CAWTHRON

REPORT NO. 3719

# INVESTIGATING A METHOD OF NONPARAMETRIC CHANGEPOINT ANALYSIS OF WATER QUALITY

# INVESTIGATING A METHOD OF NONPARAMETRIC CHANGEPOINT ANALYSIS OF WATER QUALITY

ERIC GOODWIN, CRAIG DEPREE

Prepared for DairyNZ Ltd

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:
Joanne Clapcott

ISSUE DATE: 17 November 2021

RECOMMENDED CITATION: Goodwin E, Depree C 2021. Investigating a method of nonparametric changepoint analysis of water quality Prepared for DairyNZ. Cawthron Report No. 3719. 12 p.

# 1. ASSESSMENT OF LONG-TERM CONSISTENCY, EVALUATION OF POTENTIAL CHANGE

## 1.1. Preamble

River water quality is a dynamic aspect of the natural world, reflecting the ebbs and flows of climate, geology and biological succession and activity. It fluctuates at various time scales, with stable periods and periodic disturbances. On top of this background of constant change are anthropogenic effects, the impacts of residential, agricultural, commercial, or industrial developments made by society in the environment, or improvements brought about through mitigating initiatives.

In the management and monitoring of society's activities in the environment, it can be difficult to isolate the specific changes due to these activities from the fluctuations caused by factors that are 'external' to anthropogenic land use pressure.

The NPCP (**N**on**P**arametric **C**hange**P**oint) webapp aims to provide a means to evaluate recent measures of water quality against an established baseline. Specifically, it investigates impact by comparing measures taken after an event or during an activity, against measures before that event or activity. The comparison of the 'before period' and the 'after period' is intended to explicitly acknowledge existing natural variability, allowing for this fluctuation in two ways. The app allows the user to evaluate whether there has been relevant change in the central tendency (median value) of measures of an attribute before and after a chosen date.

First, variability in water quality measurements is smoothed by calculating 5-year medians of monthly samples. Second, it is then acknowledged that variability will remain in these 5-year medians, and the range of this variability defines a 'compliance interval'. Five-year medians of monthly measurements from the period after the event or during the activity are evaluated for their appearance within this interval, or outside it.

There are two ways the compliance interval can be generated, one with reference only to the current site's data, the other based on measures of the same attribute, but from all available sites nationwide. Furthermore, when assessing the site using only site-specific statistics, it is possible to restrict the historic data used to characterise the before period. These options and alternatives are described in sections below.

The NPCP webapp described in this report is accessible at http://www.goodwin.shinyapps.io/NPCPapp. It was co-developed by Craig Depree (DairyNZ) and Eric Goodwin (Cawthron Institute), programmed by Eric Goodwin, and funded by DairyNZ. Data used in the app were downloaded from the Environmental Monitoring and Reporting Project's Land, Air, Water Aotearoa website accessible at www.lawa.org.nz. Data are covered by 'CC BY 4.0' and by terms of use listed on the

LAWA website. This app and the use of the LAWA dataset are intended as a means of evaluating the suitability of the embodied assessment methodology, which is described below. This methodology is intended to explicitly acknowledge the natural variability in monitoring timeseries datasets, when assessing for ongoing maintenance or change (i.e. improvement or deterioration) against a baseline. As a means of assessing the suitability of the embodied methodology, users of the current app are anticipated to be scientists or researchers with familiarity with both water quality management, and statistical or data science methods. A future app could be retargeted at a more general audience to include farmers, landowners and other stake holders.

## 1.2. Site and attribute selection

Operation of the app consists of first selecting a monitoring site from those displayed on the national map. The user can zoom in or out and pan to identify the site of choice. Hovering over the point will display the site's name. Once a site is selected, the drop-down below the map will list the attributes that have been measured at that site, which can be analysed for long-term consistency.

One of the attributes is selected by default, and its data will be displayed on a timeseries scatter plot (Figure 1). The user can switch attributes from the drop-down list. They are sorted in order of data abundance. Switching sites retains the attribute selection established for the previous site. When the site selection is changed, or a different attribute selected, the timeseries of measurements is replotted.

The timeseries is reduced to monthly representative (median) values where necessary (when there are multiple samples per month), and it may have gaps if monitoring was not consistent across time. Two derived statistics are plotted as lines over the monthly points: a 1-year moving median and a 5-year moving median. These are calculated for every month in the series, and reflect the median of the previous 12 or 60 monthly samples, respectively. Where gaps exist in the monthly samples, the gaps would be spanned by the median lines, and medians may be based on fewer than 12 or 60 points. Because there will never be enough data points at the start of a timeseries, the 1-year and 5-year median lines are not plotted until 1 year or 5 years after the first data point.

The 1-year median exhibits less range (less variation in magnitude) than the monthly measurements, and the 5-year median exhibits less range again, as is consistent with the central limit theorem[1]. The user can toggle the y axis to be log-scaled or naturally scaled. This setting persists through changes of attribute or site.

[1] Sums or averages derived from samples of independent variables exhibit less variation than the raw variables.

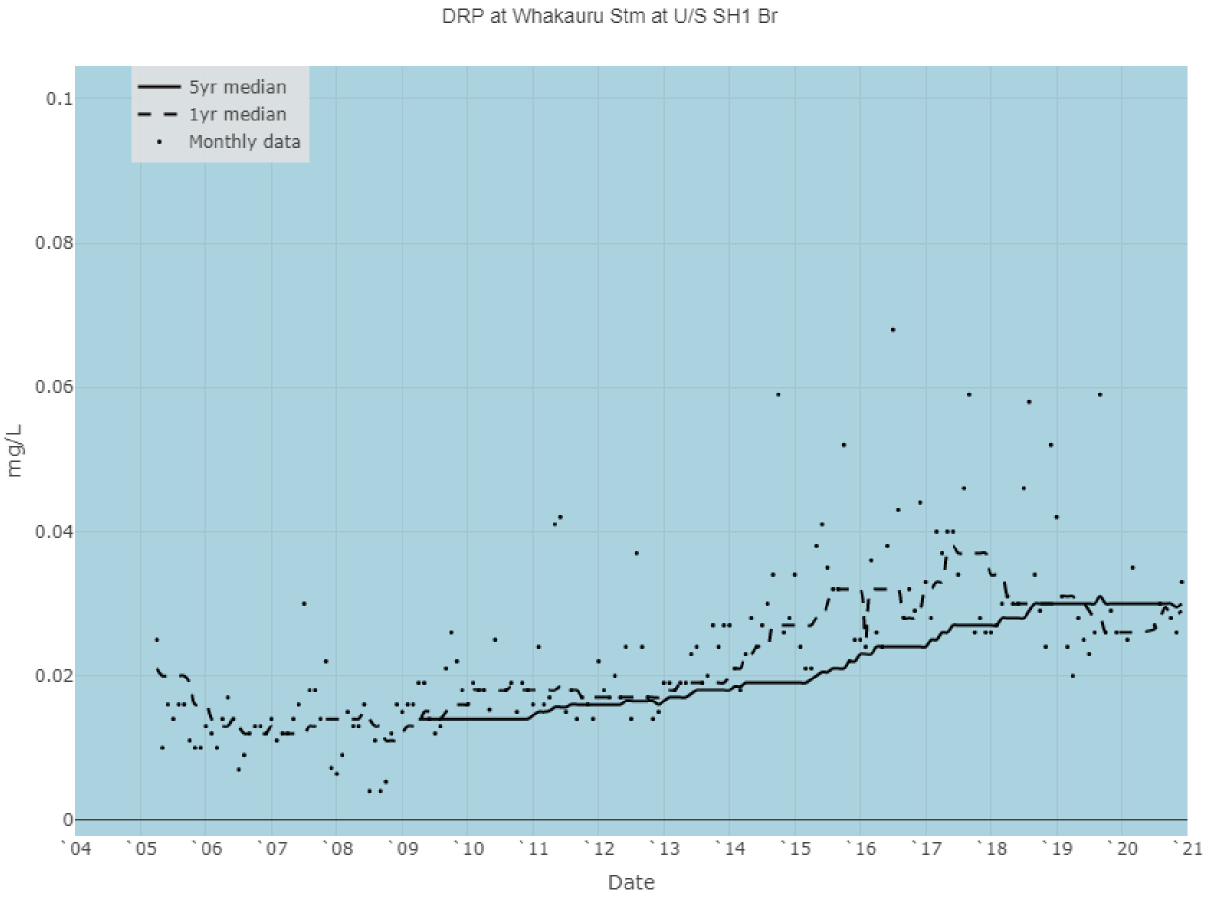

Figure 1. Monthly DRP plotted against time (points), overlaid with 1-year and 5-year medians (dashed and solid lines, respectively).

### *1.2.1. Censored data*

The data available on any site may include 'censored' data. These are data reported as being less than a certain value, such as '< 0.05 mg/L', which can occur when a nutrient (for instance) is present at levels below the detection sensitivity of the instrument or laboratory protocol for measuring it. Where laboratory methods are refined over time and become more sensitive, the detection level may drop, such as to '< 0.01 mg/L'. If these censored values were taken at face value (e.g. 0.05 and 0.01), the improved sensitivity of the laboratory method could be interpreted as a decreasing trend over time. This would however be an artefact of the measurement method, and not necessarily a real-world decrease. To avoid this artefact, any values less than or equal to the highest censored value are treated as equal, and set to half the highest censored value. This option can be switched off (and for the Visual Clarity attribute is switched off by default) by the HiCensor toggle in the user interface.

## 1.3. Comparison before and after a user-defined assessment date

A date of a known historic or current plan change or development in the catchment is selected by clicking or moving a slider below the timeseries plot. A vertical solid line is added to the plot to show the boundary between before and after periods, and tests then occur instantaneously, with results displayed on the plot as well as in a table below. Positioning the slider (and consequently the time boundary) at a year (e.g. 2018) means all data to the end of that year (e.g. to December 2018) are used to define and characterise the before period.

Where there are sufficient data in the before period, an interval is constructed from site-specific statistics of this before period. It is centred around the median of 5-year medians, and has width typical of the site's historic variability. Where there is insufficient data (fewer than 72 monthly measures yielding two 5-year medians) in the before period (or the user selects to base the test on generic characteristics of the attribute at all sites) the interval is still centred around the site-specific median, but it is the median of monthly data, and the width of the interval is determined by site-generic attribute variability.

The after period is then evaluated against whichever interval was constructed.

Results added to the plot at this point (Figure 2) include points at the 5-year medians and horizontal green lines defining an interval derived from the medians before the selected date. Median points are then colour-coded to indicate whether they were consistent with or atypical of the range of conditions seen in the before period. Monthly measurements involved in the characterisation of the before period are coloured blue, monthly measurements within five years after the selected date are coloured salmon, and where monthly measurements are beyond five years after the selected date, they are coloured cyan (e.g. see Figure 3). Early points excluded from characterisation of the before period, for reasons discussed below, are coloured grey. Legend entries are added as these new features are added to the plot.

Selection of a different attribute or site will clear these results from the plot. Selection of a different date will re-analyse the current dataset, with newly-defined before and after periods.

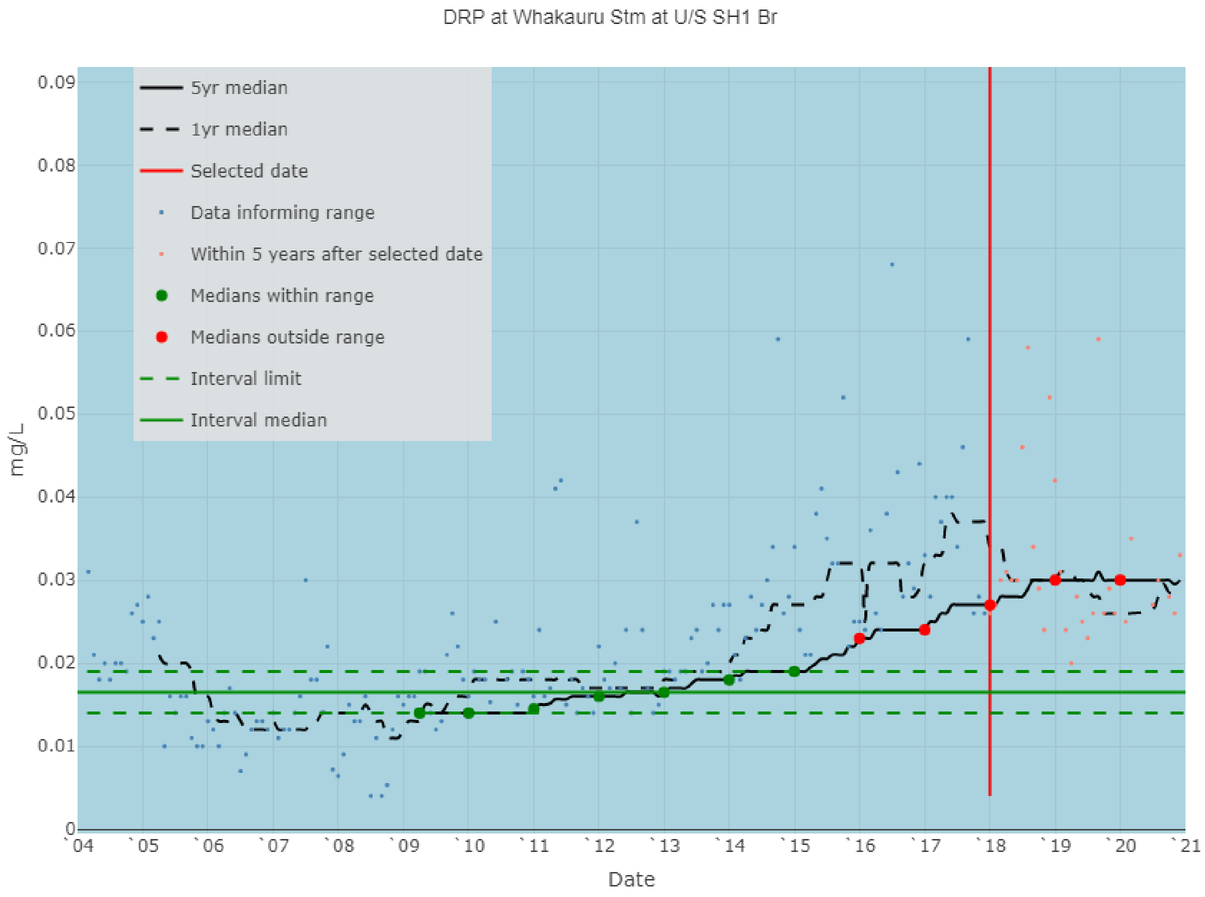

Figure 2. Here the user has selected a date in 2018 (red vertical line). Yearly 5-year medians characterise the before period. The median of these medians is bisected horizontally by a solid green line, flanked by dashed lines at a distance equal to their median difference from the median (the MAD). Red points indicate 5-year medians higher than the upper limit of this interval.

## 1.4. Characterising the before period, establishment of a compliance interval

The before period is characterised by the median of its annual 5-year medians (M5M), and the variation in these 5-year medians. The annual 5-year medians are derived by a sliding window starting from the user-selected date and moving backward in time. From these statistics a compliance interval is constructed, to be used to assess the after period's 5-year medians. This compliance interval defines a range about the M5M based on a user-defined percentile compliance statistic. If an after 5-year median is within the compliance interval, then this would indicate that the after period water quality state is consistent with being maintained. Conversely, if the after period 5-year median is outside the compliance interval, then this may indicate a change in water quality.

Under a null hypothesis of no change in the measures before and after the assessment date, a similar proportion of 5-year medians of data before and after, would be expected to fall within this interval. The central limit theorem implies that a change in the variability in monthly data, but with no change in its average value, would not be detectable, by this method.

It should be noted that even within a consistent and stationary before period, it is statistically predictable that 50 per cent of 5-year medians appear outside an interval constructed with the default width (25% below the interval, 50% within it, and 25% above it), and so the appearance of a similar proportion in the after period does not *ipso facto* identify a worsening state. Suspicion of degrading water quality requires a greater proportion of observations outside the interval in the after period, than in the before period.

Control chart theory includes a set of rules (e.g. Nelson Rules[2]) for identifying deteriorating sets of results relative to a compliance interval. These have not been implemented in the current development, but a binomial test is run based on the proportion of after period 5-year medians that are non-compliant with the interval. This test gives a p value, being the probability of observing this number of (or more) non-compliant medians if there had been no change in water quality. Note that a key assumption of the binomial test, that the set of trials (in this case, the set of 5-year medians) is independent of one another, is not met in this context. Not only are timeseries data often serially autocorrelated, but this is then exacerbated by the derivation of sliding-window (overlapping) 5-year medians.

## 1.5. Site-specific interval construction

When the before period is long enough (greater than 72 monthly values, yielding greater than two 5-year medians), it is characterised by calculating moving-window 5-year medians at one-year spacings, of the monthly measurements taken before the assessment date. The resulting yearly-spaced 5-year median values are based on data sets that overlap; they share 80% of their contributing data (i.e. 4 out of 5 years' data) with neighbouring medians. These yearly 5-year medians are added to the plot (Figure 3), colour coded to indicate whether they are typical or atypical of the interval described below.

The set of 5-year medians itself has a median value (M5M) as well as variability. This variability is characterised by a statistic called the 'Percentile Absolute Deviation' (PAD), as a variation on the established Median Absolute Deviation (MAD).

[2] https://en.wikipedia.org/wiki/Nelson_rules

The MAD is defined as the median of the absolute value of the differences between a sample's individual values and the median of all the sample's values.

$$MAD = median(|x_i - median(x)|)$$

By extension, we define the PAD as a (user-specified) percentile of the same set of absolute differences, between a sample's values and the sample's median. Thus the $\mathrm{PAD}_{50}$ (the $50^{th}$ percentile of these absolute differences), would be equal to the MAD.

The interval is constructed by the median of the 5-year medians (M5M) ± the PAD of the 5-year medians:

$$CI = M5M \ \pm PAD$$

The PAD can be calculated for any percentile value chosen through the app interface which imposes limits of the 50th and 80th percentile.

The M5M and the interval extremes are added to the plot as horizontal green solid and dashed lines, respectively.

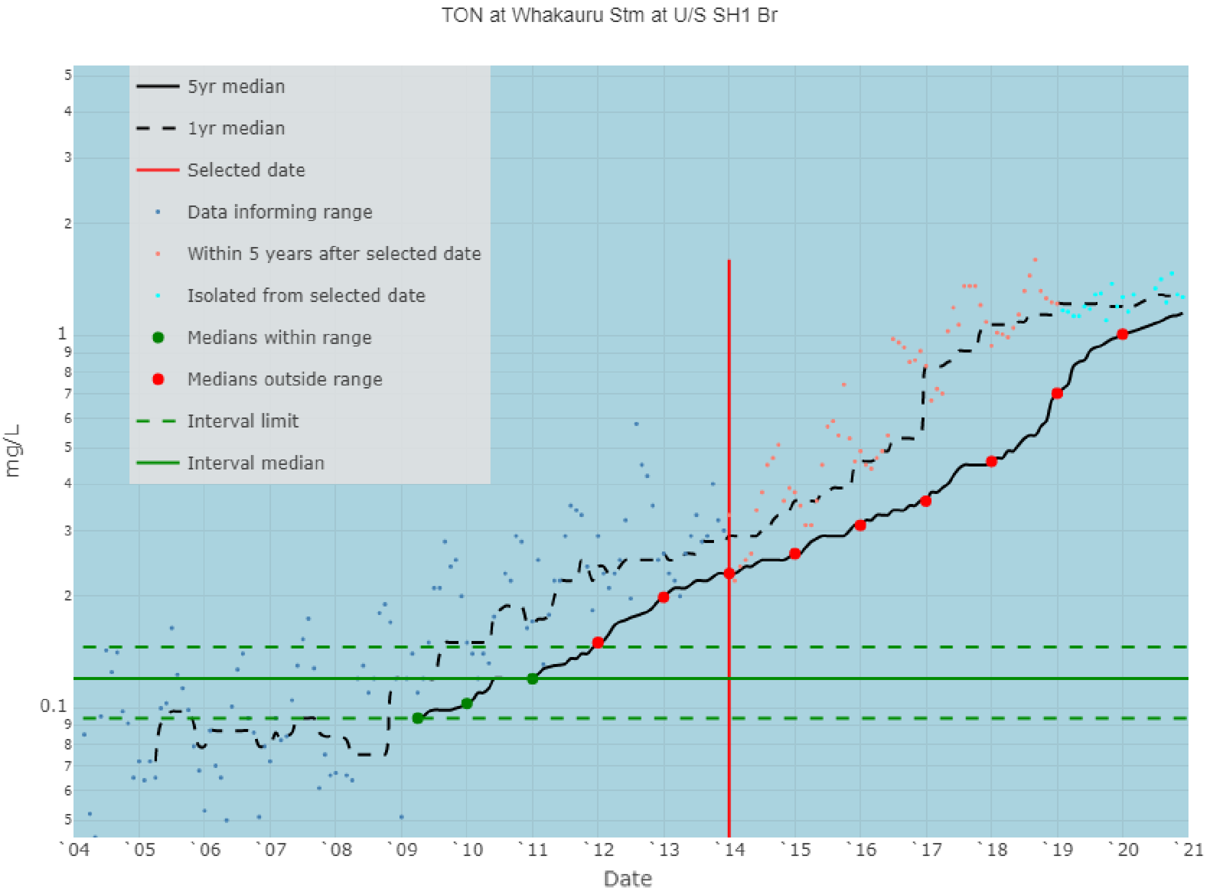

Figure 3. 5-year medians are dots colour coded green or red, according to their inclusion within the interval (shown as green horizontal lines). In this figure the y axis has log-scaling.

## 1.6. Site-generic interval construction

If the before period is too short (if it yields fewer than two 5-year medians, which happens if there is less than 7 years of before-period sampling), or if the user has selected the site-generic evaluation option, the compliance interval is constructed differently. It is centred around the *median of available monthly measures* for the selected site in its before period, and its width is informed by pre-calculated variability statistics drawn from sites nationwide. An example is shown in Figure 4.

Sites with fewer than 60 monthly measurements were excluded from this variability-characterising set. Other inclusion/exclusion criteria could be implemented. Pre-calculation involved deriving the set of 5-year medians for the full available history of each site (no division into before/after periods), and then determining the median and variability of each site's set of 5-year medians. The variability was defined by the MAD (note the limitation that no PAD can be specified under this option), divided by the median, making it a relative MAD (rMAD).

$$rMAD = \frac{MAD}{median(x)}$$

$$rMAD = \frac{\left(median(|x_i - median(x)|)\right)}{median(x)}$$

This gave a set of rMAD values for each site, for each water quality attribute. Median values of the rMAD values for each attribute are shown in Table 1.

Table 1. Attribute-specific statistics drawn from all sites with greater than 60 monthly measurements.

| Attribute | Number of sites | Median of sites' M5M | Median of sites' MAD5 | Median of sites' rMAD5 (%) |
|---|---|---|---|---|
| NH4 (mg/L) | 966 | 0.0072 | 0 | 0 |
| ECOLI (/100 mL) | 962 | 130 | 17.5 | 14.5 |
| DRP (mg/L) | 953 | 0.011 | 0.00100 | 6.67 |
| TON (mg/L) | 921 | 0.27 | 0.0215 | 8.5 |
| TN (mg/L) | 911 | 0.5 | 0.0288 | 5.56 |
| TP (mg/L) | 910 | 0.0269 | 0.0015 | 6.25 |
| TURB (NTU) | 909 | 2.5 | 0.267 | 11.5 |
| BDISC (m) | 778 | 1.5 | 0.101 | 7.36 |
| NO3N (mg/L) | 709 | 0.249 | 0.0225 | 9.52 |
| DIN (mg/L) | 632 | 0.340 | 0.0285 | 8.91 |

In the app, the rMAD value for the selected attribute is multiplied (as a proportion) by the selected site's median of historic monthly values, to set the width of the interval used for evaluating the 5-year medians of the after period:

$$Compliance\ Interval = median(x) \pm rMAD5 * median(x)$$

The median of monthly values, and the extremes of the interval, are added to the plot as horizontal green solid and dashed lines, respectively (Figure 4).

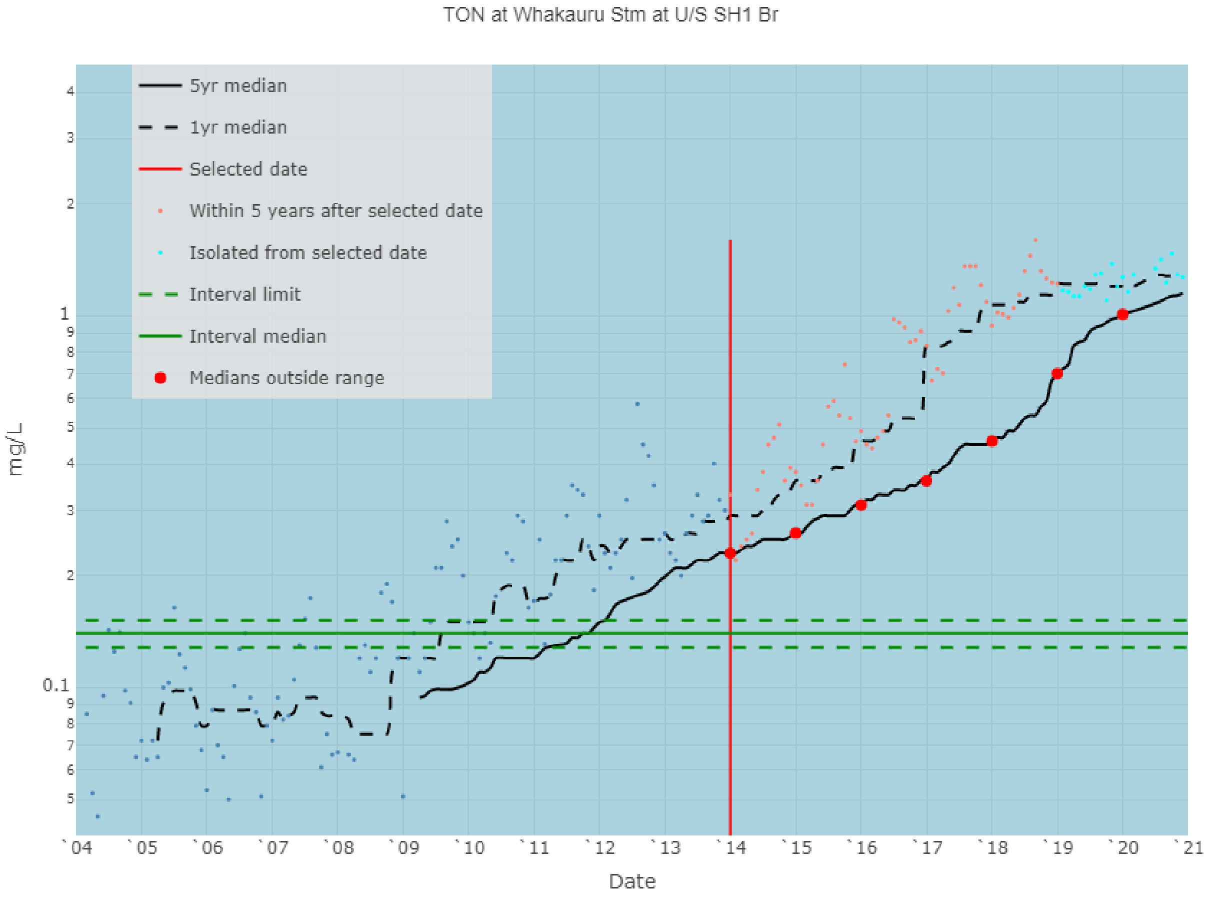

Figure 4. The site's historic medians are not plotted or colour coded when site-generic data (from all sites nationwide) have been used to characterise variability in the attribute.

## 1.7. Testing for historic shift

When using site-specific data to construct the interval, there is an option to restrict the data contributing to characterisation of the historic period. This can be toggled by the 'Test for historic shift' control.

The width of the interval used to test the after-period medians is determined by the range in the before-period medians, so the test would be inappropriately lenient if the before period featured an atypical excursion from baseline values. For instance, if a steady low baseline were interrupted by a short period of high values due to an acute environmental perturbation, the resulting high range exhibited by the medians would allow for ongoing high deviations from a more normal range. The occurrence of previous acute excursions from normal baseline may not justify ongoing excursions, so the user has the option, based on their familiarity with the site, of excluding historic periods identified as statistically significantly different from the most recent state.

To evaluate the consistency, or stationarity of the before period, it is divided into 5-year epochs, starting from the assessment date, and stepping back to the start of the available data. Vertical dashed lines added to the plot indicate the epoch boundaries. Some 5-year epochs may have atypical variation in them, if they differ substantially from the characteristics of the most recent 5-year epoch.

To identify any 5-year epochs unsuitable for the characterisation of the historic before state, the *monthly* data of each historic 5-year epoch is compared against that of the most recent 5-year epoch before the assessment date. The resolution of this evaluation, between 5-year epochs, limits the time scale of fluctuation that can be detected as change within the historic period. Inter-epoch comparison is based on a two-sample Wilcoxon rank sum test (also known as a Mann-Whitney test). This is a non-parametric rank-based sample comparison of 'location'. That is, it assesses for a partitioning in overall ranking between two groups.

Under the assumption of no significant change in the time-series data up to the assessment date, these tests should be able to step back through the available epochs and find no significant differences, establishing that it is reasonable to use all available data before the assessment date to characterise the before period.

A significant Wilcoxon test result ($p < 0.01$, subject to all the criticisms of the use of p value thresholds to make decisions), identifying a shift between epochs, excludes the older epoch from the characterisation of the before period. An intermediate 5-year epoch may be excluded, with older 5-year epochs still contributing. Significantly different epochs, excluded from the characterising set, are indicated by the addition of magenta line segments, from the centre of the most recent epoch to the centre of each significantly different epoch (Figure 5). By toggling the option, the user could visually identify differing epochs, but include their data in the before/after assessment if they choose.

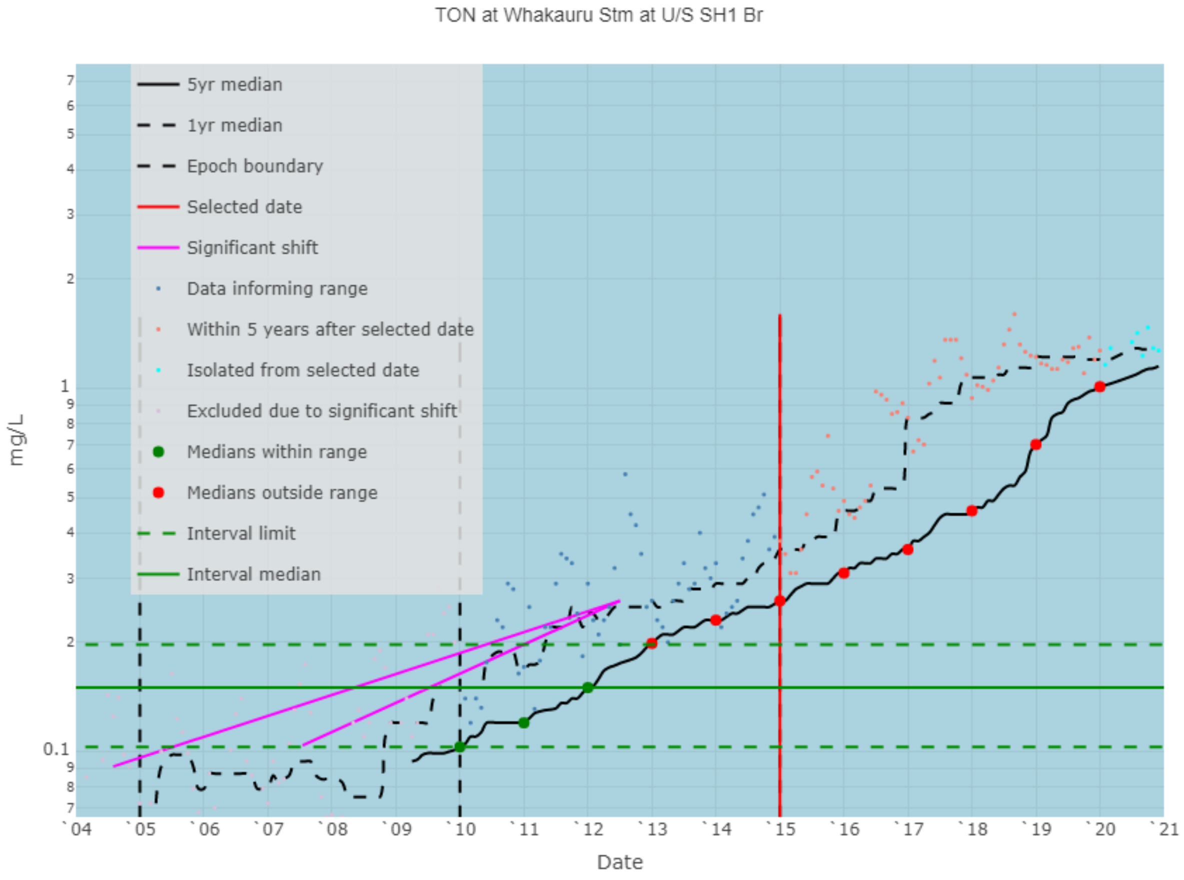

Figure 5. Testing for historic shift reveals that the two earliest epochs differ significantly ($p < 0.01$) from the most recent 5-year epoch. Their monthly data are coloured grey, and a magenta line is added to highlight the difference in epoch medians. The blue points used to characterise the before period are limited to only the most recent epoch.

The 5-year median values from the representative epochs (or all epochs, if ‘test for historic shift’ is not selected) are used to characterise the before period. Note that early 5-year medians in each epoch will be influenced by monthly values from the previous epoch, which may have been identified as significantly different. The 5-year medians contributing to the characterisation are plotted and colour coded.

## 1.8. Assessment of the after period

The assessment of the after period is simpler than the characterisation of the before period, in that no epoch-based comparison is made. Annual 5-year medians of monthly data after the assessment date are assessed for inclusion within the (site-specific (PAD, Section 1.5) or site-generic (MAD, Section 1.6)) interval, or deviation beyond it. Depending on the attribute being inspected, medians either higher or lower than the interval would represent non-compliance.

With the exception of visual clarity, 5-year medians of the after period that plot above the interval indicate potential non-compliance. For visual clarity, 5-year medians that plot below the interval indicate potential non-compliance. It is expected that 25% of 5-year medians would appear above a MAD-based interval, and ½*(100-p)% of 5-year medians would appear above a pth percentile PAD-based interval, e.g. ½*(100-80) = 10% of values above a $\text{PAD}_{80}$-based interval.

The probability of the number of observed non-compliant 5-year medians is calculated using a binomial test. The binomial test is characterised by a fixed number of trials (here the 5-year medians) with constant probability of passing or failing a test applied to each (here, non-compliance with the interval). The requirement of the binomial test that each trial is independent of others is not met in this context, as time-series data often exhibit serial autocorrelation, which is exacerbated by the calculation of sliding-window 5-year medians. The p value reported should therefore be treated with some caution, and not relied on exclusively as evidence of changing water quality.

Although the after period is not subdivided into epochs, a distinction is sometimes shown within this set. Because the evaluation is made on 5-year medians, data within five years of the selected change point are coloured to indicate that 5-year medians in this period will be influenced by conditions prior to the change.

The number of 5-year medians used to characterise the before period is displayed below the plot. This number will be lower, if the ‘Test for historic shift’ is active and has excluded any epochs. The median of the before period's 5-year medians (or of monthly values) is stated, along with the upper and lower bounds, which were generated by median ± PAD of site-specific 5-year medians, or median ± MAD of site-generic 5-year medians.

When the historic period is too short to characterise the site from its own data, and the interval has been constructed from generic variability of all sites, this is reflected in the results output. Instead of specifying the number of 5-year medians informing historic characterisation, the output specifies the number of monthly datapoints informing the historic median. Instead of specifying the M5M, it specifies the median of the monthly data points. Instead of specifying the PAD of 5-year medians, it specifies the generic rMAD used for the selected attribute, and lays out the use of this value in construction of the interval width.

# 2. CAVEATS AND CONSIDERATIONS

The evaluation and analysis of timeseries data is a rich topic, with a number of existing methodologies and established approaches. The addition to the field proposed in the app described herein is customised to address a specific query. It has a number of remaining limitations or caveats, which we acknowledge.

Timeseries analysis often explicitly addresses the issue of serial autocorrelation. No explicit allowance has been made for the potential of this confounding phenomenon.

The analysis embodied in the app described herein relies on rolling 5-year medians. Apart from the semantic confusion of discussing the ensuing medians of medians, the use of a 5-year window has some consequences which should be discussed. It is important to note that the 5-year median plotted at time t on the figure, is influenced by monthly data from 5 years before time t. So the first 5-year median in the after period (and in fact the first four such 5-year medians) is influenced by monthly data collected during the before period. It is not until five years after the selected date that the 5-year medians reflect purely the conditions that existed after that date.

There are conditions that can cause false positives, and likely conditions that can cause false negatives in this test. For example, where a timeseries has been steadily degrading, or steadily improving during the historic time before the assessment date, this situation will yield a set of 5-year medians with high variability reflected in a high PAD value. This will then generate a wide compliance interval, and tolerate a large range of values in the evaluated period after the assessment date. It would not in this case identify when conditions are not being maintained, and nor would it identify that conditions were being improved or degraded. That is, existing high variability in the before period, compromises the discriminatory power of the test method in the app.

While there is the option to exclude 5-year period that exhibit significant shift relative to the most recent 5-year period, exercising this option should be based on knowledge of what caused the shift in values. If the significant difference was in fact part of natural variability during the before period, then it should be included in that characterizing data. If it was due to an acute environmental shock, then it could be removed. The test for historic shift is based on the p value from a statistical test, with an arbitrary threshold for exclusion of a historic epoch. The use of p values in binary tests is fraught with problems, and in this case the requirements for the statistical test are not fully met. It is expected that the user of the app will use the results as an indication and guidance only, and not rely on them exclusively. Likewise, the p value associated with the observed number of non-compliant 5-year medians in the after period, is based on a binomial statistical test, which relies on a degree of independence between the 5-year medians, that they do not exhibit.

The development of the app could be continued to refine or add functionality. It would be possible to add colour bands over the timeseries plots to show how data sat relative to national guidelines defined in the National Policy Statement for Freshwater Management, where these exist for the selected attribute. It would be possible to offer other assessment methodologies, evaluations of the same timeseries data sets, for instance to illustrate trend methodologies used in other contexts, or to explore the potential of alternatives. It would also be possible to add additional water quality parameters for display and analysis.

## 3. CITATIONS

Software was developed using RStudio and R (R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.), in a Shiny architecture [Winston Chang, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert and Barbara Borges (2021). shiny: web application framework for R. R package version 1.7.1. https://CRAN.R-project.org/package=shiny], using the Leaflet package [Joe Cheng, Bhaskar Karambelkar and Yihui Xie (2021). leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.0.4.1. https://CRAN.R-project.org/package=leaflet] for the map component and the plotly package [C. Sievert (2020) Interactive web-based data visualization with R, plotly, and shiny. Chapman and Hall/CRC Florida] for the interactive data plot.

Data were acquired from Land Air Water Aotearoa (www.lawa.org.nz), licensed under CC BY 4.0.

This project was funded by DairyNZ.

# A1. CONSISTENCY OF VARIABILITY

During pre-calculation of the all-site relative MAD values for each parameter, we investigated the appropriateness of using a single representative value for all sites, and looked for consistent differences between sites in different land cover, as a potential source of systematic effects.

Figure A1.1 shows a comparison of relative MAD values between three landcover types, for 11 water quality parameters. There is little difference in MAD values between landcover types for most water quality parameters. For TN, TP and DRP, native forest appears to have lower relative MAD, but this difference has not been statistically tested. For most parameters, urban sites appear to have higher MAD, except for DIN, NO3N and TON. Again, the statistical significance of these observations has not been tested.

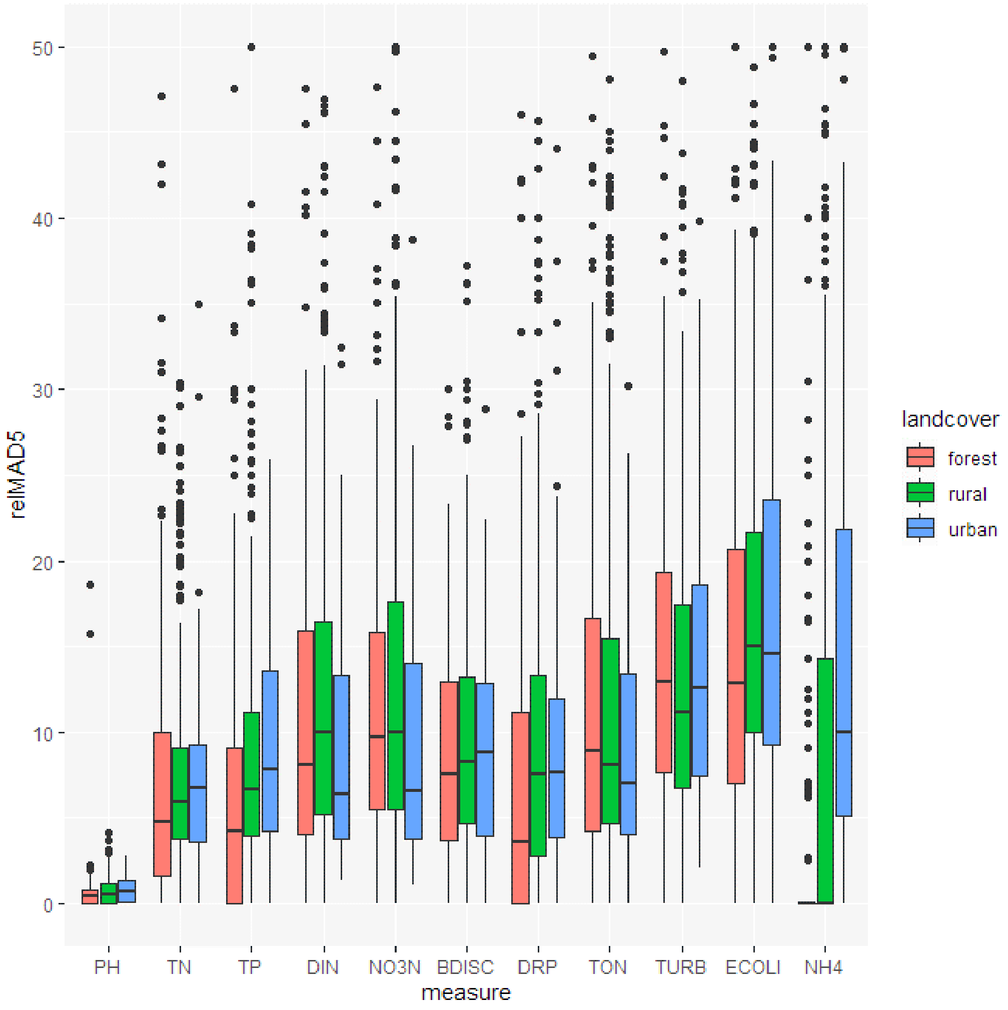

Figure A1.1. Distribution of relative MAD values for 11 water quality parameters, compared among three landcover types.