Water Research 309 (2027) 127143

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

Data-driven machine learning approach for water quality anomaly detection and NPS-FM non-compliance risk prediction: A study with five New Zealand rivers☆

Tanishqa Goyal, Parul Tiwari*

Department of Mathematical Sciences, School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, 55 Wellesley Street East, Auckland, 1010, New Zealand

ARTICLE INFO

Dataset link: https://www.lawa.org.nz

Keywords:
Machine learning
Anomaly detection
Freshwater compliance
Water quality monitoring
Land cover

ABSTRACT

The National Policy Statement for Freshwater Management (NPS-FM) 2020 sets national water quality objectives for New Zealand's freshwater, but compliance is mostly assessed after the fact, and tools for anticipating which sites may be at greatest risk of future non-compliance remain limited. This study presents a machine learning pipeline that detects abnormal water quality events and quantifies NPS-FM non-compliance risk, applied to five South Island rivers of New Zealand spanning four land cover types. Anomaly detection combined Isolation Forest method with seasonal interquartile range thresholds, with observations flagged by both methods classified as confirmed anomalies. A weighted risk score (0–100) was computed from NPS-FM exceedance rate, confirmed anomaly rate, 90th percentile severity, and Mann–Kendall trend slope. This risk score is then validated using temporal hold-out, literature benchmarking, and seasonal pattern analysis. The stability of the applied pipeline is assessed by using alternative anomaly detection algorithms and by varying the quality-control filter. Risk scores followed a clear land cover gradient: the urban site scored 34.12 (High risk), with 96% of total nitrogen and 72% of *E. coli* readings exceeding NPS-FM limits, while native vegetation sites scored 14.24–16.24 (Medium risk). A paired catchment comparison isolated land use as the dominant driver, with the pasture site recording total nitrogen concentrations 0.3–0.5 mg/L above the upstream native site. Multi-indicator anomaly co-occurrence was 10.6–14.0 times higher than expected under random noise, and temporal hold-out validation passed at four of five sites. Site rankings and risk grades were unchanged when One-Class SVM or Local Outlier Factor was substituted for Isolation Forest. This research produces interpretable, indicator-level risk scores distinguishing episodic events from persistent contamination, consistent with established land use signatures. The framework is transferable to other regulatory contexts where long-term monitoring data and threshold-based standards exist.

1. Introduction

Freshwater degradation is among the most pressing global environmental challenges of the twenty-first century, with consequences for human health, food security and ecosystem function (van Vliet et al., 2023; Damania et al., 2019). More than 1.8 billion people rely on water sources contaminated with faecal matter (Bain et al., 2014), and severe pathogen pollution has been documented in roughly one-third of rivers across Africa, Asia and Latin America (United Nations Environment Programme, 2016). The dominant driver is agricultural intensification: in many high-income countries agricultural runoff has overtaken industrial and municipal discharges as the leading source of inland water pollution (Mateo-Sagasta et al., 2018), and global nitrogen delivery to surface freshwaters approximately doubled over the twentieth century (Beusen et al., 2016), fuelling riverine and coastal eutrophication, hypoxic dead zones and harmful algal blooms (Diaz and Rosenberg, 2008). Urban expansion compounds these pressures through impervious-surface runoff and ageing wastewater infrastructure, and climate change is projected to amplify both high-flow mobilisation and low-flow concentration across most major river basins (van Vliet et al., 2023).

This pattern is reflected in New Zealand, where freshwater deterioration is among the most comprehensively documented in the Southern Hemisphere. Researchers assessed over 900 national monitoring sites and established that nutrient concentrations and bacterial loads have

☆ This article is part of a Special issue entitled: 'AI for SWM' published in Water Research.

* Corresponding author.
E-mail addresses: tanishqagoyal12@gmail.com (T. Goyal), parul.tiwari@aut.ac.nz (P. Tiwari).

https://doi.org/10.1016/j.watres.2026.127143
Received 30 May 2026; Received in revised form 27 September 2026; Accepted 13 October 2026
Available online 17 October 2026
0043-1354/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

increased significantly in pastoral and urban catchments, with no evidence of broad-scale improvement (Larned et al., 2016). Authors in literature (Julian et al., 2017) confirmed based on a 26-year analysis that this deterioration is directly attributable to land use intensity, and a national reassessment using data ending in December 2024 found these patterns persist across nearly 1000 monitored sites (Booker et al., 2026). South Island rivers are particularly exposed by some of the country's most intensive pastoral farming sites alongside growing urban centres.

New Zealand's regulatory response is the National Policy Statement for Freshwater Management 2020, amended in October 2024 (Ministry for the Environment, 2020), which grades each water quality indicator from Band A (excellent) to Band D (poor), with Band D representing the bottom line below which sites are required to improve. Key Band D thresholds include *E. coli* at 540 cfu/100 mL, nitrate nitrogen at 6.9 mg/L, dissolved reactive phosphorus at 0.018 mg/L, and turbidity at 5.5 NTU. Widespread exceedance has been documented across urban streams (Gadd et al., 2020) and pastoral catchments (Cookson et al., 2024) confirming that non-compliance is systemic rather than localised. Compliance monitoring, however, remains fundamentally retrospective: exceedances are identified after they occur through periodic sampling, typically monthly, and reported in arrears. No operational system exists to predict which sites carry the greatest risk of future non-compliance. Authors in their research demonstrated that land use effects on water quality are statistically robust and detectable over long-term records (Snelder et al., 2021a), while a companion study (Snelder et al., 2021b) showed that short-term variability is dominated by climate and seasonal drivers, reinforcing the need to detect anomalies relative to seasonal norms rather than annual averages.

Machine learning methods have shown considerable promise for anomaly detection in long-term monitoring datasets. Detection techniques are broadly divided into statistical methods that flag observations outside an expected distribution, distance and density-based methods, and isolation or tree-based methods (Chandola et al., 2009). Environmental monitoring records present a distinctive combination of challenges. Pollution events are rare, unlabelled, and embedded within strong seasonal cycles, which rules out conventional supervised classification. Authors in Essamlali et al. (2024) identified unsupervised detection as the most effective direction for early warning systems suitable for these conditions. Among unsupervised approaches, Isolation Forest (Liu et al., 2008, 2012) has emerged as a benchmark for environmental time series, and authors in Lamrini et al. (2024) identified it as one of the most effective methods for river and lake outlier detection. Deep learning alternatives typically require dense high-frequency records unavailable in monthly state-of-environment programmes (El-Shafeiy et al., 2023), while Buchhorn et al. (2024) validated percentile-based seasonal thresholds as anomaly boundaries, providing methodological support for the seasonal layer. For longer-term trajectories the nonparametric Mann–Kendall test and its seasonal extension (Mann, 1945; Kendall, 1975; Hirsch and Slack, 1984) remain the standard, being robust to non-normality, missing values and censored data. For risk classification, researchers showed that weighted feature combinations outperform single-indicator approaches (Nasir et al., 2022), while authors in Chen et al. (2024) identified the absence of indicator-level explainability as a key limitation.

Despite this growing literature, anomaly detection and regulatory compliance assessment are typically treated as a single task. Water quality index frameworks aggregate threshold exceedance into a composite quality dimension (Canadian Council of Ministers of the Environment, 2001; Lumb et al., 2011; Sutadian et al., 2016), while anomaly detection studies commonly treat threshold breach as one detection signal alongside statistical criteria. Under either arrangement, a site that is chronically above a legal limit and a site that is usually clean but occasionally departs sharply from its own baseline are represented on the same scale, and the physical distinction between them is lost. This study takes the opposite approach. Threshold comparison is excluded from anomaly detection entirely and repositioned as the compliance component of the risk score, so that the statistical layers measure departure from each site's own history while the exceedance rate measures legal non-compliance. The methodological contribution is therefore not the individual detection algorithms, which are established, but the separation of persistent from episodic non-compliance within a single interpretable score. To our knowledge this is also the first site-level NPS-FM non-compliance risk ranking produced from New Zealand's national river monitoring data.

The rest of this article is organised as follows. Section 2 describes the materials and methods used in the current study. Section 3 presents the results of the analysis, and Section 4 reports the validation of those results. The contribution of this study with discussion and conclusions is summarised in Section 5, and Section 6 sets out the limitations, implications and future work.

2. Materials and methods

2.1. Study sites and data

Fig. 1 summarises the full pipeline, from raw data through anomaly detection, exceedance classification, weighted risk scoring, and validation. Water quality monitoring data were obtained from the Land, Air, Water Aotearoa (LAWA) platform, which compiles state of the environment monitoring records from New Zealand regional councils under open data licensing. The South Island dataset covers 11 water quality indicators measured at 270 monitoring sites across Canterbury, Marlborough, Nelson, Tasman, Otago, Southland, and the Chatham Islands regions over the period January 2004 to December 2024. Five sites (Fig. 2) were selected based on three criteria: availability of a near-complete 2004–2024 monitoring record, presence of all 11 indicators, and representation of all four major land cover types in the dataset. These criteria are considerably more restrictive than the size of the source dataset suggests. Of the 270 South Island sites, few carry a near-complete record for all 11 indicators across the full 2004–2024 window, since most regional council programmes measure a reduced indicator panel or commenced monitoring after 2004. Eleven candidate sites satisfying the record-length and indicator-completeness criteria were cleaned and assessed; five were retained to give one site per land cover type together with the paired catchment, the remaining candidates being duplicates of land cover types already represented. Site selection was therefore driven by data availability and by the requirement for land-cover coverage, not by any property of the water quality records themselves, and no site was excluded on the basis of its exceedance rates, anomaly counts or risk score. Site characteristics are summarised in Table 1. The selected sites span native vegetation, exotic forest, pasture, and urban land cover across two regions, with two sites located on the same river 15 km apart to enable a paired catchment comparison isolating land use as the primary variable. Sampling frequency was approximately monthly across all sites, consistent with regional council state of the environment monitoring protocols. The 11 indicators analysed and NPS-FM Band D limits available for eight of these indicators and the thresholds applied in this study are listed in Table 2.

2.2. Data cleaning and preparation

Raw data were processed through a standardised cleaning pipeline applied consistently across all five sites. Datetime fields were parsed and a season variable was assigned to each observation based on Southern Hemisphere convention: Summer (December–February), Autumn (March–May), Winter (June–August), and Spring (September–November). Observations recorded as below or above the detection limit were retained and flagged rather than removed, accounting for 1.6%–10.5% of records depending on site. Each censored observation

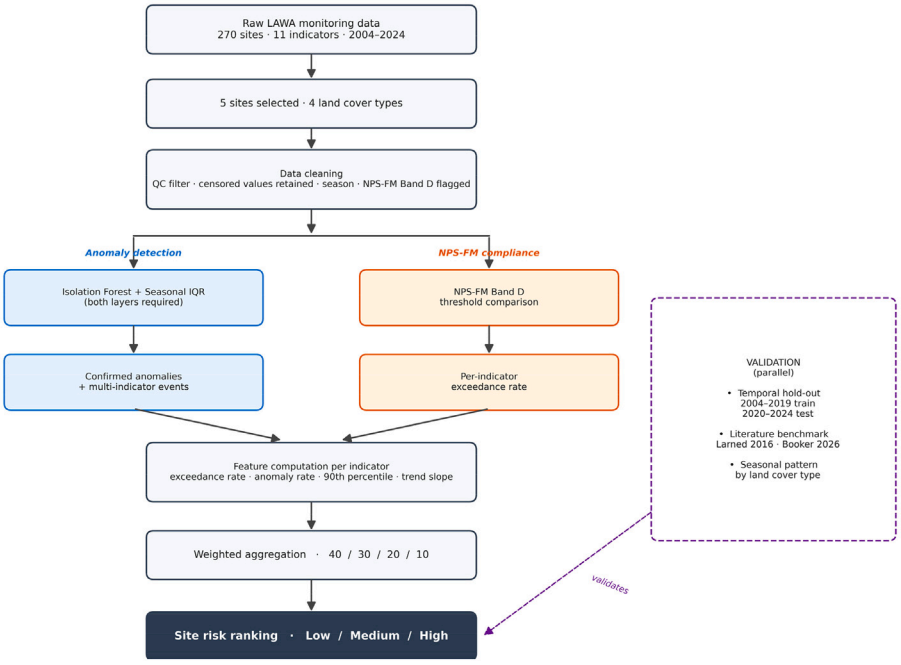


Fig. 1. End-to-end workflow of the anomaly detection and NPS-FM non-compliance risk scoring pipeline. Raw LAWA monitoring records are cleaned and passed through two independent anomaly detection layers (Isolation Forest and seasonal IQR), whose intersection defines confirmed anomalies. In parallel, observations are classified against NPS-FM Band D thresholds. Per-indicator features (exceedance rate, confirmed anomaly rate, 90th-percentile severity, and Mann–Kendall trend slope) are combined into a weighted risk score, which is then assessed through three independent validation methods.

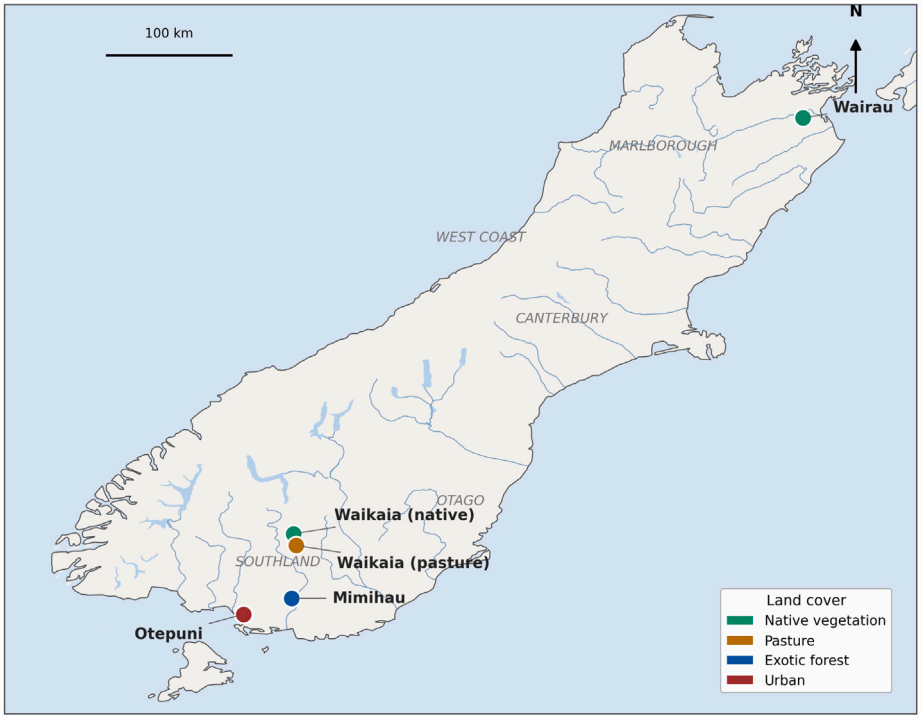


Fig. 2. Location of the five South Island river monitoring sites analysed in this study, coloured by REC land cover type. Wairau (Marlborough, native vegetation) is paired with four Southland sites covering the full land-cover gradient: Waikaia native (native vegetation), Mimihau (exotic forest), Waikaia pasture (sheep and beef pasture), and Otepuni (urban). The Waikaia native and Waikaia pasture sites are located on the same river system 15 km apart and provide a paired-catchment comparison of native vegetation versus pasture water quality. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

enters the analysis at its reported detection limit, with no substitution applied; for below-detection readings, this is an upper bound on the true concentration, so exceedance rates for low-concentration indicators are conservative in the direction of overstating rather than understating risk. Physical implausibility checks removed negative values arising from instrument artefacts. Quality control filtering removed

Table 1
Monitoring site characteristics for the five South Island rivers, ordered by land cover gradient. All sites cover the period 2004–2024. Short labels are used in subsequent tables, figures and text; official LAWA monitoring station names are given in full.

Short label	Official site name	Land cover	Region	Coordinates	Number of observations n
Wairau	Wairau River at Tuamarina	Native vegetation	Marlborough	−41.44, 173.97	2387
Waikaia-N[a]	Waikaia River at Waikaia	Native vegetation	Southland	−45.63, 168.96	2050
Mimihau	Mimihau Stream Trib. at Venlaw Forest	Exotic forest	Southland	−46.28, 169.00	1885
Waikaia-P[a]	Waikaia River at Waipounamu Bridge	Pasture	Southland	−45.69, 168.88	2042
Otepuni	Otepuni Creek at Nith Street	Urban	Southland	−46.42, 168.35	1994

[a] Paired catchment: Waikaia-N (upstream, native vegetation) and Waikaia-P (downstream, pasture) on the same river ~15 km apart.

Table 2
Band D boundary values from the NPS-FM attribute tables applied at observation level as a screening proxy rather than a formal compliance test. Eight indicators with fixed thresholds contribute to the weighted risk score; three indicators are excluded from the score (with reasons in notes) but retained for anomaly detection and trend analysis.

Indicator	Threshold (Band D)	Breach when	Attribute protected	In risk score
Included (n = 8)				
E. coli	540 cfu/100 mL	Above	Human contact, recreation	Yes
Nitrate nitrogen	6.9 mg/L	Above	Aquatic toxicity	Yes
Dissolved reactive phosphorus	0.018 mg/L	Above	Ecosystem (algal growth)	Yes
Total phosphorus	0.05 mg/L	Above	Ecosystem health	Yes
Turbidity	5.5 NTU	Above	Visual clarity	Yes
Clarity (BDISC)	1.0 m	Below	Light penetration	Yes
Ammoniacal nitrogen	1.3 mg/L	Above	Aquatic toxicity	Yes
Total nitrogen	1.0 mg/L	Above	Ecosystem health	Yes
Excluded (n = 3); retained for anomaly detection and trend analysis				
pH	6.0–9.0	Outside range	Aquatic life	–[a]
Dissolved inorganic nitrogen	Not set	–	Ecosystem health	–[b]
Total oxidised nitrogen	Not set	–	Ecosystem health	–[c]

[a] pH never breaches the acceptable range at any of the five sites across the 2004–2024 record; inclusion would not contribute to score variation.
[b] The NPS-FM 2020 does not set a value for band D for dissolved inorganic nitrogen; regional limits vary under clause 3.13 (nutrient outcome setting).
[c] Total oxidised nitrogen is the sum of nitrate and nitrite; the nitrogen signal is already captured by nitrate nitrogen.

Table 3
Number of valid observations per indicator per site following quality-control filtering. Water clarity at Wairau ends in 2018; all other series run to 2024.

Indicator	Wairau	Waikaia-N	Mimihau	Waikaia-P	Otepuni
E. coli	229	179	167	178	175
Nitrate nitrogen	246	88	69	88	75
Ammoniacal nitrogen	248	180	167	178	176
Total nitrogen	246	252	241	250	250
Dissolved reactive phosphorus	243	171	157	169	168
Total phosphorus	246	171	157	170	168
Turbidity	246	192	179	191	187
Clarity (BDISC)	153	205	173	212	193

observations assigned NEMS code 200, denoting data whose quality has not been verified. This removed 19.2% of raw records across the five sites, with the exclusion concentrated at the four Southland sites; the sensitivity of the reported results to this choice is assessed in Section 3.5. NPS-FM Band D exceedance status was flagged at the observation level during cleaning so that compliance information was available to all downstream analytical steps without recalculation. Following cleaning, usable record counts ranged from 1885 to 2387 observations per site, with counts for each indicator at each site given in Table 3.

2.3. Anomaly detection

Two independent anomaly detection methods, Isolation Forest and seasonal IQR were applied to each observation per indicator. An observation was classified as a confirmed anomaly if flagged by both methods simultaneously. Dates on which two or more indicators were simultaneously confirmed as anomalous were recorded separately as multi-indicator events, representing the highest-confidence signal of physically real pollution or hydrological events.

The two criteria are not redundant. A high summer reading may be unremarkable against the annual record yet extreme for summer, while a moderate reading may be typical for its season yet structurally isolated in the value–month feature space. Requiring both raises the evidential bar for the confirmed anomaly rate, which carries 30% weight in the risk score. The seasonal layer additionally applies a directional rule for water clarity, flagging only values below the lower fence, so that unusually high clarity flagged by the two-sided Isolation Forest layer is excluded by the intersection (see Fig. 3).

2.3.1. Isolation forest

Isolation Forest is an unsupervised ensemble algorithm that detects anomalies by isolating observations through recursive random partitioning of the feature space (Liu et al., 2008, 2012). Because anomalous observations are few and structurally different from the bulk of the data, they require fewer partitions to isolate than normal observations and consequently receive higher anomaly scores. The algorithm constructs an ensemble of binary isolation trees, each grown on a random sub-sample of the data, and records for each observation x, the path length $h(x)$ traversed from the root of a tree to the terminating node. The anomaly score for x in a sample of size n is defined as

$$s(x,n) = 2^{-\frac{E[h(x)]}{c(n)}} \tag{1}$$

where $E[h(x)]$ is the expected path length of x over the ensemble and $c(n)$ is a normalising constant equal to the average path length of an unsuccessful search in a Binary Search Tree of n points:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}, \qquad H(i) \approx \ln(i) + \gamma, \tag{2}$$

with $\gamma \approx 0.5772$ denoting the Euler–Mascheroni constant (Liu et al., 2012). Scores approach 1 for clear anomalies, fall well below 0.5 for normal points, and equal 0.5 when the average path length matches that of a random point in the data.

The algorithm was applied independently to each indicator time series rather than to a combined multi-indicator matrix, for three reasons. First, a multivariate model requires complete observation vectors, and

Table 4
Summary statistics for the cleaned monitoring dataset (2004–2024). Each cell shows the median value followed by the inter-quartile range (Q25–Q75) in parentheses, computed across all valid observations at that site for that indicator.

Indicator	Unit	Wairau (Native)	Waikaia-N (Native)	Mimihau (Forest)	Waikaia-P (Pasture)	Otepuni (Urban)
E. coli	cfu/100 mL	9.0 (3.0–27.0)	130.0 (70.0–385.0)	18.0 (10.0–52.0)	130.0 (70.0–260.0)	1300.0 (500.0–3000.0)
Nitrate-N	mg/L	0.102 (0.041–0.184)	0.137 (0.103–0.179)	0.184 (0.147–0.240)	0.625 (0.440–0.770)	1.190 (0.930–1.680)
Ammoniacal-N	mg/L	0.0040 (0.0020–0.0050)	0.0100 (0.0100–0.0100)	0.0100 (0.0100–0.0100)	0.0100 (0.0100–0.0100)	0.0570 (0.0257–0.0960)
Total nitrogen	mg/L	0.159 (0.094–0.244)	0.250 (0.210–0.320)	0.350 (0.260–0.480)	0.680 (0.542–0.840)	2.050 (1.560–2.700)
DRP	mg/L	0.0030 (0.0020–0.0053)	0.0060 (0.0040–0.0070)	0.0110 (0.0080–0.0140)	0.0068 (0.0050–0.0080)	0.0120 (0.0090–0.0163)
Total phosphorus	mg/L	0.0070 (0.0050–0.0120)	0.0140 (0.0100–0.0210)	0.0200 (0.0150–0.0250)	0.0140 (0.0110–0.0208)	0.0480 (0.0338–0.0710)
Turbidity	NTU	1.41 (0.71–4.88)	2.10 (1.23–3.82)	2.10 (1.31–3.40)	2.40 (1.17–4.10)	4.60 (3.80–7.75)
Clarity (BDISC)	m	2.58 (1.01–4.49)	1.95 (1.13–2.66)	1.50 (1.20–1.54)	1.81 (0.99–2.54)	0.89 (0.63–1.17)

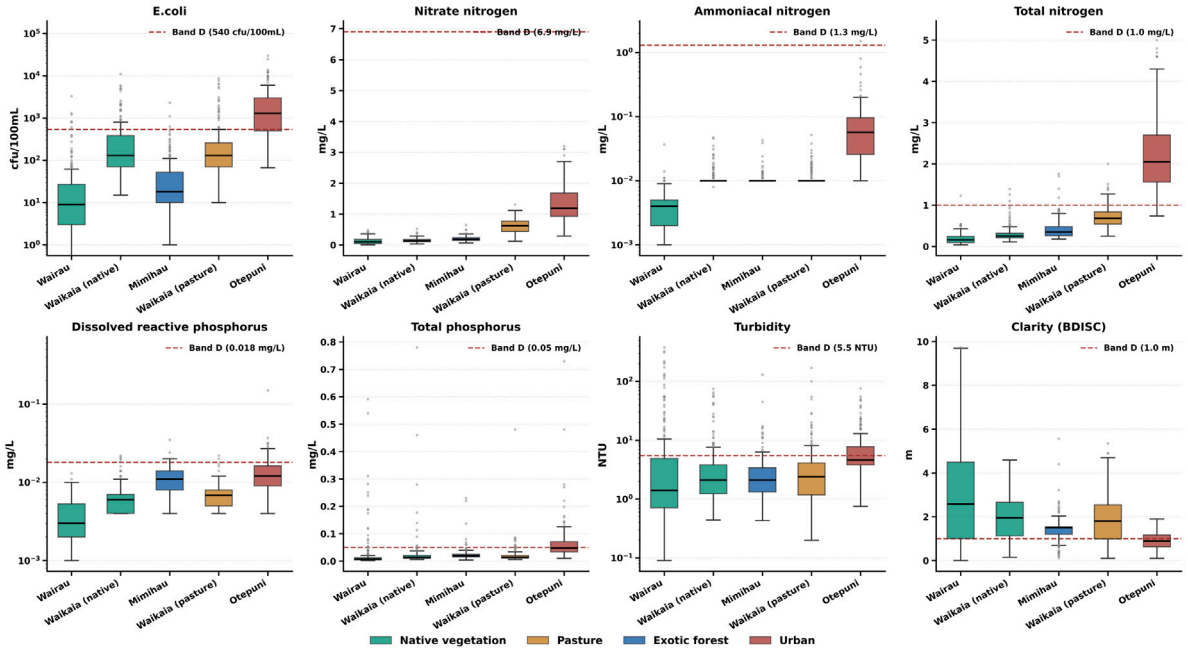


Fig. 3. Distribution of cleaned values (2004–2024) for the eight NPS-FM scored indicators across the five South Island monitoring sites. Boxes show the inter-quartile range; horizontal bars are medians; whiskers extend to 1.5×IQR; outliers are plotted as small grey dots. Box fill colour denotes REC land cover type. Dashed red lines mark the NPS-FM Band D thresholds. The land-cover-driven gradient in median concentrations is visible directly in the raw data prior to any modelling. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

only 24%–35% of sampling dates at the Southland sites record all 11 indicators simultaneously (3% at Wairau); restricting the analysis to complete cases would discard the majority of the record. Second, the risk score is constructed at indicator level, so a multivariate detector could not attribute a flagged observation to a specific contaminant, removing the indicator-level interpretability the framework is designed to provide. Third, cross-indicator dependence is instead captured through the multi-indicator event analysis described in Section 2.3.3. Two features were supplied to each model: the indicator value and the month of measurement.

$$m_{\sin} = \sin\left(\frac{2\pi m}{12}\right), \qquad m_{\cos} = \cos\left(\frac{2\pi m}{12}\right), \tag{3}$$

ensuring that December ($m = 12$) and January ($m = 1$) are treated as adjacent in feature space rather than as opposite ends of a linear scale. The contamination parameter was set to 0.05, calibrating each model to flag approximately the top 5% of most unusual readings per indicator per site as the operational anomaly threshold. Models were trained on the full 2004–2024 record for each site and indicator. A separate model was fitted for each site and each indicator; no model was fitted across sites or across indicators. Series with fewer than 20 valid observations were not modelled, a condition met by all scored indicators at all five sites. No hyperparameter tuning was performed, the ensemble size and contamination parameter were fixed a priori and applied uniformly across all series, and the random seed was fixed so that the reported anomaly flags are exactly reproducible. Sensitivity of the results to the contamination parameter was assessed by re-running the complete pipeline at values of 0.02, 0.05 and 0.10. Risk grades are unchanged at all five sites across this range and the site ranking is preserved exactly (Spearman $\rho = 1.000$ at every level), with site scores varying by approximately one point between the extremes (Table 5). The confirmed anomaly rate scales with the parameter, rising from 1.4%–1.8% at 0.02 to 3.9%–5.3% at 0.10, but remains below the nominal contamination level at every site because the intersection with the seasonal layer excludes observations that the Isolation Forest layer flags alone.

Table 5
Site risk scores and confirmed anomaly rates under three Isolation Forest contamination settings. The published analysis uses 0.05. Grades are medium at all sites except Otepuni, which is high under every setting.

Site	Site risk score (0–100)			Confirmed anomaly rate (%)		
	0.02	0.05	0.10	0.02	0.05	0.10
Wairau	13.75	14.24	14.82	1.55	3.18	5.07
Waikaia-N	15.54	16.24	16.94	1.76	3.46	5.32
Mimihau	15.01	15.60	16.43	1.75	3.29	5.15
Waikaia-P	18.56	19.02	19.62	1.37	2.45	3.87
Otepuni	33.67	34.12	34.54	1.55	3.11	4.26
Spearman ρ vs. published ranking				1.000	1.000	1.000

2.3.2. Seasonal IQR

The seasonal interquartile range method flags observations that deviate substantially from the typical range of values recorded in the same season. For each indicator and season combination, the interquartile range was computed from all available readings in that season across the full 2004–2024 record. An observation was flagged as seasonally anomalous if its value exceeded $Q_3 + 1.5 \times \text{IQR}$ for indicators where high values represent degradation, or fell below $Q_1 - 1.5 \times \text{IQR}$ for water clarity, the conventional outlier fence (Tukey, 1977). A minimum of five observations per season per indicator was required before the threshold was applied. Season boundaries followed the Southern Hemisphere convention defined in Section 2.2.

2.3.3. Multi-indicator event co-occurrence

For each site, the co-occurrence of confirmed anomalies on multi-indicator event dates was compared against the rate expected if anomaly flags were distributed independently. The observed rate was computed as the proportion of confirmed anomaly observations falling on multi-indicator dates, and the expected rate as the proportion of all unique sampling dates that were multi-indicator dates. The ratio of observed to expected co-occurrence served as an internal consistency check on the anomaly detection pipeline: a ratio of two or above was taken as evidence that the detected anomalies cluster substantially more than those would arise from independent statistical noise. This expectation is grounded in established catchment hydrology of storm-driven contaminant mobilisation, in which suspended sediment, faecal indicator bacteria, and dissolved and particulate nutrients are exported simultaneously through surface runoff and resuspension of bed-sediment stores (Davies-Colley et al., 2008; Gall et al., 2022), supporting interpretation of confirmed multi-indicator dates as physically real pollution or hydrological events rather than measurement artefacts.

2.4. NPS-FM exceedance classification

Each observation was compared directly against the Band D for its indicator as defined in the NPS-FM 2020 (amended October 2024). An exceedance was recorded when a reading exceeded the threshold for indicators where high values represent degradation, fell below the threshold for water clarity, or fell outside the acceptable range of 6.0–9.0 units for pH. Exceedance status was assigned at the observation level during data cleaning as described in Section 2.2. The NPS-FM exceedance rate for each indicator at each site was computed as the proportion of all readings over the 2004–2024 record that breached the Band D limit. Three indicators were excluded from the risk score: dissolved inorganic nitrogen, for which the NPS-FM does not set band D limits; total oxidised nitrogen, excluded to avoid double-counting the nitrogen signal already captured by nitrate nitrogen; and pH, which was assessed against its acceptable range but recorded no breaches at any of the five sites across the full monitoring record. These three indicators were retained for anomaly detection as described in Section 2.3. NPS-FM Band D thresholds, where set, are listed in Table 2 and summary statistics for eight selected indicators are listed in Table 4.

2.5. Risk score calculation

A quantitative risk score was computed for each site by aggregating indicator-level features into a weighted composite measure of NPS-FM non-compliance risk. Five summary features were derived from the full record for each of the eight threshold indicators per site: median value, 90th percentile value, NPS-FM exceedance rate, confirmed anomaly rate, and Mann–Kendall trend slope. The median served as a contextual descriptor of baseline conditions and did not enter the weighted score. The per-indicator risk score is a weighted aggregation of four features defined on the $[0, 1]$ scale. The NPS-FM exceedance rate E_i is the proportion of valid observations that breach the Band D threshold T_i:

$$E_i = \frac{1}{n_i} \sum_{t=1}^{n_i} \mathbf{1}\{\delta_i (x_{i,t} - T_i) > 0\}, \tag{4}$$

where $x_{i,t}$ is the observed value at time t, n_i is the number of valid observations, and $\delta_i \in \{+1, -1\}$ encodes the breach direction: $\delta_i = +1$ for indicators where breach is defined as $x > T$ (*E. coli*, both nitrogen species, both phosphorus species, ammoniacal nitrogen, turbidity), and $\delta_i = -1$ for water clarity (BDISC), where breach is defined as $x < T$.

The confirmed anomaly rate A_i is the proportion of observations flagged simultaneously by the Isolation Forest layer ($\mathcal{A}_i^{IF}$) and the seasonal interquartile range layer ($\mathcal{A}_i^{IQR}$) defined in Sections 2.3.1 and 2.3.2:

$$A_i = \frac{\left|\mathcal{A}_i^{IF} \cap \mathcal{A}_i^{IQR}\right|}{n_i}. \tag{5}$$

The severity feature S_i expresses worst-case observed concentrations relative to the legal threshold and is capped at unity:

$$S_i = \min\left(\rho_i,\ 1\right), \quad \rho_i = \begin{cases} P_{90}(x_{i,t})/T_i & \text{if } \delta_i = +1, \\ T_i / P_{10}(x_{i,t}) & \text{if } \delta_i = -1, \end{cases} \tag{6}$$

where ρ_i is the worst-case-to-threshold ratio: for indicators where breach is defined as $x > T$, the 90th percentile of the observed values divided by the threshold; for water clarity, the threshold divided by the 10th percentile of observed values, capturing the most severely under-threshold readings.

The trend feature τ_i^+ uses the Mann–Kendall rank correlation τ_i computed over the full 2004–2024 record (Mann, 1945; Kendall, 1975; Hirsch and Slack, 1984), with only worsening trajectories contributing positively to risk:

$$\tau_i^+ = \max\left(\delta_i \cdot \tau_i,\ 0\right), \tag{7}$$

ensuring that improving or stable trends contribute zero regardless of indicator direction.

The per-indicator risk score at site s is the weighted sum of the four features:

$$r_{i,s} = w_E E_i + w_A A_i + w_S S_i + w_\tau \tau_i^+, \tag{8}$$

with weights $w_E = 0.40$, $w_A = 0.30$, $w_S = 0.20$, $w_\tau = 0.10$ as listed in Table 6, and $w_E + w_A + w_S + w_\tau = 1$. The weights are based on expert judgement, not empirical calibration. The site risk score R_s is the

Table 6
Feature weights for the site-level risk score. Per-indicator scores are computed as the weighted sum of four normalised features on a 0–1 scale, then averaged across the eight included indicators and rescaled to 0–100. Sites are graded as High (≥30), Medium (10–29), or Low (<10) risk based on the resulting site score.

Feature	Weight	Rationale
NPS-FM exceedance rate	40%	Most policy-relevant; directly reflects legal non-compliance frequency
Confirmed anomaly rate	30%	Captures episodic event frequency beyond threshold breaches
90th percentile (normalised)	20%	Worst-case severity, normalised against the NPS-FM Band D limit
Trend slope (Mann–Kendall τ)	10%	Forward-looking signal; a worsening 21-year trend adds risk

mean of per-indicator scores across the 8 NPS-FM threshold indicators, rescaled to $[0, 100]$:

$$R_s = \frac{100}{N} \sum_{i=1}^{N} r_{i,s}. \tag{9}$$

Risk grades are assigned as High for $R_s \geq 30$, Medium for $10 \leq R_s < 30$, and Low for $R_s < 10$. The 0–100 score range with banded risk grades follows the design principle adopted in established water-quality index frameworks, including the Canadian Council of Ministers of the Environment Water Quality Index (Canadian Council of Ministers of the Environment, 2001) and the New Zealand NPS-FM banded system (Ministry for the Environment, 2020), under which a continuous quality scale is partitioned into discrete management-relevant tiers. These boundaries are indicative management tiers rather than validated classification thresholds. No analytical conclusion in this study depends on their placement: the risk score is continuous and the site ranking, not the banding, carries the substantive result. No site in this study falls below 10, so the Low–Medium boundary is untested by these data, and calibration against a larger site population would be required before the bands are used operationally (Lumb et al., 2011; Sutadian et al., 2016).

The four features were combined using fixed weights chosen to reflect their relative roles in capturing NPS-FM non-compliance risk. The NPS-FM exceedance rate received the dominant weight (40%) as it is the only feature directly tied to the definition of a breach under the Band D limits. The confirmed anomaly rate (30%) was weighted next to capture episodic events that exceedance rates alone underestimate at sites with clean baselines, followed by the 90th percentile (20%) to incorporate worst-case severity normalised against each indicator's threshold. The Mann–Kendall trend slope received the lowest weight (10%), reflecting both the higher noise inherent in 21-year trend estimates and its role as a forward-looking adjustment rather than a measure of current state. The weights were set by judgement informed by this regulatory ordering, not derived by optimisation: without labelled pollution events there is no outcome against which weights could be fitted, and fitting them to the observed scores would be circular.

Sensitivity of the site ranking to the weighting choice was tested by recomputing risk scores under three alternative schemes: equal weighting (25/25/25/25), exceedance-heavy weighting (50/25/15/10), and zero-trend weighting (45/35/20/0), with results shown in Fig. 4. Otepuni retained its high risk grade and Wairau remained the lowest-scoring site under all four schemes, and all four Medium sites retained Medium grades under every scheme tested. The ranking is preserved exactly under two of the three alternatives (Spearman $\rho = 1.000$); the only order change was a position swap between Waikaia-N and Mimihau (ranks 3 and 4) under equal weighting ($\rho = 0.900$), where their separation falls below one point (18.5 vs. 17.7). Under equal weighting, Otepuni's score fell to 31.7, remaining above the high threshold but by a smaller margin. The relative ordering of the features therefore matters to the result; their precise magnitudes do not.

2.6. Validation

Three independent validation methods were applied to assess the reliability of the risk scores, each testing a different aspect of pipeline performance. Two conventional thresholds are used across these checks. An indicator is treated as elevated risk in the temporal hold-out if its training-period exceedance rate exceeds 5%, and a site is recorded as PASS where 70% or more of the compared indicators meet the relevant criterion. Neither value is derived: they are working conventions chosen to give a consistent reporting rule across three checks whose underlying quantities are not directly comparable. Per-site consistency percentages are reported in full in Table 12, so the classification can be recovered at any other threshold, and the review results are interpreted on their substance rather than on the label.

2.6.1. Temporal hold-out

Risk features were computed on data from 2004 to 2019 only. An indicator was classified as elevated risk in the training period if its NPS-FM exceedance rate exceeded 5%. Validation assessed whether elevated-risk indicators recorded actual NPS-FM breaches in the 2020–2024 test period, and whether indicators not classified as elevated risk remained compliant in the test period. A site was considered to pass this validation if 70% or more of indicators were correctly classified.

2.6.2. Literature benchmarking

Site medians for each indicator were compared against published national medians for matching REC land cover types from two sources: Larned et al. (2016) for *E. coli*, Turbidity and Total nitrogen, and Booker et al. (2026) for Total phosphorus and DRP as listed in Table 7. A site median was considered consistent with the benchmark if it fell within a factor of two of the reference value (ratio 0.5–2.0). A site was considered to pass if 70% or more of the compared indicators met this criterion. The factor-of-two tolerance is consistent with the approach used in national reference-condition work (McDowell et al., 2013) and accommodates legitimate within-class variation arising from climate, catchment hydrology, and land-use intensity differences not captured by REC land cover alone.

The interpretation of consistency or departure at each site was further informed by land-use-specific New Zealand literature representative of each site's catchment characteristics. For native vegetation sites (Wairau, Waikaia-N), Julian et al. (2017) documents the relationship between native land cover and low-intensity baseline water quality across a 26-year national record. For the exotic forest site (Mimihau), Baillie and Neary (2015) reviews water quality across the planted-forest cycle in New Zealand, including the elevated sediment and nutrient signatures expected during harvest operations. For the pasture site (Waikaia-P), Snelder et al. (2021a) provides attribution of nutrient and bacterial trends to agricultural land use intensity, with the caveat that most New Zealand pastoral comparators reflect intensive dairy systems rather than the sheep and beef regime that characterises Waikaia-P. For the urban site (Otepuni), Gadd et al. (2020) and Cookson et al. (2024) document the structural contamination patterns of New Zealand urban streams, including the persistent elevated *E. coli* and nutrient signatures characteristic of this land cover. Findings are further contextualised against the recent national reassessment (Booker et al., 2026), which confirms that the land-cover-driven gradients reported by Larned et al. (2016) persist in data through 2024.

2.6.3. Seasonal pattern validation

For each indicator at each site, the season with the highest exceedance rate was identified and compared against the expected window for the site's land cover. A site was considered to pass if 70% or more of indicators showed peaks consistent with the expected land cover signature. For urban sites, where no seasonal peak is expected, consistency was assessed using a flatness ratio defined as the highest seasonal exceedance rate divided by the lowest, with values below

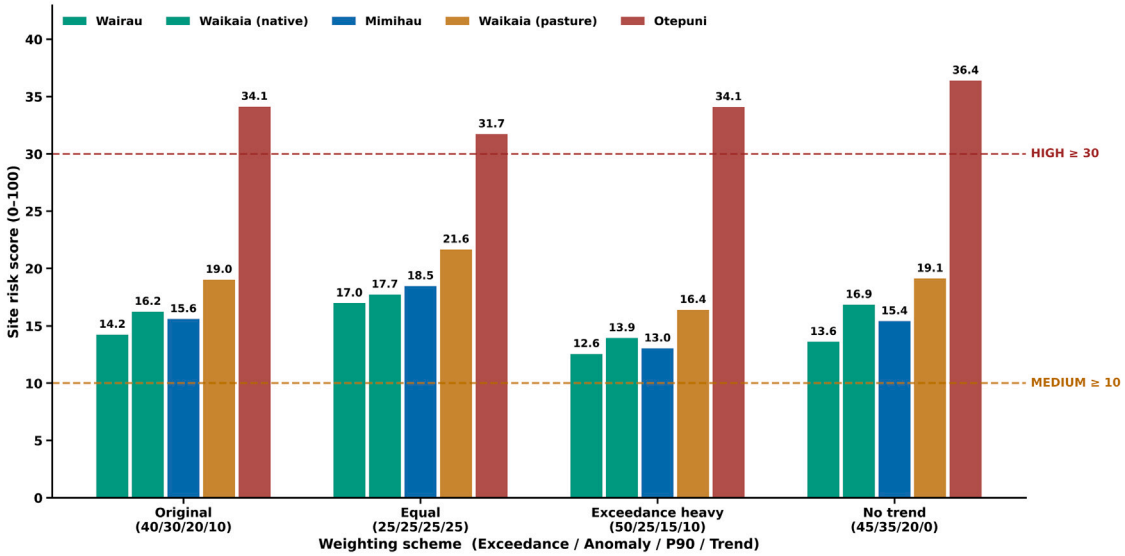


Fig. 4. Site risk scores recomputed under three alternative weighting schemes for the four risk-score features (exceedance rate, anomaly rate, 90th-percentile severity, trend slope). The original (40/30/20/10) scheme is shown alongside equal weighting, an exceedance-heavy variant, and a trend-free variant. Otepuni retains High-risk classification under all four schemes and the four other sites retain Medium classification in every case, demonstrating that the site risk rankings are robust to the choice of weighting.

Table 7
Published national medians by REC land cover used as benchmark reference values for the literature benchmarking validation (Section 2.6.2). *E. coli*, Turbidity and Total nitrogen benchmarks from Larned et al. (2016); Total phosphorus and DRP benchmarks from Booker et al. (2026).

Indicator	Unit	Native veg	Exotic forest	Pasture	Urban
E. coli	cfu/100 mL	10	30	126	490
Turbidity	NTU	1.5	2.5	5.0	6.0
Total nitrogen	mg/L	0.20	0.35	0.80	2.0
Total phosphorus	mg/L	0.015	0.020	0.040	0.060
DRP	mg/L	0.005	0.007	0.012	0.020

2.0 taken as indicating no dominant seasonal peak. This threshold is likewise a working convention.The seasonal distribution of NPS-FM exceedances at each site was compared against the seasonal contamination signatures expected for its dominant land cover type. This check works as an external consistency test because the timing of pollutant delivery to surface waters is governed by physical processes specific to each land use (Larned et al., 2016), and a site whose observed exceedance peaks align with these established processes provides independent evidence that the detection pipeline is identifying real environmental signals rather than measurement artefacts.

Native vegetation sites were expected to peak in Summer or Autumn, when contamination arrives episodically through intense rainfall events that mobilise small loads of sediment and organic material from otherwise clean catchments lacking persistent point or diffuse sources (Julian et al., 2017). Pasture sites were expected to peak in Winter or Spring, driven by saturated-soil runoff, direct livestock access to waterways during wet ground conditions, and the application of effluent and fertiliser onto wet soils that have limited capacity to retain dissolved nutrients (Wilcock et al., 2013; Monaghan et al., 2007; Davies-Colley et al., 2008). Exotic forest sites were expected to peak in Winter or Spring during and immediately following harvest operations, which are scheduled to avoid the summer fire risk window and which expose mineral soil and disturbed organic horizons to wet-season rainfall, producing pulses of suspended sediment, particulate phosphorus and ammoniacal nitrogen (Baillie and Neary, 2015). Urban sites were expected to show near-uniform year-round contamination driven by the structural features of urban catchments — impervious surfaces that bypass natural filtration, ageing stormwater infrastructure, combined sewer overflows during rainfall, and dry-weather contributions from cross-connections and broken pipes — which deliver nutrients, sediment and bacterial loads to receiving waters regardless of season (Gadd et al., 2020; Walsh et al., 2005).

For each indicator at each site, the season with the highest exceedance rate was identified and compared against the expected window for the site's land cover. A site was considered to pass if 70% or more of indicators showed peaks consistent with the expected land cover signature. For urban sites, where no seasonal peak is expected, consistency was assessed using a flatness ratio defined as the highest seasonal exceedance rate divided by the lowest, with values below 2.0 taken as indicating no dominant seasonal peak. This threshold is likewise a working convention.

3. Results

Following the cleaning pipeline described in Section 2.2, usable record counts ranged from 1885 observations at Mimihau to 2387 at Wairau. All five sites retain observations across the full 2004–2024 window and all 11 indicators, but coverage within that window is uneven. The records removed by the quality-control filter are not scattered individual readings: they are whole sampling occasions, on which every indicator measured that day was coded as unverified and therefore excluded together. At the four Southland sites this removes four years of record for most scored indicators, leaving gaps in the middle of the series rather than at either end. Total nitrogen is unaffected at every site, and water clarity at Wairau ends in 2018. No values were imputed to fill these gaps: each site's features are computed from the observations that remain, and Section 3.5 reports what happens when the excluded records are retained instead. NPS-FM exceedance rates revealed a clear land cover gradient before any modelling was applied (Fig. 5). Otepuni (Urban) exceeded the Band D limit for Total nitrogen in 96% of readings and the *E. coli* limit in 72% of readings over the 2004–2024 record. Turbidity at the same site exceeded the 5.5 NTU limit in 41% of readings. These rates are consistent with year-round structural contamination of an urban catchment, where impervious surfaces, ageing stormwater infrastructure, and probable combined sewer overflows deliver nutrient and bacterial loads to the creek regardless of

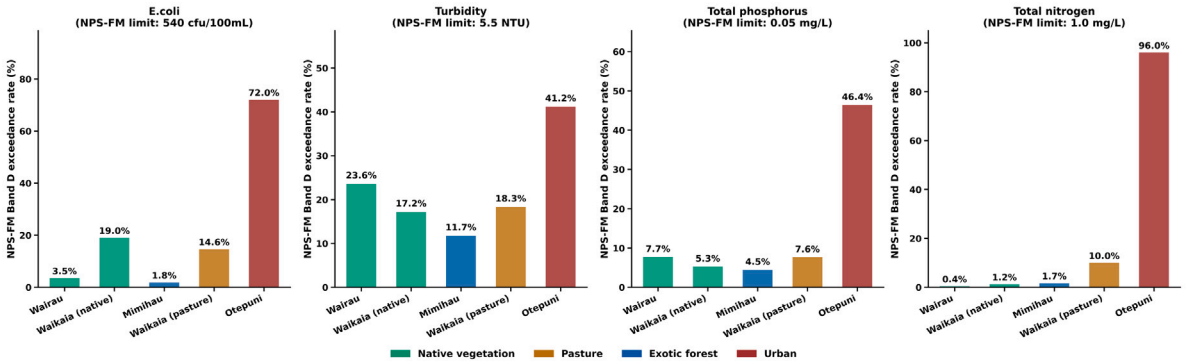


Fig. 5. NPS-FM Band D exceedance rates for the four headline indicators (*E. coli*, Turbidity, Total phosphorus, Total nitrogen) at the five South Island monitoring sites over 2004–2024. Sites are ordered by land cover gradient. Bar colours indicate land cover type. The land cover gradient is visible directly in the raw exceedance data before any modelling is applied. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

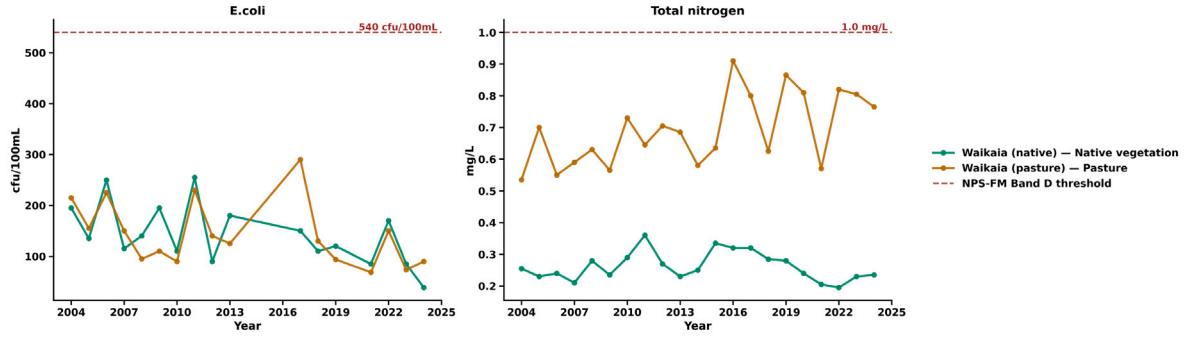


Fig. 6. Paired catchment comparison on the Waikaia River, with the upstream native vegetation site (es-00042) and downstream pasture site (es-00025) located 15 km apart. Geology, climate, and altitude are effectively identical between the two sites; land use is the only systematic variable. Annual median values for *E. coli* (left) and Total nitrogen (right) over 2004–2024 are shown, with the NPS-FM Band D threshold marked. Total nitrogen at the pasture site is consistently 0.3–0.5 mg/L higher than at the upstream native site.

season or rainfall (Gadd et al., 2020; Cookson et al., 2024). The near-flat seasonal distribution of *E. coli* and total nitrogen exceedances at Otepuni, discussed further in Section 4.3, supports an infrastructure-driven contamination pathway for these contaminants rather than the weather-driven episodic pattern observed at less-modified sites; turbidity at the same site retains a Winter peak consistent with rainfall-driven sediment mobilisation.

Native vegetation sites recorded substantially lower exceedance rates for chemical and bacterial indicators. *E. coli* exceeded the legal limit in only 3.5% of readings at Wairau and Total nitrogen in only 0.4%. Turbidity at Wairau, however, exceeded the 5.5 NTU limit in 23.6% of readings, the highest turbidity exceedance rate among the four non-urban sites. This counterintuitive result reflects the physical hydrology of the Wairau catchment rather than land-use pressure: the river is a large braided system draining geologically active terrain in the Marlborough region, and natural sediment mobilisation from erodible greywacke and schist during high-flow events plausibly generates turbidity spikes that exceed the legal threshold episodically even in the absence of anthropogenic inputs.

The Waikaia paired catchment provides the cleanest available isolation of land-use effects within the dataset. The two monitoring points sit 15 km apart on the same river, with the upstream site under native vegetation and the downstream site under pasture; geology, climate, and altitude are effectively identical. Across the full 2004–2024 record, the pasture site recorded Total nitrogen concentrations 0.3–0.5 mg/L higher than the upstream native site (Fig. 6). The most plausible attribution for this consistent elevation is fertiliser leaching from sheep and beef farming on pasture soils adjacent to the river, combined with direct nutrient inputs from livestock urine and faeces, transported to the river via subsurface flow and overland runoff during rainfall events (Snelder et al., 2021a). Regional source tracking supports this attribution, with wildfowl a recognised faecal source under base-flow conditions and ruminant markers becoming more prominent following rainfall (Environment Southland, 2024).

The exotic forest site (Mimihau) recorded intermediate exceedance rates for nutrients and turbidity, with notably low *E. coli* exceedances confirming the absence of livestock as a bacterial source and distinguishing forestry from pastoral land use despite their broadly similar physical disturbance signatures. Elevated dissolved reactive phosphorus and total nitrogen at this site are plausibly attributable to nutrient mobilisation associated with the forestry harvest cycle: clear-felling, soil disturbance during replanting, and decomposition of harvest debris in and near waterways release dissolved nutrients over several years following each operation, while logging access roads contribute sediment and nutrient runoff during rainfall.

3.1. Anomaly detection and multi-indicator events

The two-layer anomaly detection framework described in Section 2.3 produced confirmed anomaly counts in the range of 50 to 76 observations per site, corresponding to between 2.4% and 3.5% of all observations across the five sites (Table 8). These rates are computed against the per-series observation counts reported in Table 3, which range from 69 observations (nitrate nitrogen at Mimihau) to 252 (total nitrogen at Waikaia-N). The consistency of confirmed anomaly rates across sites of very different land cover types is a meaningful structural feature of the pipeline rather than an artefact. Isolation Forest and seasonal IQR both detect deviations relative to each site's

Table 8
Confirmed anomaly counts, rates and multi-indicator event counts per site over 2004–2024. Confirmed anomalies are observations flagged simultaneously by both the Isolation Forest layer and the seasonal interquartile range layer described in Section 2.3. Rate is computed as confirmed anomalies as a percentage of all valid observations at that site. Multi-indicator events are dates on which two or more indicators were simultaneously confirmed as anomalous.

Site	Land cover	Region	Total observations	Confirmed (%)	Multi-indicator events
Wairau	Native vegetation	Marlborough	2387	76 (3.18%)	15
Waikaia (native)	Native vegetation	Southland	2050	71 (3.46%)	15
Mimihau	Exotic forest	Southland	1885	62 (3.29%)	11
Waikaia (pasture)	Pasture	Southland	2042	50 (2.45%)	14
Otepuni	Urban	Southland	1994	62 (3.11%)	16

own historical baseline, so a chronically polluted site such as Otepuni does not register a disproportionate number of confirmed anomalies despite recording the highest NPS-FM exceedance rates in the dataset. Persistent contamination is captured by the NPS-FM exceedance rate, which feeds the risk score with 40% weight, while the two anomaly detection layers identify episodic events that depart from the site's normal behaviour and would otherwise be invisible to a threshold-only assessment.

Multi-indicator events (dates on which two or more indicators were simultaneously confirmed as anomalous) are the highest confidence physical signal in the anomaly detection output. Across the five sites, between 11 and 16 multi-indicator events were recorded over the 2004–2024 period (Table 8). Otepuni recorded the most events (16), reflecting the combined effect of a chronically elevated baseline with episodic stormwater overflows during heavy rain. The native vegetation site at Wairau recorded 15 events, consistent with its hydrological responsiveness as a large braided river that produces multi-indicator runoff signatures during intense rainfall.

The single most extreme event in the dataset occurred on 14 June 2017, when nine indicators were simultaneously confirmed as anomalous at the Waikaia native vegetation site and six at the Waikaia pasture site located 15 km downstream. The temporal and spatial coincidence of these signals is consistent with a catchment-wide flood event affecting the entire river system simultaneously, sweeping sediment, nutrients, and bacterial loads through both monitoring points within hours. The detection of this event at multiple indicators at both paired sites, on the same date, is exactly the kind of physical signal that a multi-indicator approach is designed to surface and that single-indicator threshold monitoring would partially miss.

The co-occurrence ratio described in Section 2.3 — the rate at which confirmed anomalies cluster on the same dates across multiple indicators, expressed as a multiple of the rate expected under random noise — ranged from 10.6 to 14.0 across the five sites. Under the null hypothesis that anomaly flags are independent statistical artefacts, co-occurrence on the same date would occur at a rate driven only by sampling chance, so the consistent elevation observed across all five sites is strong evidence that the pipeline is detecting physically real events rather than statistical noise.

3.2. Risk ranking and per-indicator scores

The site-level risk scores produced by the weighted aggregation method (Section 2.5) preserve the land cover gradient visible in the raw exceedance data while incorporating the additional signal contributed by anomaly frequency, worst-case severity, and 21-year trend direction. Otepuni Creek received the only High risk grade in the dataset, with an overall score of 34.12. The remaining four sites all received Medium risk grades, with scores ranging from 14.24 at Wairau to 19.02 at the Waikaia pasture site (Table 9). The Otepuni risk score is driven primarily by Total nitrogen (per-indicator score 0.592), *E. coli* (0.500), water clarity (0.434), and total phosphorus (0.402). The simultaneous elevation across nearly all eight scored indicators is characteristic of structural urban contamination rather than a single dominant pollution pathway.

The Total nitrogen score in particular reflects an exceedance rate of 96%, indicating that the site sits in regulatory non-compliance under almost every monitoring observation. Site-specific evidence supports a structural rather than episodic pathway. The monitoring reach is channelised between concrete banks over a degraded gravel and sediment bed, and is described by the monitoring programme as locally affected by urban stormwater inputs (Land, Air, Water Aotearoa, 2025). Invercargill operates a separated stormwater network discharging to the Otepuni and four other receiving waters through 147 outfalls; during consenting of that network, contamination of the stormwater discharge by human sewage was identified as the principal issue, and the consent granted in 2017 requires the network operator to identify and remedy such contamination where it is found (Environment Southland, 2017). Faecal source tracking in Southland has identified human markers among the contamination sources at urban monitoring sites (Environment Southland, 2024). Cross-connection of foul sewers into a separated stormwater network delivers contamination continuously rather than during rainfall alone, which is consistent with the near-uniform seasonal distribution of *E. coli* and total nitrogen exceedances reported in Section 4.3. The absence of natural filtration capacity in a channelised reach with limited riparian vegetation compounds these inputs. The persistence of nitrogen and bacterial contamination across both wet and dry seasons, discussed further in Section 4.3, supports an infrastructure-driven pathway for these contaminants rather than weather-dependent episodic loading.

Mimihau (Exotic forest) scored 15.60, below the Waikaia pasture site. Water clarity (0.280) and turbidity (0.262) emerged as the top-scoring indicators at this site, both consistent with sediment mobilisation from forestry operations. Dissolved reactive phosphorus (0.197) and Total nitrogen (0.146) also contribute notably, and both exceed the corresponding scores at either native vegetation site, reflecting the documented nutrient mobilisation that accompanies plantation forestry: exposure of mineral soils during clear-felling and replanting, multi-year decomposition of harvest debris in and near waterways, and chronic sediment and nutrient inputs from logging access roads during rainfall events. The notable distinguishing feature of this site is its low *E. coli* risk score (0.117), which confirms the absence of livestock as a bacterial source and separates forestry from pastoral land use despite the

Table 9
Per-indicator risk scores (0–1 scale) at the five South Island monitoring sites. Each cell is the weighted aggregation of NPS-FM exceedance rate (40%), confirmed anomaly rate (30%), 90th-percentile severity (20%), and Mann–Kendall trend slope (10%) for that indicator at that site over 2004–2024. The highest-scoring indicator at each site is shown in bold. Site risk scores in the bottom row are the mean of the eight per-indicator scores, rescaled to 0–100.

Indicator	Wairau	Waikaia-N	Mimihau	Waikaia-P	Otepuni
E. coli	0.074	**0.291**	0.117	0.274	0.500
Nitrate nitrogen	0.017	0.017	0.050	0.054	0.081
Ammoniacal nitrogen	0.060	0.017	0.018	0.018	0.045
Total nitrogen	0.081	0.101	0.146	0.262	**0.592**
Dissolved reactive phosphorus	0.093	0.137	0.197	0.125	0.295
Total phosphorus	0.191	0.169	0.179	0.200	0.402
Turbidity	**0.312**	0.284	0.262	0.287	0.381
Clarity (BDISC)	0.311	0.283	**0.280**	**0.302**	0.434
Site risk score (0–100)	**14.24**	**16.24**	**15.60**	**19.02**	**34.12**
Risk grade	Medium	Medium	Medium	Medium	High

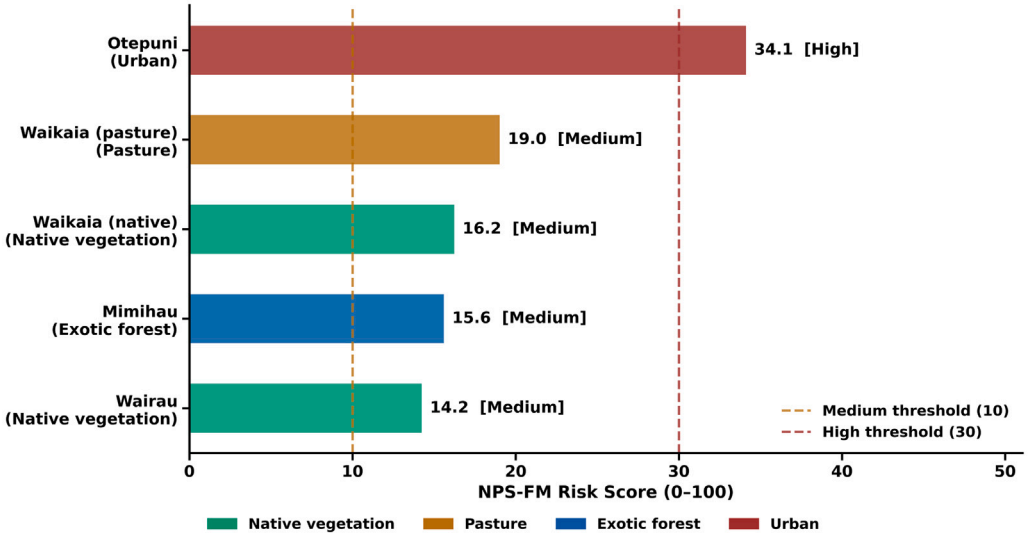


Fig. 7. NPS-FM non-compliance risk scores (0–100) for the five South Island monitoring sites, computed as the weighted aggregation of NPS-FM exceedance rate (40%), confirmed anomaly rate (30%), 90th-percentile severity (20%), and Mann–Kendall trend slope (10%). Risk grade thresholds are marked: Low (<10), Medium (10–29), High (≥30).

broadly similar nutrient and sediment signatures of the two land cover types (Fig. 7). Wairau (Native vegetation) recorded the lowest overall risk score (14.24). Turbidity is the dominant contributing indicator (0.312), but the driver of this elevation is hydrological rather than anthropogenic: the Wairau is a large braided river draining geologically active terrain in Marlborough, and natural sediment mobilisation from erodible greywacke and schist generates turbidity spikes during high-flow events even in the absence of land-use pressure. The remaining indicators score consistently low, with Total nitrogen at 0.081 and *E. coli* at 0.074, confirming that the river is in good ecological condition relative to its regulatory benchmarks (Fig. 8). A high turbidity score in a braided river system therefore carries different management implications from an identical score in a pastoral or urban catchment.

3.3. Trend analysis

Mann–Kendall trend slopes computed over the full 2004–2024 record reveal that water quality at most sites has been improving or stable over the 2004–2024 monitoring period, with site-specific exceptions warranting management attention (Fig. 9). The Waikaia native vegetation site showed seven out of eight indicators trending towards improvement (clarity essentially flat), and Otepuni and the Waikaia pasture site both showed improving trajectories for the majority of indicators. This broadly improving pattern is consistent with the findings of Snelder et al. (2021a), who documented phosphorus and bacterial improvements at many monitored sites nationally following the adoption of better land management practices. However, the pattern is heterogeneous across sites and indicators, with several notable departures from the general trend.

Otepuni Creek, the only High-risk site in the dataset, shows uniformly improving or stable trends across all eight scored indicators (τ values between −0.15 and +0.09). This confirms that the site's High risk classification is driven by persistently elevated baseline contamination rather than a worsening trajectory: the site is chronically polluted but is not deteriorating further over the monitoring period. From a management perspective the distinction matters: Otepuni is not getting worse, but the level at which its contamination has stabilised remains well above NPS-FM Band D limits across most indicators.

Wairau, by contrast, shows the largest concentration of worsening trends despite recording the lowest overall risk score in the dataset. Ammoniacal nitrogen at this site recorded the strongest worsening trend in the dataset (τ = +0.50); *E. coli*, dissolved reactive phosphorus, and nitrate nitrogen all show smaller but consistent worsening trajectories; and water clarity is declining (τ = −0.12), indicating that the river is becoming murkier over the monitoring period. The most plausible explanation is that climate-driven hydrological variability has increased the frequency or intensity of high-flow events in the Marlborough region over the monitoring period (Snelder et al., 2021b), mobilising sediment, nutrients, and bacterial loads more frequently than during the early years of the record. These trends warrant attention even at a site whose absolute risk score remains the lowest in the dataset.

Mimihau (Exotic forest) recorded the strongest worsening *E. coli* trend among the five sites (τ = +0.29), an unexpected result given the absence of livestock at a forested site. Nitrate nitrogen at this site also trends upward (τ = +0.23). These trends may reflect intensified forestry harvesting activity in the catchment over the monitoring period, with cumulative effects on soil disturbance, harvest debris loading, and access road runoff. The worsening *E. coli* signal in particular warrants further investigation as it is inconsistent with the expectation that exotic forest sites should show stable bacterial conditions (Larned et al., 2016).

The Waikaia pasture site shows the most divergent indicator-level trends. Dissolved reactive phosphorus, total phosphorus, ammoniacal nitrogen, and *E. coli* all show strongly improving trajectories (tau values between −0.14 and −0.38), consistent with regulatory and voluntary uptake of nutrient management planning in Southland from the mid-2000s onwards. However, nitrate nitrogen (τ = +0.24) and total nitrogen (τ = +0.17) trend upward despite the improving phosphorus and bacterial signals. This pattern is consistent with the well-documented lag in nitrogen response to land management changes: legacy nitrogen accumulated in soil profiles over decades of pastoral land use continues to leach to surface water even as current-management inputs decline.

Fig. 10 presents the same trend information as a stacked waveform display, isolating five indicators: *E. coli*, nitrate nitrogen, clarity (BDISC), total nitrogen, and ammoniacal nitrogen to highlight the cross-site trajectory of each contaminant class over the 2004–2024 record. Each panel shares a common x-axis ordered by the land cover gradient from native vegetation through exotic forest and pasture to urban, with the wave rising above the zero baseline where a site is worsening and falling below where it is improving. The stacked layout allows the dominant direction of each indicator to be read at a glance across all five sites simultaneously, complementing the grouped bar representation in Fig. 9 by emphasising the shape and continuity of the cross-site pattern rather than individual τ values.

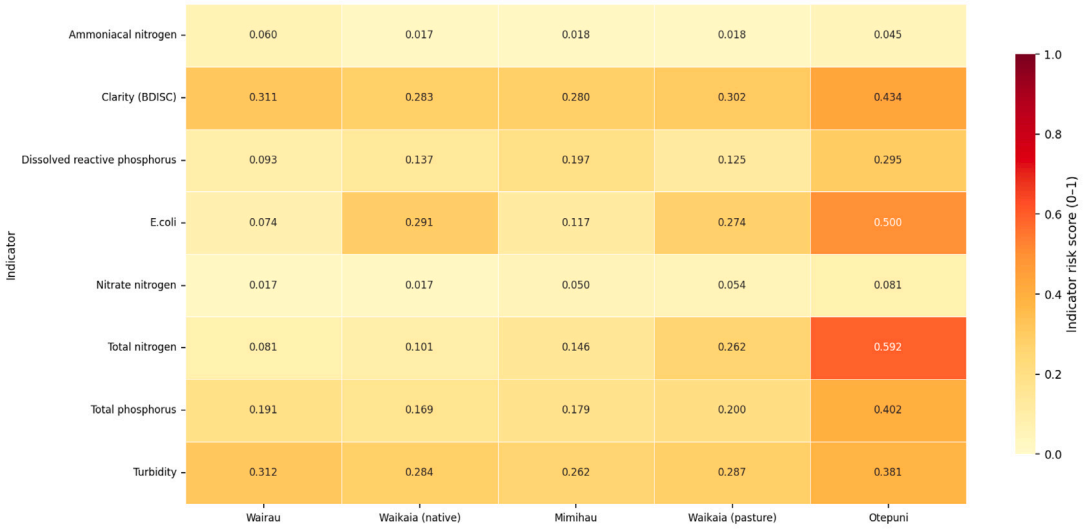


Fig. 8. Per-indicator risk scores for the eight NPS-FM threshold indicators at the five South Island monitoring sites, displayed as a heatmap. Higher values (darker red) indicate greater NPS-FM non-compliance risk for that indicator at that site. The simultaneous elevation across nearly all indicators at Otepuni is characteristic of structural urban contamination, while the four Medium-risk sites show distinct indicator-specific profiles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

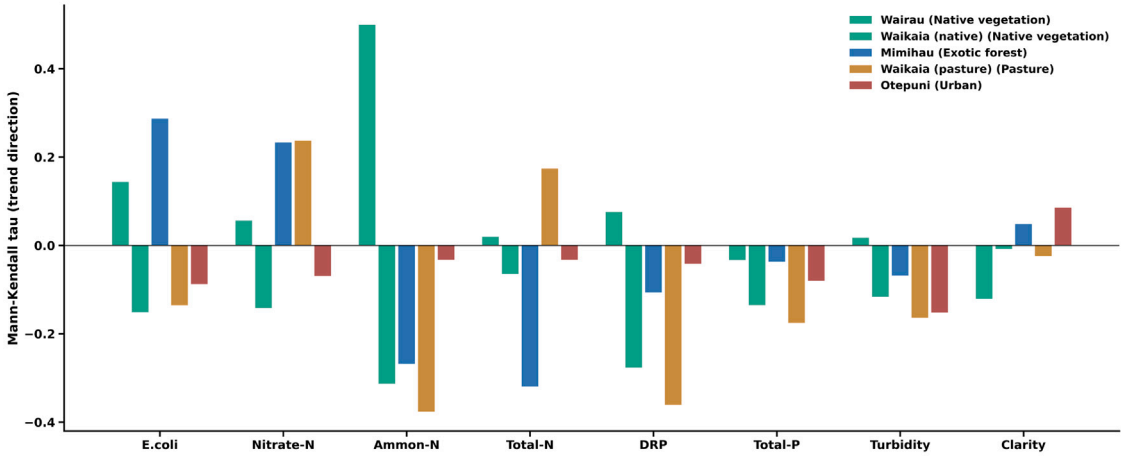


Fig. 9. Mann–Kendall trend slopes (τ) for the eight NPS-FM threshold indicators at the five South Island monitoring sites over 2004–2024. Positive values indicate worsening trends; negative values indicate improving trends. The horizontal line at $\tau = 0$ separates the two regimes. Most trends are negative (improving), with notable exceptions including Wairau Ammoniacal nitrogen ($\tau = +0.50$, the strongest worsening trend in the dataset) and Mimihau *E. coli* ($\tau = +0.29$). Clarity τ is plotted unadjusted; negative values indicate declining clarity.

The most visually distinctive panel is ammoniacal nitrogen, where Wairau's strong worsening signal ($\tau = +0.50$) sits in sharp contrast to the broadly improving trajectories at the remaining four sites, producing the most pronounced wave amplitude in the figure. The *E. coli* panel identifies Mimihau as the only other site with a notable worsening trend ($\tau = +0.29$), while the nitrate nitrogen panel captures the divergence within the Waikaia paired catchment, where the native site is improving and the pasture site trends upward ($\tau = +0.24$). The clarity and total nitrogen panels are broadly flat or negative across most sites, consistent with the stable-to-improving characterisation. Collectively, the waveform display reinforces that the land cover gradient predicts general risk level but not direction of change, and that the most management-relevant trend signals — Wairau ammoniacal nitrogen and Mimihau *E. coli* — arise at sites that would otherwise be classified as lower overall risk.

3.4. Comparison with alternative anomaly detection algorithms

To assess whether the framework's conclusions depend on the choice of anomaly detection algorithm, One-Class SVM and Local Outlier Factor were each substituted for Isolation Forest in the layer-2 position of the pipeline. The seasonal interquartile range layer, the intersection rule, and all risk score weights were held fixed, so that the detection algorithm is the sole variable. Both alternatives operate on distance-based criteria and were therefore applied to log-transformed and standardised inputs; Isolation Forest, which partitions on rank order, was additionally run on the same transformed inputs to confirm that any differences are attributable to the algorithm rather than to preprocessing. Each alternative was matched to the published Isolation Forest flag count on every site–indicator series, so that agreement statistics reflect which observations are selected rather than how many.

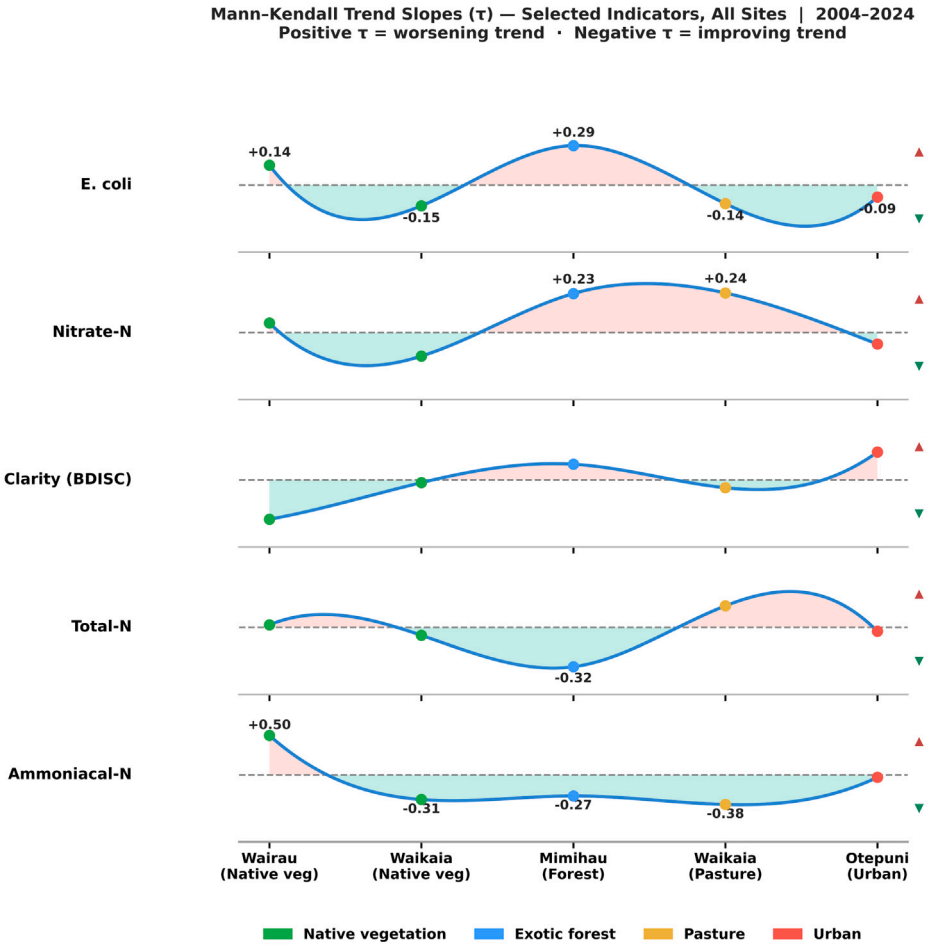


Fig. 10. Mann–Kendall trend slopes (τ) for five NPS-FM threshold indicators at the five South Island monitoring sites over 2004–2024, shown as per-indicator panels. Sites are ordered along the horizontal axis by land cover gradient, from native vegetation through exotic forest and pasture to urban. Positive values indicate worsening trends and negative values indicate improving trends, separated by the $\tau = 0$ reference line. Markers are coloured by REC land cover, with the two native-vegetation sites distinguished by position along the gradient. Clarity τ is plotted unadjusted; negative values indicate declining clarity. The strongest worsening trend in the dataset is Wairau ammoniacal nitrogen ($\tau = +0.50$), followed by Mimihau *E. coli* ($\tau = +0.29$); most other trajectories sit at or below zero. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The results are given in Table 10. Risk grades are identical across all three detectors at all five sites, and the site ranking is preserved exactly, with Spearman rank correlation of 1.000 against the published ordering in each case. The largest change in any site score is 0.84 points, at Waikaia-N; Otepuni retains its High grade with a margin of at least 3.60 points above the classification boundary.

Agreement with the published flag set varies considerably between the alternatives. Local Outlier Factor overlaps at Jaccard indices of 0.47 to 0.62 across the five sites, whereas One-Class SVM overlaps at only 0.17 to 0.24. That a detector agreeing on approximately one fifth of flagged observations nonetheless produces near-identical risk scores is consistent with the intersection design, in which the seasonal layer requires independent support before an observation enters the confirmed anomaly rate.

Isolation Forest applied to transformed inputs agreed with the published flags at Jaccard 0.71 to 0.89 rather than unity, indicating that the algorithm is not fully invariant to monotone transformation at these sample sizes, as sub-sampling and split selection interact with the transformed spacing of values. The effect on site scores is negligible, at most 0.06 points.

A note on parameterisation is warranted. The One-Class SVM parameter ν bounds the fraction of training errors and does not fix the proportion of observations flagged, unlike the contamination parameter of Isolation Forest. Applied natively at $\nu = 0.05$, One-Class SVM flagged 166 to 239 observations per site against a nominal expectation of approximately 100, a factor of 1.6 to 2.4. The matched-count protocol used in Table 10 removes this discrepancy from the comparison.

Autoencoder-based and density-clustering methods were not evaluated. The series analysed contain 69 to 252 observations per site–indicator combination, insufficient for reconstruction-based neural models, which in comparable water quality applications are trained on high-frequency sensor records (El-Shafeiy et al., 2023). DBSCAN requires a neighbourhood radius specified per indicator and offers no direct control over the flagged proportion, reintroducing the parameter sensitivity that this comparison is intended to characterise.

3.5. Sensitivity to quality-control exclusion

The quality-control filter described in Section 2.2 removes 19.2% of raw records across the five sites, with the exclusion concentrated at the four Southland sites. To establish whether this materially affects the reported results, the complete pipeline was re-run on an unfiltered dataset in which all records coded NEMS 200 were retained, increasing the Southland series by approximately 30%. Risk scores, grades and rankings were then compared against the published values.

The effect is negligible. Risk grades are unchanged at all five sites and the site ranking is preserved exactly (Spearman $\rho = 1.000$). The largest change in any site score is 0.44 points, at Waikaia-P, and the

Table 10
Site risk scores and grades under substitution of the layer-2 anomaly detection algorithm. The seasonal interquartile range layer, the intersection rule, and all risk score weights are held fixed. IF_t denotes Isolation Forest applied to log-transformed and standardised inputs. Each alternative detector is matched to the published Isolation Forest flag count on every site–indicator series. Jaccard indices express agreement with the published layer-2 flag set. Grades are High (≥30), Medium (10–29), Low (<10).

Site	Published	Site risk score (0–100)			Risk grade		
	(Table 9)	IF_t	OCSVM	LOF	IF_t	OCSVM	LOF
Wairau	14.24	14.19	13.82	14.26	Medium	Medium	Medium
Waikaia-N	16.24	16.22	15.40	16.21	Medium	Medium	Medium
Mimihau	15.60	15.58	14.86	15.63	Medium	Medium	Medium
Waikaia-P	19.02	18.96	18.41	19.02	Medium	Medium	Medium
Otepuni	34.12	34.06	33.60	34.05	High	High	High
Spearman ρ vs. published		1.000	1.000	1.000			

Site	Jaccard index vs. published layer-2 flags		
	IF_t	OCSVM	LOF
Wairau	0.826	0.212	0.605
Waikaia-N	0.893	0.225	0.582
Mimihau	0.849	0.174	0.620
Waikaia-P	0.730	0.235	0.500
Otepuni	0.711	0.241	0.471

Table 11
Site risk scores with records coded NEMS 200 excluded (as published) and retained. Retention increases the four Southland series by approximately 600 records each. Grades are High (≥30), Medium (10–29), Low (<10).

Site	Published	NEMS 200 retained	Difference	Grade
Wairau	14.24	14.24	+0.00	Medium
Waikaia-N	16.24	16.25	+0.01	Medium
Mimihau	15.60	15.50	−0.10	Medium
Waikaia-P	19.02	19.46	+0.44	Medium
Otepuni	34.12	34.06	−0.06	High
Spearman ρ vs. published ranking				1.000

remaining four sites move by 0.10 points or less (Table 11). Per-indicator changes are small and run in both directions, the largest being total phosphorus at Waikaia-P (+0.027) and at Otepuni (−0.022), which is consistent with the excluded records representing ordinary observations rather than a systematically biased subset. The temporal hold-out result at Mimihau falls from 62.5% to 50.0% under retention, remaining a REVIEW result in both cases; no other validation outcome changes status. A related question is whether the missing years distort the Mann–Kendall trend component. Imposing the same four-year gap on the Wairau series, which retains those years, changes τ by at most 0.045 across the eight scored indicators, with no trend reversing sign.

The exclusion is therefore retained as published. Records coded NEMS 200 carry no assessed quality, and their inclusion in an assessment of regulatory compliance would introduce observations of unknown reliability; the analysis above establishes that this choice does not affect any conclusion drawn in this study.

4. Validation outcomes

The three independent validation methods described in this section are applied to assess the reliability of the site-level risk scores obtained in Section 3. Two of the five sites passed all three validation methods cleanly; the remaining three recorded one or two REVIEW results, each of which proved to be physically interpretable when examined against site-specific land use and seasonal characteristics rather than indicating a failure of the pipeline. The REVIEW results at Waikaia-P and Mimihau are explained by site-specific land use characteristics; the Otepuni result reflects divergent seasonality between infrastructure-delivered and runoff-delivered contaminants.

4.1. Temporal hold-out

The temporal hold-out test, in which risk features were computed on 2004–2019 data and tested against actual 2020–2024 NPS-FM compliance, passed at four of the five sites with consistency rates between 85.7% and 100%. Indicators classified as elevated risk in the training period reliably recorded NPS-FM breaches in the test period, and indicators not classified as elevated risk remained compliant. The single REVIEW result was recorded at Mimihau (62.5% consistency), where three indicators were misclassified, two of them with training exceedance rates near the 5% classification threshold in the test period. This borderline behaviour reflects the inherent sensitivity of any classification scheme to indicators that hover near the decision threshold rather than a systematic failure of the predictive model.

4.2. Literature benchmark

Site medians for each indicator were compared against published national medians for matching REC land cover types from Larned et al. (2016) (*E. coli*, Turbidity, Total nitrogen) and Booker et al. (2026) (Total phosphorus, DRP). Four of the five sites passed at 80% or higher: the native vegetation sites at Wairau and Waikaia-N matched published native vegetation medians (80% consistency each), Mimihau matched published exotic forest medians (100% consistency), and Otepuni matched published urban medians for Turbidity, Total nitrogen, Total phosphorus and DRP (80% consistency), with *E. coli* the single departure at 2.65 times the published urban median. The Waikaia pasture site recorded 60% consistency (REVIEW): site medians for Turbidity and Total phosphorus both fell below the national pastoral median by more than a factor of two. This result is methodologically interpretable rather than a pipeline failure — the published national pastoral benchmarks are dominated by intensive dairy sites, while Waikaia is a sheep and beef catchment with substantially lower phosphorus loadings and less sediment disturbance, a distinction also reported nationally by Snelder et al. (2021a). The site-specific result is therefore a meaningful finding about Waikaia's land use intensity rather than evidence of pipeline error (see Table 13).

4.3. Seasonal pattern validation

Observed seasonal exceedance peaks were compared against the expected seasonality for each land cover type. Three sites recorded 100% consistency: Summer and Autumn peaks at Wairau and Waikaia-N reflected rainfall-driven episodic runoff in unmodified catchments, and Mimihau's Winter and Spring peaks aligned with forestry harvest cycle expectations. Otepuni recorded 66.7% consistency (REVIEW), with the urban flat-distribution expectation met for two of three indicators. *E. coli* (flatness ratio 1.27) and total nitrogen (1.07) show the near-uniform year-round distribution expected of an infrastructure-driven catchment, but turbidity does not (3.52), peaking in Winter at 62.5%. This is physically interpretable rather than a pipeline failure: faecal and nutrient loads at Otepuni are delivered continuously through stormwater infrastructure, cross-connections and ageing pipes, whereas suspended sediment is mobilised by rainfall and retains a seasonal signal even in a heavily modified catchment. The departure from a uniform urban signature is therefore confined to the indicator for which a weather-driven pathway would be expected.

The Waikaia pasture site recorded 33% consistency (turbidity peaked in Spring as expected). *E. coli* peaked in Summer and total nitrogen in Autumn rather than the expected Winter or Spring. The seasonal expectation for pastoral catchments derives from studies of intensive dairy systems, in which effluent application and stock access during saturated ground conditions produce a Winter peak (Wilcock et al., 2013; Monaghan et al., 2007). Waikaia-P is a sheep and beef catchment, where stock are grazed on hill country through summer. The validation therefore does not indicate that the detection pipeline has failed;

Table 12
Validation summary for the site risk scores. Each cell reports the status (PASS or REVIEW) and the percentage of indicators meeting the validation criterion for that site. Sites with ≥70% consistency pass; below 70% are flagged as REVIEW. Two sites pass all three methods; the two REVIEW results at Waikaia-P and the one at Mimihau are explained by site-specific land use characteristics.

Site	Temporal hold-out	Literature benchmark	Seasonal pattern	Methods passed
Wairau	PASS (85.7%)	PASS (80.0%)	PASS (100.0%)	3/3
Waikaia-N	PASS (100.0%)	PASS (80.0%)	PASS (100.0%)	3/3
Mimihau	**REVIEW** (62.5%)	PASS (100.0%)	PASS (100.0%)	2/3
Waikaia-P	PASS (87.5%)	**REVIEW** (60.0%)	**REVIEW** (33.3%)	1/3
Otepuni	PASS (100.0%)	PASS (80.0%)	REVIEW (66.7%)	2/3

Table 13
Per-site literature benchmark comparison. Site medians for five compliance-relevant indicators were compared against published national medians for matching REC land cover types from Larned et al. (2016) (*E. coli*, Turbidity, Total nitrogen) and Booker et al. (2026) (Total phosphorus, DRP). A site median is considered consistent (✓) if its ratio to the published median falls within 0.5–2.0; otherwise it is flagged for review (R). A site passes if 70% or more of compared indicators meet the consistency criterion.

Site (land cover)	Indicator (unit)	Site median	Benchmark median	Ratio	Result
Wairau (Native vegetation) – Marlborough					
	E. coli (cfu/100 mL)	9.0	10	0.90	✓
	Turbidity (NTU)	1.41	1.5	0.94	✓
	Total N (mg/L)	0.159	0.20	0.79	✓
	Total P (mg/L)	0.007	0.015	0.47	R (below)
	DRP (mg/L)	0.003	0.005	0.60	✓
Site consistency:					**4/5 (80%) ✓**
Waikaia (native) (Native vegetation) – Southland					
	E. coli	130.0	10	13.00	R (above)
	Turbidity	2.1	1.5	1.40	✓
	Total N	0.250	0.20	1.25	✓
	Total P	0.014	0.015	0.93	✓
	DRP	0.006	0.005	1.20	✓
Site consistency:					**4/5 (80%) ✓**
Mimihau (Exotic forest) – Southland					
	E. coli	18.0	30	0.60	✓
	Turbidity	2.1	2.5	0.84	✓
	Total N	0.350	0.35	1.00	✓
	Total P	0.020	0.020	1.00	✓
	DRP	0.011	0.007	1.57	✓
Site consistency:					**5/5 (100%) ✓**
Waikaia (pasture) (Pasture) – Southland					
	E. coli	130.0	126	1.03	✓
	Turbidity	2.4	5.0	0.48	R (below)
	Total N	0.680	0.80	0.85	✓
	Total P	0.014	0.040	0.35	R (below)
	DRP	0.0068	0.012	0.57	✓
Site consistency:					**3/5 (60%) R**
Otepuni (Urban) – Southland					
	E. coli	1300.0	490	2.65	R (above)
	Turbidity	4.6	6.0	0.77	✓
	Total N	2.05	2.0	1.03	✓
	Total P	0.048	0.060	0.80	✓
	DRP	0.012	0.020	0.60	✓
Site consistency:					**4/5 (80%) ✓**

it indicates that the site does not conform to a seasonal signature derived from a different farming system. The same distinction appears in the literature benchmark, where Waikaia-P's turbidity and total phosphorus medians fall well below national pastoral values dominated by dairy sites (Fig. 11).

4.4. Synthesis of the validation methods

The validation results provide consistent support for the site risk rankings. Two sites (Wairau and Waikaia-N) pass all three validation methods, and the internal consistency check reported in Section 2.3.3 shows co-occurrence ratios between 10.6 and 14.0 across all five sites. Mimihau passes two of three validations, with the temporal hold-out REVIEW attributable mainly to indicators near the classification boundary. Waikaia-P passes only the temporal hold-out, but its REVIEW results in the literature benchmark and seasonal pattern checks both reflect a site-specific land use signature (sheep and beef rather than dairy) that is physically interpretable. Otepuni passes the temporal hold-out and literature benchmark, with its seasonal REVIEW arising from turbidity alone, for which a runoff-driven seasonal peak is expected on physical grounds. The pipeline therefore produces a reliable risk ranking, with the small number of REVIEW results functioning as informative findings about site characteristics rather than as evidence of methodological failure.

5. Discussion and conclusion

The pipeline presented here integrates three components that have typically been addressed in isolation in the existing literature: unsupervised anomaly detection (Essamlali et al., 2024; Lamrini et al., 2024), NPS-FM regulatory compliance assessment, and predictive risk scoring (Nasir et al., 2022; Chen et al., 2024). Four methodological choices

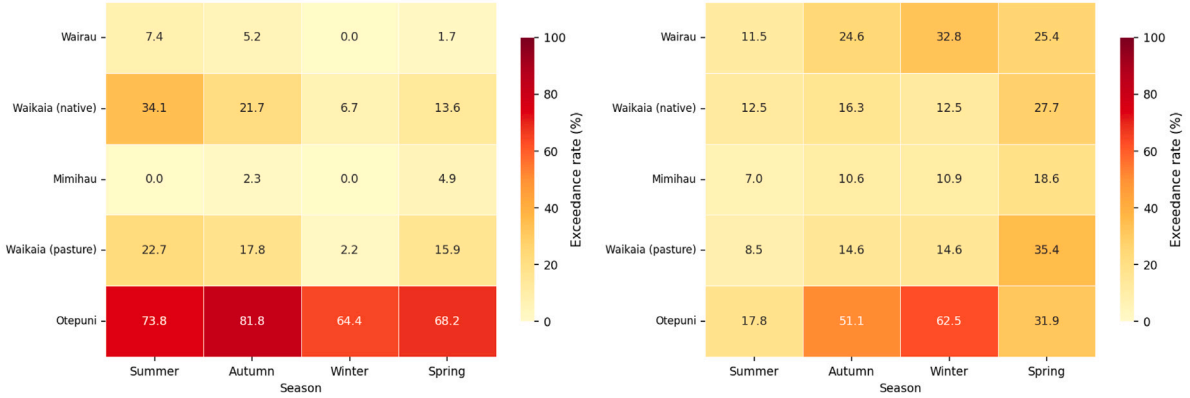


Fig. 11. Seasonal NPS-FM Band D exceedance rates (%) for *E. coli* (left) and Turbidity (right) at the five South Island monitoring sites. Sites are ordered top to bottom by land cover gradient. The near-flat seasonal distribution of *E. coli* at Otepuni reflects year-round stormwater-driven contamination, whereas turbidity at the same site peaks in Winter, consistent with rainfall-driven sediment mobilisation. The Waikaia pasture Autumn peak for *E. coli* reflects sheep and beef grazing seasonality distinct from the winter-dominated dairy pattern.

distinguish this work from related approaches. First, the strict requirement that confirmed anomalies be flagged by both Isolation Forest and seasonal IQR, rather than the more permissive 2-of-3 confirmation rule used in some pipelines, reduces false positives and strengthens the co-occurrence support for detected events. Second, risk scores are computed using a transparent weighted aggregation of four interpretable features (exceedance rate, anomaly rate, 90th percentile severity, and trend slope), directly addressing the indicator-level explainability gap identified by Chen et al. (2024). Third, the weighting choice was subjected to sensitivity testing under three alternative schemes; site rankings and risk grades proved robust to the choice of weights, with only one position swap observed under equal weighting. Fourth, the framework's output was shown to be insensitive to the choice of anomaly detection algorithm: substituting One-Class SVM or Local Outlier Factor for Isolation Forest leaves every risk grade unchanged and preserves the site ranking (Section 3.4). Together with the three-method validation procedure, these choices produce a risk ranking whose interpretability and stability can both be defended explicitly rather than asserted.

The two components of the risk score answer different questions, and reading them together distinguishes site conditions that a single measure conflates. The exceedance rate asks whether a site breaches a legal limit and how often; the confirmed anomaly rate asks whether a site departs from its own recorded behaviour. Because the statistical layers calibrate against each site's own history, a chronically polluted site does not accumulate anomalies simply by being polluted, and a clean site is not exempt from them. Four combinations are therefore possible and each carries a distinct management implication. High exceedance with low anomaly frequency indicates persistent structural non-compliance, which describes Otepuni, where 96% of total nitrogen readings breach Band D while the confirmed anomaly rate (3.11%) is unremarkable against the other four sites; remediation here is a question of infrastructure rather than event response. Low exceedance with elevated anomaly frequency indicates an otherwise compliant site subject to episodic departures, where the appropriate response is investigation of the events themselves. High values on both indicate a site that is both chronically impaired and actively unstable; low values on both indicate compliance with no evidence of episodic disturbance.

The Waikaia paired catchment illustrates why indicator-level scores matter. The native vegetation site records a higher *E. coli* per-indicator score (0.291) than the pasture site 15 km downstream (0.274), consistent with its higher *E. coli* exceedance rate (19.0% against 14.6%), whereas total nitrogen shows the reverse, with a score of 0.262 at the pasture site against 0.101 at the native site. A single site-level score, or an aggregate over indicators alone, would hide this divergence, and the two sites would be read as differing in degree rather than in which contaminants drive their risk. Whether the episodic or the persistent reading is more consequential is a management judgement rather than a statistical one, and the framework is designed to present both rather than reduce them to a single number.

This study applied a two-layer machine learning pipeline to twenty one years of LAWA monitoring data from five South Island rivers spanning four land cover types, producing a quantitative risk score for NPS-FM Band D non-compliance at each site. Otepuni Creek (Urban) recorded the only high risk grade with a score of 34.12, driven primarily by Total nitrogen, *E. coli*, water clarity, and total phosphorus. The four remaining sites all received Medium risk grades, with scores between 14.24 and 19.02 broadly following the land cover gradient. The site ranking proved robust to the choice of risk score weights under sensitivity testing. The methodological contribution lies in the integration of unsupervised anomaly detection, regulatory threshold comparison, weighted risk scoring, and multi-method validation into a single transparent pipeline. Per-indicator risk scores expose which contaminants drive each site's risk, enabling targeted rather than generic management responses. The distinction between persistent contamination (captured by the NPS-FM exceedance rate) and episodic contamination (captured by confirmed anomaly events) provides a more nuanced characterisation than threshold comparison alone, with co-occurrence ratios between 10.6 and 14.0 times higher than chance supporting the physical reality of detected events.

6. Limitations, implications and further work

This study has certain limitations. The site sample is small ($n = 5$), chosen to span all four major land cover types and to include a paired catchment for land-use isolation, but limiting statistical generalisation. The five-site design supports demonstration of the framework and isolation of land-use effects through the paired catchment, but does not support inference about the distribution of compliance risk across South Island rivers generally. The risk score is computed independently at each site from that site's own record, so the pipeline scales without modification to the full 270-site dataset; what has not been demonstrated is whether the grade boundaries, the land-cover gradient, or the validation pass rates hold across a larger site population. The monthly LAWA sampling frequency means sub-monthly pollution events between observations are not captured; the pipeline detects what has been measured, not what occurred. Risk score weights, although sensitivity-tested, reflect expert judgement rather than data-driven calibration. The analysis is restricted to the South Island; transferability to North Island catchments is plausible but not yet demonstrated empirically. The NPS-FM Band D thresholds applied here are static, so the framework

will require re-application as regulatory standards evolve, although no methodological change is required for this.

Several sources of uncertainty qualify the risk estimates. Quality-control filtering left the four Southland series thinner in the middle of the record than the site totals suggest, which bears on the temporal hold-out validation in particular, since its training and test windows are affected to differing degrees. Measurement uncertainty is not propagated, as LAWA records report single values without error bounds. Censored readings enter at the detection limit, biasing low-concentration indicators upward. Climate-driven interannual variability affects both the anomaly detection and the trend estimates, and Snelder et al. (2021b) show this to be a substantial component of observed short-term change in New Zealand rivers. The framework characterises risk relative to each site's own recorded history and does not separate these sources, so the risk scores should be read as ordinal indicators of relative compliance risk rather than absolute probabilities of future breach.

The absence of labelled pollution events also means supervised classification metrics cannot be computed: no independent record exists of which observations correspond to genuine contamination episodes. Validation is therefore indirect, resting on temporal consistency, agreement with published land-cover benchmarks, seasonal plausibility, and the demonstrated insensitivity of the results to detector choice and parameter settings (Sections 3.4 and 3.5).

The framework addresses a specific gap in New Zealand freshwater monitoring practice: compliance is currently assessed retrospectively, with no operational system providing a forward-looking risk signal. The distinction between persistent contamination (exceedance rate, weighted 40%) and episodic contamination (confirmed anomaly rate, weighted 30%) has direct management implications: an Otepuni-type site requires infrastructure remediation, while a Wairau-type site requires understanding of hydrological drivers rather than land-use intervention. The framework is regulator-agnostic, requiring only fixed numerical thresholds and a multi-year monitoring record, and is therefore transferable to any jurisdiction operating threshold-based water quality standards, including the U.S. Clean Water Act and the EU Water Framework Directive. The minimum requirements are a numerical threshold per scored indicator, a record long enough to estimate seasonal distributions and a monotonic trend, and sampling that resolves the seasonal cycle; at approximately monthly sampling these conditions are met by records of roughly five years, although the 2004–2024 records used here give substantially more stable estimates.

Two adaptations may be needed in jurisdictions structured differently from the NPS-FM. Where standards are expressed as multi-year statistics rather than observation-level limits, the exceedance rate could be redefined as the proportion of assessment periods failing the standard. Where limits are set regionally with no fixed national threshold, the severity feature would lose its normalising denominator, and a catchment-specific reference value could be substituted, although this would reduce comparability between catchments. Extension to all 270 LAWA South Island sites, integration of climate covariates, and adaptation to high-frequency monitoring data represent immediate avenues for further development. The pipeline is implemented in Python with all random seeds fixed, and the cleaning, detection, scoring and validation stages are separated so that thresholds, weights and detection parameters can each be substituted without modification to the remaining stages.

CRediT authorship contribution statement

Tanishqa Goyal: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Parul Tiwari:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Investigation, Validation, Supervision, Resources, Writing – review & editing, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors wish to thank the Mathematical Modelling and Analytics Research Centre at Auckland University of Technology for providing resources and technical support towards the completion of this manuscript.

Data availability

Monitoring data analysed in this study are publicly available from the Land, Air, Water Aotearoa (LAWA) platform (https://www.lawa.org.nz) under open data licensing. Analysis code is available from the corresponding author on reasonable request.

References

Baillie, B.R., Neary, D.G., 2015. Water quality in New Zealand's planted forests: A review. N. Z. J. For. Sci. 45 (1), 7. http://dx.doi.org/10.1186/s40490-015-0040-0.

Bain, R., Cronk, R., Wright, J., Yang, H., Slaymaker, T., Bartram, J., 2014. Fecal contamination of drinking-water in low- and middle-income countries: A systematic review and meta-analysis. PLOS Med. 11 (5), e1001644. http://dx.doi.org/10.1371/journal.pmed.1001644.

Beusen, A.H.W., Bouwman, A.F., Van Beek, L.P.H., Mogollón, J.M., Middelburg, J.J., 2016. Global riverine N and P transport to ocean increased during the 20th century despite increased retention along the aquatic continuum. Biogeosciences 13 (8), 2441–2451. http://dx.doi.org/10.5194/bg-13-2441-2016.

Booker, D.J., Smith, R., Wood, D., Fraser, C., Snelder, T., 2026. Water Quality State and Trends in New Zealand Rivers: Analysis of National Data Ending in 2024. Client Report 2025331CH, New Zealand Institute for Earth Science Limited (Earth Sciences New Zealand), Christchurch, New Zealand, Prepared for Ministry for the Environment.

Buchhorn, K., Santos-Fernandez, E., Mengersen, K., Salomone, R., 2024. Graph neural network-based anomaly detection for river network systems. F1000Research 12, 991. http://dx.doi.org/10.12688/f1000research.136097.2, Version 2.

Canadian Council of Ministers of the Environment, 2001. Canadian Water Quality Guidelines for the Protection of Aquatic Life: CCME Water Quality Index 1.0, Technical Report. Technical Report, Canadian Council of Ministers of the Environment, Winnipeg, Manitoba, Canada.

Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: A survey. ACM Comput. Surv. 41 (3), 15:1–15:58. http://dx.doi.org/10.1145/1541880.1541882.

Chen, W., Xu, D., Pan, B., Zhao, Y., Song, Y., 2024. Machine learning-based water quality classification assessment. Water 16 (20), 2951.

Cookson, A.L., Devane, M., Marshall, J.C., Moinet, M., Gardner, A., Collis, R.M., Rogers, L., Biggs, P.J., Pita, A.B., Cornelius, A.J., et al., 2024. Population structure and pathogen interaction of *Escherichia coli* in freshwater: Implications of land-use for water quality and public health in Aotearoa New Zealand. Environ. Microbiol. Rep. 16 (4), e13319.

Damania, R., Desbureaux, S., Rodella, A.-S., Russ, J., Zaveri, E., 2019. Quality Unknown: The Invisible Water Crisis. World Bank, Washington, DC, http://dx.doi.org/10.1596/978-1-4648-1459-4.

Davies-Colley, R., Nagels, J., Lydiard, E., 2008. Stormflow-dominated loads of faecal pollution from an intensively dairy-farmed catchment. Water Sci. Technol. 57 (10), 1519–1523. http://dx.doi.org/10.2166/wst.2008.257.

Diaz, R.J., Rosenberg, R., 2008. Spreading dead zones and consequences for marine ecosystems. Science 321 (5891), 926–929. http://dx.doi.org/10.1126/science.1156401.

El-Shafeiy, E., Alsabaan, M., Ibrahem, M.I., Elwahsh, H., 2023. Real-time anomaly detection for water quality sensor monitoring based on multivariate deep learning technique. Sensors 23 (20), 8613.

Environment Southland, 2017. Invercargill City Council stormwater discharge consent (APP-20168843). https://www.es.govt.nz/environment/consents/notified-consents/2016/invercargill-city-council. (Accessed 13 August 2026).

Environment Southland, 2024. *E. coli*. https://www.es.govt.nz/environment/water/whats-in-our-waterways/e-coli. (Accessed 13 August 2026).

Essamlali, I., Nhaila, H., El Khaili, M., 2024. Advances in machine learning and IoT for water quality monitoring: A comprehensive review. Heliyon 10 (6), e27920. http://dx.doi.org/10.1016/j.heliyon.2024.e27920.

Gadd, J., Snelder, T., Fraser, C., Whitehead, A., 2020. Current state of water quality indicators in urban streams in New Zealand. N. Z. J. Mar. Freshw. Res. 54 (3), 354–371.

Gall, M.P., Davies-Colley, R., Milne, J., Stott, R., 2022. Suspended sediment and faecal contamination in a stormflow plume from the Hutt River in Wellington Harbour, New Zealand. N. Z. J. Mar. Freshw. Res. http://dx.doi.org/10.1080/00288330.2022.2088569.

Hirsch, R.M., Slack, J.R., 1984. A nonparametric trend test for seasonal data with serial dependence. Water Resour. Res. 20 (6), 727–732. http://dx.doi.org/10.1029/WR020i006p00727.

Julian, J.P., De Beurs, K.M., Owsley, B., Davies-Colley, R.J., Ausseil, A.-G.E., 2017. River water quality changes in New Zealand over 26 years: response to land use intensity. Hydrol. Earth Syst. Sci. 21 (2), 1149–1171.

Kendall, M.G., 1975. Rank Correlation Methods, fourth ed. Charles Griffin, London.

Lamrini, M., Ben Mahria, B., Chkouri, M.Y., Touhafi, A., 2024. Towards reliability in smart water sensing technology: Evaluating classical machine learning models for outlier detection. Sensors 24 (13), 4084.

Land, Air, Water Aotearoa, 2025. Otepuni Creek at Nith Street — river quality monitoring site. https://www.lawa.org.nz/explore-data/southland-region/river-quality/waihopai-stream/otepuni-creek-at-nith-street. (Accessed 13 August 2026).

Larned, S., Snelder, T., Unwin, M., McBride, G., 2016. Water quality in New Zealand rivers: current state and trends. N. Z. J. Mar. Freshw. Res. 50 (3), 389–417.

Liu, F.T., Ting, K.M., Zhou, Z.-H., 2008. Isolation forest. In: Proceedings of the 8th IEEE International Conference on Data Mining. ICDM 2008, IEEE, pp. 413–422. http://dx.doi.org/10.1109/ICDM.2008.17.

Liu, F.T., Ting, K.M., Zhou, Z.-H., 2012. Isolation-based anomaly detection. ACM Trans. Knowl. Discov. Data 6 (1), 3:1–3:39. http://dx.doi.org/10.1145/2133360.2133363.

Lumb, A., Sharma, T.C., Bibeault, J.-F., 2011. A review of genesis and evolution of water quality index (WQI) and some future directions. Water Qual. Expo. Health 3 (1), 11–24. http://dx.doi.org/10.1007/s12403-011-0040-0.

Mann, H.B., 1945. Nonparametric tests against trend. Econometrica 13 (3), 245–259. http://dx.doi.org/10.2307/1907187.

Mateo-Sagasta, J., Marjani Zadeh, S., Turral, H. (Eds.), 2018. More people, more food, worse water? A global review of water pollution from agriculture. Food and Agriculture Organization of the United Nations (FAO) and International Water Management Institute (IWMI), Rome.

McDowell, R.W., Snelder, T.H., Cox, N., Booker, D.J., Wilcock, R.J., 2013. Establishment of reference or baseline conditions of chemical indicators in New Zealand streams and rivers relative to present conditions. Mar. Freshwater Res. 64 (5), 387–400. http://dx.doi.org/10.1071/MF12153.

Ministry for the Environment, 2020. National policy statement for freshwater management 2020 (Amended October 2024). https://environment.govt.nz/acts-and-regulations/national-policy-statements/national-policy-statement-for-freshwater-management/. Amended October 2024.

Monaghan, R.M., Hedley, M.J., Di, H.J., McDowell, R.W., Cameron, K.C., Ledgard, S.F., 2007. Nutrient management in New Zealand pastures — recent developments and future issues. N. Z. J. Agric. Res. 50 (2), 181–201. http://dx.doi.org/10.1080/00288230709510290.

Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., Al-Shamma'a, A., 2022. Water quality classification using machine learning algorithms. J. Water Process. Eng. 48, 102920.

Snelder, T., Fraser, C., Larned, S., Monaghan, R., De Malmanche, S., Whitehead, A., 2021a. Attribution of river water-quality trends to agricultural land use and climate variability in New Zealand. Mar. Freshwater Res. 73 (1), 1–19.

Snelder, T., Larned, S., Fraser, C., De Malmanche, S., 2021b. Effect of climate variability on water quality trends in New Zealand rivers. Mar. Freshwater Res. 73 (1), 20–34.

Sutadian, A.D., Muttil, N., Yilmaz, A.G., Perera, B.J.C., 2016. Development of river water quality indices—a review. Environ. Monit. Assess. 188 (1), 58.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, Massachusetts.

United Nations Environment Programme, 2016. A Snapshot of the World's Water Quality: Towards a Global Assessment. Technical Report, United Nations Environment Programme, Nairobi, Kenya.

van Vliet, M.T.H., Jones, E.R., Flörke, M., Franssen, W.H.P., Hanasaki, N., Wada, Y., Yearsley, J.R., 2023. Global river water quality under climate change and hydroclimatic extremes. Nat. Rev. Earth Environ. 4, 687–702. http://dx.doi.org/10.1038/s43017-023-00472-3.

Walsh, C.J., Roy, A.H., Feminella, J.W., Cottingham, P.D., Groffman, P.M., Morgan, II, R.P., 2005. The urban stream syndrome: Current knowledge and the search for a cure. JNABS 24 (3), 706–723. http://dx.doi.org/10.1899/04-028.1.

Wilcock, R.J., Monaghan, R.M., Quinn, J.M., Srinivasan, M.S., Houlbrooke, D.J., Duncan, M.J., Wright-Stow, A.E., Scarsbrook, M.R., 2013. Trends in water quality of five dairy farming streams in response to adoption of best practice and benefits of long-term monitoring at the catchment scale. Mar. Freshwater Res. 64 (5), 401–412. http://dx.doi.org/10.1071/MF12155.