Don't Swim in Data: Real-Time Microbial Forecasting for New Zealand Recreational Waters

Asif Juzar Cheena[a,*], Katharina Dost[b], Theo Sarris[d], Nina Straathof[c], Jörg Simon Wicker[a]

[a]*School of Computer Science, The University of Auckland, Auckland, New Zealand*

[b]*Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia*

[c]*Statistics & Data, Environment Canterbury, Christchurch, New Zealand*

[d]*Institute for Environmental Science and Research, Christchurch, New Zealand*

Abstract

Traditional water quality monitoring, which relies on infrequent sampling and 48-hour laboratory delays, fails to capture rapid contamination fluctuations, leaving recreational water users exposed to health risks. To address this critical gap, we developed two novel machine learning frameworks for real-time forecasting of Enterococci concentrations in Canterbury, New Zealand.

The **Probabilistic Forecasting Framework** employs an ensemble of quantile regression models (covering the 5th to 98th percentiles), a gradient boosting meta-learner, and Conformalized Quantile Regression (CQR) to produce both accurate point forecasts and calibrated 90% prediction intervals. This approach captures the full range of contamination scenarios, enabling proactive, risk-based water quality management.

The **Matrix Decomposition Framework** uses Non-negative Matrix Factorization (NMF) to separate complex spatio-temporal water quality data into interpretable latent factors, which are then modeled with multi-target Random Forests. This method enhances interpretability and generalization, particularly for new monitoring sites with limited historical data.

Evaluated on a comprehensive dataset (2021–2024, 15 sites, 1047 samples, 100 exceedance events), the Probabilistic Framework achieved an overall exceedance sensitivity of 67.0% (rising to 75.7% in 2023–2024), a precautionary sensitivity of 77.0%, and a specificity of 92.3%, with a WMAPE of 17.2% during exceedance events. The Matrix Decomposition Framework delivered comparable performance, with an exceedance sensitivity of 61.0%, a precautionary sensitivity of 74.0%, a specificity of 90.6%, and a WMAPE of 20.3%. Together, these

frameworks not only exceed USGS guidelines but also outperform traditional operational methods and standard ML benchmarks (e.g., linear regression, logistic regression, decision trees, and multi-layer perceptrons), while displaying highly competitive performance relative to state-of-the-art systems such as Auckland's Safeswim.

SHAP analysis confirmed that short-term rainfall and wind conditions are the primary drivers of contamination, aligning with hydrological principles. A complete forecasting system—comprising a real-time data pipeline with automated validation and an interactive analytics dashboard—has been deployed in a staging environment, demonstrating both operational feasibility and the potential for broader applications in environmental risk management.

Keywords: water quality forecasting, Enterococci, probabilistic forecasting, quantile regression, matrix decomposition, Non-negative Matrix Factorization, NMF, machine learning, SHAP, explainable AI, real-time monitoring, uncertainty quantification, environmental modeling, time series, spatial analysis, gradient boosting, LightGBM, Random Forest, multi-target regression, Canterbury, New Zealand

1. Introduction

Coastal recreational waters offer immense value but can pose significant public health risks when contaminated. Traditional monitoring—based on infrequent sampling and delayed analysis—often fails to capture rapid changes in Enterococci levels, a key indicator of fecal contamination.

To address this, we propose two novel forecasting frameworks for real-time prediction of Enterococci concentrations. The *Probabilistic Forecasting Framework* employs an ensemble of quantile regression models with a meta-learning strategy to produce risk-aware point forecasts and calibrated prediction intervals. In contrast, the *Matrix Decomposition Framework* uses Non-negative Matrix Factorization (NMF) to separate spatial and temporal patterns, enhancing interpretability and generalization.

Our evaluation using extensive field data from Canterbury demonstrates that both frameworks outperform conventional methods by achieving higher sensitivity to exceedance events, robust uncertainty quantification, and improved interpretability (via SHAP analysis). Notably, the Probabilistic Framework excels in detecting high-risk events, while the Matrix Decomposition Framework captures transferable spatial patterns effectively.

By integrating advanced machine learning with domain insights, our work paves the way for proactive water quality management. The remainder of this paper is organized as follows:

*Corresponding author: ache234@aucklanduni.ac.nz

Section 3 details the methodological foundations and model architectures; Section 5 presents experimental evaluations; and Section 10 discusses implications and future directions.

2. Literature Review

Monitoring recreational water quality is critical for public health, yet traditional culture-based methods suffer from 24–48-hour delays, infrequent sampling, and low detection rates for high-risk events [24, 2, 21]. In contrast, machine learning (ML) offers real-time predictions using environmental data but faces challenges in sensitivity, uncertainty quantification, interpretability, and handling new sites.

2.1. Traditional Water Quality Monitoring

Traditional methods rely on laboratory analysis of *Enterococci*, but are limited by delayed results, infrequent sampling, high resource demands, and an inability to capture dynamic environmental conditions [21, 11, 2].

2.2. Machine Learning for Water Quality Prediction

ML models, including Gradient Boosting Machines and neural networks, have been applied for dynamic forecasting with mixed results—moderate sensitivity and issues with interpretability or data requirements [33, 5, 6, 32]. Common challenges include skewed data distributions, poor separation of spatial and temporal drivers, lack of probabilistic outputs, and cold-start problems [22, 5, 27, 14, 21].

2.3. Probabilistic Forecasting and Matrix Factorization

Probabilistic methods, such as quantile regression and conformal prediction, provide calibrated uncertainty estimates essential for risk-based decisions but are underutilized in water quality forecasting [4, 15]. Matrix Factorization (MF) techniques decompose high-dimensional data into latent factors, effectively disentangling spatial and temporal patterns and improving generalizability and performance on new sites [9, 36, 35].

2.4. Research Gaps and Opportunities

Current models inadequately quantify uncertainty, disentangle spatiotemporal effects, and adapt to new monitoring sites. Future work should integrate empirically validated prediction intervals, leverage MF for robust latent factor extraction, and enhance interpretability with tools like SHAP, aligning predictions with hydrological principles.

3. Methodology

This chapter outlines our systematic approach to nowcasting microbial contamination in recreational waters by integrating hydrological insights with advanced machine learning. Our methodology centers on two novel forecasting frameworks, supported by concise data handling, evaluation, and deployment strategies.

3.1. Dataset Overview and Preprocessing

We utilize a comprehensive dataset from 15 recreational water sites in Lyttelton and Akaroa harbours, Canterbury, New Zealand, spanning 11 summer sampling seasons (2013–2024). The dataset includes Water Quality Data (4100 samples measuring Enterococci in MPN/100mL during peak recreational hours), Hydro-Meteorological Data (hourly rainfall, wind speed/direction, air temperature, and tide predictions), and Site Metadata (geographical, catchment, and usage characteristics).

Preprocessing involves several steps: Target Transformation using logarithmic or square root transforms to mitigate the right-skewed, log-normal distribution of Enterococci; Missing Data Imputation by applying Last Observation Carried Forward (LOCF) for gaps up to 6 hours and a 12-week rolling mean for longer gaps; and Categorical Encoding and Scaling using sine/cosine transforms for temporal cyclic features, one-hot encoding for other categorical data, and min-max or standard scaling as required by the model.

A summary table (Table 1) details the key time-varying and site-static features used for forecasting.

3.2. Forecasting Frameworks

Our contributions focus on two frameworks: a *Probabilistic Forecasting Framework* and a *Matrix Decomposition Forecasting Framework*.

3.2.1. Probabilistic Forecasting Framework

This framework overcomes the limitations of traditional point forecasts by modeling the full conditional distribution of Enterococci levels. It consists of three streamlined stages:

Stage 1: Quantile Modeling: An ensemble of LightGBM [18] quantile regression models estimates ten non-uniformly spaced quantiles (e.g., $\tau \in \{0.05, 0.30, \ldots, 0.98\}$) using the asymmetric pinball loss to emphasize accurate detection of high-risk events.

Category	Feature Name	Type	Time Varying
Rainfall	Accumulated Rainfall (3/6/12/24/48/72H)	Numerical	Yes
	Rainfall Intensity (3/6/12/24/48/72H)	Numerical	Yes
	Consecutive Rainfall Hours	Numerical	Yes
	Consecutive Dry Hours	Numerical	Yes
Wind	Average Wind Speed (3/6/12/24H)	Numerical	Yes
	Average Wind Direction (3/6/12/24H)	Numerical	Yes
	Wind-Shore Type (3/6/12/24H)	Categorical	Yes
Tide	Tidal State	Binary	Yes
	Tidal Height	Numerical	Yes
	Time to Nearest Tide	Numerical	Yes
Temperature	Daily Temperature Range	Numerical	Yes
	Peak Daily Temperature	Numerical	Yes
Temporal	Hour of Day	Cyclic	Yes
	Day of Week	Cyclic	Yes
	Month	Cyclic	Yes
	Week of Year	Cyclic	Yes
	Weekend	Binary	Yes
	Holiday Period	Binary	Yes
	Sampling Period	Categorical	Yes
	Season	Categorical	Yes
Historical	Site Season Average	Numerical	Yes
	Site Historical Exceedance Rate	Numerical	Yes
Physical	Site Identifier	Categorical	No
	Harbour Identifier	Categorical	No
	Beach Orientation	Cyclic	No
	Site Coordinates	Numerical	No
	Catchment Slope	Categorical	No
	Bay/Valley Width	Categorical	No
	Water Depth	Categorical	No
Environmental	Soil Type	Categorical	No
	Land Cover	Categorical	No
	Waterway Inputs	Numerical	No
	Stormwater Outlets	Binary	No
	Sewage Discharge	Binary	No
	Beach Agriculture	Binary	No
Usage	Site Popularity	Categorical	No
	Watercraft Activity	Binary	No

Table 1: Model Feature Set Overview

Stage 2: Point Forecast and Calibration: A secondary Gradient Boosting Machine (meta-learner) synthesizes out-of-fold quantile predictions into a robust point estimate. Simultaneously, Conformalized Quantile Regression (CQR) refines prediction intervals to ensure 90% coverage.

Stage 3: Post-Processing: A monotonic sorting step reorders the quantiles and adjusts the point forecast to lie within the calibrated interval.

Figure 1 illustrates the streamlined pipeline.

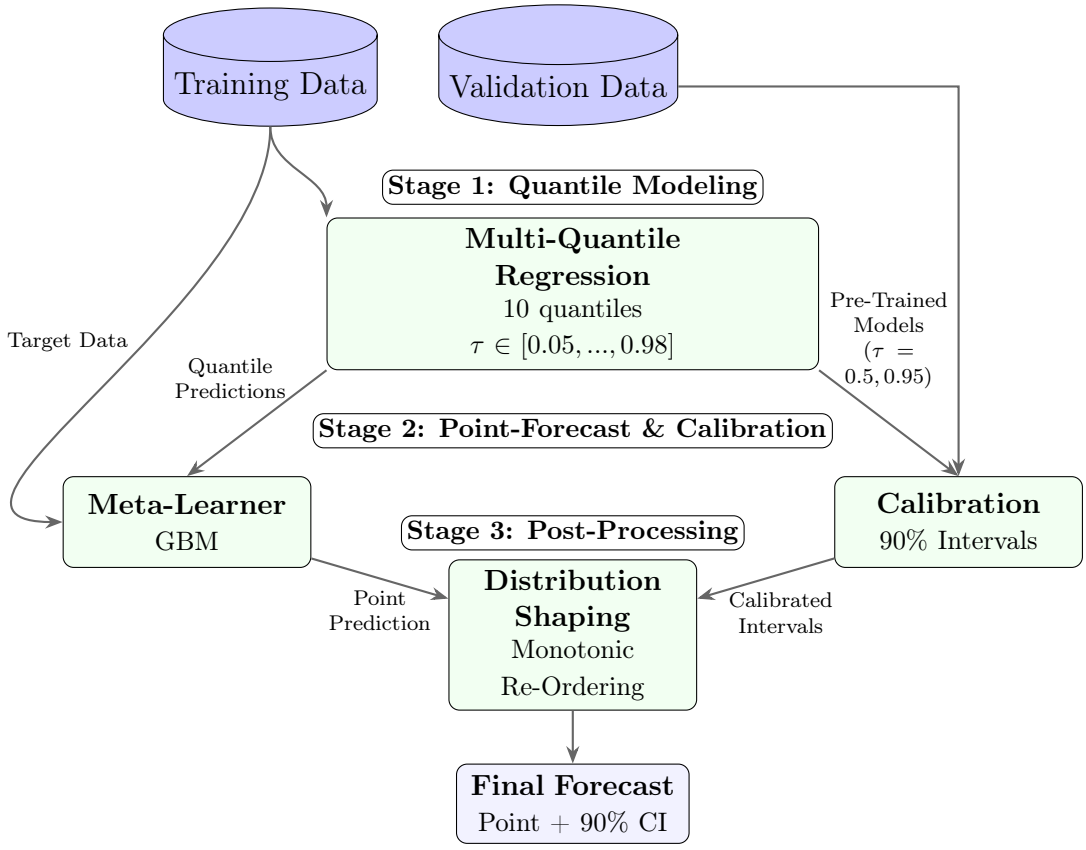

Figure 1: Probabilistic Forecasting Framework Pipeline

3.2.2. Matrix Decomposition Forecasting Framework

To capture complex spatio-temporal dynamics and address cold-start challenges, we decompose the high-dimensional Enterococci data matrix $X \in \mathbb{R}^{m \times n}$ using Non-negative Matrix Factorization (NMF):

$$X \approx WH,$$

where $W \in \mathbb{R}^{m \times k}$ and $H \in \mathbb{R}^{k \times n}$ capture temporal and spatial patterns, respectively.

This framework comprises three stages:

Stage 1: Decomposition: NMF decomposes X into W and H with a non-negativity constraint for interpretability.

Stage 2: Latent Factor Learning: Two multi-target Random Forest models predict the latent factors from time-varying features (F_{time}) and site-static features (F_{sites}), respectively.

Stage 3: Reconstruction: The predicted factors $\hat{W}$ and $\hat{H}$ are multiplied to reconstruct the forecasted concentration matrix $\hat{X}$, which is then transformed back to the original scale.

A detailed schematic is provided in Figure 2.

3.3. Benchmark Models and Hyperparameter Optimization

For comparison, we implement a suite of benchmark models including naïve statistical baselines that mimic current operational methods, Generalized Linear Models (both linear and logistic regression), a decision tree, and a Multi-Layer Perceptron. Hyperparameters for all models are optimized using the Tree-structured Parzen Estimator (TPE) via Optuna [3], employing nested cross-validation (time-series and leave-one-site-out) to ensure robust generalization. Detailed search spaces and configurations are provided in the supplement.

3.4. Evaluation Framework

Our evaluation framework combines multiple performance metrics, including:

- **Regression Metrics:** RMSE, Normalized RMSE, R-squared, and WMAPE.
- **Classification Metrics:** Accuracy, Sensitivity, Specificity, and Precautionary Sensitivity.
- **Quantile Regression Metrics:** Quantile (pinball) loss, Prediction Interval Coverage Probability (PICP), and Mean Prediction Interval Width (MPIW).

We apply specialized cross-validation strategies—nested time-series (TSCV) and Leave-One-Site-Out (LOSOCV)—and assess statistical significance using the Wilcoxon signed-rank test with Bonferroni correction.

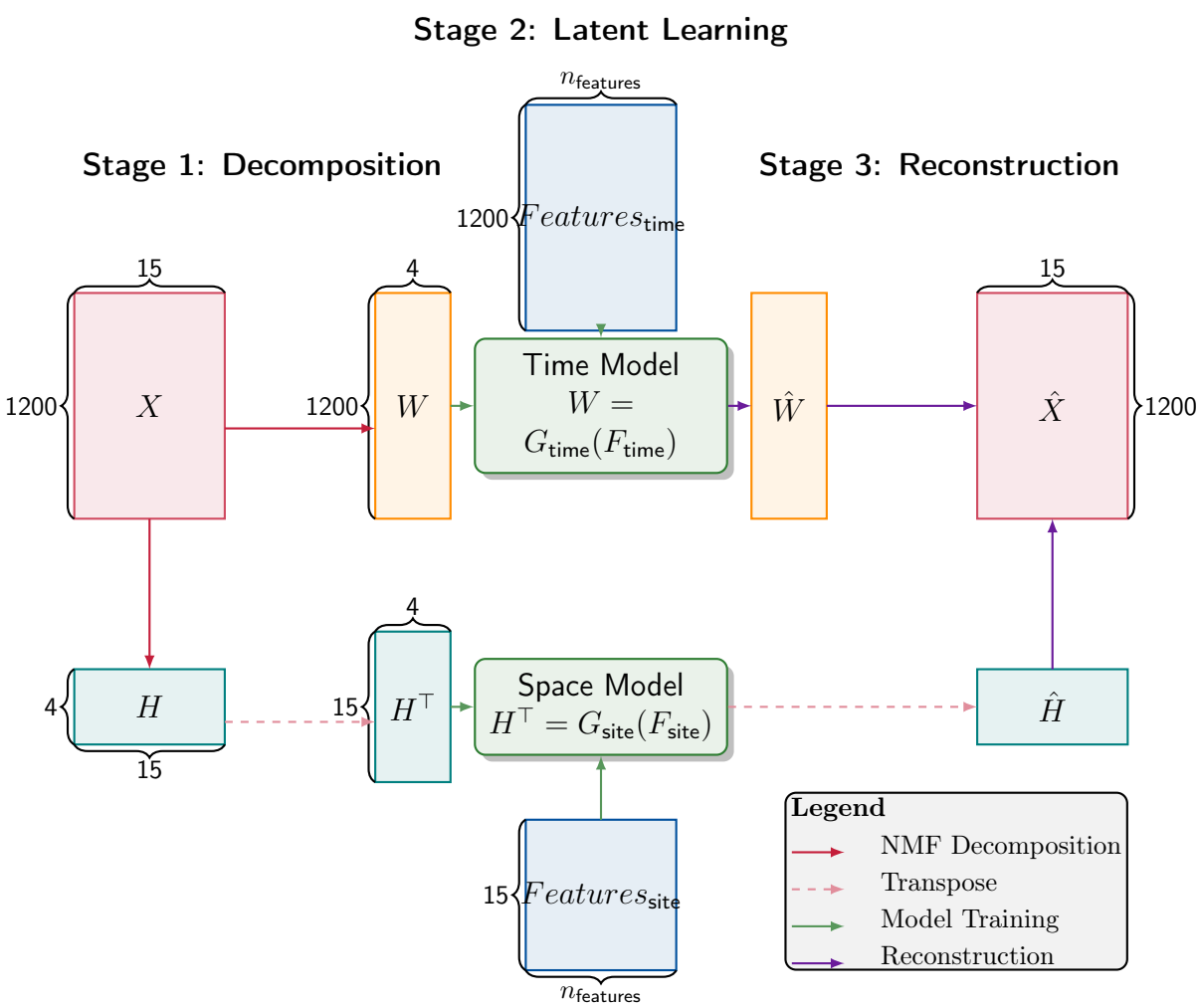

Figure 2: Matrix Decomposition Framework

3.5. Model Interpretability

To build trust and validate our models, we use SHapley Additive exPlanations (SHAP) [28] for:

- **Global Interpretation:** KernelExplainer [29] assesses overall feature importance.
- **Component-level Analysis:** TreeExplainer [30] provides detailed insights for the tree-based models within each framework.

Additional analyses (local prediction, temporal/spatial variation, and latent factor contributions) further elucidate model behavior; full details are in the supplement.

3.6. Deployment Considerations

A streamlined real-time deployment pipeline has been implemented on Heroku [10]. The system features an automated CI/CD process [8], a robust data validation pipeline (using ADWIN drift detection [1] and missing data checks), and a fallback mode for handling anomalies. An interactive Streamlit [31] dashboard provides real-time nowcasts and performance metrics, demonstrating the system's operational feasibility.

4. Experiments

This section presents our experimental evaluation of the forecasting frameworks. We describe our streamlined experimental setup, summarize key data preprocessing and feature engineering experiments, outline model development and validation procedures, report comprehensive performance and interpretability analyses, and validate the deployment of our system.

4.1. Experimental Setup

Experiments were conducted on an Apple MacBook Air (M2 chip, 8-core CPU, 10-core GPU, 16 GB memory) using Python 3.9 in a Conda environment. Key libraries included scikit-learn [26], LightGBM [18], MAPIE [17], Pandas [25], NumPy [23], scikit-multiflow [1], Streamlit [31], and SHAP [28]. Historical data were stored in CSV files, and real-time data were acquired via APIs from NIWA [20], LINZ [12], and the Lyttelton Port Company [16]. Version control and experiment tracking were managed with Git [7], MLflow [19], and GitHub Actions [8], and the final system was deployed on Heroku [10].

4.2. Evaluation Implementation

We employed nested time-series cross-validation (TSCV) and Leave-One-Site-Out Cross-Validation (LOSOCV) to assess temporal and spatial generalization, respectively. Three evaluation periods (2021–2022, 2022–2023, 2023–2024) were defined for TSCV, and performance metrics—such as RMSE, WMAPE, sensitivity, specificity, pinball loss, Prediction Interval Coverage Probability (PICP), and Mean Prediction Interval Width (MPIW)—were computed. Statistical significance was evaluated using the Wilcoxon signed-rank test with Bonferroni correction (p-value < 0.05).

4.3. Data Preprocessing and Feature Engineering Experiments

We compared three target transformations (raw, logarithmic, and square root) to select the one that best balanced improved statistical normality with predictive performance. Our feature engineering experiments focused on two key analyses. First, Feature Group Ablation was conducted by systematically removing rainfall, wind, tide, temporal, and site features to determine their individual impact on forecasting accuracy. Second, Lag Window Optimization was performed by evaluating multiple lag intervals for rainfall (3, 6, 12, 24, 48, 72, and 96 hours) and wind (3, 6, 12, and 24 hours) using correlation analysis and SHAP-based feature importance. Detailed results are provided in the Supplement.

4.4. Model Development and Validation Experiments

All forecasting frameworks and benchmark models (naïve baselines, generalized linear models, decision tree, MLP, and Virtual Beach (MLR) [34]) were developed on the same preprocessed dataset. Hyperparameters were optimized using the Tree-structured Parzen Estimator (TPE) over 100 trials with early stopping. For the Matrix Decomposition Framework, a single multi-target Random Forest model was selected (after comparison with independent forests and regressor chains) based on superior performance and efficiency.

4.5. Comprehensive Performance Evaluation

We evaluated model performance across several dimensions. Overall accuracy was assessed by comparing RMSE, sensitivity, specificity, and WMAPE across all models. In addition, we analyzed performance under critical conditions, including extreme contamination events, dry exceedances, and periods of high usage. Temporal generalization was evaluated by testing models across different sampling seasons and using an expanding window analysis, while spatial generalization was assessed via leave-one-site-out cross-validation (LOSOCV) and cross-harbor experiments. For the Probabilistic Framework, uncertainty quantification was examined by analyzing prediction intervals through metrics such as PICP and MPIW. Statistical tests confirmed significant performance differences between our frameworks and benchmarks.

4.6. Model Interpretability Experiments

To elucidate model behavior, we conducted SHAP analyses. Global feature importance was determined using KernelExplainer [29], providing an overall view of the influence of each feature. Local analysis was performed on selected cases, such as an exceedance event and a recovery period, to decompose individual predictions and reveal the contribution of specific features. Additionally, we conducted a quantile-specific analysis, where SHAP [28] values for each quantile model highlighted variations in feature contributions across the predicted distribution. For the Matrix Decomposition Framework, TreeExplainer [30] was employed to link input features to each latent factor, thereby enhancing our understanding of how environmental and site-specific factors drive the model outputs.

5. Results

This chapter presents the main findings of our study, highlighting the performance, generalizability, interpretability, and operational feasibility of our two proposed forecasting frameworks—the Probabilistic Forecasting Framework and the Matrix Decomposition Framework—relative to established benchmarks.

5.1. Data Exploration and Preprocessing Outcomes

Our analysis of the historical Enterococci dataset confirms that the raw data is highly right-skewed with a heavy tail and significant outliers (mean = 93 MPN/100mL, median = 5 MPN/100mL, skewness = 25, kurtosis = 732. This distribution motivated our investigation into target transformations. Experimental comparisons showed that while a logarithmic transformation substantially reduced skewness and kurtosis, retaining raw values in the Probabilistic Framework maximized sensitivity to exceedance events. In contrast, the Matrix Decomposition Framework and benchmark models benefitted from log-transformation to improve model stability and convergence of the matrix factorization optimization. Correlation analyses further revealed that rainfall features exhibit strong positive associations with contamination up to 72 hours, whereas wind variables are most relevant within 24 hours. These insights drove our final feature selection, with rainfall emerging as the dominant predictor.

5.2. Comparative Model Performance

Overall performance was evaluated using both regression (RMSE, WMAPE) and classification metrics (accuracy, sensitivity, specificity) under a nested time-series and spatial cross-validation framework. As summarized in Table 2, our Probabilistic Framework achieved an RMSE of 872 MPN/100mL, with the highest exceedance sensitivity (67.0%) and strong

specificity (92.3%). The Matrix Decomposition Framework delivered a slightly lower RMSE of 870 MPN/100mL, although its sensitivity (61.0%) was marginally lower. In comparison, conventional models—such as Linear Regression, Decision Trees, and the industry-standard Virtual Beach—consistently exhibited higher prediction errors and lower sensitivity. Statistical testing confirmed that the improvements offered by our frameworks are significant ($p < 0.05$).

	RMSE	WMAPE (%)		Classification (%)			
Model	MPN/100mL	Safe	Exc.	Acc.	Sens.	Sens.(P)	Spec.
Long-Term Grading	999	102.4	21.3	61.6 (645/1047)	45.0 (45/100)	**85.0 (85/100)**	63.4 (600/947)
Last-Observation	1376	50.5	58.7	84.2 (882/1047)	16.0 (16/100)	16.0 (16/100)	91.4 (866/947)
Virtual Beach	934	48.2	42.4	87.7 (918/1047)	38.0 (38/100)	44.0 (44/100)	92.9 (880/947)
Linear Reg.	867	**42.6**	38.3	88.8 (930/1047)	40.0 (40/100)	45.0 (45/100)	**93.9 (890/947)**
Logistic Reg.	-	-	-	82.9 (868/1047)	35.0 (35/100)	35.0 (35/100)	87.9 (833/947)
Decision Tree	1002	45.3	32.6	87.9 (920/1047)	46.0 (46/100)	51.0 (51/100)	92.3 (874/947)
MLP	893	115.3	20.4	88.3 (925/1047)	56.0 (56/100)	63.0 (63/100)	91.8 (869/947)
Prob. Framework	872	52.3	**17.2**	**89.9 (941/1047)**	**67.0 (67/100)**	76.0 (76/100)	92.3 (874/947)
Matrix Decomp.	**870**	54.2	20.3	87.8 (919/1047)	61.0 (61/100)	74.0 (74/100)	90.6 (858/947)

Note: Best performance shown in **bold**. RMSE in MPN/100mL. Sens.(P) indicates sensitivity to precautionary alerts.

Table 2: Comparative Model Performance

5.3. Critical Event Analysis

Our analysis of extreme contamination events (exceedances $>$ 5392 MPN/100mL) demonstrates that both proposed frameworks outperform benchmark models in capturing high-risk conditions. While all methods tend to underpredict the magnitude of extreme events, the Probabilistic Framework shows superior sensitivity, particularly for rainfall-driven events. However, both frameworks exhibit near-zero sensitivity for dry exceedances, underscoring a limitation in capturing events not driven by rainfall.

In periods of high site usage (e.g., weekends), the frameworks achieve sensitivity as high as 71.4–85.7% with low WMAPE for exceedance events. Conversely, during holiday periods—likely influenced by non-environmental factors—both frameworks struggle to detect exceedances. These results suggest that while our models effectively capture environmental signals, incorporating additional predictors may be necessary for comprehensive risk assessment during atypical usage periods.

5.4. Probabilistic Forecasting Framework Analysis

Within the Probabilistic Framework, an ensemble of ten LightGBM quantile regression models was trained to predict a range of percentiles (5th to 98th). The ensemble's adaptive behavior is evident: lower quantile models, characterized by low error and high specificity, contrast with higher quantile models that improve exceedance detection at the expense

of increased error. To synthesize these predictions, a gradient boosting meta-learner was employed. As shown in Table 3, the meta-learner substantially improved point forecast accuracy (RMSE of 872 MPN/100mL) and balanced sensitivity and specificity better than simpler combination methods. Monotonic sorting was applied as a post-processing step to ensure consistent quantile predictions.

Method	RMSE	WMAPE (%)		Classification (%)		
		Safe	Exc.	Sensitivity	Sensitivity(P)	Specificity
Meta-Learner	**872**	52.3	**17.2**	**65.0**	**77.0**	96.2
Single Quantile (0.8)	874	50.7	20.3	60.0	73.0	90.2
Quantile Averaging	980	52.9	22.0	61.0	72.0	92.6
Interval Mid-Point	942	**36.7**	27.7	46.0	67.0	**96.6**

Note: Best performance for each metric shown in **bold**. Sens.(P) indicates sensitivity to precautionary alerts.

Table 3: Comparison of point prediction methods within the Probabilistic Framework.

6. Matrix Decomposition Framework Analysis

For the Matrix Decomposition Framework, Non-negative Matrix Factorization (NMF) was used to decompose the high-dimensional Enterococci data into latent temporal and spatial factors. We found that a latent dimensionality of $k = 4$ offered the best trade-off between reconstruction fidelity and forecasting performance (Table 4). Subsequent analysis using multi-target Random Forests to predict these factors yielded an overall RMSE of 870 MPN/100mL and provided interpretable spatial and temporal patterns. SHAP analyses confirmed that the spatial factors were largely driven by site characteristics (e.g., beach orientation, soil type), while the temporal factors captured distinct rainfall dynamics over varying time scales. Although some redundancy was observed among the temporal factors, the framework effectively separates spatio-temporal drivers, enhancing both interpretability and predictive performance.

7. Generalization Capability

Temporal generalizability was assessed by evaluating model performance over three distinct sampling seasons (2021–2022, 2022–2023, and 2023–2024) using an expanding window approach. As shown in Table 5, the Probabilistic Framework improved its exceedance sensitivity from 44.8% in 2021–2022 to 75.7% in 2023–2024, while both frameworks reached performance plateaus after approximately eight years of training data.

Spatial generalizability was evaluated through two complementary approaches. In the cross-harbor evaluation (Table 7), models trained on one harbor were tested on the other, demon-

	Framework Performance				Decomposition Quality	
Factors (k)	RMSE (MPN/100mL)	WMAPE Exc. (%)	Sens. (%)	Spec. (%)	Reconstruction Error (RMSE)	Exceedance Recall (Sensitivity)
1	720	26.6	43.0	96.8	358.5	0.33
2	780	24.8	47.0	95.0	287.1	0.39
3	825	22.4	52.0	93.5	226.2	0.57
4	**870**	**20.3**	**59.0**	**92.3**	186.6	0.58
5	900	20.8	57.0	92.1	160.3	0.62
6	930	21.4	55.0	92.0	137.3	0.68
7	945	21.6	54.0	92.0	127.7	0.74
8	980	21.8	53.0	92.0	**112.2**	**0.75**

Note: Best framework performance metrics shown in **bold**. Reconstruction RMSE and Exceedance Recall measure matrix factorization quality independent of prediction accuracy.

Table 4: Impact of Latent Factor Dimensionality

strating that both frameworks maintain high specificity (above 85%) with acceptable sensitivity despite local variations. Additionally, leave-one-site-out cross-validation (LOSOCV) results in Table 6 indicate that the Probabilistic Framework generally achieves higher sensitivity and specificity at individual sites, whereas the Matrix Decomposition Framework tends to yield lower RMSE values, suggesting more precise magnitude predictions.

Together, these results confirm that our forecasting frameworks generalize well both temporally and spatially, underscoring their robustness and potential for real-world deployment in diverse water quality monitoring settings.

		Regression			Classification			
Period	Model	RMSE (MPN/100mL)	WMAPE (%) Safe	Exc.	Acc. (%)	Sens. (%)	Sens.(P) (%)	Spec. (%)
2021-22	Linear	505	**41.0**	34.8	**87.8 (224/255)**	34.5 (10/29)	34.5 (10/29)	**94.7 (214/226)**
	Prob.	454	43.0	21.5	**87.8 (224/255)**	44.8 (13/29)	55.2 (16/29)	93.4 (211/226)
	Matrix	**450**	47.4	**21.1**	86.3 (220/255)	**55.2 (16/29)**	**62.1 (18/29)**	90.3 (204/226)
2022-23	Linear	1269	**50.3**	37.3	86.2 (287/333)	41.2 (14/34)	47.1 (16/34)	**91.3** (273/299)
	Prob.	1355	53.9	**18.2**	**88.0 (293/333)**	**76.5 (26/34)**	**82.4 (28/34)**	89.3 (267/299)
	Matrix	**1138**	55.7	19.9	85.9 (286/333)	67.6 (23/34)	79.4 (27/34)	88.0 (263/299)
2023-24	Linear	637	**45.2**	40.3	91.3 (419/459)	43.2 (16/37)	51.4 (19/37)	**95.5 (403/422)**
	Prob.	569	52.2	**14.3**	**92.4 (424/459)**	**75.7 (28/37)**	**86.5 (32/37)**	93.8 (396/422)
	Matrix	**553**	53.3	17.3	90.0 (413/459)	59.5 (22/37)	78.4 (29/37)	92.7 (391/422)

Note: Best performance for each period and metric shown in **bold**. Sens.(P) indicates sensitivity to precautionary alerts.

Table 5: Performance metrics for both frameworks across three time-series cross-validation folds (2021-2024). Metrics include regression measures (RMSE, WMAPE) and classification performance indicators (accuracy, sensitivity, specificity) for safe and exceedance predictions.

Harbor	Site	Probabilistic Framework			Matrix Framework		
		RMSE (MPN/100mL)	Sens. (%)	Spec. (%)	RMSE (MPN/100mL)	Sens. (%)	Spec. (%)
Akaroa	Akaroa Beach	623	66.7 (4/6)	**92.5 (62/67)**	**597**	50.0 (3/6)	89.6 (60/67)
	Duvauchelle Bay	843	**75.0 (6/8)**	**93.8 (61/65)**	**817**	62.5 (5/8)	92.3 (60/65)
	French Farm Bay	132	**100.0 (1/1)**	**94.2 (49/52)**	**127**	**100.0 (1/1)**	92.3 (48/52)
	Glen Bay	248	**100.0 (3/3)**	**92.8 (64/69)**	**242**	**100.0 (3/3)**	92.8 (65/70)
	Takamatua Bay	**103**	**100.0 (2/2)**	**94.3 (50/53)**	118	**100.0 (2/2)**	92.5 (49/53)
	Tikao Bay	904	**63.6 (7/11)**	89.7 (52/58)	**883**	54.5 (6/11)	**86.2 (50/58)**
	Wainui Beach	2587	**80.0 (4/5)**	**92.8 (64/69)**	**2494**	60.0 (3/5)	91.3 (63/69)
	Overall	777	**75.0 (27/36)**	**92.8 (402/433)**	**754**	69.4 (25/36)	91.2 (395/433)
Lyttelton	Cass Bay	403	**25.0 (1/4)**	**92.9 (65/70)**	**378**	**25.0 (1/4)**	88.6 (62/70)
	Charteris Bay	298	66.7 (2/3)	96.3 (52/54)	**283**	**100.0 (3/3)**	90.7 (49/54)
	Church Bay	302	**100.0 (4/4)**	**96.4 (53/55)**	**276**	**100.0 (4/4)**	87.3 (48/55)
	Corsair Bay	647	**61.5 (8/13)**	93.1 (67/72)	**623**	53.8 (7/13)	**94.4 (68/72)**
	Diamond Harbor	897	**66.7 (6/9)**	**93.5 (58/62)**	**852**	55.6 (5/9)	93.5 (58/62)
	Governors Bay	806	**36.4 (4/11)**	**88.0 (44/50)**	**784**	**36.4 (4/11)**	86.0 (43/50)
	Purau Bay	347	**60.0 (6/10)**	**92.5 (74/80)**	**324**	50.0 (5/10)	92.5 (74/80)
	Rāpaki Bay	703	**60.0 (6/10)**	**90.0 (63/70)**	**677**	50.0 (5/10)	88.6 (62/70)
	Overall	550	**57.8 (37/64)**	**92.8 (476/513)**	**525**	53.1 (34/64)	90.4 (464/513)

Note: Results show framework performance under leave-one-site-out cross-validation. Best performance for each metric shown in **bold**.

Table 6: Site-level generalization performance of the Probabilistic and Matrix Decomposition Frameworks, evaluated using Leave-One-Site-Out Cross-Validation (LOSOCV). The table shows RMSE, sensitivity, and specificity for each site, along with sample counts.

Harbor	Site	Probabilistic Framework			Matrix Framework		
		RMSE (MPN/100mL)	Sens. (%)	Spec. (%)	RMSE (MPN/100mL)	Sens. (%)	Spec. (%)
Akaroa	Akaroa Beach	**654**	50.0 (3/6)	**95.5 (64/67)**	681	**66.7 (4/6)**	92.5 (62/67)
	Duvauchelle Bay	**861**	50.0 (4/8)	**92.3 (60/65)**	881	50.0 (4/8)	89.2 (58/65)
	French Farm Bay	**128**	**100.0 (1/1)**	**88.5 (46/52)**	152	**100.0 (1/1)**	**88.5 (46/52)**
	Glen Bay	**259**	66.7 (2/3)	**92.8 (64/69)**	331	**100.0 (3/3)**	87.0 (60/69)
	Takamatua Bay	**106**	**100.0 (2/2)**	**94.3 (50/53)**	126	**100.0 (2/2)**	90.6 (48/53)
	Tikao Bay	**941**	45.5 (5/11)	**94.8 (55/58)**	942	**54.5 (6/11)**	93.1 (54/58)
	Wainui Beach	2785	60.0 (3/5)	**89.9 (62/69)**	**2764**	**80.0 (4/5)**	85.5 (59/69)
	Overall	**819**	55.6 (20/36)	**92.6 (401/433)**	840	**66.7 (24/36)**	89.4 (387/433)
Lyttelton	Cass Bay	**422**	**25.0 (1/4)**	**90.0 (63/70)**	473	**25.0 (1/4)**	82.9 (58/70)
	Charteris Bay	316	**66.7 (2/3)**	**96.3 (52/54)**	**279**	33.3 (1/3)	92.6 (50/54)
	Church Bay	317	**100.0 (4/4)**	**96.4 (53/55)**	**218**	75.0 (3/4)	90.9 (50/55)
	Corsair Bay	665	46.2 (6/13)	**93.1 (67/72)**	**650**	**53.8 (7/13)**	87.5 (63/72)
	Diamond Harbor	962	55.6 (5/9)	**93.5 (58/62)**	**841**	**66.7 (6/9)**	88.7 (55/62)
	Governors Bay	849	**36.4 (4/11)**	80.0 (40/50)	**777**	**36.4 (4/11)**	**84.0 (42/50)**
	Purau Bay	342	40.0 (4/10)	**92.5 (74/80)**	**313**	**60.0 (6/10)**	91.3 (73/80)
	Rāpaki Bay	730	50.0 (5/10)	**88.6 (62/70)**	**758**	**60.0 (6/10)**	84.3 (59/70)
	Overall	575	48.4 (31/64)	**91.4 (469/513)**	**539**	**53.1 (34/64)**	87.7 (450/513)

Note: Results show framework performance when trained on one harbor and tested on the other. Best performance per site shown in **bold**

Table 7: Cross-harbor generalization performance of the Probabilistic and Matrix Decomposition Frameworks, showing RMSE, sensitivity, and specificity. Models were trained on one harbor (Lyttelton or Akaroa) and tested on the other. Sample counts are included for each site and harbor.

8. Model Interpretability and Feature Analysis

Transparency is crucial for operational trust, and our SHAP-based analysis provides insights into the decision-making processes of our models. Global interpretability results indicate that rainfall and wind are the dominant predictors across both frameworks. Local analyses of representative exceedance and recovery events reveal that the Probabilistic Framework dynamically emphasizes mid-range quantile predictions during high-risk events and shifts weighting during recovery, while the Matrix Decomposition Framework leverages site-static features to account for spatial variability. Furthermore, pairwise feature interactions—such as between 12-hour rainfall and 6-hour wind speed—were identified, reinforcing the models' alignment with known physical processes.

9. Operational Feasibility and Deployment

Our forecasting system was deployed in a real-time staging environment on Heroku. Despite modest increases in training times relative to simpler models, both frameworks achieve sub-second inference times (0.1–0.2 s) and maintain low memory footprints (24–30 MB). Over a twelve-month operational simulation, the real-time data pipeline effectively managed data drift (with ADWIN detecting drift events corresponding to significant rainfall) and handled missing data using a LOCF approach. The integrated analytics dashboard provided stakeholders with transparent nowcasts, while operational validation confirmed that the system consistently generated timely and reliable forecasts under realistic conditions.

In summary, the experimental results demonstrate that our proposed forecasting frameworks not only outperform traditional and benchmark models in terms of accuracy and sensitivity but also offer robust uncertainty quantification and enhanced interpretability. These qualities, combined with proven generalizability and operational feasibility, underscore the potential of our approaches to support proactive, real-time water quality management.

10. Discussion

In this study, we introduced two novel forecasting frameworks for real-time water quality prediction—the Probabilistic Framework and the Matrix Decomposition Framework. Both approaches significantly improve the detection of high-risk Enterococci exceedance events and provide interpretable outputs with robust uncertainty estimates. In this section, we discuss the innovations and performance characteristics of our methods, their operational implications, limitations, and potential future directions.

10.1. Framework Innovations and Performance

The Probabilistic Framework leverages an ensemble of gradient boosting quantile regression models alongside an adaptive meta-learner to generate calibrated point forecasts and 90% prediction intervals. This design effectively addresses the challenges posed by the non-normal, heteroscedastic nature of Enterococci data. Our SHAP analyses confirm that the model adaptively shifts its focus—emphasizing mid-range quantiles under normal conditions and higher quantiles during exceedance events—resulting in an exceedance sensitivity of 67.0% and a precautionary sensitivity of 77.0%.

In contrast, the Matrix Decomposition Framework employs Non-negative Matrix Factorization (NMF) to disentangle spatial and temporal influences, allowing the model to capture site-specific vulnerabilities and dynamic environmental patterns. Although the NMF process better captures rainfall-driven trends than non-linear wind effects, its integration with site-static features compensates by enhancing spatial generalizability. Overall, this framework achieved an RMSE of 870 MPN/100mL and demonstrated superior cross-harbor generalization, highlighting its potential to extrapolate latent spatial characteristics to new environments.

10.2. Operational Implications

Beyond predictive accuracy, both frameworks meet operational requirements. Resource utilization tests indicate that forecast generation occurs in under one second on modest hardware, making real-time deployment feasible. The integration of uncertainty margins—calibrated via Conformalized Quantile Regression—provides decision-makers with a nuanced risk profile that supports proactive interventions. Furthermore, SHAP-based explainability enhances stakeholder trust by ensuring that model predictions align with established domain knowledge, thereby facilitating targeted responses in water quality management.

10.3. Limitations and Future Work

Despite their strengths, our frameworks have limitations. Both models underpredict the magnitude of extreme contamination events and struggle with dry exceedances, suggesting that key drivers such as localized point-source pollution or human usage patterns may not be fully captured. Additionally, the granularity of site-static features limits the Matrix Decomposition Framework's ability to distinguish subtle environmental differences between locations.

Future work should focus on incorporating additional data sources—including high-resolution site metadata, real-time sensor measurements for point-source discharges, and direct indicators of recreational activity—to enhance model sensitivity, especially for dry exceedances.

Exploring non-linear extensions of NMF or alternative deep learning architectures, such as quantile regression neural networks, may further improve performance while maintaining interpretability.

10.4. Broader Impact

Although this research targets Enterococci forecasting in Canterbury, the underlying methodologies have broader applicability in environmental monitoring. The Probabilistic Framework's robust uncertainty quantification and the Matrix Decomposition Framework's interpretable spatio-temporal separation can be adapted to other domains such as air quality or coastal erosion management. Ultimately, our work demonstrates the potential of advanced machine learning techniques to support proactive, data-driven public health and environmental risk management.

11. Conclusion

We introduced and evaluated two novel machine learning frameworks for real-time forecasting of Enterococci concentrations in recreational waters. The Probabilistic Forecasting Framework combines an ensemble of quantile regression models with a meta-learner to produce risk-aware point forecasts and calibrated uncertainty intervals, while the Matrix Decomposition Framework uses Non-negative Matrix Factorization to decouple spatial and temporal influences into interpretable latent factors.

Experiments using data from Canterbury's recreational waters demonstrate that both frameworks substantially outperform existing operational practices and conventional benchmark models. The Probabilistic Framework achieved an exceedance sensitivity of 67.0% (with a precautionary sensitivity of 77.0%), and the Matrix Decomposition Framework exhibited superior generalizability across harbor environments. Moreover, both approaches provide reliable uncertainty quantification and transparent SHAP-based explanations that align with known hydrological processes.

Although challenges remain—such as underpredicting extreme contamination events and detecting non-rainfall-driven (dry) exceedances—our results highlight the potential for further improvement through enhanced feature engineering and the integration of additional data sources.

In summary, this work represents a significant advancement toward proactive, risk-based water quality management. By combining improved predictive accuracy, robust uncertainty quantification, and enhanced interpretability, our proposed frameworks offer a promising foundation for operational forecasting systems that can better safeguard public health and optimize the management of coastal resources.

12. Author Contribution

Asif Cheena led the project and was responsible for conceptualization, methodology development, software implementation (including coding and data analysis), and manuscript writing. Katharina Dost and Jörg Wicker provided supervisory oversight by critically reviewing the work, offering constructive feedback, and contributing key ideas throughout the research process. Nina Straathof from Environment Canterbury facilitated the study by providing essential datasets, background information, and foundational support. This work was conducted as part of Asif Cheena's Master's thesis, and no external funding was required.

13. Data Availability

The data associated with this study include both raw and processed datasets, detailed as follows:

Raw Data: The raw sensor data consist of hourly measurements from the monitoring stations. These data are proprietary and are provided by the respective organizations (NIWA [20], LPC [16], and LINZ [12]). Due to restrictions imposed by these data providers, the raw sensor data are not publicly available. The sources of these raw data have been cited in the manuscript.

Enterococci Sampling Data: Enterococci sampling data are publicly available on the LAWA website [13]. For this study, these data have been curated and structured into a comprehensive dataset suitable for modelling. The curated dataset is available upon reasonable request from the corresponding author.

Processed Data: The processed data include derived features such as weather lag variables, site-specific attributes, and the corresponding Enterococci sample readings used in our analyses. These processed datasets can be obtained by contacting the corresponding author.

Code and Dashboard: The code supporting this study is available on GitHub. A public dashboard providing live predictions is also hosted online; however, please note that the dashboard may experience intermittent downtime due to hosting and maintenance costs.
Github: `[https://github.com/asif-jc/Dont-Swim-in-Data.git]`
Dashboard: `[https://microbial-forecasting-ml-app-1a0204792e39.herokuapp.com/]`

Licensing and Usage: The processed data and code are released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. This license permits unrestricted use, distribution, and reproduction in any medium, provided that the original work is properly cited. Researchers reusing these materials are kindly requested to cite the original data sources as referenced in the manuscript.

References

[1] *ADWIN*. Accessed: 2025-01-16. scikit-multiflow. 2025. URL: https://scikit-multiflow.readthedocs.io/en/stable/api/generated/skmultiflow.drift_detection.ADWIN.html.

[2] U.S. Environmental Protection Agency. *Recreational Water Quality Criteria and Methods*. Accessed: 2025-02-19. 2012. URL: https://www.epa.gov/wqc/recreational-water-quality-criteria-and-methods.

[3] Takuya Akiba et al. *Optuna: A hyperparameter optimization framework*. https://optuna.org. Accessed: 2024-11-02. 2019.

[4] Alex J. Cannon. "Non-crossing, monotone quantile regression neural networks: An application to rainfall extremes". In: *Stochastic Environmental Research and Risk Assessment* 32.11 (2018), pp. 3207–3225. DOI: 10.1007/s00477-018-1573-6. URL: https://doi.org/10.1007/s00477-018-1573-6.

[5] Daniela Džal et al. "Modelling Bathing Water Quality Using Official Monitoring Data". In: *Water* 13.21 (2021), p. 3005. DOI: 10.3390/w13213005. URL: https://doi.org/10.3390/w13213005.

[6] Engineers Australia. *Safeswim: Managing Water Quality at Beaches and Rivers*. https://www.engineersaustralia.org.au/sites/default/files/Safeswim.pdf. Accessed: 2024-07-11.

[7] *Git Documentation*. Accessed: 2025-01-25. Git. 2025. URL: https://git-scm.com/doc.

[8] *Github Actions*. Accessed: 2025-01-16. Github. 2025. URL: https://github.com/features/actions.

[9] A. Hannachi, I. T. Jolliffe, and D. B. Stephenson. "Empirical Orthogonal Functions and Related Techniques in Atmospheric Science: A Review". In: *International Journal of Climatology* 27.9 (2007), pp. 1119–1152. DOI: 10.1002/joc.1499.

[10] *Heroku*. Accessed: 2025-01-25. Heroku. 2025. URL: https://www.heroku.com/.

[11] Claudia Condé Lamparelli et al. "Are fecal indicator bacteria appropriate measures of recreational water risks in the tropics: A cohort study of beach goers in Brazil?" In: *Water Research* 87 (2015), pp. 59–68. DOI: 10.1016/j.watres.2015.09.001. URL: https://doi.org/10.1016/j.watres.2015.09.001.

[12] Land Information New Zealand. *Tide Predictions*. Accessed: 2024-11-13. URL: https://www.linz.govt.nz/products-services/tides-and-tidal-streams/tide-predictions.

[13] LAWA (Land, Air, Water Aotearoa). *Canterbury Region Water Quantity*. https://www.lawa.org.nz/explore-data/canterbury-region/water-quantity. Accessed: 2024-07-28.

[14] Lingbo Li et al. "Interpretable tree-based ensemble model for predicting beach water quality". In: *Water Research* 211 (2022), p. 118078. ISSN: 0043-1354. DOI: 10.1016/j.watres.2022.118078. URL: https://www.sciencedirect.com/science/article/pii/S0043135422000410.

[15] Tianxiang Liu et al. "Ensemble water quality forecasting based on decomposition, sub-model selection, and adaptive interval". In: *Environmental Research* 237 (2023), p. 116938. DOI: 10.1016/j.envres.2023.116938. URL: https://www.sciencedirect.com/science/article/pii/S0013935123017425.

[16] Lyttelton Port Company. *Lyttelton Port Company*. https://www.lpc.co.nz/. Accessed: 2024-06-23.

[17] *MAPIE - Model Agnostic Prediction Interval Estimator*. Accessed: 2025-01-10. MAPIE. 2025. URL: https://mapie.readthedocs.io/en/latest/.

[18] Microsoft. *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. https://github.com/microsoft/LightGBM. Accessed: 2024-07-06. 2024.

[19] *MLflow Documentation*. Accessed: 2025-01-25. MLflow. 2025. URL: https://mlflow.org/docs/latest/index.html.

[20] National Institute of Water and Atmospheric Research. *NIWA*. https://niwa.co.nz/. Accessed: 2024-06-23.

[21] Meredith B. Nevers, Muruleedhara N. Byappanahalli, and Richard L. Whitman. "Choices in Recreational Water Quality Monitoring: New Opportunities and Health Risk Trade-Offs". In: *Environmental Science Technology* 47.6 (2013), pp. 3073–3081. DOI: 10.1021/es304408y. URL: https://pubs.acs.org/doi/10.1021/es304408y.

[22] Meredith B. Nevers and Richard L. Whitman. "Efficacy of monitoring and empirical predictive modeling at improving public health protection at Chicago beaches". In: *Water Research* 45.4 (2011), pp. 1659–1668. DOI: 10.1016/j.watres.2010.12.010. URL: https://www.sciencedirect.com/science/article/pii/S0043135410008377.

[23] *NumPy*. Accessed: 2025-01-25. NumPy Developers. 2025. URL: https://numpy.org/.

[24] World Health Organization. *Guidelines for Safe Recreational Water Environments. Volume 2: Swimming Pools and Similar Environments*. Geneva, Switzerland: World Health Organization, 2006. ISBN: 92 4 154680 8. URL: https://www.who.int/publications/i/item/9241546808.

[25] *pandas*. Accessed: 2025-01-25. pandas Developers. 2025. URL: https://pandas.pydata.org/.

[26] *Scikit-Learn: Machine Learning in Python*. Accessed: 2025-02-12. Scikit-Learn Developers. 2025. URL: https://scikit-learn.org/stable/.

[27] Akram Seifi, Majid Dehghani, and Vijay P. Singh. "Uncertainty analysis of water quality index (WQI) for groundwater quality evaluation: Application of Monte-Carlo method for weight allocation". In: *Ecological Indicators* 117 (2020), p. 106653. DOI: 10.1016/j.ecolind.2020.106653. URL: https://www.sciencedirect.com/science/article/pii/S1470160X20305902.

[28] SHAP Documentation. *SHAP (SHapley Additive exPlanations)*. https://shap.readthedocs.io/en/latest/. Accessed: 2024-08-12.

[29] SHAP Documentation. *SHAP KernelExplainer*. https://shap.readthedocs.io/en/latest/generated/shap.KernelExplainer.html. Accessed: 2024-08-12.

[30] SHAP Documentation. *SHAP TreeExplainer*. https://shap-lrjball.readthedocs.io/en/latest/generated/shap.TreeExplainer.html. Accessed: 2024-08-12.

[31] *streamlit*. Accessed: 2025-01-16. streamlit. 2025. URL: https://streamlit.io/.

[32] W. Thoe et al. "Predicting water quality at Santa Monica Beach: Evaluation of five different models for public notification of unsafe swimming conditions". In: *Water Research* 67 (2014), pp. 105–117. DOI: 10.1016/j.watres.2014.09.001. URL: https://www.sciencedirect.com/science/article/pii/S0043135414006241.

[33] Wei Thoe and Jiyoung Lee. "Daily forecasting of marine beach water quality in Hong Kong using models developed from beach-specific data". In: *Water Research* 92 (2016), pp. 228–238. DOI: 10.1016/j.jher.2012.05.003. URL: https://doi.org/10.1016/j.jher.2012.05.003.

[34] U.S. Environmental Protection Agency. *Virtual Beach (VB) Software for Modeling and Predicting Beach Water Quality*. https://www.epa.gov/hydrowq/virtual-beach-vb. Accessed: 2024-10-10.

[35] Ruixue Yuan et al. "Bayesian non-negative matrix factorization with Student's t-distribution for outlier removal and data clustering". In: *Engineering Applications of Artificial Intelligence* 132 (2024), p. 107978. DOI: 10.1016/j.engappai.2024.107978. URL: https://www.sciencedirect.com/science/article/pii/S0952197624001362.

[36] Peijie Zhang et al. "Autoregressive matrix factorization for imputation and forecasting of spatiotemporal structural monitoring time series". In: *Mechanical Systems and Signal Processing* 169 (2022), p. 108718. DOI: 10.1016/j.ymssp.2021.108718. URL: https://www.sciencedirect.com/science/article/pii/S0888327021010372.