

Chapter

Hybrid Modelling of Water Quality Dynamics: Data Assimilation with Machine Learning for Enhanced Predictions

Parul Tiwari, Channa Rajanayaka and Jing Yang

Abstract

Predicting *Escherichia coli* concentrations in recreational waters is essential for safeguarding public health and ensuring water quality compliance. This study applies time series analysis to forecast *E. coli* levels at six sites in New Zealand using historical data from 2005 to 2020. The goal is to develop a reliable predictive model that helps in proactive water management and early contamination warnings. Initially, an autoregressive integrated moving average (ARIMA) model was applied with parameters selected through a stepwise fitting approach. However, ARIMA demonstrated limitations in accurately capturing *E. coli* variability due to external environmental factors. Then the seasonal autoregressive integrated moving average with exogenous regressors (SARIMAX) model was applied for better predictive performance using water quality parameters and climate variables as input predictors. Results showed that no single water quality parameter consistently predicted *E. coli* across all sites, though total phosphorus emerged as a key predictor in five locations. The four-year forecasts showed patterns aligned with historical trends, suggesting reasonable predictive capability. However, forecast accuracy varied across sites, likely due to site-specific hydrological conditions. This study highlights the importance of site-specific modelling, real-time environmental data integration, and advanced machine learning techniques to improve water quality predictions. A refined forecasting approach can support early warning systems and risk-based decision-making, ultimately reducing health risks associated with microbial contamination in recreational waters.

Keywords: water quality dynamics, climate data, *Escherichia coli*, machine learning, data assimilation

1. Introduction

Water is a natural resource and a priceless gift of nature to all living organisms and is a constant threat of pollution by life itself [1]. It is essential for survival as it carries nutrients to all cells in our body and oxygen to our brain. Clean water is vital to our health, communities, and economy, which is essential to function and flourish

and will ultimately lead towards a more sustainable future. Four major categories of water-related diseases are waterborne, water-based, water-related, and water-scarce diseases [2]. Water quality is a complex interplay of physical, chemical, and biological parameters [3]. Several factors are responsible for degradation in water quality [4]. These include the concentration of microscopic pathogens such as bacteria and viruses, quantities of pesticides, insecticides, heavy metals and several other contaminants. Additionally, water quality dynamics are influenced by natural processes such as precipitation, sedimentation, and seasonal temperature variations, as well as anthropogenic activities like wastewater discharge, agricultural runoff, and industrial effluents.

Mathematical models have emerged as indispensable tools for understanding, predicting, and managing water quality [5, 6]. These models provide a quantitative framework for simulating the behaviour of pollutants, assessing the impacts of various stressors and developing management strategies. One widely used model is the advection-diffusion-reaction equation, which describes the transport and transformation of pollutants in water bodies [7]. This model incorporates terms for advection (movement of pollutants with water flow), diffusion (spreading due to concentration gradients), and reaction (chemical and biological transformations). A recent study has applied this model to simulate the dispersion of contaminants in rivers and estuaries, enabling precise predictions of pollution hotspots and their impacts on aquatic ecosystems [8].

The regulatory and advisory committees around the world set up a threshold limit on the microbial quality of water based on the concentrations of *E. coli* [9]. *E. coli*, a type of bacteria commonly found in the intestines of humans and animals, often enters water bodies through sewage discharge, agricultural runoff, or stormwater drainage. When contamination levels are high, water becomes unsafe for public use, posing serious health risks. Exposure to *E. coli*-contaminated water can lead to stomach cramps, diarrhoea, nausea, and even more severe infections [10]. There are various strains of this bacteria. A few of them are enteroaggregative *E. coli* (EAEC), enterohaemorrhagic *E. coli* (EHEC), and enterotoxigenic *E. coli* (ETEC). A specific strain of *E. coli*, called *E. coli* O157:H7, causes severe intestinal infections in humans and can lead to life-threatening conditions like kidney failure [11].

E. coli is ranked third among the 12 antibiotic-resistant priority pathogens identified by the World Health Organisation (WHO). *E. coli* contamination does not just affect human health; it also harms aquatic ecosystems. High levels of *E. coli* often mean there are other pollutants in the water, like nutrients from farms, which can cause algal blooms. These blooms use oxygen in the water, making it hard for fish and other aquatic animals to survive. This chain reaction harms biodiversity, damages ecosystems, and affects the overall health of rivers, lakes, and oceans. Understanding the dynamics of water quality is essential for assessing the state of water bodies and implementing effective management strategies. Accurate prediction of *Escherichia coli* contamination in surface water is challenging due to considerable uncertainty in the physical, chemical, and biological variables that control *E. coli* occurrence and sources in surface waters [11].

Collecting the samples and then culturing and incubating them until bacteria growth is visible in laboratories takes several hours [12]. Within this time gap, the actual levels of *E. coli* may change notably. For effective water quality management, fast and efficient alternative methods are needed in addition to conventional laboratory methods to estimate the levels of these bacteria. Data-driven predictive models have the capability to collaborate with the hard work of researchers and all contributors in the water industry towards achieving the standards of water quality. Machine

learning-based models have the potential to detect and understand unexpected gradual changes in the physical, chemical and biological qualities of water. Based on turbidity, temperature and pH, the authors in [13] applied several machine learning algorithms to estimate water quality and developed a deep neural network with 93% accuracy. The results are tested according to the World Health Organisation (WHO) standards. A combination of single feed-forward and multiple neural networks is used in [14] to get an R^2 and MSE score of 0.9270 and 0.1200, respectively. The authors used 25 features as the input and applied backwards elimination and forward selection combination methods.

The authors in Ref. [15] suggested that the existence of *E. coli* contamination and its prediction in rivers and reservoirs is a major concern because of its adverse effects on human life and food safety. Despite the very low survival rate outside of the host organism of *E. coli*, water resources have often been found to be contaminated [16]. The authors in Ref. [17] found that *E. coli* levels could be predicted using the amounts of algal pigments like chlorophyll and phycocyanin. *E. coli* interacts with algae by giving them organic nutrients and blocking sunlight.

A robust predictive model can be developed with careful consideration of input parameters. It is important to note here that in accounting for all these factors and heterogeneity present in *E. coli* fate and transport, great care is needed to generalise the results. Also, in case of non-point source pollution, concentrations of *E. coli* in waste and manure vary widely, and the dismissal of livestock waste is not well defined. Researchers observed that *E. coli* loads were notably in the high range at high flows compared to low flows [9]. The peaks of turbidity and of *E. coli* loads are compared in [18]. The authors in Ref. [18] observed that *E. coli* peaks are always preceded by turbidity peaks with similar timings, which resulted in a non-linear relationship between *E. coli* and turbidity.

Process-based models using mass conservation principles [19, 20] have been developed. However, hydrodynamic models are effective for water quality dynamics using *E. coli* as a microbial pollutant indicator [21]. A statistical and machine learning model has been developed for predicting *E. coli* loads using water quality, hydrodynamic data and a few other input variables [22]. Several multiple regression models have been developed to predict *E. coli* loads using several input features including turbidity, wave height, and chlorophyll [23–25]. A comparison study between artificial neural network (ANN) and support vector regression (SVR) is done to predict the concentration of *E. coli* at two recreational beaches [26].

In recent years, researchers have applied various data-driven regression models to elaborate the influences of various physicochemical water quality parameters on concentrations of faecal indicator organisms (FIOs) and artificial intelligence methods such as artificial neural network (ANN) and support vector machines [22, 23, 27]. The wide range of applications of machine learning algorithms and computational intelligence as decision support tools has made them indispensable tools [28–30]. Unlike traditional multivariate regression, which assumes variables are independent and models only linear relationships, machine learning uses advanced techniques to find patterns and connections in system data [31, 32]. It can handle complex, non-linear relationships between noisy and interdependent variables. This allows machine learning to make predictions and decisions that are often too complex for traditional methods [33, 34].

Laboratory analysis and the use of artificial intelligence for predictive models enable us to be proactive and to take corrective action for a contaminated site [35]. The applicability of ML methods in water quality management depends on the choice and response of different kernel functions in predictive models [36]. Recreational

water activities, such as swimming, kayaking, paddleboarding, and boating, have become increasingly important in today's fast-paced world. These activities provide people with a chance to unwind, reconnect with nature, and engage in physical exercise, which is essential for maintaining good health [37]. Beyond the physical benefits, such activities also play a key role in improving mental well-being, reducing stress, and fostering social connections. Families and friends often bond through shared water experiences, and the sense of calm that comes from being near or on the water can be deeply therapeutic [38]. Additionally, recreational water activities contribute to local economies by attracting tourists and supporting businesses like boat rentals, water parks, and beach resorts.

The aim of this study is to investigate the suitability of different machine learning models to predict the presence of *E. coli* for the New Zealand river dataset as per the recreational and bathing standards. The raw data collected from different sites of rivers in New Zealand provide valuable information in support of utilities to enable them to make better decisions in conjunction with machine learning techniques.

2. Materials and methods

2.1 Site description

This study utilised six different water quality measurement sites across New Zealand – three in the North Island and three in the South Island (**Figure 1**). The North Island sites include Mohaka (HV6), Whanganui (TU1), and Tongariro (TU2), while the South Island sites are Clutha (DN4), Wairau (NN3), and Maitara (DN5).

The Clutha River, the longest in the South Island and the second longest in New Zealand, stretches 338 kilometres (210 miles) from Lake Wānaka in the Southern Alps to the Pacific Ocean, flowing south-southeast through Central and South Otago. The river boasts the largest catchment and outflow in New Zealand, draining a basin of approximately 21,960 square kilometres (8480 square miles) and discharging a mean flow of 614 cubic metres per second.

The Wairau River is in the northeast of the South Island, in the Marlborough region. It originates in the Spenser Mountains and flows northeast to the Cook Strait.

The Maitara River is a significant waterway in the Southland Region of New Zealand. Stretching 240 kilometres (150 miles) from its source in the Eyre Mountains to its mouth at Toetoes Bay on the Pacific Ocean, the Maitara River is renowned for its natural beauty and recreational opportunities. The river faces issues with declining water quality due to intensive farming.

The Tongariro River, located in the North Island of New Zealand, is a vital waterway that flows north from the Central Plateau into Lake Taupō. In its lower reaches, the river's minimum flow ranges from 16 to 21 cubic metres per second. Renowned for its world-class trout fishing, it is often referred to as the "Mecca of trout fishing."

The Mohaka River is a significant waterway in the east-central North Island of New Zealand, spanning 172 km (107 miles) with a catchment area of 2444 square kilometres (910 square miles). Its mean annual flow, measured at Raupunga in the lower catchment, is 78.1 cubic metres per second.

The Whanganui River, one of the longest in the North Island at 290 kilometres (180 miles), flows northwest before turning southwest at Taumarunui, eventually reaching the Tasman Sea. Known for its scenic beauty, it is popular for jet boating, canoeing, and multi-day river journeys.

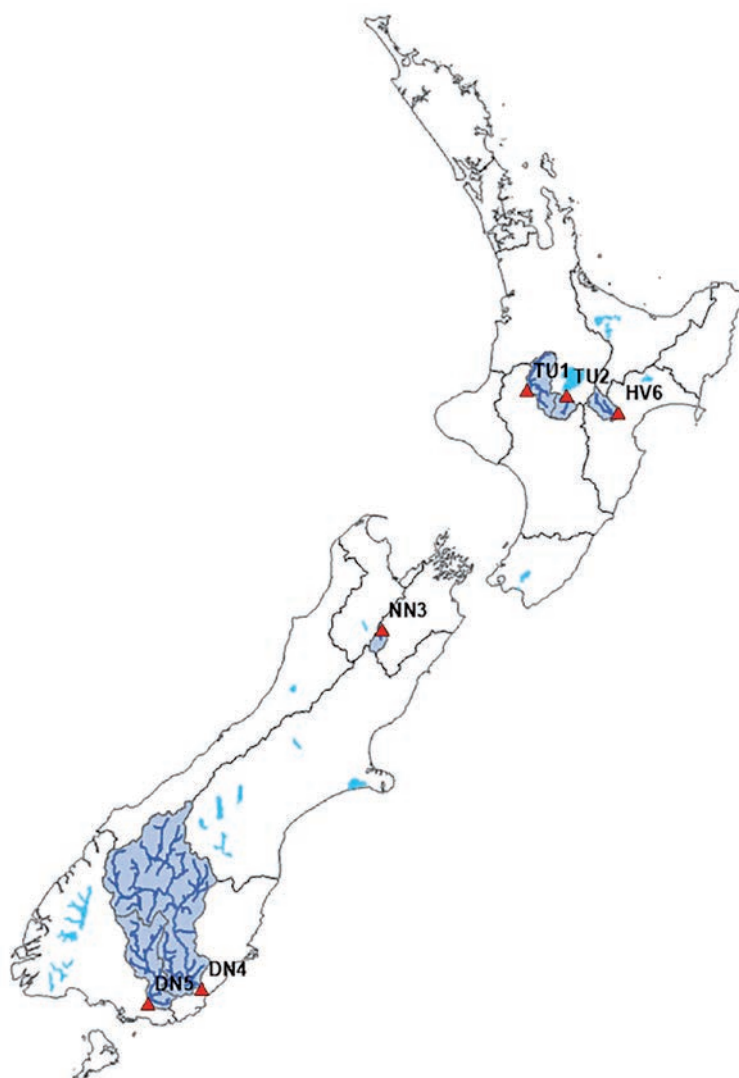


Figure 1.
 Location of the sites under study (red triangles: sampling sites; shaded polygons: upstream catchments; and blue lines: upstream river network).

2.2 Data collection

Water quality and streamflow data: Water quality and streamflow data from 1989 to 2020 were obtained from the National Institute for Water and Atmospheric Research's (NIWA) National River Water Quality Network (NRWQN) in New Zealand [39]. This dataset provides extensive information on key physical, chemical, and biological water quality parameters for 35 major rivers, collectively draining approximately half of New Zealand's total land area. Monthly data were collected from 58 active river sites, including 13 water quality variables and two biomonitoring variables, using a combination of *in situ* measurements and grab sampling.

Climate data: Climate data were sourced from NIWA’s Virtual Climate Station Network (VCSN) [40], a high-resolution gridded dataset that spans all of New Zealand. The VCSN dataset provides daily weather data including precipitation and temperature at a 5 km x 5 km resolution, offering comprehensive climate coverage across the study area.

Upstream catchment characteristics: Data on upstream catchment characteristics, including climate, geomorphology, geology, and land cover, were obtained from the River Environmental Classification [41] and the Freshwater Ecosystems of New Zealand database [42]. **Table 1** shows the description of sites and the units of water quality and climate variables used in this study.

2.3 Data analysis

Exploratory data analysis (EDA) is performed to understand the distribution of water quality parameters, trends, seasonal patterns, and correlations between *E. coli* levels and other water quality parameters.

Site Description	
Full name	Abbreviation
Mataura	DN5
Clutha	DN4
Wairau	NN3
Mohaka	HV6
Whanganui	TU1
Tongariro	TU2
Water Quality and Stream Flow Variables	
Parameter	Unit
Clarity (CLAR)	m
Dissolved reactive phosphorus (DRP)	mg/l
Turbidity (TURB)	NTU
Total nitrogen (TN)	mg/l
Total phosphorus (TP)	mg/l
Ammoniacal nitrogen (NH4N)	mg/l
Nitrate-nitrite-nitrogen (NNN)	mg/l
<i>Escherichia coli</i> (ECOLI)	cfu/100 ml
Climate Variables	
Parameter	units
Daily flow	m ³ /s
Rainfall	mm
Temperature	degC

Table 1.
Description of the sites and input parameters.

The mean concentration of *E. coli* is higher at sites DN5 and DN4 than the other four sites. This indicates that these locations have consistently elevated *E. coli* levels, which may suggest a persistent contamination source, such as sewage discharge, agricultural runoff, or reduced water flow leading to stagnation. Descriptive statistics of *E. coli* values (in cfu/100 ml) for all six sites are given in **Table 2**.

Sites DN5 and TU1 also show a higher standard deviation in *E. coli* concentrations. A higher standard deviation means that the *E. coli* levels at these sites fluctuate more significantly over time. This variability could be due to irregular contamination events, seasonal changes, varying pollution sources, or differences in water dynamics affecting bacterial distribution. Among all the sites, NN3 has the lowest mean concentration of *E. coli*, and its maximum recorded value is also the lowest compared to the other sites. This suggests that NN3 is the least contaminated site, likely benefiting from cleaner water sources, better dilution, or effective natural self-purification processes.

The density plots for these sites provide a visual representation of the distribution of *E. coli* concentrations and help to further interpret the statistical findings. Sites DN5 (Mataura), DN4 (Mohaka), and TU1 (Whanganui) frequently experience higher levels of contamination. In contrast, the density plot for NN3 (Wairau) is much narrower and concentrated at lower *E. coli* values, reflecting its lowest mean concentration and the fact that even its maximum recorded value is lower than that of the other sites.

Additionally, the density plots for DN5 and TU1 likely exhibit a wider spread, indicating a greater variability in *E. coli* concentrations. This broader distribution corresponds with the higher standard deviation observed at these sites, suggesting that their *E. coli* levels fluctuate more significantly over time. The density plots showing the distribution of *E. coli* values are plotted in **Figure 2**.

We calculated skewness and kurtosis to analyse the variability in data across all sites. **Table 3** shows these statistical measures across all sites. The skewness and kurtosis values for Wairau are the highest among all the sites. This high value suggests that its distribution is heavily skewed, likely due to the presence of very low values with occasional slightly higher readings. This indicates that *E. coli* concentrations at NN3 are generally low, with a few isolated instances of relatively higher values.

Statistic	Site name					
	DN5	DN4	NN3	HV6	TU2	TU1
Count	185	185	185	160	185	185
Mean	979.27	197.74	23.73	90.31	55.47	336.68
Std	2149.34	471.26	138.50	324.28	182.56	552.29
Min	2	2	0	1	0	0
25%	172.2	18	1	9.7	10.9	59.8
50%	359	37.9	3.2	17.7	17	114.5
75%	878	90.9	9	37.75	32.3	365.4
Max	17,328	2613	1597	3076	1935	3255

Table 2.
 Descriptive measures for *E. coli* values across six sites under study.

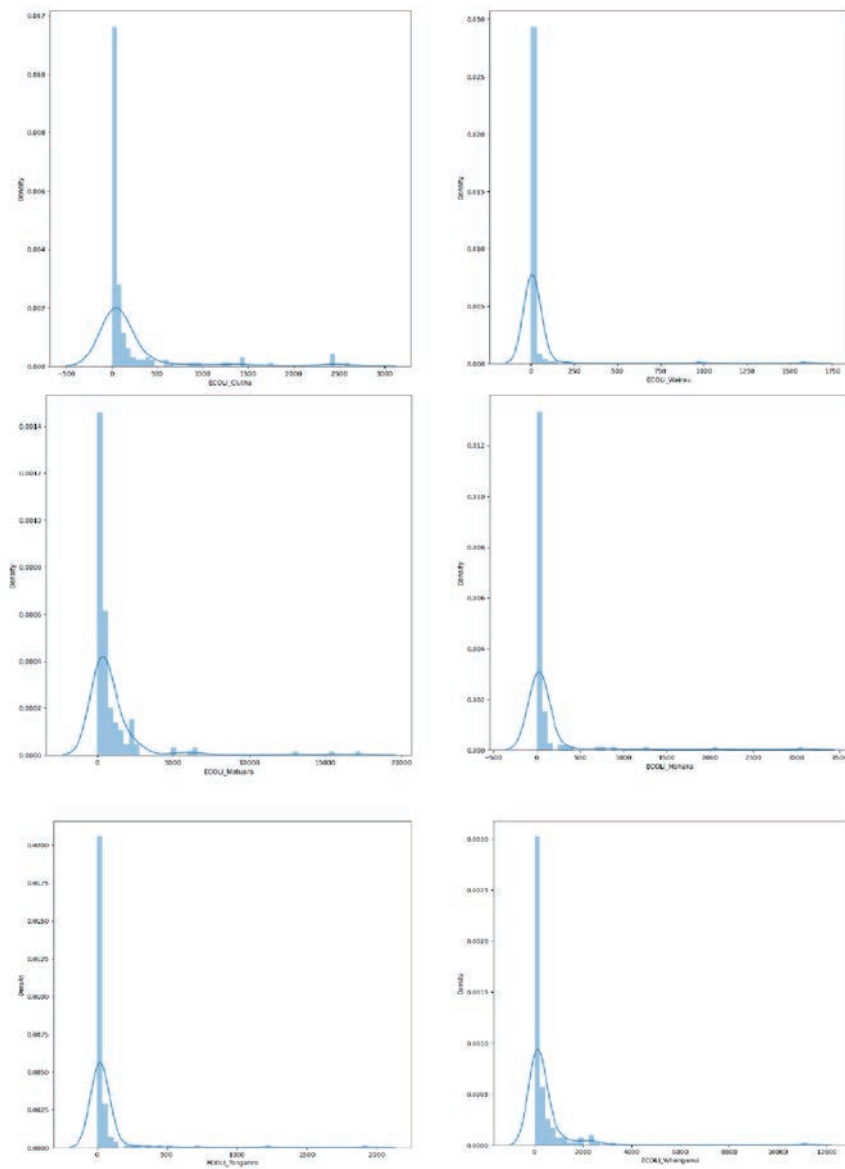


Figure 2.
Density plots for E. coli values across different sites.

Statistical measures	Sites					
	Matauara	Clutha	Wairau	Mohaka	Tongariro	Whanganui
Skewness	5.46	3.68	9.92	6.91	7.83	8.12
Kurtosis	34.44	13.70	103.61	54.38	70.62	85.39

Table 3.
Statistical analysis of data variability across sites.

Kurtosis, on the other hand, measures the “tailedness” of the distribution. A higher kurtosis value for Wairau means that its *E. coli* concentration data is more peaked with heavier tails. This suggests that most of the readings are clustered around very low values, with fewer occurrences of extreme values compared to other sites. The high kurtosis also indicates that *E. coli* levels at this site are more stable, with rare fluctuations in contamination levels.

We also plotted facets of input water quality parameters to visualise and compare trends across different monitoring sites. The facets for all input water quality parameters including *E. coli* for the Clutha River are shown in **Figure 3**.

All sites have variations in water quality parameters. *E. coli* values in some of the sites (Wairau and Tongariro) are very low compared to other sites (Mataura and Clutha). **Figure 4** shows the distribution of *E. coli* values (cfu/100 ml) for all six sites. It is clear from the graph that each site bears very high peaks. It could be due to unexpected or sudden changes in some parameters. To study this, we performed a time series analysis for all the sites.

We created box plots for all sites to visually analyse the distribution, spread, and potential outliers in the data. Box plots help us identify the central tendency, variability, and skewness of the data while highlighting any extreme values. This visualisation allows for a quick comparison across different sites, ensuring a deeper understanding of data consistency and detecting any anomalies or irregular patterns. Using box plots, we observed that site DN5 (Mataura) has high variability and includes a

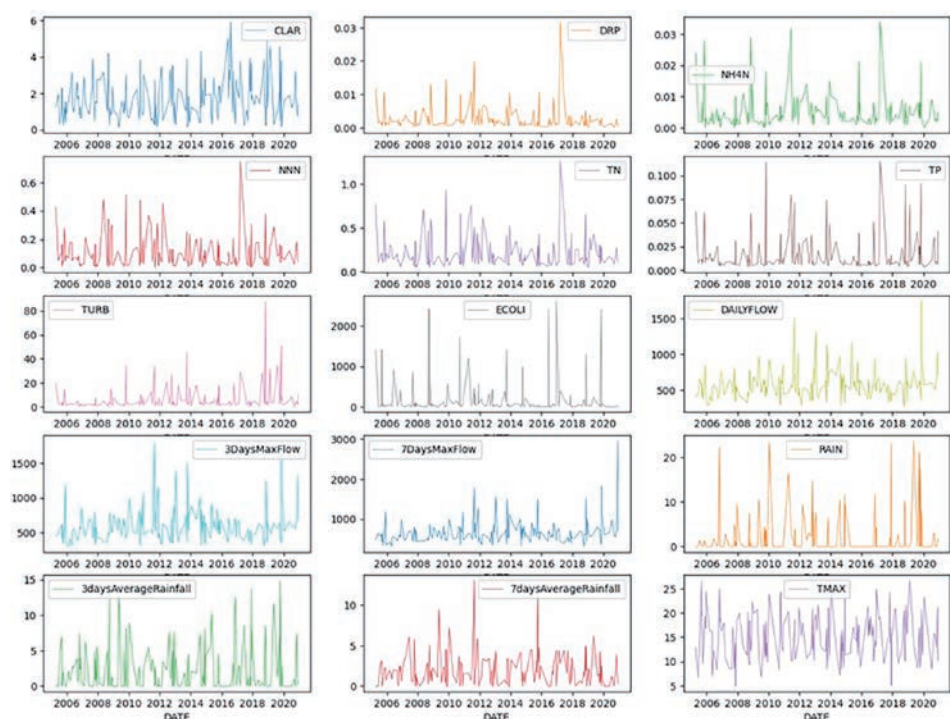


Figure 3.
 Facets for water quality parameters for the Clutha river. Units of the variables are given in **Table 1**.

maximum number of outliers. **Figure 5** illustrates that values of turbidity, *E. coli* and total phosphorus have more variability in river Mataura.

Karl Pearson's coefficient of correlation is calculated to observe the relationship between water quality parameters. Correlation heatmaps for the sites are shown in **Figure 6**.

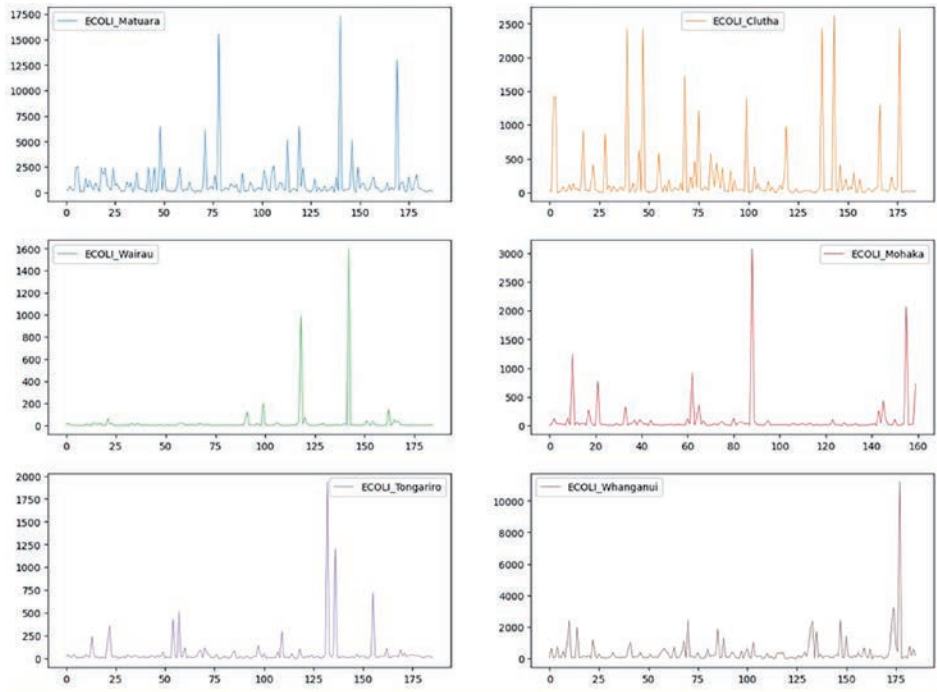


Figure 4.
Monthly *E. coli* concentrations (cfu/100 ml) at six selected sites.

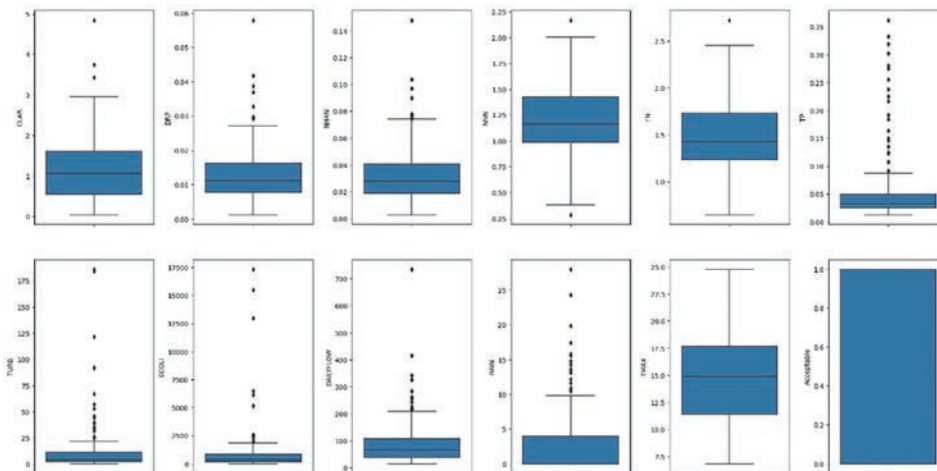


Figure 5.
Box plot for water quality and climate data distribution in the Mataura river.

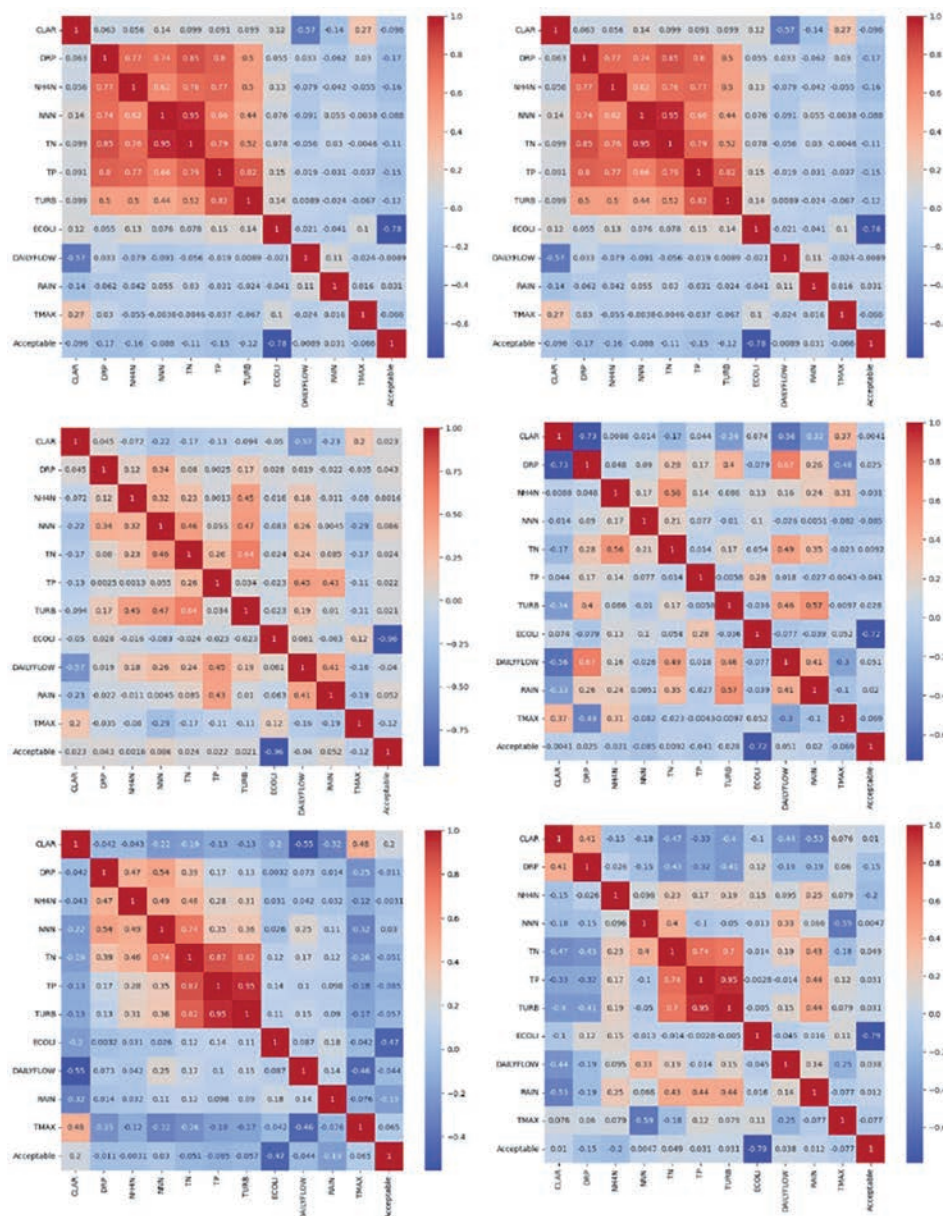


Figure 6.
Heat map for all sites using Pearson's correlation coefficient of the observed water quality parameters.

Land Air Water Aotearoa (LAWA) displays the results of water contamination regularly at recreational sites. **Figure 7** shows the acceptable threshold value of *E. coli* for recreational activities suggested by LAWA.

2.4 Classification algorithms

To check the suitability of water quality for recreational and bathing activities for these sites, we applied five classification algorithms including Logistic Regression

Mode	Trigger level		Management response
	Beach: Enterococci / 100mL	River/Lake: <i>E. coli</i>/100 mL	
 Surveillance	Equal to or less than 140 Enterococci / 100 mL	Equal to or less than 260 <i>E. coli</i> / 100 mL	Routine monitoring.
 Alert	More than 140 Enterococci / 100 mL	More than 260 <i>E. coli</i> / 100 mL	Increase monitoring and investigate source.
 Action	More than 280 Enterococci / 100 mL	More than 540 <i>E. coli</i> / 100 mL	Public warnings if required, increased monitoring and investigation of contaminant source.

Figure 7.

Standard threshold value for *E. coli* concentration suitable for recreational activities in New Zealand. Source: <https://www.lawa.org.nz/learn/factsheets/can-i-swim-here/coastal-and-freshwater-recreational-monitoring>.

(LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost).

Logistic regression is a fundamental statistical and machine learning technique used for binary classification problems. It predicts the probability that an instance belongs to a particular class. The core of logistic regression is the logistic (sigmoid) function:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (1)$$

where z is the input (linear combination of features), e is the base of the natural logarithm, and $\sigma(z)$ outputs a value between 0 and 1.

The logistic regression model can be represented as:

$$P(y = 1|x) = \sigma(w^0 + w^1x^1 + w^2x^2 + \dots + w_nx_n) \quad (2)$$

where $P(y = 1|x)$ is the probability that $y = 1$ given features x . Here w_0 is the bias term (intercept), $w_1, w_2, \dots, w_n$ are the weights and $x_1, x_2, \dots, x_n$ are the feature values. The Gradient Descent method is used for optimisation. The weights are updated using gradient descent as

$$w := w - \frac{\alpha \partial J(w)}{\partial w} \quad (3)$$

where α is the learning rate and $\partial C(w) / \partial w$ is the gradient of the cost function $C(w)$. The learning rate (α) is crucial for successful training. Larger values of (α) may overshoot the minimum values of the function, whereas much smaller values will

result in slow convergence. We chose adaptive learning rates in which we started with a larger learning rate of 0.1 and decreased the rate to 0.01 as training progressed.

In the Decision Tree classifier, we used the Classification and Regression Tree (CART) Gini index to achieve maximum information.

Gini index $G(T)$ measures the impurity of a set of classes and is calculated as

$$G(T) = 1 - \sum_1^c p_i^2 \quad (4)$$

where T is the current state and p_i is the probability of class i .

Random forest algorithm aggregates several decision trees to form predictions. This involves constructing a multitude of decision trees, where each tree is trained on a random collection of the training dataset, and a random selection of features is used to split the data at every level. When predicting, all the trees in the forest provide their output, and the average (in case of regression) or majority vote (in case of classification) is taken.

Mathematically, if we have N decision trees, the final prediction for classification is given by:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_N(x)\} \quad (5)$$

where $T_i(x)$ represents the prediction from the i th tree. For regression, the prediction is computed as

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N T_i(x) \quad (6)$$

where the predictions from all trees are averaged.

Based on the classification report, the total number of acceptable (good for recreational activities) and not acceptable (not good for recreational activities) data points for the six sites are shown in **Figure 8**.

Table 4 presents model selection results using five classification algorithms. The RF algorithm and XGBoost produced the best performance across all metrics. These models are superior in performance metrics including accuracy, F1 score, Precision and Recall. Another ensemble method, DT, provided comparable performance with an accuracy of 98.41%. The remaining two algorithms, LR and SVM, show good predictive performances but lower performance than the other algorithms.

2.5 Model development

2.5.1 Time series analysis

Time series analysis was conducted to examine the temporal patterns, trends, and residual variations in the *E. coli* values across all sites. This analysis allowed us to decompose the data into three key components: trend, seasonality, and residuals. Trend shows the long-term trend in the data, indicating whether values increase, decrease, or remain stable over time. Seasonality focuses on repeating patterns

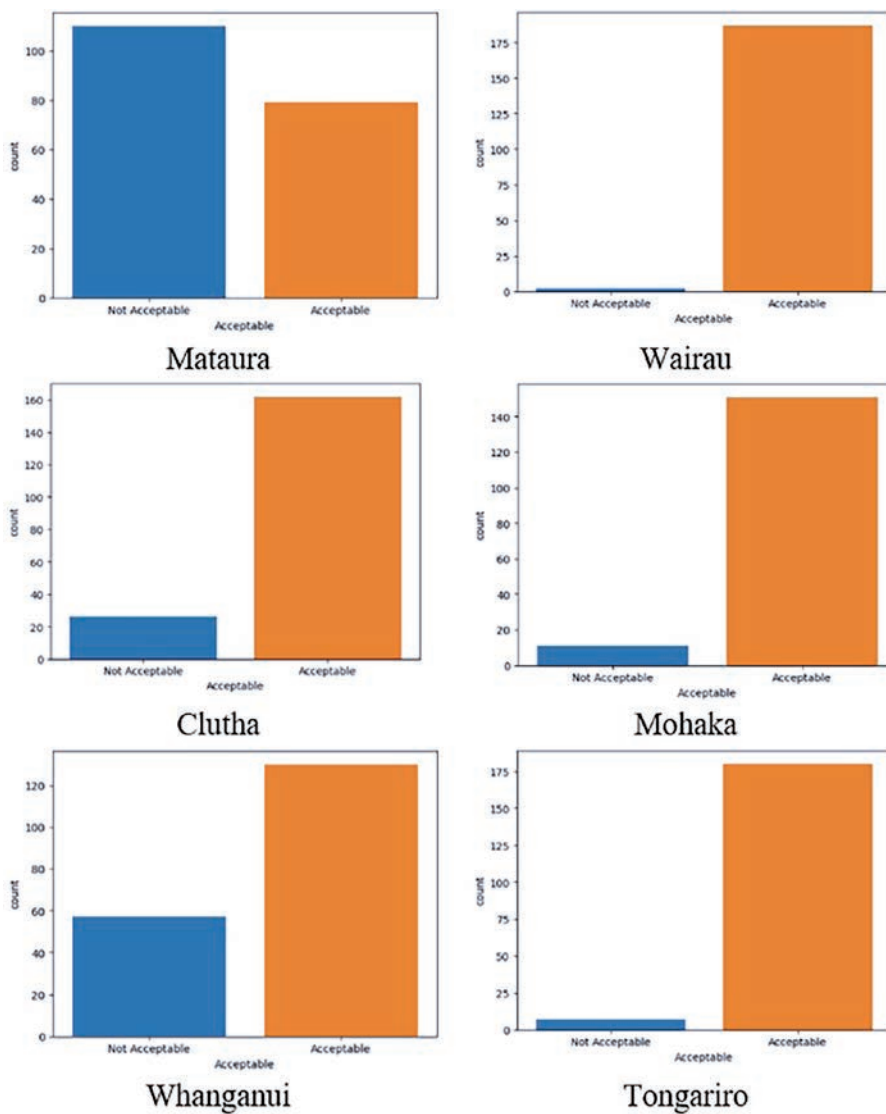


Figure 8.
Bar plot for classification results based on *E. coli* concentration present across all sites.

observed at regular intervals, suggesting periodic fluctuations in the data, and residuals capture the irregular variations that remain after accounting for trend and seasonal effects, helping us assess noise and unexpected changes.

Trends were plotted from year 2005 to year 2020. **Figure 9** illustrates the temporal variations in *E. coli* values across all six sites. From the analysis, we found that seasonal patterns were evident, indicating that the data includes recurring fluctuations at specific time intervals. This suggests that external factors, such as temperature, rainfall and river flow, played a significant role in shaping the data trends. The residuals, while present, showed no dominant pattern, confirming that the primary variations were well captured by the trend and seasonal components. Identifying seasonality helped in predicting future patterns, which is particularly useful for planning and forecasting.

Classification algorithm	Prediction	Precision	Recall	F1-score	Accuracy
Logistic Regression (LR)	0	0.80	0.89	0.84	0.8095
	1	0.83	0.70	0.76	
Decision Tree (DT)	0	1.00	0.97	0.99	0.9841
	1	0.96	1.00	0.98	
Random Forest (RF)	0	1.00	1.00	1.00	1.00
	1	1.00	1.00	1.00	
SVM	0	0.69	0.81	0.74	0.6825
	1	0.67	0.52	0.58	
XGBoost	0	1.00	1.00	1.00	1.00
	1	1.00	1.00	1.00	

Table 4.
 Performances of different classification algorithms for *E. coli* concentration.

E. coli concentrations showed a clear seasonal variation, with higher values observed during warmer months, likely due to increased bacterial growth, runoff, and recreational water use. The three sites, Mataura, Clutha, and Whanganui, have high residuals compared to other sites, which shows that these sites have experienced unexpected changes. These changes are observed due to the past few days (3 days) max flow in river during those instances.

2.5.2 Stationarity analysis

We used the Augmented Dickey-Fuller (ADF) test to determine stationarity. A stationary time series has constant mean, variance, and autocovariance over time, making it easier to model and forecast. The test showed that all the time series under study are stationary. We also examined Autocorrelation function (ACF) and Partial Autocorrelation Function (PACF) as shown in **Figure 10**. The ACF plot helped us identify whether past values influenced future values and to what extent. High autocorrelation at certain lags suggested seasonality in *E. coli* values and helped us in choosing the model parameters. In addition, the PACF plot helped us understand the direct relationship between observations at different lags, excluding the influence of intermediate points. This was particularly useful for selecting the appropriate order for autoregressive (AR) components in our model.

The site Mohaka has no lags, which means that past *E. coli* values at this location do not significantly influence future values. This is due to high variability in rainfall and temperature data.

2.5.3 ARIMA model

For model development, we chose the ARIMA model initially. We used the stepwise fit function to determine the optimal parameters for the ARIMA model. These parameters varied for different sites due to differences in local environmental conditions, pollution sources, hydrological patterns, and data trends unique to each location. Despite applying the ARIMA model, the forecasting results were not highly accurate. This could be attributed to several factors, including the high variability of *E. coli* levels, external

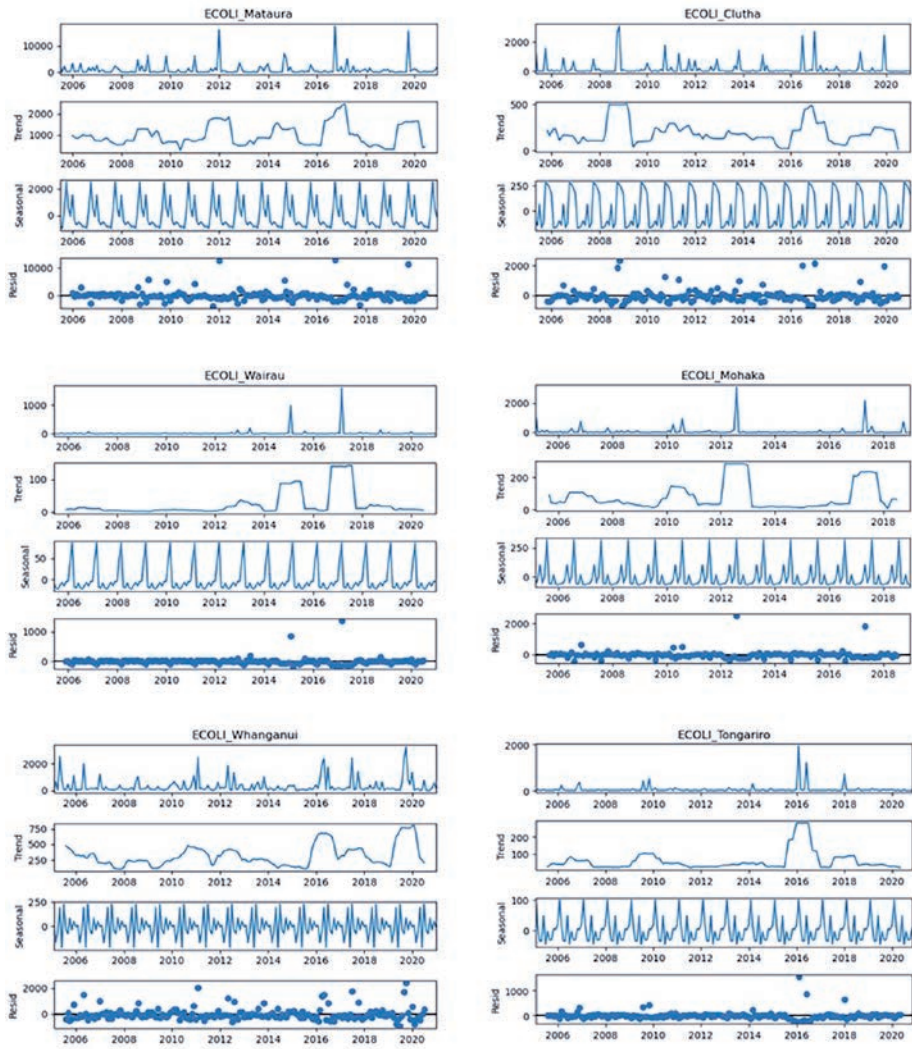


Figure 9.
Time series decomposition of *E. coli* concentration across all sites.

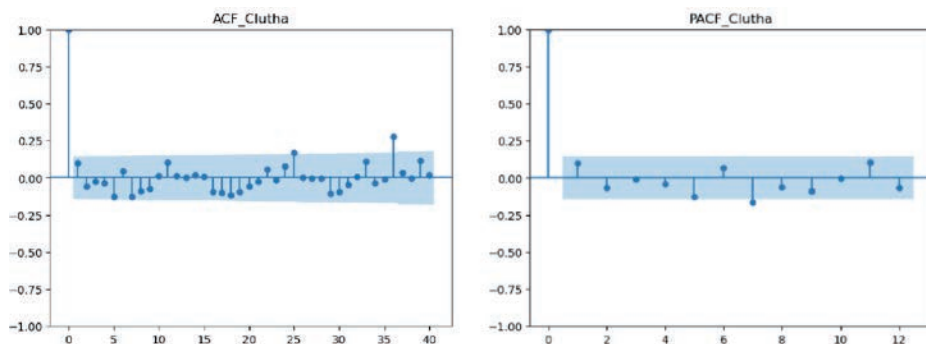


Figure 10.
Autocorrelation function (ACF) and partial autocorrelation function (PACF) for the Clutha river.

environmental influences such as rainfall and temperature variations, and possible data limitations. Additionally, *E. coli* concentrations may be influenced by non-linear and complex interactions that ARIMA, a linear model, may not fully capture.

2.5.4 SARIMAX model

Since the ARIMA model did not provide accurate results, we implemented the Seasonal Auto Regressive Integrated Moving Average with Exogenous variables (SARIMAX) model. SARIMAX extends ARIMA by incorporating seasonality and external regressors, making it more effective in capturing environmental influences on *E. coli* levels. We used other water quality parameters such as rainfall, 3-day average rainfall, total phosphorus, total nitrogen, turbidity, as an external regressor and found that no one water quality parameter accurately predicts *E. coli* levels. It is different for each site, but total phosphorus is common for the five sites except Whanganui. These external variables can help improve the accuracy of forecasts by accounting for additional information that might impact the time series. SARIMAX models combine the concepts of SARIMA models with the ability to include exogenous variables. The main components of SARIMAX models are as follows:

Seasonal Autoregressive (SAR) Component (P): Captures the relationship between the current value and past values at the same seasonal lag.

Seasonal Integrated (SI) Component: Involves differencing the series at the seasonal interval to achieve seasonal stationarity.

Seasonal Moving Average (SMA) Component (Q): Models the relationship between the current value and past error terms at the same seasonal lag.

Exogenous (X) Variables: These are external factors that might influence the time series. Including exogenous variables helps the model capture additional patterns and relationships beyond the inherent time series components.

SARIMAX model is represented as $SARIMAX(p, d, q) \times (P, D, Q, s)$, where p is the order of the non-seasonal AutoRegressive (AR) component, d is the degree of non-seasonal differencing, q is the order of the non-seasonal moving average (MA) component. In addition, D is the degree of seasonal differencing and s is the number of time steps in each seasonal period. Fitting a SARIMAX model with exogenous variables involves not only identifying the appropriate orders for the seasonal and non-seasonal components but also selecting and incorporating relevant exogenous variables.

3. Results and discussion

Using time series analysis, we predicted *E. coli* concentrations for the next 4 years based on historical data from 2005 to 2020. The predictions provided valuable insights into future trends. In this analysis, we found that total phosphorus (TP) serves as an exogenous variable for all sites except the Mohaka, for which 3-day average rainfall is the exogenous variable. The predicted *E. coli* values followed similar seasonal and long-term patterns observed in the historical dataset. This indicates that our model successfully captured the periodic nature of *E. coli* fluctuations, including seasonal peaks and dips. The model treated the extreme values as outliers. Predicted *E. coli* values along with historic data are plotted in **Figure 11**.

Sites with seasonal variations in past data showed similar trends in future projections, reinforcing the model's effectiveness in detecting cyclic behaviour. The

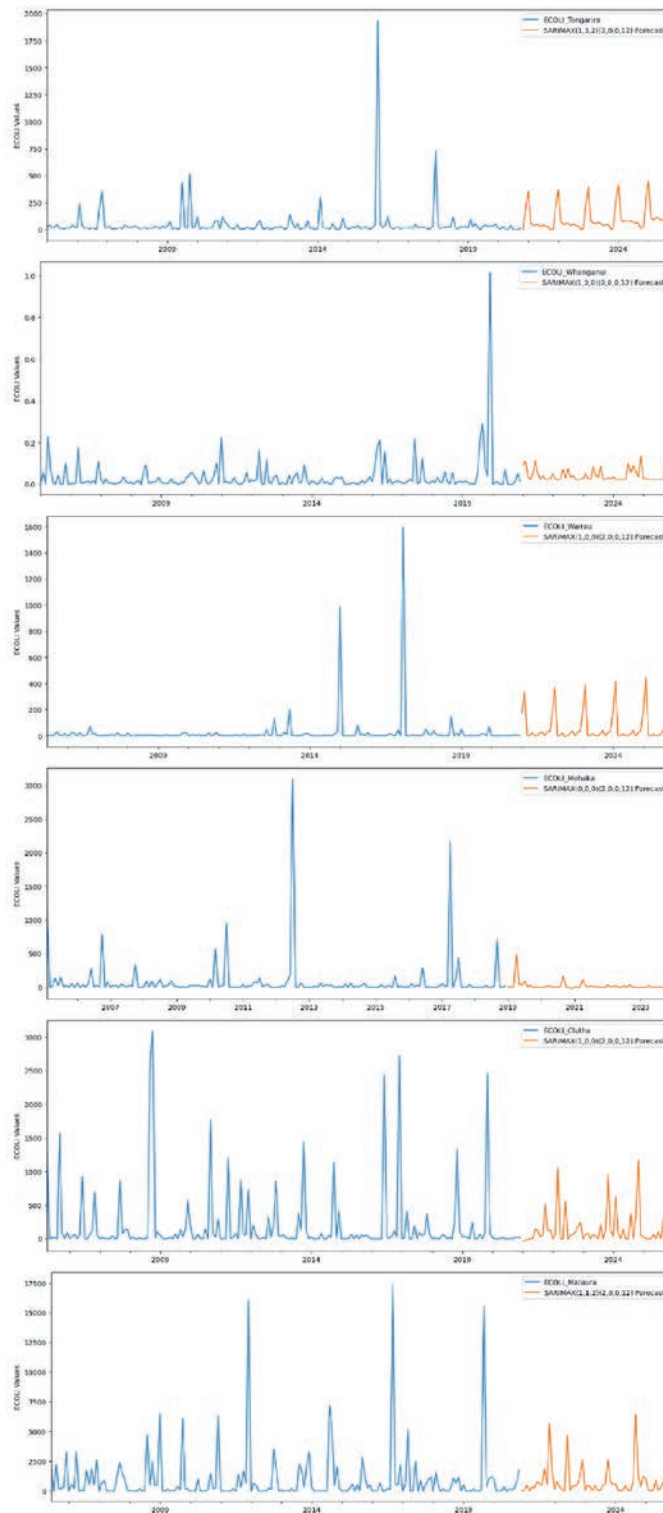


Figure 11.
Forecasting for E. coli concentration across all sites from year 2021 to year 2024.

forecasted values were evaluated using statistical metrics, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). These metrics indicated an accuracy of 86% for the model, confirming that this model effectively captured the underlying patterns of *E. coli* variations. Cross-validation with historical data showed that the model was able to reproduce past trends with minimal error. Furthermore, the projected *E. coli* concentrations did not show unrealistic spikes, suggesting that the model did not overfit to noise in the data. The predicted values remained within the expected range observed in past records, making them more reliable for practical use in water quality monitoring and risk assessment.

However, an important point to note here is that while overall trends were consistent, the rate of increase or decrease in *E. coli* concentrations varied across different sites. This highlights the influence of site-specific environmental factors, which are rainfall, temperature, river flow and chemical variables such as TN, TP, and NH₄N. The Clutha River and the Whanganui River showed greater variability, while others exhibited more stable forecasts, suggesting the need for localised calibration and continuous monitoring.

4. Limitations

Traditional ARIMA and SARIMAX models primarily rely on historical patterns and assume that future trends will follow similar behaviour, which may not always hold true in highly variable environmental systems such as heavy rainfall, land usage and pollution events. *E. coli* concentrations in natural water bodies are affected by multiple external factors, including temperature, precipitation, agricultural runoff, and wastewater discharge. Although we incorporated temperature, predictions were not consistent across all sites.

5. Conclusion

This study provided valuable insights into predicting *E. coli* concentrations across six sites in New Zealand. By analysing historical data from 2005 to 2020, we applied ARIMA and SARIMAX models to forecast *E. coli* levels for the next 4 years. These findings demonstrated that while the models captured overall trends and seasonal variations, their accuracy was site-dependent, and external environmental factors played a significant role in influencing contamination patterns.

The presence of elevated *E. coli* levels in recreational waters poses health risks, as it indicates possible faecal contamination, which can lead to waterborne illnesses. Reliable forecasting models can help authorities implement timely interventions, such as issuing public health advisories, monitoring pollution sources, and mitigating contamination risks. This study highlighted some key challenges.

No single water quality parameter, except total phosphorus and 3-day average rainfall, could consistently predict *E. coli* levels across all sites, indicating the complexity of microbial contamination. While historical trends were useful in forecasting, external factors, such as rainfall and temperature, introduced significant variability. Site-specific differences in *E. coli* behaviour suggest that a one-size-fits-all approach is not ideal for water quality forecasting. More localised models or regional calibration techniques may be necessary to improve forecast reliability for specific locations. Regional councils may utilise real-time sensor technology and remote

sensing data to capture rapid changes in water quality, and the frequency of water sampling can be increased to improve model resolution.

Future research should aim to enhance predictive models by integrating real-time environmental data, using advanced modelling techniques, and tailoring approaches to site-specific conditions. These improvements will strengthen forecasting accuracy, making it a valuable tool for safeguarding recreational water quality and minimising health risks from waterborne pathogens.

Conflict of interest

The authors declare no conflict of interest.

Author details

Parul Tiwari^{1*}, Channa Rajanayaka² and Jing Yang²

1 Department of Mathematical Sciences, Auckland University of Technology,
New Zealand

2 National Institute of Water and Atmospheric Research, Christchurch, New Zealand

*Address all correspondence to: parul.tiwari@aut.ac.nz

IntechOpen

© 2025 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] He C, Liu Z, Wu J, Pan X, Fang Z, Li J, et al. Future global urban water scarcity and potential solutions. *Nature Communications*. 2021;**12**:1. Available from: <https://www.nature.com/articles/s41467-021-25026-3>
- [2] Australian Drinking Water Guidelines. Version 3.8. Australia: National Health and Medical Research Council (NHMRC); Sept 2022. Available from: <https://www.nhmrc.gov.au/about-us/publications/australian-drinking-water-guidelines>
- [3] Kumar D, Kumar R, Sharma M, Awasthi A, Kumar M. Global water quality indices: Development, implications, and limitations. *Total Environment Advances*. 2024;**9**:200095
- [4] World Health Organization. Guidelines for safe recreational water. Volume 1. Coastal and fresh waters. Geneva. 2003;**1**:219. Available from: http://www.who.int/water_sanitation_health/bathing/srwe2full.pdf
- [5] Ziemińska-Stolarska A, Skrzypski J. Review of mathematical models of water quality. *Ecological Chemistry and Engineering S*. 2012;**19**(2):197-211
- [6] Tiwari PK, Singh RK, Khajanchi S, Kang Y, Misra AK, Tiwari PK, et al. A mathematical model to restore water quality in urban lakes using Phoslock. *Discrete and Continuous Dynamical Systems – B*. 2021;**26**(6):3143-3175. Available from: <https://www.aims sciences.org/en/article/doi/10.3934/dcdsb.2020223>
- [7] Yadav RR, Roy J. Analytical solutions of one-dimensional scale dependent advection-dispersion equations for finite domain solute transport. *Groundwater for Sustainable Development*. 2022;**16**:100712
- [8] Sanskrityayn A, Suk H, Chen JS, Park E. Generalized analytical solutions of the advection-dispersion equation with variable flow and transport coefficients. *Sustainability*. 2021;**13**(14):7796. Available from: <https://www.mdpi.com/2071-1050/13/14/7796/html>
- [9] Abimbola OP, Mittelstet AR, Messer TL, Berry ED, Bartelt-Hunt SL, Hansen SP. Predicting *Escherichia coli* loads in cascading dams with machine learning: An integration of hydrometeorology, animal density and grazing pattern. *Science of The Total Environment*. 2020;**722**:137894
- [10] Kapinusova G, Lopez Marin MA, Uhlik O. Reaching unreachable: Obstacles and successes of microbial cultivation and their reasons. *Frontiers in Microbiology*. 2023;**14**:1089630. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10027941/>
- [11] Ahmed U, Mumtaz R, Anwar H, Shah AA, Irfan R, García-Nieto J. Efficient water quality prediction using supervised machine learning. *Water*. 24 Oct 2019;**11**(11):2210, 14
- [12] Wen X, Chen F, Lin Y, Zhu H, Yuan F, Kuang D, et al. Microbial indicators and their use for monitoring drinking water quality—A review. *Sustainability*. 2020;**12**(6):2249. Available from: <https://www.mdpi.com/2071-1050/12/6/2249/html>
- [13] Ebomah KE, Adefisoye MA, Okoh AI. Pathogenic *Escherichia coli* strains recovered from selected aquatic resources in the Eastern Cape,

South Africa, and its significance to public health. International Journal of Environmental Research and Public Health. 2018;**15**(7). Available from: <https://pubmed.ncbi.nlm.nih.gov/30018212/>

[14] Delair Z, Schoeman M, Reyneke B, Singh A, Barnard TG. Assessing the impact of *Escherichia coli* on recreational water safety using quantitative microbial risk assessment. Journal of Water and Health. 2024;**22**(10):1781-1793. Available from: <http://iwaponline.com/jwh/article-pdf/22/10/1781/1499046/jwh2024081.pdf>

[15] Efting AA, Snow DD, Fritz SC. Cyanobacteria and microcystin in the Nebraska (USA) Sand Hills lakes before and after modern agriculture. Journal of Paleolimnology. 2011;**46**(1):17-27

[16] Pandey PK, Kass PH, Soupir ML, Biswas S, Singh VP. Contamination of water resources by pathogenic bacteria. AMB Express. 2014;**4**(1):1-16. Available from: <https://amb-express.springeropen.com/articles/10.1186/s13568-014-0051-x>

[17] Stocker MD, Pachepsky YA, Hill RL. Prediction of *E. Coli* concentrations in agricultural pond waters: Application and comparison of machine learning algorithms. Front. Artificial Intelligence. 2022;**4**:768650. Available from: www.frontiersin.org

[18] Dwivedi D, Mohanty BP, Lesikar BJ. Estimating *Escherichia coli* loads in streams based on various physical, chemical, and biological factors. Water Resources Research. 2013;**49**(5):2896. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC3914718/>

[19] Baffaut C, Benson VW. A bacteria TMDL for Shoal Creek using SWAT modeling and DNA source tracking. In:

Total Maximum Daily Load (TMDL) Environmental Regulations II. American Society of Agricultural and Biological Engineers; 2003. p. 1

[20] Coffey R, Benham B, Wolfe ML, Cummins E. Predicting the effects of environmental change on microbial transport. In: ASABE 1st Climate Change Symposium: Adaptation and Mitigation. 2015. pp. 219-221

[21] Eregno FE, Tryland I, Tjomsland T, Kempa M, Heistad A. Hydrodynamic modelling of recreational water quality using *Escherichia coli* as an indicator of microbial contamination. Journal of Hydrology. 2018;**561**:179-186

[22] Kuroki S, Ogata R, Sakamoto M. Predicting the presence of *E. coli* in tap water using machine learning in Nepal. Water and Environment Journal. Aug 2023;**37**(3):402-411

[23] Cyterski M, Shanks OC, Wanjugi P, McMinn B, Korajkic A, Oshima K, et al. Bacterial and viral fecal indicator predictive modeling at three Great Lakes recreational beach sites. Water Research. 2022;**223**:118970. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9724166/>

[24] Francy DS, Stelzer EA, Duris JW, Brady AMG, Harrison JH, Johnson HE, et al. Predictive models for *Escherichia coli* concentrations at inland lake beaches and relationship of model variables to pathogen detection. Applied and Environmental Microbiology. 2013;**79**(5):1676-1688

[25] Herrig IM, Böer SI, Brennholt N, Manz W. Development of multiple linear regression models as predictive tools for fecal indicator concentrations in a stretch of the lower Lahn River, Germany. Water Research. 2015;**85**:148-157. Available from: <https://pubmed.ncbi.nlm.nih.gov/26318647/>

- [26] Park Y, Cho KH, Park J, Cha SM, Kim JH. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. *Science of The Total Environment*. 2015;**502**:31-41
- [27] van der Meulen ES, Tertienko A, Blauw AN, Sutton NB, van de Ven FHM, Rijnaarts HHM, et al. A review of prediction models for *E. Coli* in urban surface waters. *Urban Water Journal*. 2024;**21**(5):539-548. Available from: <https://www.tandfonline.com/doi/abs/10.1080/1573062X.2024.2313634>
- [28] Bhardwaj P, Tiwari P, Olejar K, Parr W, Kulasiri D. A machine learning application in wine quality prediction. *Machine Learning with Applications*. 2022;**8**:100261
- [29] Tyagi A, Tiwari P, Bhardwaj P, Chawla H. Prognosis of sexual dimorphism with unfused hyoid bone: Artificial intelligence informed decision making with discriminant analysis. *Science and Justice*. 2021;**61**(6):789-796
- [30] Tiwari P, Bhardwaj P, Keprate A, Tyagi A. Breast cancer survival prediction using machine learning. *Studies in Computational Intelligence*. 2022;**1016**:143-158
- [31] Zhu M, Wang J, Yang X, Zhang Y, Zhang L, Ren H, et al. A review of the application of machine learning in water quality evaluation. *Eco-Environment and Health*. 2022;**1**(2):107. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10702893/>
- [32] Tiwari P, Bhardwaj P, Somin S, Parr WV, Harrison R, Kulasiri D. Understanding quality of pinot noir wine: Can modelling and machine learning pave the way? *Food*. 2022;**11**(19):3072. Available from: <https://www.mdpi.com/2304-8158/11/19/3072/html>
- [33] Brooks W, Corsi S, Fienen M, Carvin R. Predicting recreational water quality advisories: A comparison of statistical methods. *Environmental Modelling and Software*. 2016;**76**:81-94
- [34] Vidal V, Sampognaro L, de León F, Kruk C, Perera G, Crisci C, et al. A critical review of model construction and performance for nowcast systems for faecal contamination in recreational beaches. *Science of the Total Environment*. 2024;**954**:176233. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0048969724063897>
- [35] Olawade DB, Wada OZ, Ige AO, Egbewole BI, Olojo A, Oladapo BI. Artificial intelligence in environmental monitoring: Advancements, challenges, and future directions. *Hygiene and Environmental Health Advances*. 2024;**12**:100114
- [36] Shams MY, Elshewey AM, El-kenawy ESM, Ibrahim A, Talaat FM, Tarek Z. Water quality prediction using machine learning models based on grid search method. *Multimedia Tools and Applications*. 2024;**83**(12):35307-35334. Available from: <https://link.springer.com/article/10.1007/s11042-023-16737-4>
- [37] Farrell ML, Joyce A, Duane S, Fitzhenry K, Hooban B, Burke LP, et al. Evaluating the potential for exposure to organisms of public health concern in naturally occurring bathing waters in Europe: A scoping review. *Water Research*. 2021;**206**:117711
- [38] Russo GS, Eftim SE, Goldstone AE, Dufour AP, Nappier SP, Wade TJ. Evaluating health risks associated with exposure to ambient surface waters during recreational activities: A systematic review and

meta-analysis. Water Research. 2020;**176**:115729. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10287035/>

[39] Smith DG, Maasdam R. New Zealand's national river water quality network, design and physico-chemical characterisation. New Zealand Journal of Marine and Freshwater Research. 1994;**28**(1):19-35

[40] Tait A, Henderson R, Turner R, Zheng XG. Thin plate smoothing spline interpolation of daily rainfall for New Zealand using a climatological rainfall surface. International Journal of Climatology. 2006;**26**:2097-2115

[41] Snelder TH, Biggs BJF, Woods RA. Improved eco-hydrological classification of rivers. River Research and Applications. 2005;**21**:609-628. DOI: 10.1002/rra.826

[42] Leathwick JR, West D, Chadderton L, Gerbeaux P, Kelly D, Robertson H. Freshwater Ecosystems of New Zealand (FENZ) Geodatabase: Version One – August 2010 – User Guide. New Zealand: Department of Conservation; 2010. pp. 1-51