CAWTHRON

REPORT NO. 3948

ALTERNATIVE APPROACHES TO CALCULATING LAWA TRENDS AND VISUALISING STATE AND TREND COMBINATIONS

ALTERNATIVE APPROACHES TO CALCULATING LAWA TRENDS AND VISUALISING STATE AND TREND COMBINATIONS

FINNBAR LEE, DARRYL MACKENZIE, ROGER YOUNG

Prepared for Ministry for the Environment

CAWTHRON INSTITUTE
98 Halifax Street East, Nelson 7010 | Private Bag 2, Nelson 7042 | New Zealand
Ph. +64 3 548 2319 | Fax. +64 3 546 9464
www.cawthron.org.nz

REVIEWED BY:
Paula Casanovas

APPROVED FOR RELEASE BY:
Robin Holmes

ISSUE DATE: 26 September 2023

RECOMMENDED CITATION: Lee F, MacKenzie D, Young R. 2023. Alternative approaches to calculating LAWA trends and visualising state and trend combinations. Nelson: Cawthron Institute. Report 3948. Prepared for Ministry for the Environment.

EXECUTIVE SUMMARY

Land, Air, Water Aotearoa (LAWA) leverages a database of water quality and ecological health indicators to report on the state and trend of these attributes at individual sites and at a national scale. Trends are calculated for indicators over the last 5, 10 and 15 years using a two-stage process, via a Mann–Kendall or Seasonal Kendall statistical test.

Concerns have been expressed that while this method results in a large number of sites with likely trends, some of these trends appear relatively weak and it is difficult to identify potential causes for the trends observed. In addition, there have been instances at some sites where a trend in one direction has been indicated by the data, but the following year a trend in the opposite direction has been identified. While trends may change over time at sites due to climatic cycles or implementation of management actions, regular trend switching does not foster confidence in trend detection approaches and makes it hard to determine appropriate management responses.

While cases of trend switching have been identified, the scale and frequency with which this occurs is unclear across the country. To investigate this, we quantified the consistency of trends calculated on a rolling 8-year dataset over a 10-year period. We also compared the current LAWA approach to trend detection with an alternative approach (quantile regression; QR) for each of five water quality and ecological indicators: black disc (BDISC), dissolved reactive phosphorus (DRP), *Escherichia coli* (ECOLI), Macroinvertebrate Community Index (MCI) and ammoniacal nitrogen (NH4N).

Across all five indicators less than 40% of sites showed consistent trend categories for all three datasets (1–8 years, 2–9 years and 3–10 years), although most changes in trend category were relatively small and less than 35% of sites had a change in trend direction across the three rolling datasets. While changes in trend strength and direction at a site over time are not unexpected, we found that up to 15% of sites changed from a ‘very likely’ trend in one direction to either ‘likely’ or ‘very likely’ in the reverse direction over the three rolling datasets. At a national scale the proportion of sites in each of the trend categories was relatively consistent (less than 10% difference) between the default LAWA method and an alternative approach, where the median of the three rolling datasets was used.

Results of the comparison between the LAWA and QR method showed that the latter was consistently more conservative than the LAWA method for all five water quality indicators. A higher percentage of sites were classified as likely improving, likely degrading and indeterminate with the QR method than with the LAWA method, whereas significantly fewer sites were classified as very likely improving or very likely degrading (5–45% for the LAWA method vs 3–32% for the QR method).

Finally, state and trend scores are currently given as two distinct metrics in LAWA and there is a desire to simplify how they are presented, ideally as a single metric or value. We looked

at two possible approaches for combining state and trend into a new ‘condition’ attribute, where each site was classified as ‘good’, ‘at risk’, ‘recovering’ or ‘bad’.

The proportion of sites in each classification varied by indicator. For example, less than 5% of sites were classified as ‘bad’ for NH4N for the 5-year trend period, whereas more than 45% were classified as ‘bad’ for ECOLI over the same period. One limitation was the requirement that for a given site both state and trend scores must be available (many sites have one or the other, but not both) before condition can be calculated.

The results presented here highlight that trend scores are likely to change from year to year. Changes in trends over time are not unexpected, but a change from a very likely trend in one direction to a very likely trend in another direction, resulting from only a few new data points, raises concerns about the robustness of the trend detection.

While we present some possible solutions to reduce the frequency of trends switching directions with the addition of small amounts of data, ultimately we suggest that more work is needed to identify the best approach to dealing with this issue. We suggest that a simulation-based assessment of the approaches be undertaken, where datasets with various trends, noise levels and censored values are simulated, trends calculated and results compared, with a specific focus on trend consistency. Within the current methodological framework we see four areas that warrant deeper examination: 1) using the rolling window approach, and the effect of using different-sized windows; 2) changing the trend attribution cut-off to be more conservative; 3) increasing the number of values above (or below) the detection limit that are required before a trend can be calculated; and 4) implementing trend slope thresholds that represent ecological significance. Each of these changes has the potential to reduce trend switching, without a wholesale change in methodology. QR could then be looked at in more detail using the simulation-based assessment outlined above. Finally, additional analyses looking at methods that account for spatial autocorrelation (e.g. state-space modelling) and methods that account for larger-scale natural environmental (e.g. climatic) cycles would also be valuable, as these have been identified as potential trend drivers.

The combination of the state and trend scores into a new ‘condition’ attribute seems relatively straightforward and could be used to generate simplified national-scale visualisations and summaries. The limitation of requiring both state and trend attributes to be present for a given site to calculate ‘condition’ could be overcome by continuing to also present state and trend data separately, so no information is lost.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. BACKGROUND

Land, Air, Water Aotearoa (LAWA) leverages a database of water quality and ecological health indicators to report on the state and trend of these attributes at individual sites and at a national scale. Data are derived from monitoring of more than 1,000 sites around Aotearoa New Zealand, with this network maintained by regional councils and the National Institute of Water and Atmospheric Research (NIWA). The 'state' describes the attribute in terms of National Policy Statement for Freshwater Management (NPS-FM) bands (A, B, C, D or E), with 'A' being indicative of 'good' water quality, and 'D' and 'E' being indicative of poor water quality. The 'trend' describes the attribute in terms of temporal changes (very likely improving, likely improving, indeterminate, likely degrading and very likely degrading). Trends are calculated for water quality and ecological indicators over the last 5, 10 and 15 years, and are used to infer changes and evaluate relationships between water quality and ecological indicators and drivers (Ministry for the Environment and StatsNZ 2023).

LAWA trend analysis was previously done with a *p*-value, null hypothesis (slope = zero) approach. A criticism of the *p*-value-based approach was the relatively small number of sites where trends were identified, with the majority of sites in the 'indeterminant' category. One of the reasons for changing to the current method 4–5 years ago was that it would enable a larger number of sites with possible trends to be identified. More importantly, the current method builds on the perspective of McBride (2019), who argues that a traditional statistical null hypothesis test approach to trend analysis does not make sense. It is very unlikely that a long-term dataset would have a slope of exactly zero – so why should the null hypothesis be set as slope = zero when this outcome is very unlikely? In LAWA, trends are currently identified using a two-stage process focused on confidence in trend direction, rather than determining whether there is evidence to reject the null hypothesis that there is no trend. First, trend direction and the confidence in trend direction are calculated using either the Mann–Kendall assessment or, where seasonality is identified in the observations, the Seasonal Kendall assessment (Snelder et al. 2021). Second, trend rate and the confidence in trend rate are evaluated using non-parametric Sen slope regressions of attribute observations through time (Helsel et al. 2020; Snelder et al. 2021). This two-stage process is simplified and reported as a single score, where a continuous measure of confidence is split into discrete categories (very likely improving, likely improving, indeterminate, likely degrading and very likely degrading [hereafter referred to as the 'LAWA method']; see Snelder et al. 2021 for details). It is worth noting the original null hypothesis method and the current 'likelihood' methods are mathematically similar, but philosophically different in how they discern whether a trend exists and the confidence in any trends identified.

While these methods are typically robust, they do have limitations. Many of the likely trends that are identified are relatively weak and it is difficult to determine possible causes of the trends. For example, at the likely improvement and likely degrading levels the confidence for this classification is relatively low (down to 66% confidence), and therefore it is sometimes hard to 'see' the trends in the data and it does not take many additional data points to change the trend direction. There are some situations where trends identified in one direction switch to trends identified in another direction with just a small amount of new data. In addition, some trends have such a low slope that the change over time is likely to be ecologically meaningless. Given that regional councils may need to alter their management strategies depending on these results, it is important to have a robust approach to trend detection that focuses attention on trends that are ecologically meaningful and is not overly sensitive to small amounts of new data.

Within the existing framework there are some options to be more stringent when assigning a trend, such as altering the confidence levels at which a given trend is assigned, or increasing the number of measurements above (or below) detection limits (currently set to five) that are required before a trend can be robustly identified. One potential alternative approach to calculating trends is quantile regression (QR), which is also robust to outlier values and censored values, does not make assumptions about underlying distributions and can also quantify trends for multiple quantiles. Being a regression-based method, QR can potentially be used where the temporal resolution of the water quality dataset changes (e.g. quarterly to monthly measurements). It also allows the evidence for a trend to be evaluated after accounting for the effects of other factors (e.g. management actions, environmental changes) through including appropriately defined predictor variables.

There has also been criticism that there is a focus in LAWA on determining whether there are linear trends in datasets, while non-linear patterns are often observed in datasets over time. Non-linear trends can be assessed using QR, visually, or using other approaches, such as generalised additive models (Morton and Henderson 2008). Here, however, we consider only linear trends.

Currently, LAWA presents state and trend summaries as two distinct metrics. This can make it difficult to quickly evaluate the combined condition of sites. For example, a very likely degrading trend will be of interest for different reasons for sites that have a state of A or D. Similarly, management response at a site with no evidence of a trend could be quite different depending on whether the current state is 'good' or 'bad'. Therefore, there are benefits to combining the state and trend metrics into a single new metric that communicates both of these results at once.

Here, we address the three issues outlined above. Specifically, we:

1. quantified the consistency of trends across a rolling 8-year period to determine if trend switching over time is a common feature
2. evaluated an alternative method (QR) for calculating trends in water quality attributes

3. presented an approach to combining both state and trend estimates into a new metric of site ‘condition’.

2. TREND CONSISTENCY

A criticism of the current LAWA method is that trends are prone to changing direction from year to year. While trends may change over time at sites due to climatic cycles or implementation of management actions, regular trend switching does not foster confidence in trend detection approaches and makes it hard to determine appropriate management responses. For example, situations have been identified where, for a given site, the 5-year (2015–19) trend is 'likely improving', then the following year (2016–20) the trend changes to 'likely degrading'. This issue arises in part from the discretisation of a continuous variable 'confidence level' (McBride 2019; Snelder et al. 2021). The 'likely' designation / category includes a 0.66–0.90 confidence of trend direction, which is relatively low compared to the typical $\alpha = 0.05$ ($p = 0.95$) level that is often used to determine statistical significance. In other words, a site with a 67% confidence of a 'likely improving' trend also has a 33% confidence of a degrading trend. It is plausible that a few additional data points may switch the balance and provide increased confidence in the alternative trend direction. Although trend switching over time has been reported for some sites, the scale and frequency of these switches is not clear.

To determine the frequency with which trends change from year to year, we evaluated the consistency of LAWA trends across rolling 8-year periods within the last 10 years. This was achieved by taking the last 10 years of data at all LAWA sites and recalculating trends for the periods 1–8 years, 2–9 years and 3–10 years for each of five water quality and ecological indicators – black disc (BDISC), dissolved reactive phosphorus (DRP), *Escherichia coli* (ECOLI), Macroinvertebrate Community Index (MCI) and ammoniacal nitrogen (NH4N). We then calculated the percentage of sites that had changed trend direction, or had changed trend category but stayed with the same direction (e.g. 'very likely improving' to 'likely improving'). Finally, we compared the original 10-year trend to the median of the three rolling trends (e.g. if a site had rolling trends of 'likely improving', 'indeterminate' and 'likely degrading', it would be classified as 'indeterminate'). While we chose an 8-year window, this decision was arbitrary and the window size could be increased or decreased.

Across all five indicators, less than 40% of sites showed consistent trend categories for all three datasets (1–8 years, 2–9 years and 3–10 years; Figure 1, top), although less than 35% of sites had a change in trend *direction* (Figure 1, bottom). Rather, it was more common to see changes from 'very likely' to 'likely' (or vice versa), but in a consistent direction (up to 34% of sites), or a change from 'indeterminate' to a trend (or vice versa). Less than 15% of sites had a change from 'very likely' in one direction to a trend in the opposite direction. While this is a relatively small proportion, it is still important. When comparing the proportion of sites in each of the trend categories, there were some small differences between the LAWA method and our approach of using the median of the three rolling datasets. Specifically, there was a decrease in the proportion of 'very likely improving' sites for BDISC, MCI, ECOLI and NH4N, whereas there was an increase in the proportion of ‘very likely improving’ sites for DRP (Figure 2). This suggests that while individual sites may change their trend from year to year, at a national summary scale there is largely consistency in the number of sites in each trend category.

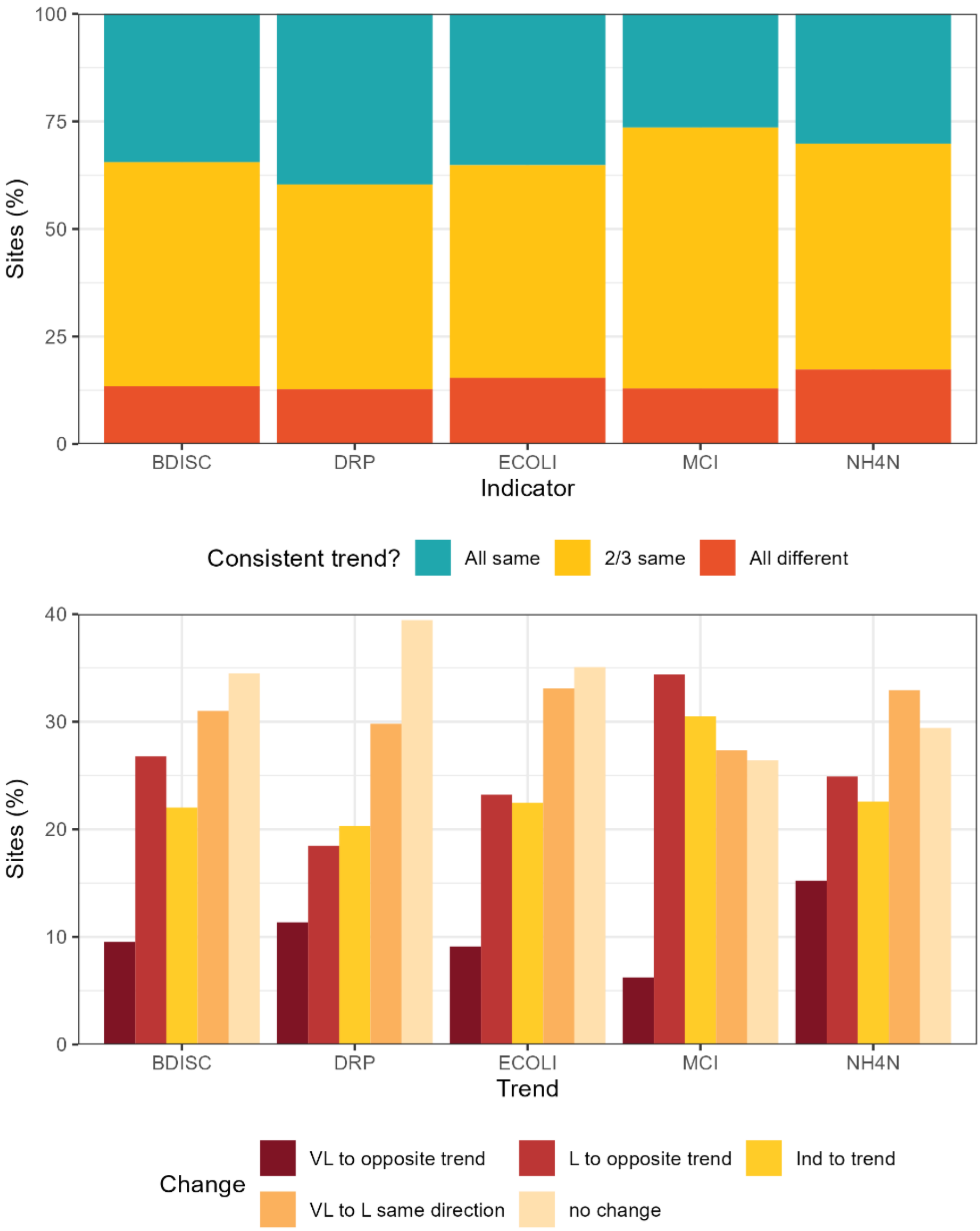

Figure 1. Consistency of rolling 8-year trends calculated across the last 10 years of monitoring data, using the LAWA method. Top: total proportion of sites where the trend category changed at least once across the three datasets. Note that this includes changes in the same direction (e.g. 'very likely improving' to 'likely improving'). Bottom: results broken down by the various trend categories. 'VL to opposite trend' indicates where a 'very likely' trend has changed to either 'likely' or 'very likely' in the reverse direction. 'L to opposite trend' indicates where a 'likely' trend has changed to either 'likely' or 'very likely' in the reverse direction. 'Ind to trend' indicates where an 'indeterminate' trend has changed to either 'likely' or 'very likely' in any direction. 'VL to L same direction' indicates where a 'very likely' trend has changed to 'likely' but in the same direction, or vice versa. 'no change' indicates a consistent trend across the three 8-year trends. Note that some sites changed across all three trend categories across all three rolling datasets and so are represented in multiple bars.

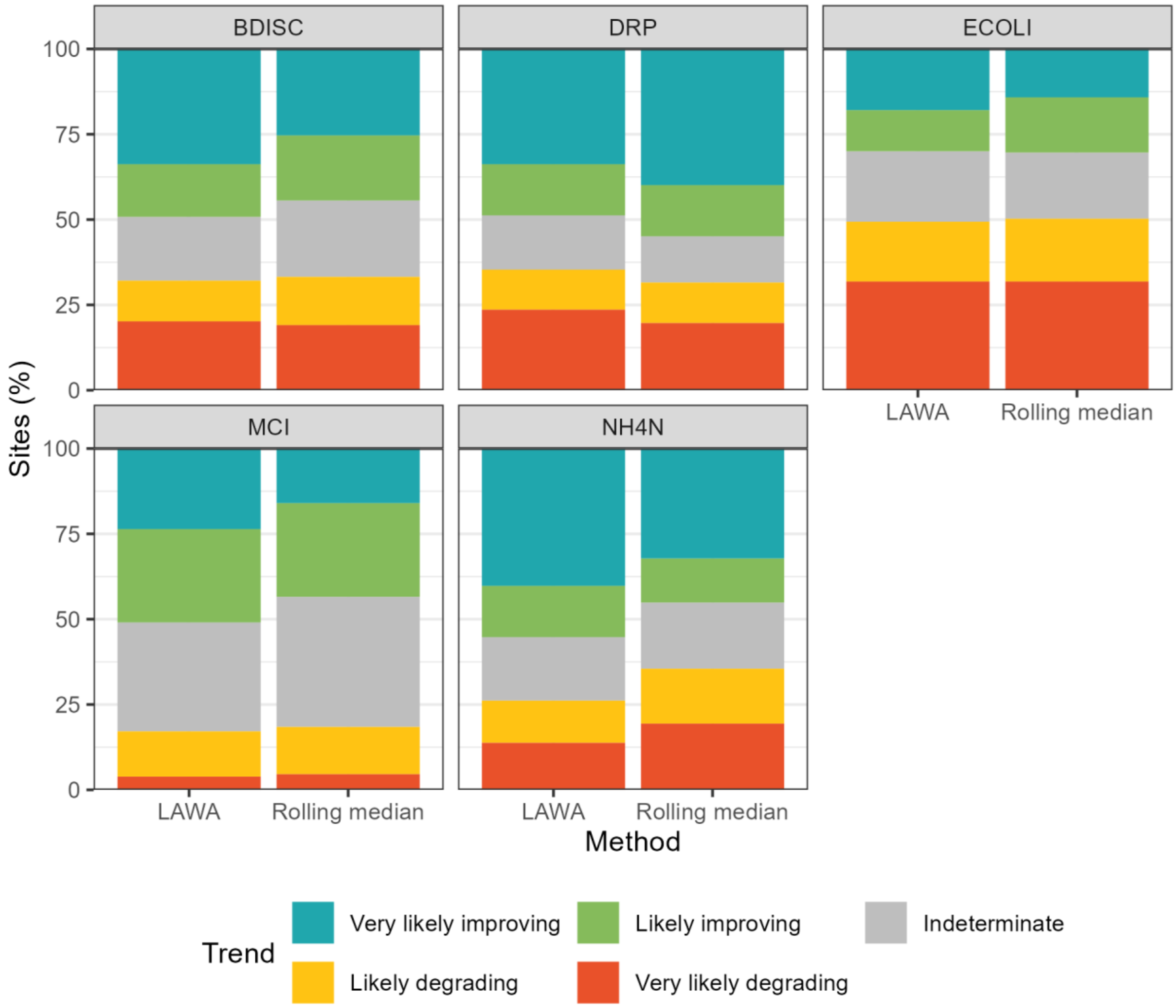

Figure 2. Consistency of trend scores calculated for rolling 8-year periods calculated across the last 10 years of monitoring data, using the LAWA method. The ‘rolling median’ is the median classification across the three 8-year periods, whereas ‘LAWA’ is the standard 10-year trend results.

3. QUANTILE REGRESSION TREND CALCULATION

In this section we 1) compare the results of the LAWA method to a QR method (0.5 quantile, or 50th percentile; Yu et al. 2003); 2) compare the default QR method described below to a second method, where we reran the QR on multiple quantiles (0.1–0.9 by 0.1 intervals) and classified each site based on the most common trend observed across the nine quantiles (Table 1); and 3) compare the default QR method to an alternative, where we assumed a 1% annual change was required before a trend could be assigned (see Section 3.1.2).

Table 1. Rules used to determine overall site trend based on multiple quantile regression (QR) analyses, calculated on quantiles 0.1–0.9 by 0.1 intervals. VLI = very likely increasing, LI = likely increasing, Ind = indeterminate, LD = likely degrading, VLD = very likely degrading.

Trend of each quantile	Number across quantiles (out of a possible 9)	Overall trend
VLI	5+	VLI
LI	5+	LI
Ind	5+	Ind
LD	5+	LD
VLD	5+	VLD
If no single trend has 5+ consistent		
VLI or LI	5+	LI
VLD or LD	5+	LD
All other permutations		Ind

3.1. Methodology

In the interest of keeping the results as comparable to the LAWA method as possible, the data preparation workflow used for the LAWA method was also applied for the QR method. Prior to running the QR analysis, censored values were dealt with, the minimum data requirements were checked and only those datasets that contained enough datapoints were kept for analysis (Snelder et al. 2021). Trends were calculated on log-transformed data.

3.1.1. Quantile regression method

Quantile regression is similar in intent to regular linear regression, except that instead of estimating the mean of the response variable (i.e. a water quality or ecological indicator) as a linear combination of predictor variables, interest is directed at a specified quantile of the response variable (e.g. the median; 0.5 quantile). Advantages of QR over linear regression include that it is more robust to non-normality of the data, and it enables analyses to be conducted on population metrics other than mean values. In the current context, the prime predictor variable of interest is time (measured in years), where the corresponding effect size is the estimated annual trend. Survey month was also included as a predictor variable (when multiple sampling periods were used within a year) to account for systematic monthly differences in an indicator during the year. As noted above, other predictor variables could be included in a QR analysis to evaluate the evidence for a trend over time after accounting for other factors that may affect the water quality or ecological indicator. Specifically, QR was performed by fitting the model:

$$y_i = \beta_0 + \sum_{j=1}^{12} \gamma_j month_{j,i} + \delta time_i$$

where y_i is the *i*th observation at a monitoring site; β_0 is the deseasonalised baseline level; γ_j are the month effects to allow for seasonal changes within the year; $month_{j,i}$ is a set of dummy variables that $= 1$ if survey i was collect in month j, and $= 0$ otherwise; δ is the annual trend; and $time_i$ is the time of the survey (in years). Note that inclusion of month effects will lessen any regular seasonal effects in the data, which can be particularly relevant if the sampling frequency has changed to allow greater comparability of temporal changes. Furthermore, the constraint that $\sum_{j=1}^{12} \gamma_j = 0$ was applied such that the β_0 parameter does not relate to any particular month. The month terms were excluded from the QR for MCI as, typically, this was measured only once per year.

Using QR, the likelihood of a trend was evaluated using 1-sided confidence intervals and the decision rules given in Table 1. Note that a 1-sided $(1-\alpha)100\%$ confidence interval is constructed by calculating a 2-sided $(1-2\alpha)100\%$ confidence interval, and considering only the relevant lower or upper bound that is of interest. That is, when constructing a 1-sided $(1-\alpha)100\%$ confidence interval, it is of interest to establish only a lower or an upper bound on the estimate, rather than both.

Table 2. Trend decision rules using 1-sided confidence intervals (CI).

Trend decision	1-sided 67% CI	1-sided 90% CI
Very likely increasing (VLI)	Lower limit > 0	Lower limit > 0
Likely increasing (LI)	Lower limit > 0	Lower limit < 0
Indeterminant (Ind)	All other cases	
Likely decreasing (LD)	Upper limit < 0	Upper limit > 0
Very likely decreasing (VLD)	Upper limit < 0	Upper limit < 0

3.1.2. Censored values

Censored values are those values that are either above or below detection limit. The default LAWA method deals with censored values by multiplying any values below the detection limit by 0.5 and multiplying values above the detection limit by 1.1 (Snelder et al. 2021). In addition, where a step-change in detection limit has occurred through time, all values below (or above, e.g. for BDISC) are shifted to the highest (or lowest) detection limit value. To be used, a given dataset must have a least five values that have not been censored (along with all the other data prerequisites; Snelder and Fraser 2018). Where a large proportion of values have been censored, this can cause issues with the QR method (and likewise with the LAWA method). For example, if a large majority of the measures are below detection limit and values * 0.5 are used, this results in a large proportion of values in the dataset having the exact same value (detection limit * 0.5). This can result in trends being estimated on a series of identical numbers plus a small number of non-censored values (e.g. Appendix 1). The resulting QR slope estimate will be very small (e.g. 0.0000001), with a tight confidence interval. Sites are then assigned ‘very likely’ trends, which given the minuscule slope, intuitively seems unreasonable. A way to deal with this issue is to require some minimal year-on-year change in the water quality attribute before a trend is considered possible – for example, an annual change of at least 1% in the given water quality attribute. This results in sites with a high proportion of censored values often being assigned ‘indeterminate’ trend scores. Alternatively, as in the LAWA approach, sites with a high proportion of censored values could simply be ‘flagged’.

3.1.3. Software

All analyses were carried out using the statistical computing software R v4.2.3 (R Core Team 2021). We used the tidyverse v2.0.0 metapackage (Wickham et al. 2019) for data manipulation, the quantreg package (Koenker et al. 2023) for QR and the future v1.32.0 package (Bengtsson 2020) for parallelisation.

3.2. Results

Results of the comparison between the LAWA and QR methods showed that the QR method (0.5 quantile) was consistently more conservative than the LAWA method (Figure 3). For all five water quality indicators and across all three time periods, the QR method resulted in a higher percentage of sites being classified as ‘likely improving’, ‘likely degrading’ and ‘indeterminate’ than the LAWA method. Likewise, significantly fewer sites were classified as ‘very likely improving’ or ‘very likely degrading’ for the QR method compared to the LAWA method (5–45% for the LAWA method vs 3–32% for the QR method). This pattern was particularly extreme for NH4N; this had relatively high levels of measures below detection limit (see Appendix 1), which contributed to the observed differences.

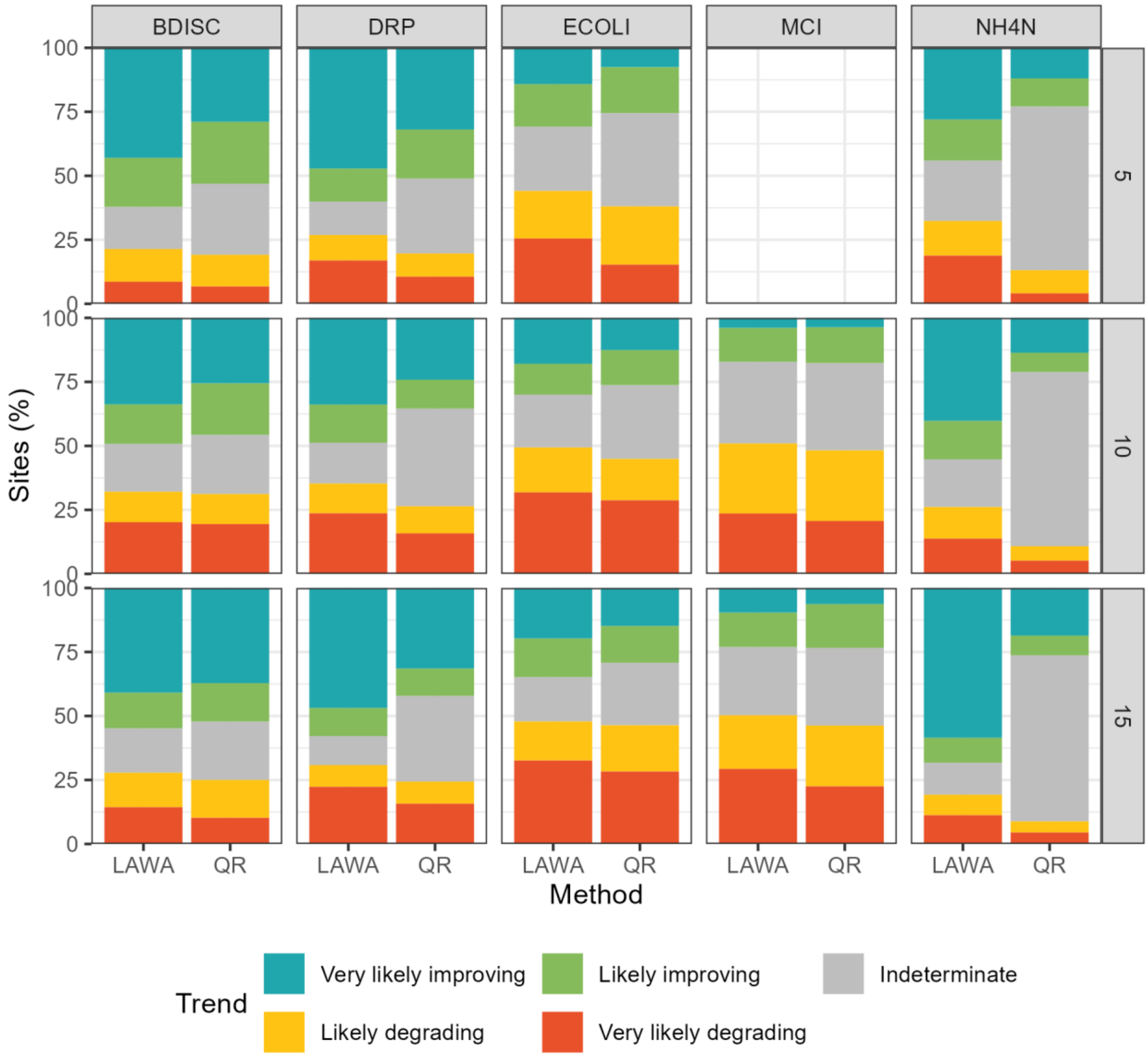

Figure 3. Comparison of LAWA and quantile regression (QR) methods for calculating trends in water quality and ecological indicators (columns), and across 5-, 10- and 15-year periods (rows). Five-year trends for Macroinvertebrate Community Index (MCI) were not calculated as MCI data are measured only on an annual basis, and so any 5-year trends would be calculated on only 5 data points (or fewer), which is not considered robust.

Comparison of the two alternative QR methods (0.5 quantile vs most common trend across quantiles) produced similar results (Figure 4). This suggests that if the QR method is to be used, the simpler, 0.5 quantile (i.e. the median) metric would be adequate to describe trends. Requiring a slope of greater than a 1% annual change in the given water quality attribute before a trend could be assigned resulted in a decrease in the proportion of sites with a 'very likely' or 'likely' trend and a subsequent increase in the proportion of sites classified as 'indeterminate'. This pattern was extreme for the ecological indicator MCI (Figure 4).

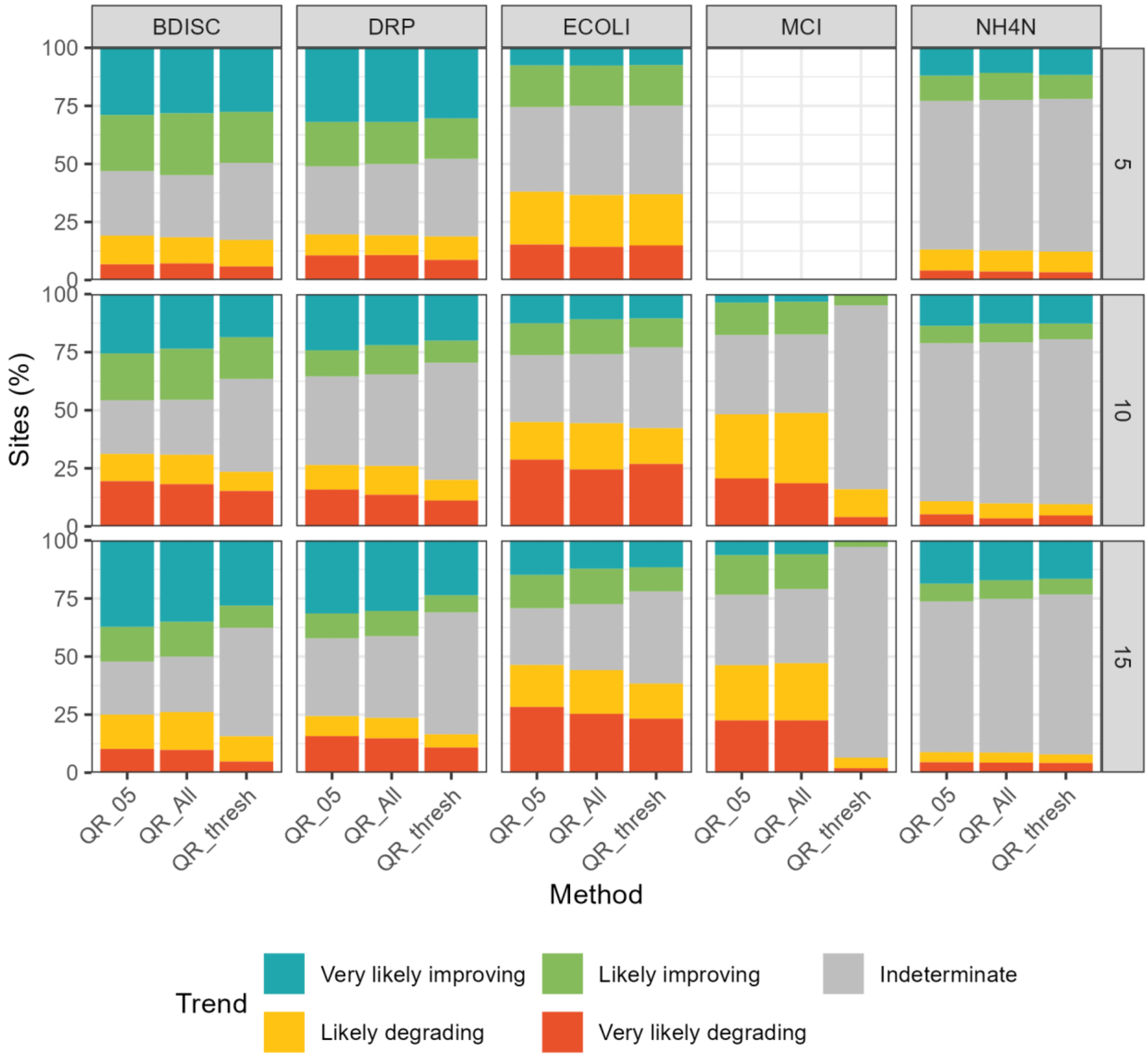

Figure 4. Consistency of quantile regression (QR) results using the 0.5 quantile (QR_05) compared to using the most common trend across a range of quantiles (QR_All; 0.1–0.9 by 0.1 intervals), and requiring a minimum slope estimate of 0.001 for increasing trends and -0.001 for decreasing trends (QR_thresh).

4. COMBINING STATE AND TREND METRICS

Site state scores can range from 'A' to 'D' for attributes DRP, NH4N, MCI and BDISC, and 'A' to 'E' for ECOLI, with 'A' being indicative of 'good' water quality and 'D' and 'E' indicative of poorer water quality. Currently, communicating the condition of a site requires descriptions of both the state and trend, as communicating one or the other leaves out important information. For example, a 'very likely degrading' trend will mean different things for sites that have a state of 'A' or a state 'D'. Here, we consider two simple methods for combining state and trend metrics into a single new metric, 'condition'. The goal of combining state and trend was to make it easier to communicate how water quality indicators are behaving at a site and to enable more meaningful, high-level summaries by being able to present the proportion of sites in different 'conditions' for each variable.

4.1. Missing data

A precondition of calculating site condition is having both state and trend data available for a given site. For many sites, either state or trend data are not available, or the data are available but there are not enough data points to calculate a trend and so it is classified as 'not determined' (Figure 5). Across the water quality metrics, the proportion of sites for which both state and trend were measured decreased as the trend period increased (5, 10 and 15 years). Across all possible sites and time periods, 35–75% of sites had both trend and state metrics. Ignoring sites that had neither state nor trend (i.e. the given water quality metric was not measured at the site), this increased to more than 55% (BDISC 15-year period) to 99% (BDISC 5-year period) of sites.

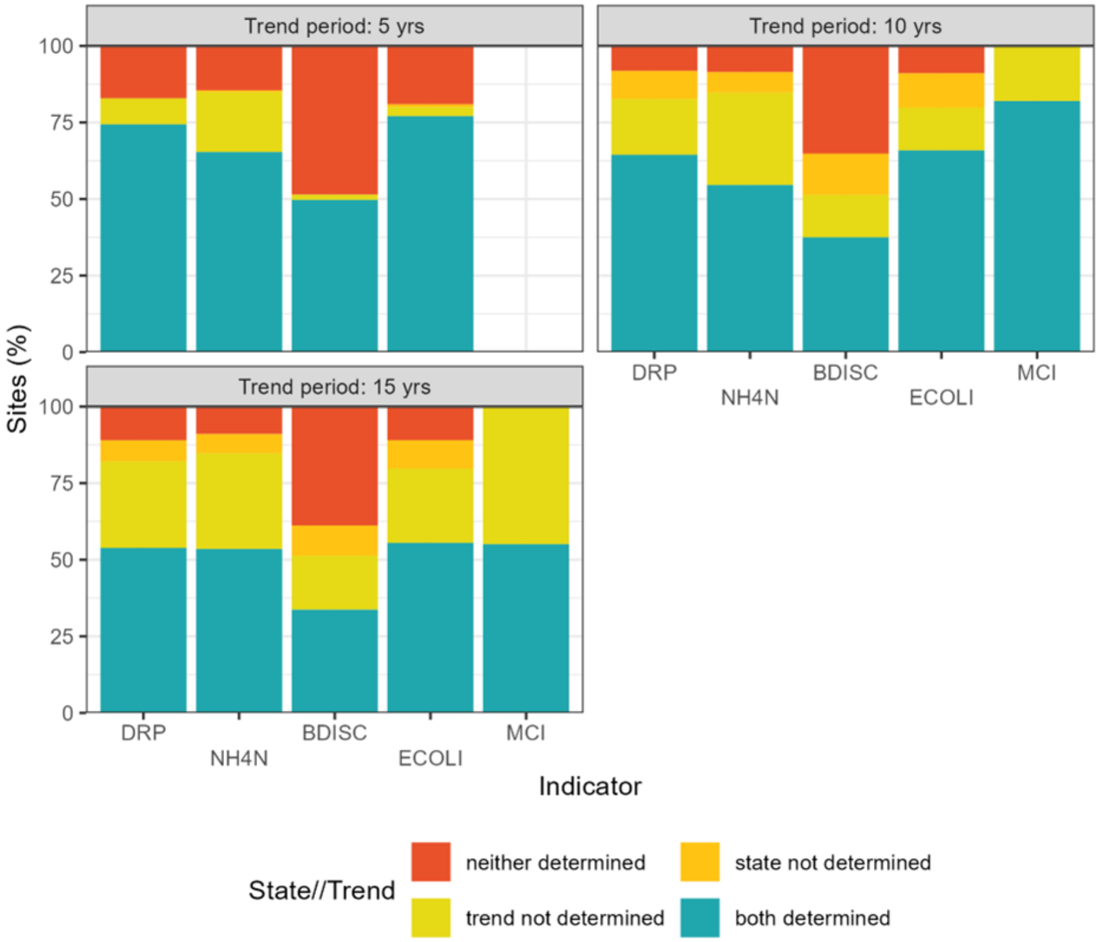

Figure 5. Percentage of sites that have either state or trend data, or both, missing for each of the water quality indicators and time periods. Macroinvertebrate Community Index (MCI) was not calculated for the 5-year trend.

4.2. Classifying site condition

We tested two approaches to combining state and trend into a new metric 'condition'. For both approaches we used the 2022 state and trend data available from the LAWA website.[1] For the first approach, we classified a site as pass / fail determined based on the state attribute, and then improving / degrading based on the trend attribute (two-by-two classification, plus 'not determined' sites). The resulting classification meant a site could either be classified as 'good' (pass state, improving or indeterminate trend), 'at risk' (pass state, degrading trend), 'recovering' (fail state, improving trend), 'bad' (fail state, indeterminate or degrading trend), or 'not determined' (either state or trend not determined; Table 3). Here, we used the 'A' and 'B' bands from the NPS-FM as a pass state and the 'C', 'D' and 'E' bands as a fail state as 'proof of concept' thresholds, and the default LAWA method when calculating

[1] https://www.lawa.org.nz/download-data

trends. Ultimately, it will be up to regional councils and local communities to decide what constitutes a pass or fail. Target attribute states (TAS) would be a logical choice once these are determined.

For the second approach, we allowed differentiation between the state attribute bands. This resulted in eight possible condition scores (four-by-two classification): 'very good' (A, improving or indeterminate), 'good' (B, improving or indeterminate trend), 'at low risk' (A, degrading trend), 'at risk' (B, degrading trend), 'recovering' (C, improving trend), 'possibly recovering' (D/E, improving trend), 'bad' (C, indeterminate or degrading trend), 'very bad' (D/E, indeterminate or degrading trend), and 'not determined' (either state or trend not determined; Table 4).

Table 3. Combinations of state and trend used to calculate a new metric condition, based on a simple 2 × 2 decision matrix of pass / fail for the state attribute and improving / degrading for the trend. VLI = very likely improving, LI = likely improving, I = indeterminate, LD = likely degrading, VLD = very likely degrading, ND = not determined. Indeterminate trends are sites where it is unclear if the trend is improving or degrading. Not determined are sites that did not meet the data requirements to calculate a metric, e.g. too few data points.

State / trend	VLI	LI	I	LD	VLD	ND
A	Good	Good	Good	At risk	At risk	ND
B	Good	Good	Good	At risk	At risk	ND
C	Recovering	Recovering	Bad	Bad	Bad	ND
D	Recovering	Recovering	Bad	Bad	Bad	ND
E	Recovering	Recovering	Bad	Bad	Bad	ND
ND	ND	ND	ND	ND	ND	ND

Table 4. Combinations of state and trend used to calculate new metric condition, based on a simple 4 × 2 decision matrix of A / B / C / D / E for the state attribute, and improving / degrading for the trend. VLI = very likely improving, LI = likely improving, I = indeterminate, LD = likely degrading, VLD = very likely degrading, ND = not determined. Indeterminate trends are sites where it is unclear if the trend is improving or degrading. Not determined are sites that did not meet the data requirements to calculate a metric, e.g. too few data points.

State / trend	VLI	LI	I	LD	VLD	ND
A	Very good	Very good	Very good	At low risk	At low risk	ND
B	Good	Good	Good	At risk	At risk	ND
C	Recovering	Recovering	Bad	Bad	Bad	ND
D	Possibly recovering	Possibly recovering	Very bad	Very bad	Very bad	ND
E	Possibly recovering	Possibly recovering	Very bad	Very bad	Very bad	ND
ND	ND	ND	ND	ND	ND	ND

4.3. Classification results

The simple two-by-two classification worked well and was easy to interpret. MCI was the only indicator where more than 50% of sites were classified as in 'bad' condition. For all other indicators, 2–45% of sites were classified as 'bad', 10–25% as 'at risk', 2–30% as 'recovering' and 12–70% as 'good'. The proportion of sites in each condition band was relatively stable across the 5-, 10- and 15-year trend periods (Figure 6).

Visualising condition via national maps illustrates where data are currently lacking and some potential spatial patterns. For example, NH4N is 'good' across much of the country, except Southland and the central North Island, and MCI is 'bad' across much of the country (Figures 7–11). The increased resolution of the second approach (four-by-two classification) allowed more detail to be seen when inspecting results, but overall it did not add much, and it made summarising high-level patterns more cumbersome (Figure 12).

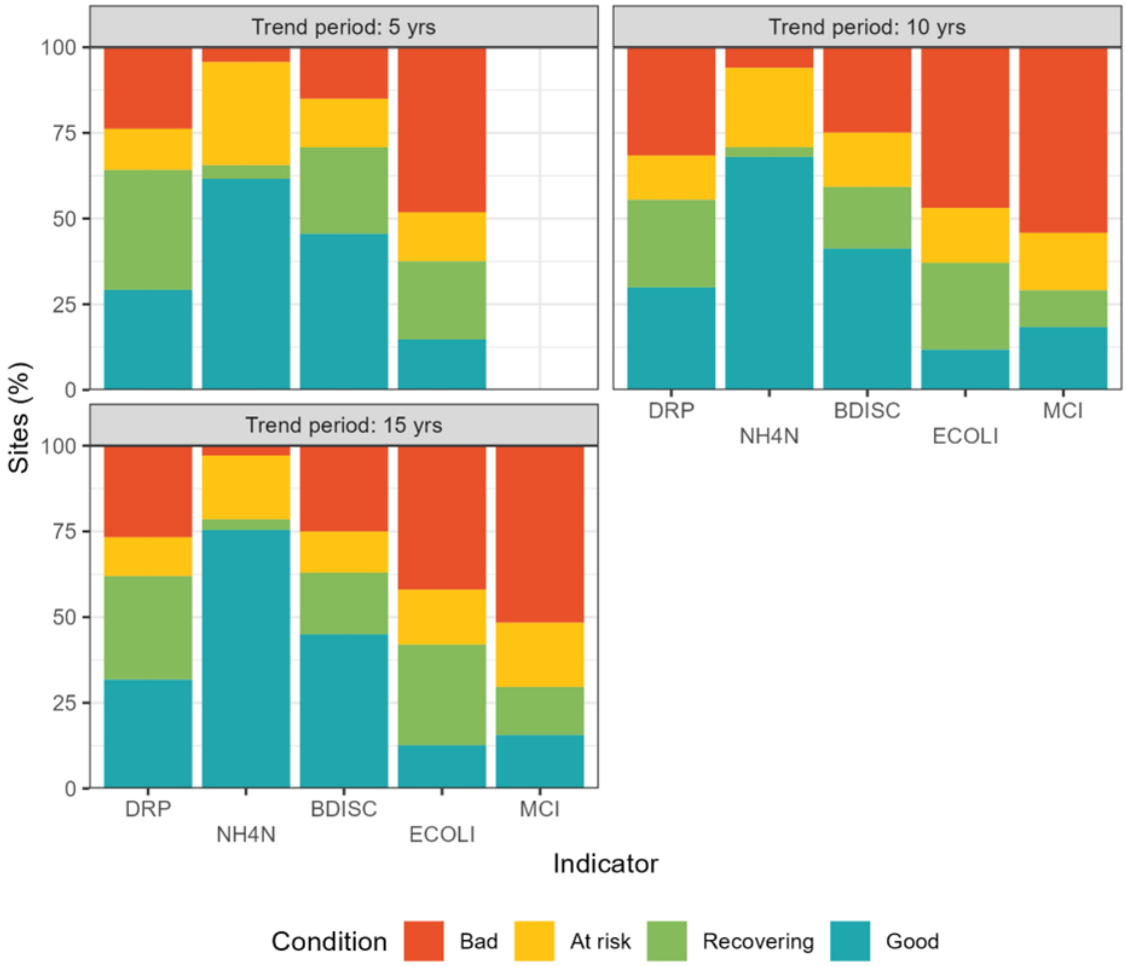

Figure 6. Results of combining state and trend attributes using a 2 × 2 decision tree (state: A / B / C / D / E; trend: improving / degrading).

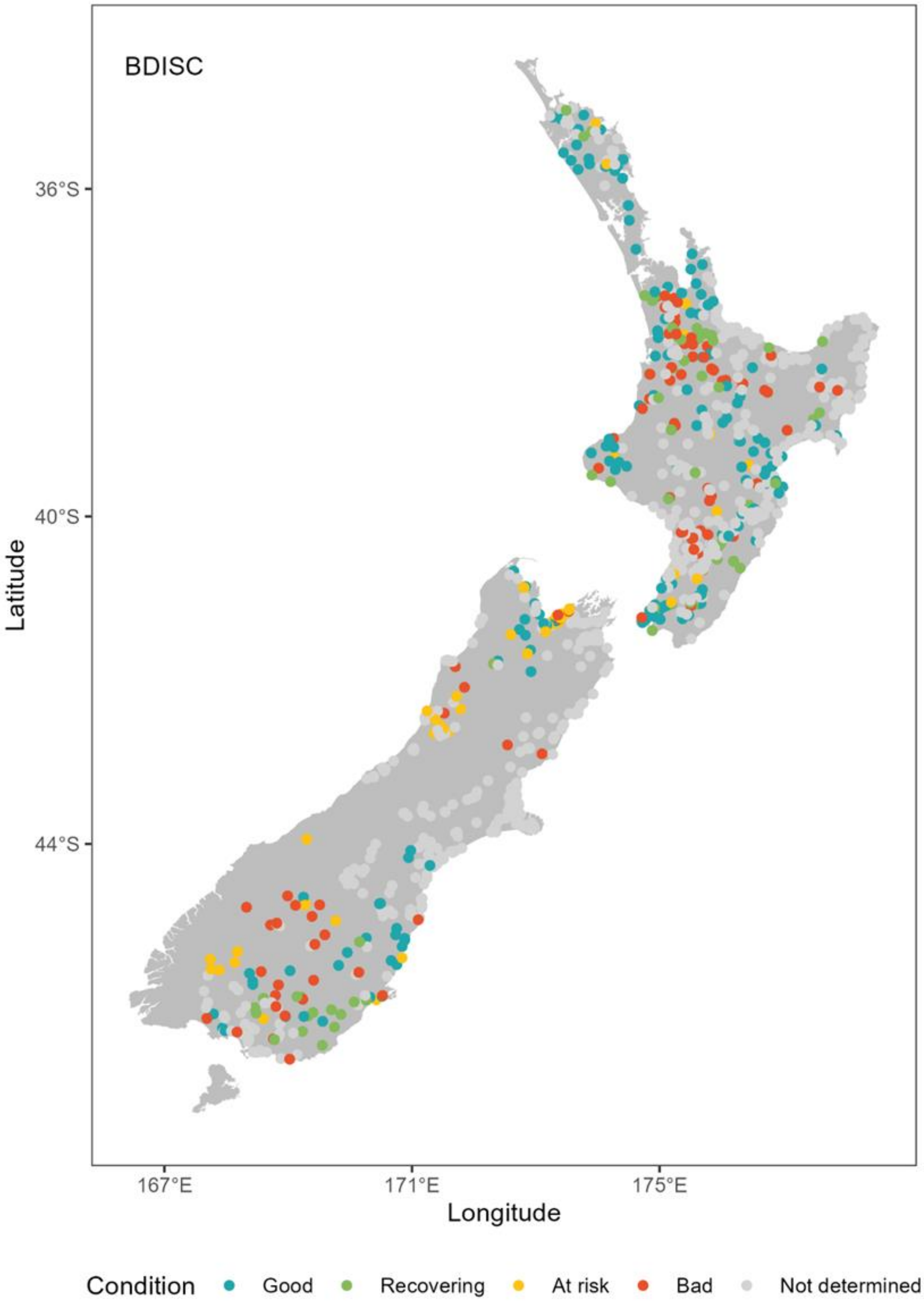

Figure 7. Condition (BDISC) of LAWA sites across Aotearoa New Zealand, based on 10-year trend data.

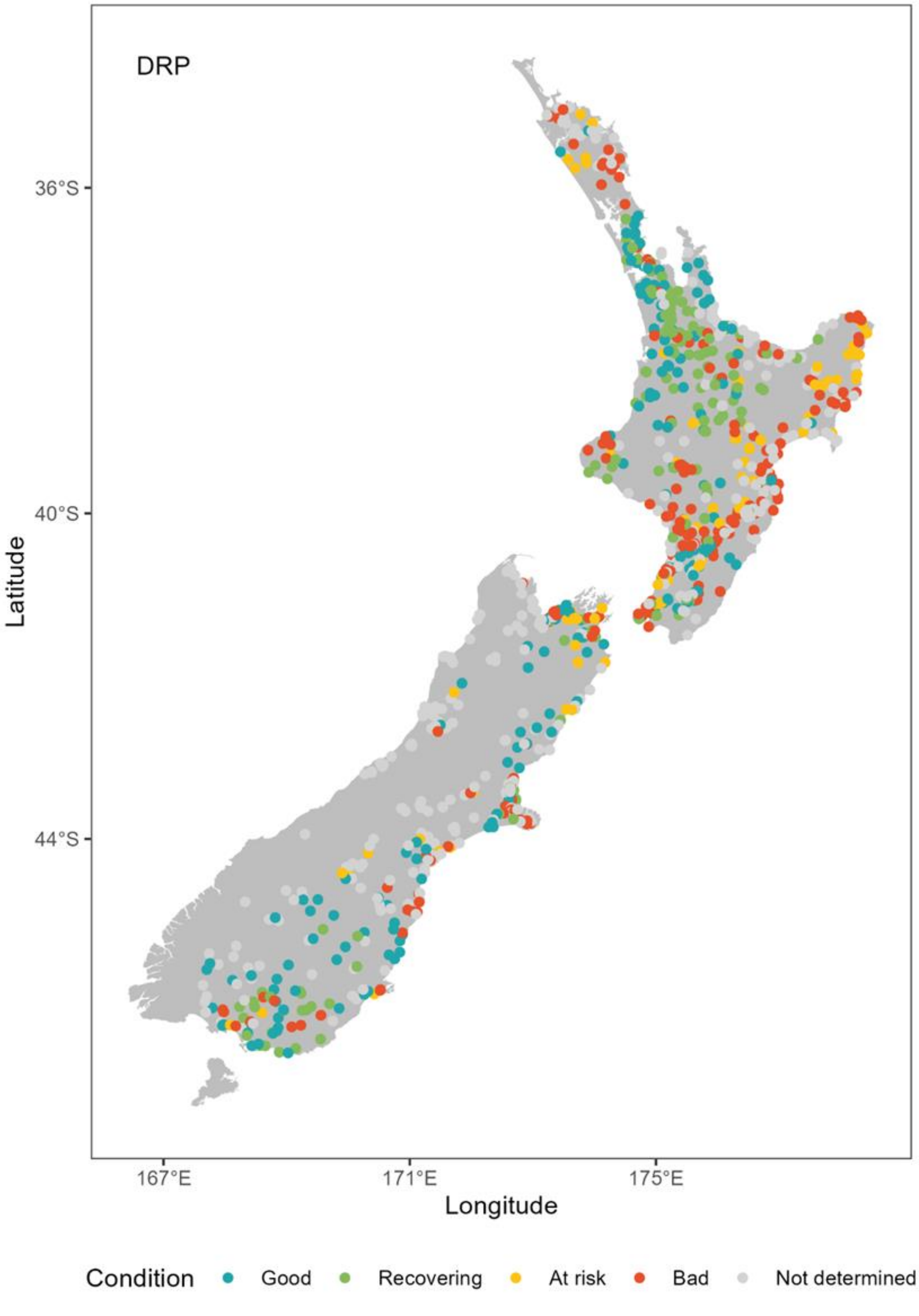

Figure 8. Condition (DRP) of LAWA sites across Aotearoa New Zealand, based on 10-year trend data.

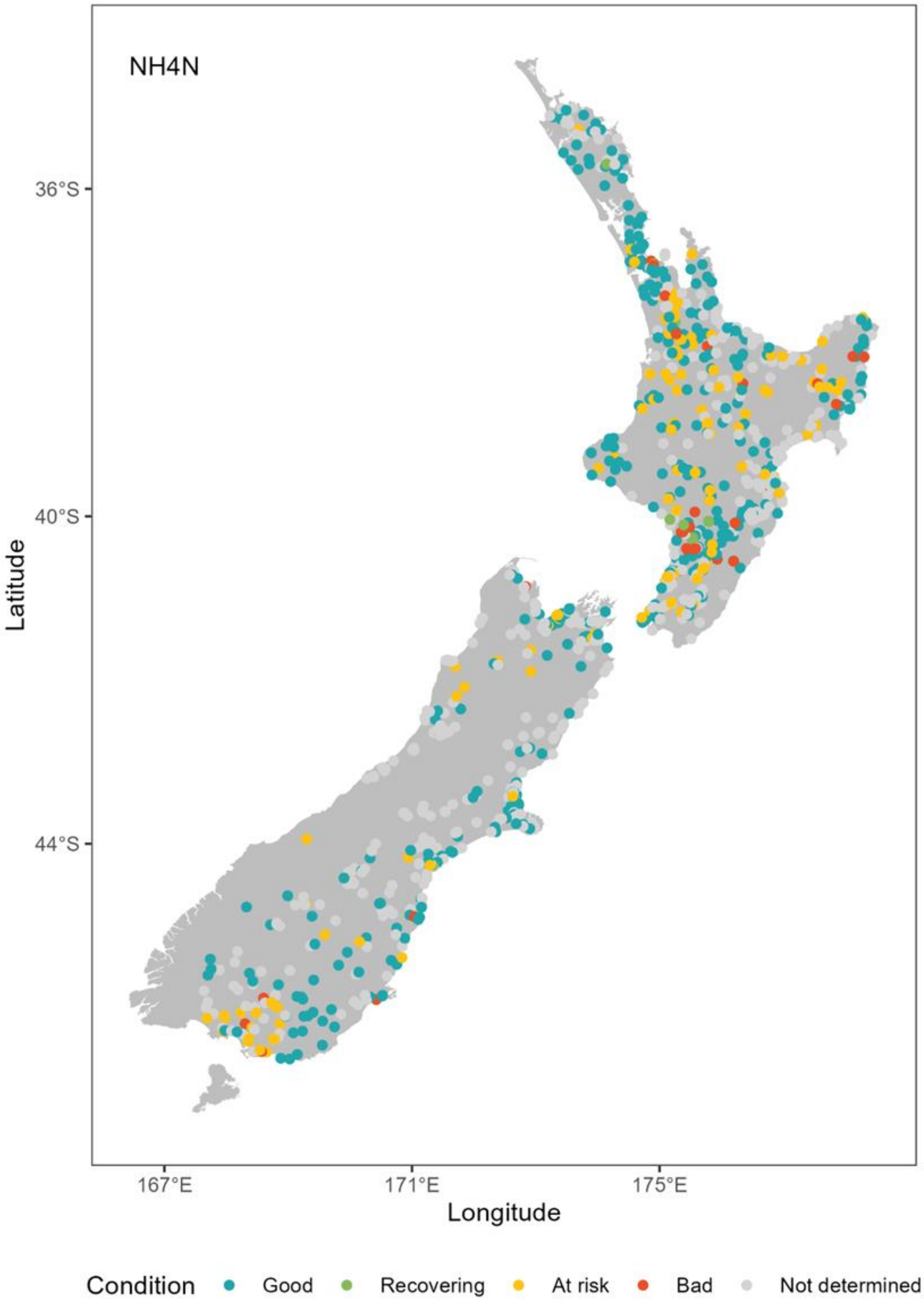

Figure 9. Condition (NH4N) of LAWA sites across Aotearoa New Zealand, based on 10-year trend data.

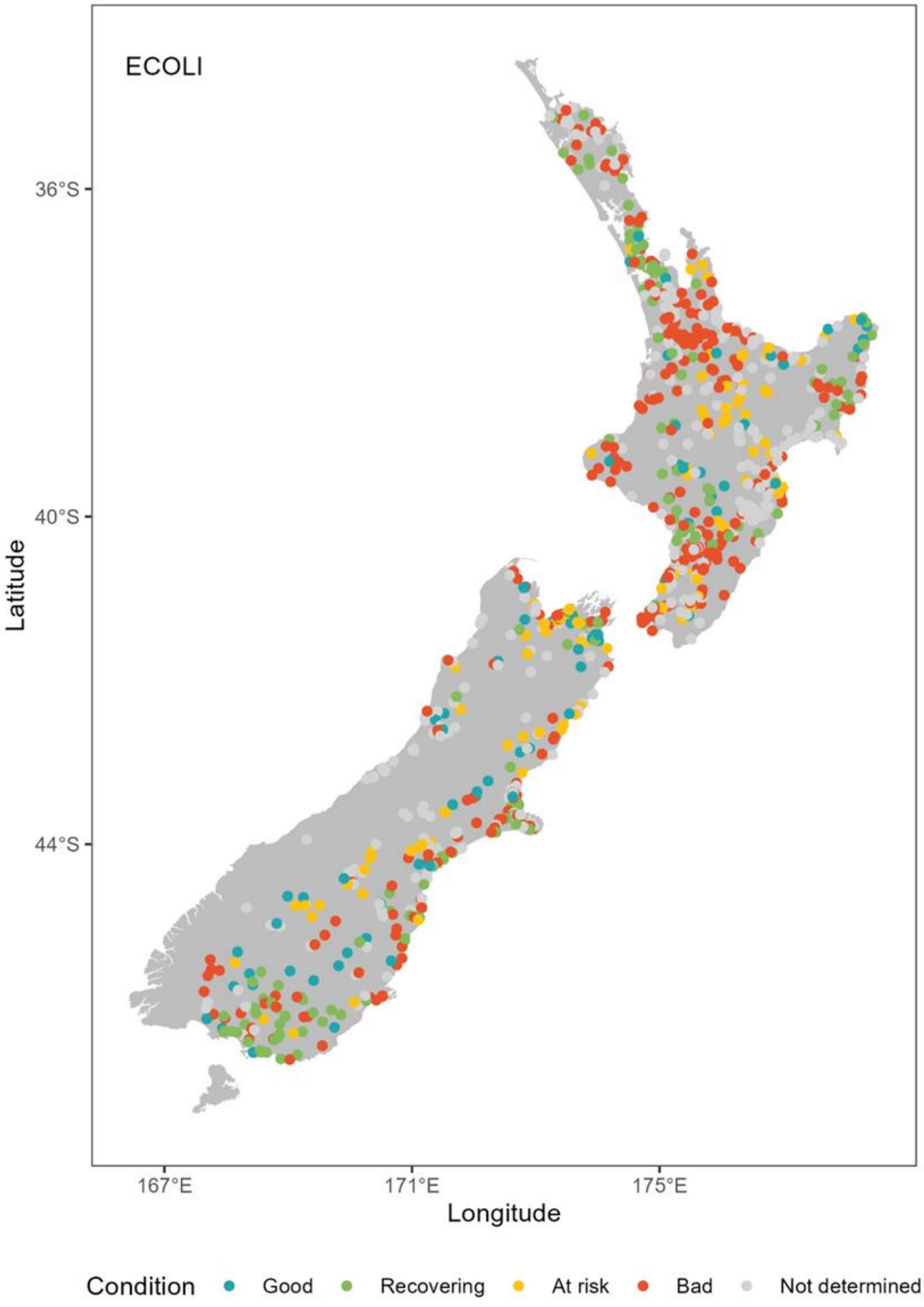

Figure 10. Condition (ECOLI) of LAWA sites across Aotearoa New Zealand, based on 10-year trend data.

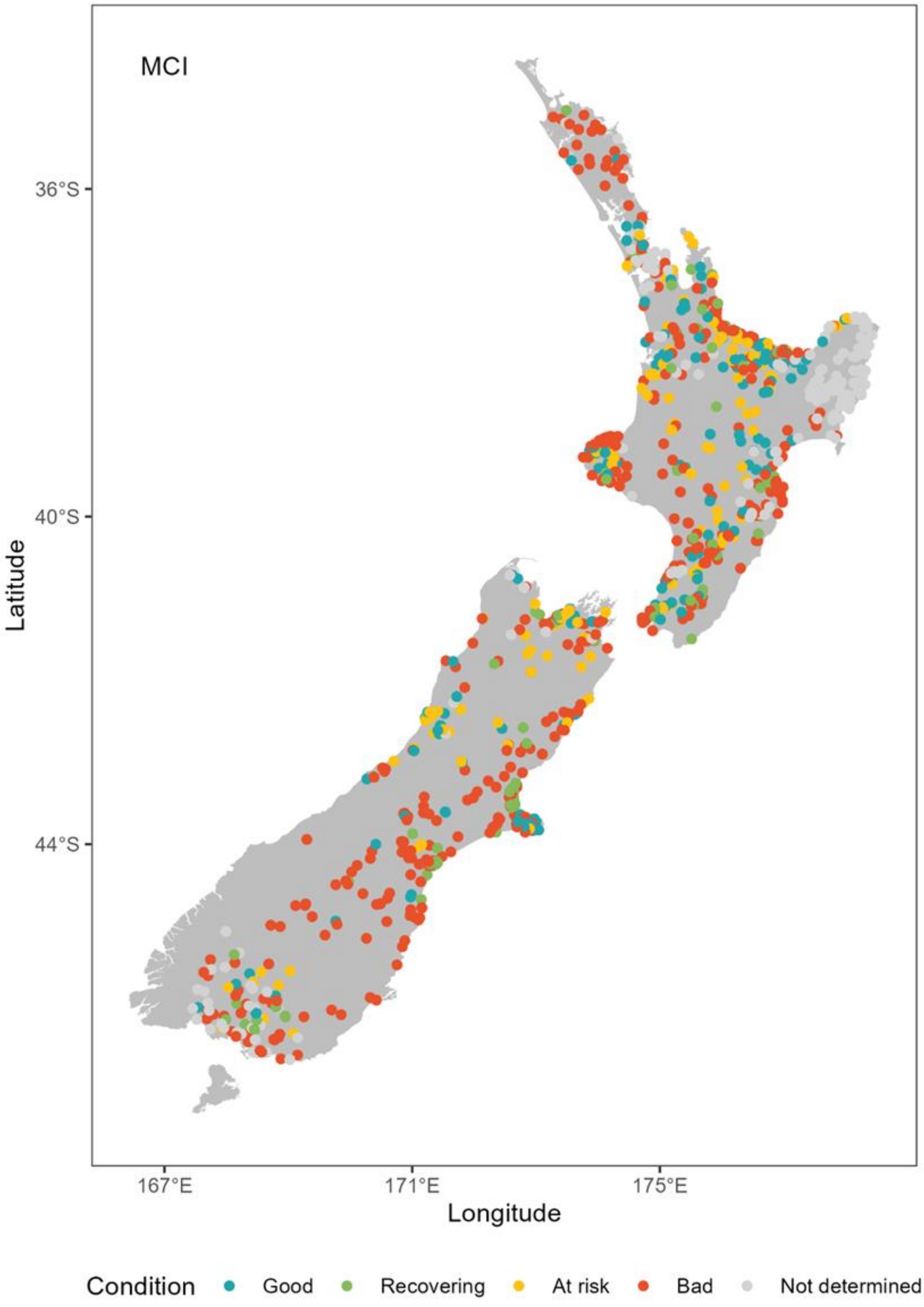

Figure 11. Condition (MCI) of LAWA sites across Aotearoa New Zealand, based on 10-year trend data.

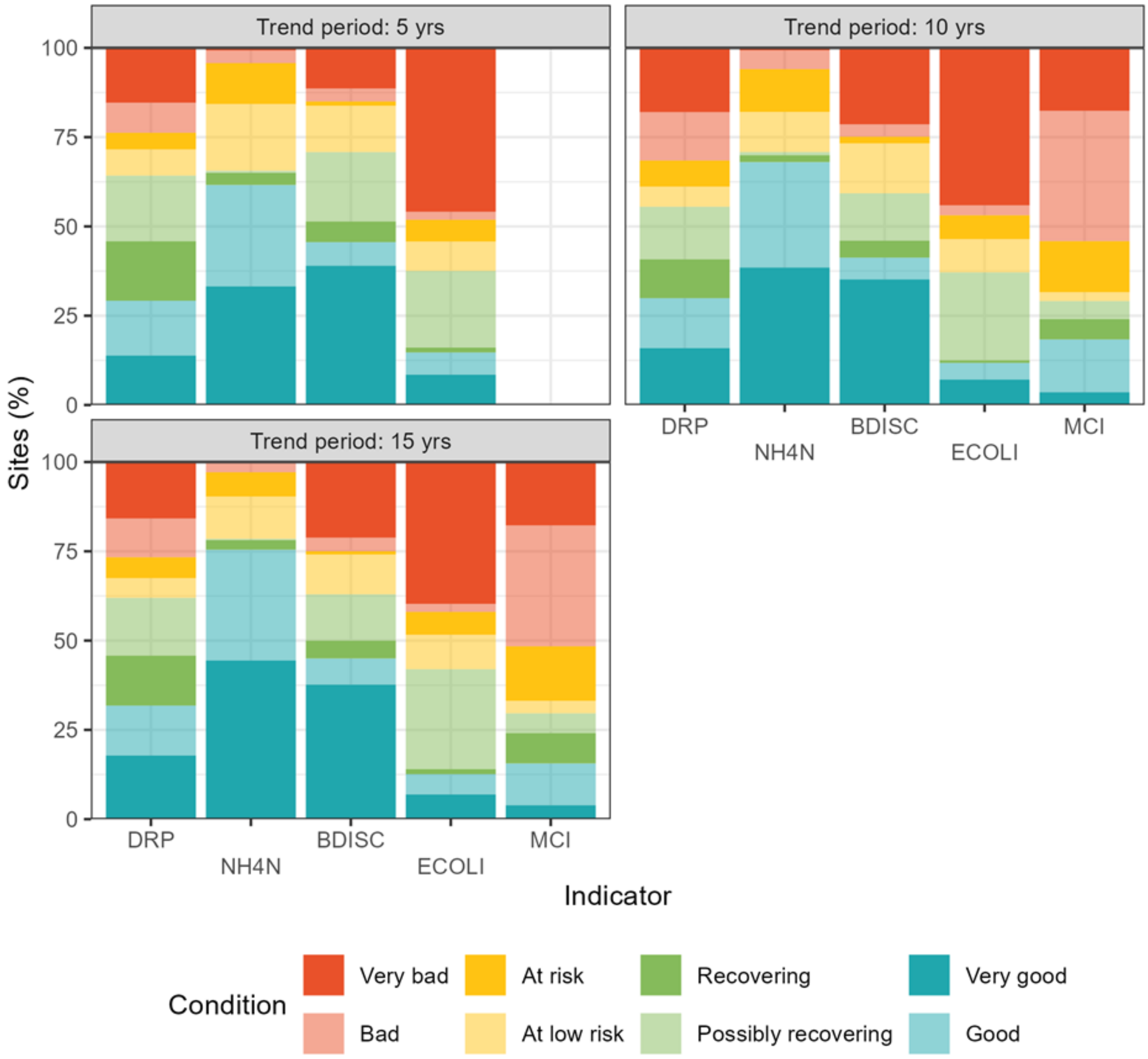

Figure 12. Results of combining state and trend attributes using a 4 × 2 decision matrix (state: A / B / C / D / E; trend: improving / degrading).

5. SUMMARY

Land, Air, Water Aotearoa (LAWA) leverages a database of water quality and ecological health indicators to report on the state and trend of these attributes at individual sites and at a national scale. Trends are calculated for indicators over the last 5, 10 and 15 years using a two-stage process, via a Mann–Kendall statistical test and Sen slope regression analysis. Concerns have been expressed that this method results in a large number of sites with likely trends, but that some of these trends appear relatively weak and it is difficult to identify potential causes for those observed. In addition, there have been instances at some sites where a trend in one direction has been indicated by the data, but the following year a trend in the opposite direction has been identified. This trend switching over time has made some people question the robustness of the approach.

Here, we explored two distinct approaches to addressing these limitations for each of five water quality and ecological indicators: black disc (BDISC), dissolved reactive phosphorus (DRP), *Escherichia coli* (ECOLI), Macroinvertebrate Community Index (MCI) and ammoniacal nitrogen (NH4N). First, we looked at the consistency of trends calculated using the LAWA method over three rolling 8-year datasets. Our results show that from year to year a given trend had a high chance of changing, with less than 40% of sites having a consistent trend category across the three rolling datasets. However, most changes in trend category were relatively small, and less than 35% of sites had a change in trend direction across the three rolling datasets. It was more common to see changes from 'very likely' to 'likely' (or vice versa) in a consistent direction, or a change from 'indeterminate' to a trend (or vice versa).

The second method we examined was a QR approach. Within this approach we considered three different options: 1) calculating regression estimates on the 0.5 quantile, 2) calculating regression estimates on nine different quantiles (0.1–0.9) and assigning the most common trend across the quantiles as the 'true' trend, and 3) requiring a 1% or greater annual change in the given water quality parameter before a trend could be assigned. Each of the three QR methods resulted in a significant increase in the proportion of sites classified as 'indeterminate', 'likely increasing' and 'likely decreasing' compared with the LAWA approach, with associated decreases in the proportion of sites classified as 'very likely increasing' and 'very likely decreasing'. This pattern was particularly extreme for the ecological indicator MCI when requiring a 1% or greater annual change. The large increase in indeterminate sites for NH4N but not other metrics when using the 0.5 quantile QR approach, and similarly for MCI when using the threshold approach, shows that not all metrics are affected equally by the different methods.

We observed differences in the proportion of sites in each of the trend categories when comparing the LAWA method to the QR method. This indicates that 1) there were some methodological choices made that resulted in the QR method assigning

more sites as having indeterminate trends, or 2) there are potential differences in how confidence levels translate into ascribed trend categories. This is an area that requires further investigation.

Given the limited scope of this analysis, a more thorough study is warranted to compare all possible options (the LAWA, QR and any other appropriate methods). We suggest using a simulation-based assessment, where datasets with various trends, noise levels and censored values are simulated, trends calculated and results compared, with a specific focus on trend consistency. Within the current methodological framework, we see four areas that warrant a deeper examination: 1) using the rolling window approach, and the effect of using different-sized windows; 2) changing the confidence level trend attribution cut-off values to be more conservative; 3) increasing the number of values above (or below) the detection limit that are required before a trend can be calculated; and 4) implementing trend slope thresholds that represent ecological significance. Each of these changes has the potential to reduce trend switching, without a wholesale change in methodology. QR could then be looked at in more detail using the simulation-based assessment outlined above. Finally, additional analyses looking at methods that account for spatial autocorrelation (e.g. state-space modelling) and methods that account for larger-scale natural environmental (e.g. climactic) cycles would also be valuable, as these have been identified as potential trend drivers (Snelder et al. 2021). A simulation-based assessment is more informative than applying alternative methods to an actual dataset as ‘truth’ is known when simulating data, so the performance properties of each method can be determined more objectively.

The combination of the state and trend scores into a new ‘condition’ attribute was straightforward to calculate, and the new metric is easy to interpret and can be readily presented alongside existing metrics. The new site ‘condition’ metric could also be generated quickly to produce simplified national-scale visualisations and summaries. The limitation of requiring both state and trend attributes to be present for a given site to calculate ‘condition’ could be overcome by continuing to also present state and trend data separately, so no information is lost. The only challenge associated with the condition metric is determining what state band (A / B / C / D / E) represents a ‘pass’ or ‘fail’ for a given site. One option could be to apply the metric only to sites where a target attribute state has been set. Alternatively, an approach similar to that applied here (A and B are a pass, and C, D and E are a fail) could be adopted until target attribute states have been implemented nationally.

6. APPENDIX

Raw NH4N data for two sites is presented in Figure A1.1, along with the LAWA and QR 10-year trends. These are two sites where the LAWA and QR methods produced contrasting results.

At 'ccc-00005' (Riccarton Drain), the LAWA trend was 'likely degrading', whereas the QR trend was 'indeterminate'. This site had only two censored values in the 10-year trend dataset.

At 'gw-00010' (Waikanae River at Greenaway Road), the LAWA trend was 'very likely improving', whereas the QR trend was 'indeterminate'. This site had a large proportion of censored values, and a step-change in the censoring level. Prior to trends being calculated, all values less than the highest censored value were replaced with the highest censored value (0.01 in this case), so that the step-change did not result in an artificial trend. In this case, there were a number of points above the detection limit between 2012 and 2014, which has resulted in a 'very likely improving' trend for the LAWA method. Whether or not 'gw-00010' passes the 'eyeball' test as a site that should be classified as a 'very likely improving' trend is subjective, but an increase in the required number of non-censored values (currently a minimum of five) would reduce the number of sites that are assigned trends.

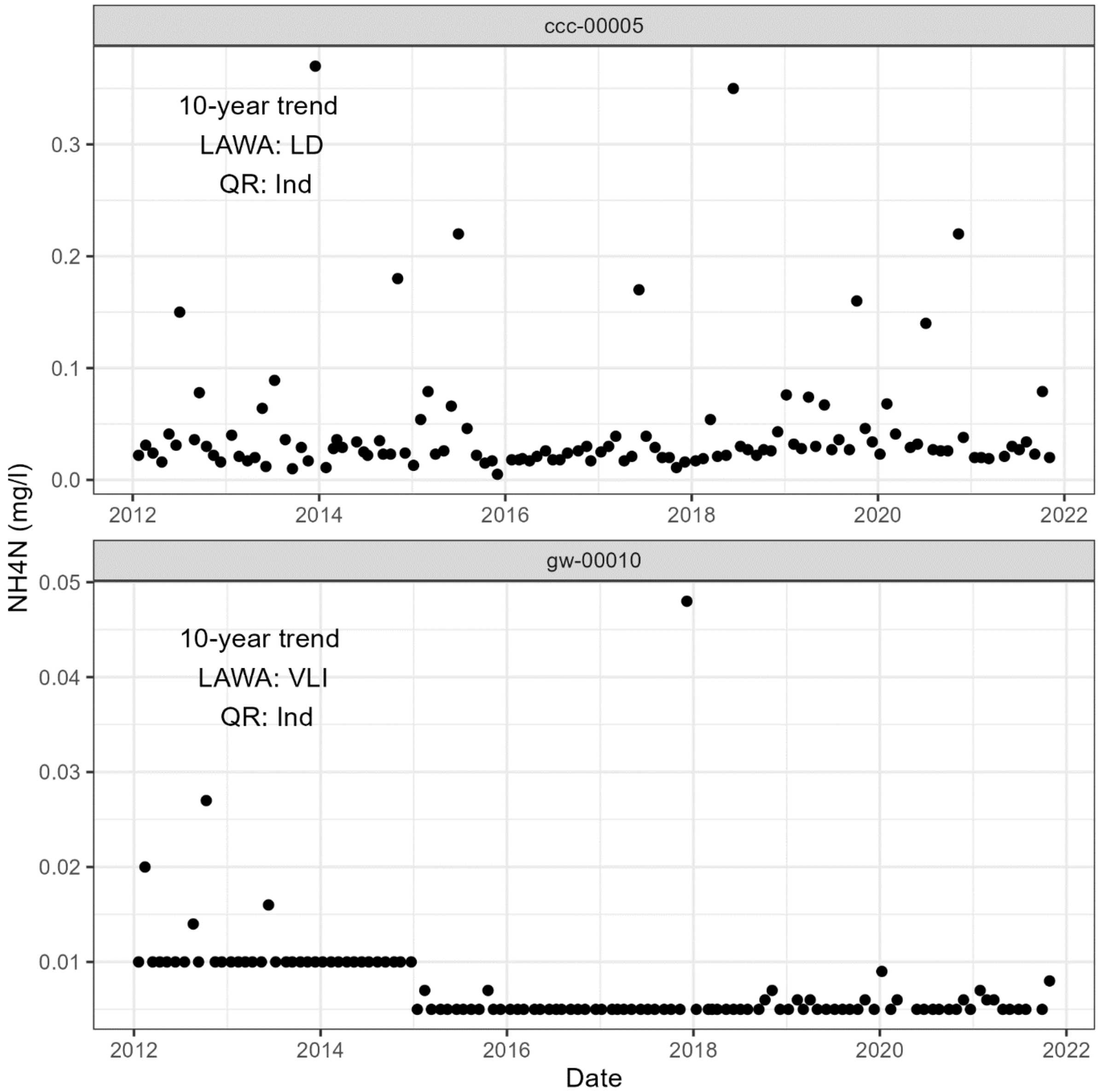

Figure A1.1. Raw NH4N values for two sites where the LAWA and QR methods produced contrasting results.

7. REFERENCES

Bengtsson H. 2020. A unifying framework for parallel and distributed processing in R using futures. [accessed 3 July 2023]. arXiv preprint arXiv:2008.00553

Helsel DR, Hirsch RM, Ryberg KR, Archfield SA, Gilroy EJ. 2020. Statistical methods in water resources: US Geological Survey techniques and methods, book 4, chapter A3. Reston (VA): US Geological Survey.

Koenker R, Portnoy S, Ng PT, Zeileis A, Grosjean P, Ripley BD. 2023. quantreg: quantile regression. version 5.95. [accessed 3 July 2023]. https://cran.r-project.org/web/packages/quantreg/index.html

McBride GB. 2019. Has water quality improved or been maintained? A quantitative assessment procedure. Journal of Environmental Quality. 48:412–420.

Ministry for the Environment and Stats NZ. 2023. New Zealand's Environmental Reporting Series: Our freshwater 2023. Wellington: Ministry for the Environment and Stats NZ.

Morton R, Henderson BL. 2008. Estimation of nonlinear trends in water quality: an improved approach using generalized additive models. Water Resource Research. 44(7): W07420.

R Core Team. 2021. R: a language and environment for statistical computing. version 4.3.0 [accessed 3 July 2023]. https://www.R-project.org

Snelder T, Fraser C. 2018. Aggregating trend data for environmental reporting. Lyttelton: LWP Ltd. LWP Client Report 2018-01. Prepared for Ministry for the Environment.

Snelder T, Fraser C, Larned S, Whitehead A. 2021. Guidance for the analysis of temporal trends in environmental data. Hamilton: National Institute of Water and Atmospheric Research. NIWA Client Report 2021017WN. Prepared for Horizons Regional Council and MBIE Envirolink.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. 2019. Welcome to the tidyverse. Journal of Open Source Software. 4(43):1686.

Yu K, Lu Z, Stander J. 2003. Quantile regression: applications and current research areas. Journal of the Royal Statistical Society. 52(3):331–350.