation analysis. The assessment focused on key water quality parameters including Escherichia coli (E. coli), phosphate, pH, and dissolved oxygen (DO). These parameters were selected due to their critical role in assessing overall water quality and their established importance in regulatory frameworks. E. coli serves as an indicator of fecal contamination and potential pathogenic threats; phosphate levels are crucial for understanding nutrient dynamics and potential eutrophication risks; pH is fundamental for aquatic life support and chemical processes in water; and dissolved oxygen is essential for aerobic aquatic organisms and overall ecosystem health (Bartram & Ballance, 1996). Moreover, these parameters align with the UK's regulatory standards set by bodies such as the Environment Agency and the UK Technical Advisory Group (UKTAG) on the Water Framework Directive, ensuring the assessment's relevance to local environmental management goals.

Step 1: Compliance Analysis

The compliance analysis involved evaluating the measured values of each parameter against the UK environmental standards. This step is crucial for understanding the overall health of Bristol's surface waters and identifying potential areas of concern.

E. coli compliance was found to be 75.71%, the lowest among all parameters. This indicates that approximately one-quarter of the samples exceeded the standard of ≤900 CFU/100ml set for good ecological status in rivers (Environment Agency, 2019). The relatively low compliance rate suggests potential issues with fecal contamination, which could pose risks to both human health and aquatic ecosystems.

Phosphate compliance was higher at 91.04%, based on the standard of ≤0.12 mg/L (as annual mean) for good ecological status in rivers (Environment Agency, 2019). This higher compliance rate suggests that nutrient levels are generally well-managed in Bristol's surface waters, although there is still room for improvement.

pH and dissolved oxygen showed excellent compliance rates of 98.50% and 96.99% respectively. These high compliance rates, based on standards of 6.0 to 9.0 for pH and ≥60% saturation (5-6 mg/L) for DO (UKTAG, 2008), indicate that Bristol's surface waters generally provide suitable conditions for aquatic life in terms of acidity/alkalinity and oxygenation.

Step 2: Temporal Trend Analysis Monthly compliance trends were analyzed to identify seasonal patterns. E. coli and Phosphate maintained consistently high compliance year-round. pH showed the most variability, with lowest compliance in May (94.5%) and highest in July and November (nearly 100%). Dissolved Oxygen (DO) exhibited clear seasonal variation, with lowest compliance in summer months (95%) and highest in winter (99%). This pattern aligns with the inverse relationship between water temperature and oxygen solubility (Wetzel, 2001).

Step 3: Parameter Correlation Analysis A correlation matrix was generated to understand relationships between water quality parameters. Key findings include:

* Temperature negatively correlated with DO (-0.20), confirming the seasonal DO pattern.
* Strong positive correlation between conductivity and salinity (0.52), both showing moderate positive correlations with temperature.
* E. coli showed weak correlations with all parameters, suggesting its presence is more related to direct contamination than environmental conditions.
* Phosphate showed a weak positive correlation with temperature (0.14) and moderate positive correlation with salinity (0.26).

Interpretation and Recommendations:

1. Seasonal Management: Implement enhanced aeration strategies for water bodies during summer months to mitigate lower DO levels. Increase frequency of pH monitoring in spring when compliance dips.
2. Source Identification: Despite high compliance, investigate potential sources of pH fluctuations in spring and phosphate increases during warmer periods. Focus on identifying direct sources of E. coli contamination year-round.
3. Integrated Monitoring: Develop an integrated monitoring approach that accounts for the interrelationships between parameters, particularly temperature, conductivity, and salinity (Vega et al., 1998).
4. Climate Change Preparedness: Given the observed temperature effects on DO and other parameters, develop long-term strategies to mitigate potential impacts of climate change on water quality.
5. Site-Specific Management: Conduct detailed investigations at sites with lower compliance rates (e.g., Site 15 with 88.74% overall compliance) to identify and address local factors affecting water quality.
6. Public Awareness: Use the seasonal trend data to inform the public about water quality variations, particularly regarding recreational water use in summer when DO levels are lower.
7. Continuous Improvement: While overall compliance is high, continue efforts to further improve water quality, focusing on the parameters and seasons with lower compliance rates.

This EIA provides a comprehensive overview of Bristol's surface water quality, highlighting both strengths in current management practices and areas for potential improvement. The multi-faceted approach allows for a nuanced understanding of water quality dynamics, enabling targeted and effective management strategies (Patil et al., 2002).

# 5 Discussion:

### 5.1 Discussion of results in relation to the research objectives:

The analysis of Bristol's surface water quality in 2023 revealed several key findings that address the study's objectives. The exploratory data analysis (EDA) showed significant temporal variations in E. coli and Total Coliform levels, with notable spikes between 2016 and 2020, indicating potential major contamination events. Temperature patterns remained relatively stable, while pH levels showed a slight increase in recent years. The spatial distribution analysis revealed a relatively uniform distribution of low E. coli and total coliform levels across the region, suggesting generally good microbial water quality. However, variations in dissolved oxygen levels were observed, with higher concentrations more prevalent in the northern part of the region.

The correlation analysis identified important relationships between water quality parameters. A strong positive correlation was found between E. coli and Total Coliforms, indicating common sources or environmental conditions promoting both bacterial populations. The moderate negative correlation between temperature and dissolved oxygen aligns with established environmental principles.

The developed Linear Regression model for predicting E. coli levels demonstrated high performance, with an R-squared value of 0.9837 on the test set. This model significantly outperformed the baseline model, reducing the RMSE by approximately 87.2%, thus meeting the objective of developing an effective predictive tool for water quality management.

The environmental impact assessment revealed varying levels of compliance with UK standards across different parameters. E. coli compliance was the lowest at 75.71%, indicating potential issues with fecal contamination. In contrast, pH and dissolved oxygen showed excellent compliance rates of 98.50% and 96.99% respectively, suggesting generally suitable conditions for aquatic life.

### 5.2 Comparison with existing Literature:

Our findings align with recent studies on urban water quality. The observed spikes in E. coli and Total Coliform levels are consistent with Wilson et al.'s (2024) study on microplastic pollution in the Bristol Channel, highlighting the prevalence of various forms of urban water contamination. While their study focused on microplastics, the presence of such pollutants often correlates with other water quality issues, including bacterial contamination. The performance of our Linear Regression model for E. coli prediction compares favorably with similar models described by Ren et al. (2020), who also found MLR effective for predicting water quality parameters in urban rivers..

The seasonal variations in dissolved oxygen levels, with lower compliance in summer months, align with the findings of Makropoulos et al. (2018) on urban water resilience, emphasizing the importance of considering seasonal factors in water quality management. Our spatial analysis results support the conclusions of Goonetilleke et al. (2019) on the importance of location-specific management strategies in urban water quality.

### 5.3 Limitations:

Despite its comprehensive approach, this study has several limitations. The analysis is based on data from a single year (2023), which may not capture long-term trends or account for year-to-year variations in water quality. The spatial coverage of sampling sites may not fully represent all water bodies in Bristol, potentially missing localized water quality issues.

The Linear Regression model, while effective, assumes linear relationships between variables, which may not capture all the complexities of water quality dynamics. More advanced machine learning techniques, such as those discussed by Rahmani et al. (2021), could potentially provide more nuanced predictions.

The study focused on a limited set of water quality parameters. Including additional parameters such as heavy metals, emerging contaminants, or microplastics could provide a more comprehensive assessment of water quality, as suggested by Patel et al. (2019).

Lastly, the reliance on secondary data means that the study is subject to any limitations or biases in the original data collection process. Future studies could benefit from integrating multiple data sources, including real-time sensor data and citizen science initiatives, to provide a more comprehensive and dynamic assessment of water quality.

# 6. Conclusion:

### 6.1 Summary of key findings

This study has provided a comprehensive analysis of surface water quality in Bristol for the year 2023, employing advanced data analysis techniques and machine learning. The key findings of this research are:

1. The analysis revealed significant temporal variations in E. coli and Total Coliform levels, with notable spikes between 2016 and 2020, indicating potential major contamination events.
2. Spatial distribution analysis showed generally good microbial water quality across the region, with variations in dissolved oxygen levels, particularly higher concentrations in the northern part of the region.
3. A strong positive correlation was found between E. coli and Total Coliforms, while a moderate negative correlation was observed between temperature and dissolved oxygen.
4. The developed Linear Regression model for predicting E. coli levels demonstrated high performance, with an R-squared value of 0.9837 on the test set, significantly outperforming the baseline model.
5. Environmental impact assessment revealed varying levels of compliance with UK standards, with E. coli compliance being the lowest at 75.71%, while pH and dissolved oxygen showed excellent compliance rates of 98.50% and 96.99% respectively.
6. Seasonal variations were observed, particularly in dissolved oxygen levels, with lower compliance in summer months.

### 6.2Contribution to the field

This study makes several significant contributions to the field of urban water quality management:

1. It demonstrates the effective application of machine learning techniques, specifically Multiple Linear Regression, to predict E. coli levels in urban surface waters, providing a valuable tool for proactive water quality management.
2. The study showcases the potential of open data initiatives in environmental science, utilizing publicly available data to generate insights for urban water management.
3. It provides a comprehensive methodology for assessing urban water quality, integrating advanced data analysis techniques with environmental impact assessment, which can be adapted for use in other urban contexts.
4. The research contributes to the understanding of spatial and temporal variations in urban water quality, highlighting the need for location-specific and season-specific management strategies.
5. The study bridges the gap between data science and environmental management, demonstrating how advanced analytical techniques can inform practical water quality management decisions.

### 6.3 Recommendations for future research

Based on the findings and limitations of this study, several recommendations for future research are proposed:

1. Extend the temporal scope of the analysis to multiple years to capture long-term trends and account for year-to-year variations in water quality.
2. Expand the range of water quality parameters analyzed to include emerging contaminants, heavy metals, and microplastics, providing a more comprehensive assessment of water quality.
3. Explore more advanced machine learning techniques, such as neural networks or random forests, to potentially capture non-linear relationships in water quality dynamics.
4. Integrate multiple data sources, including real-time sensor data and citizen science initiatives, to provide a more comprehensive and dynamic assessment of water quality.
5. Investigate the impact of urban development, land use changes, and climate change on water quality parameters to inform long-term water management strategies.
6. Conduct comparative studies with other urban areas to identify common challenges and best practices in urban water quality management.
7. Explore the potential of predictive modeling for other key water quality parameters beyond E. coli.
8. Investigate the effectiveness of various water quality improvement interventions using the developed predictive models.

# 7. References:

Bristol City Council. (2020). *Bristol One City Plan 2020.*<https://www.bristolonecity.com/wp-content/uploads/2020/01/One-City-Plan_2020.pdf>

Bristol Open Data. (2023). *Bristol Surface Water Quality Data.*<https://opendata.bristol.gov.uk/explore/?q=water+quality&sort=modified>

Bristol Water. (2022). *Annual water quality report 2021.*<https://www.bristolwater.co.uk/about-us/water-quality/water-quality-reports>

Environment Agency. (2021). *The state of the environment: water quality.*<https://www.gov.uk/government/publications/state-of-the-environment/the-state-of-the-environment-water-quality>

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions - Royal Society. Mathematical, Physical and Engineering Sciences/Philosophical Transactions - Royal Society. Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160360-. <https://doi.org/10.1098/rsta.2016.0360>

Goonetilleke, A., Liu, A., Managi, S., Wilson, C., Gardner, T., Bandala, E. R., Walker, L., Holden, J., Wibowo, M. A., Suripin, S., Joshi, H., Bonotto, D. M., & Rajapaksa, D. (2017). Stormwater reuse, a viable option: Fact or fiction? Economic Analysis and Policy, 56, 14–17. https://doi.org/10.1016/j.eap.2017.08.001

Grimm, N. B., Faeth, S. H., Golubiewski, N. E., Redman, C. L., Wu, J., Bai, X., & Briggs, J. M. (2008). Global Change and the Ecology of Cities. Science, 319(5864), 756–760. <https://doi.org/10.1126/science.1150195>

Langston, W. J., Pope, N. D., Jonas, P. J. C., Nikitic, C., Field, M. D. R., Dowell, B., Shillabeer, N., Swarbrick, R. H., & Brown, A. R. (2010). Contaminants in fine sediments and their consequences for biota of the Severn Estuary. Marine Pollution Bulletin, 61(1-3), 68-82.<https://doi.org/10.1016/j.marpolbul.2009.12.014>

Md Mashud, A. H., Pervin, M., Mishra, U., Daryanto, Y., Tseng, M.-L., & Lim, M. K. (2021). A sustainable inventory model with controllable carbon emissions in green-warehouse farms. *Journal of Cleaner Production*, *298*, 126777-. <https://doi.org/10.1016/j.jclepro.2021.126777>

Wang, X., & Yang, W. (2019). Water quality monitoring and evaluation using remote sensing techniques in China: a systematic review. *Ecosystem Health and Sustainability*, *5*(1), 47–56. <https://doi.org/10.1080/20964129.2019.1571443>

Makropoulos, C., Nikolopoulos, D., Palmen, L., Kools, S., Segrave, A., Vries, D., Koop, S., van Alphen, H. J., Vonk, E., van Thienen, P., Rozos, E., & Medema, G. (2018). A resilience assessment method for urban water systems. *Urban Water Journal*, *15*(4), 316–328. <https://doi.org/10.1080/1573062X.2018.1457166>

Oral, H. V., Carvalho, P., Gajewska, M., Ursino, N., Masi, F., van Hullebusch, E. D., Kazak, J. K., Exposito, A., Cipolletta, G., Andersen, T. R., Finger, D. C., Simperler, L., Regelsberger, M., Rous, V., Radinja, M., Buttiglieri, G., Krzeminski, P., Rizzo, A., Dehghanian, K., … Zimmermann, M. (2020). A review of nature-based solutions for urban water management in European circular cities: A critical assessment based on case studies and literature. *Blue-Green Systems*, *2*(1), 112–136. <https://doi.org/10.2166/bgs.2020.932>

Manvendra Patel, Rahul Kumar, Kamal Kishor, Todd Mlsna, Charles U. Pittman, & Dinesh Mohan. (2019). Pharmaceuticals of Emerging Concern in Aquatic Systems: Chemistry, Occurrence, Effects, and Removal Methods. *Chemical Reviews*, *119*(6), 3510–3673. <https://doi.org/10.1021/acs.chemrev.8b00299>

Patil, A. A., Annachhatre, A. P., & Tripathi, N. K. (2002). Comparison of conventional and geo-spatial EIA: A shrimp farming case study. *Environmental Impact Assessment Review*, *22*(4), 361–375. <https://doi.org/10.1016/S0195-9255(02)00013-6>

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. <https://doi.org/10.5555/1953048.2078195>.

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., & Shen, C. (2021). Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. *Environmental Research Letters*, *16*(2), 024025-. <https://doi.org/10.1088/1748-9326/abd501>

Ren, Z. H., Liu, Z. Z., Wang, Y. D., & Du, Q. L. (2014). Prediction of Water Quality Using Fuzzy Logic. *Advanced Materials Research*, *955–959*, 3310–3313. <https://doi.org/10.4028/www.scientific.net/AMR.955-959.3310>

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*, *8*, 42200–42216. <https://doi.org/10.1109/ACCESS.2020.2976199>

Shen, C. (2018). A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. *Water Resources Research*, *54*(11), 8558–8593. <https://doi.org/10.1029/2018wr022643>

Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., & Demir, I. (2020). A comprehensive review of deep learning applications in hydrology and water resources. *Water Science and Technology*, *82*(12), 2635–2670. <https://doi.org/10.2166/wst.2020.369>

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley Pub. Co.

Vega, M., Pardo, R., Barrado, E., & Debán, L. (1998). Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Research*, *32*(12), 3581–3592. <https://doi.org/10.1016/S0043-1354(98)00138-9>

Davies, P. M., Vörösmarty, C. J., Dudgeon, D., Sullivan, C. A., Glidden, S., Liermann, C. R., Prusevich, A., McIntyre, P. B., Gessner, M. O., Bunn, S. E., & Green, P. (2010). Global threats to human water security and river biodiversity. *Nature*, *467*(7315), 555–561. <https://doi.org/10.1038/nature09440>

Wilson, D. R., Godley, B. J., Haggar, G. L., Santillo, D., & Sheen, K. L. (2024). Corrigendum to “The influence of depositional environment on the abundance of microplastic pollution on beaches in the Bristol Channel, UK” [Mar. Pollut. Bull. 164 (2021) 111997]. *Marine Pollution Bulletin*, *202*, 116388–116388. <https://doi.org/10.1016/j.marpolbul.2024.116388>

Xie, X., Wu, T., Zhu, M., Jiang, G., Xu, Y., Wang, X., & Pu, L. (2021). Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land. *Ecological Indicators*, *120*, 106925-. <https://doi.org/10.1016/j.ecolind.2020.106925>

# 