LWP
landwaterpeople

Monitoring change over time

Interpreting water quality trend assessments

March 2019

Prepared By:

Ton Snelder

Caroline Fraser

For any information regarding this report please contact:

Ton Snelder

Phone: 03 377 3755
Email: ton@lwp.nz

LWP Ltd
PO Box 70
Lyttelton 8092
New Zealand

LWP Client Report Number: 2019-01

Report Date: March 2019

LWP Project: LWP Project 2019-01

Quality Assurance Statement

Version	Reviewed By	
1	Scott Larned (NIWA)	

Table of Contents

Executive Summary

This report provides supplementary information that assists with the interpretation of the recently completed assessment of trends in the water quality of New Zealand's lakes and rivers (Larned et al., 2018a; Larned et al., 2018b). Although based on the assessments and datasets from these studies, the results and conclusions in the current report are more broadly applicable to the evaluation and interpretation of water quality trends.

Using river water quality data and results of trend assessments produced by Larned et al. (2018a) we undertook analyses in three steps:

1. assessment of the factors that are associated with variation in the confidence of trend evaluations,
2. assessment of variation in trends with time-period length (e.g., a ten-year trend period from 2008 to 2017 compared to a five-year period from 2013 to 2017), and
3. assessment of variation in trends with time-period window (e.g., a ten-year trend period window from 2007 to 2016 compared to the window from 2008 to 2017).

Analyses undertaken at step one established that variation in the confidence of trend evaluations was linked to a variety of factors but most strongly linked to trend magnitude itself. Confidence (or precision) in estimated trend magnitude was negatively related to trend magnitude (described by the relative Sen slope) but confidence in trend direction (as indicated by Kendall test *p*-value) was positively related to trend magnitude. The trend period length was also moderately negatively correlated the with confidence in predicted trend direction, while the precision of the data (proportion of unique observations) for DRP, TP and TN were moderately positively correlated to the confidence in predicted trend direction. Based on these observations, we were able to approximately define the minimum detectable trend magnitude at a given level of confidence for each water quality variable, for a given time-period length. In general, as the time-period length increases, smaller trends can be detected with confidence.

The second set of analyses established that trend magnitudes tend to decrease with increasing time-period length. This indicates that trend magnitudes estimated for shorter periods are unlikely to persist over longer periods. In addition, the results indicate the need to be cautious when comparing trends evaluated over time periods of differing length.

The third set of analyses established that trends are sensitive to the time-period window. This means that, for different time-period windows, there can be large variation in both the magnitude of trends at individual sites and the proportion of sites with improving trends. In addition, our analyses showed a strong association between temporal variation in climate conditions and both individual water quality observations and fluctuations in trends between time-period windows. For example, for most water quality variables, the southern oscillation index (SOI) explained in the order of 10%, 30% and 80% of the variation in the proportion of improving site trends for time-period lengths of 5, 10 and 20-years respectively.

Our analyses were based on flow adjusted water quality data, so we assume the observed patterns are not simply due to the influence of climate on river flows. Furthermore, we showed that water quality observations at baseline (i.e., minimally impacted by human activities) and impacted sites were similarly associated with the monthly value of the SOI as at impact sites. This provides evidence for the association of patterns observed with aspects of natural climate variability (e.g., temperature and hydrological regimes) rather than with the influence of changing land management practices in association with climate variation.

The results of our study do not mean that climate is solely responsible for long term changes in water quality. However, our results indicate that climate exerts considerable influence over water quality at inter-annual time scales. These findings are consistent with one previously published New Zealand study and a limited number of studies in other countries.

Trend assessments provide no information about the causes of water quality trends. However, there is always an interest in attributing the observed water quality changes to drivers of water quality such as land use and management. Attributing trends to causes helps to understand the efficacy of management actions and contributes to the feedback part of the policy cycle. This study has shown that in order to understand anthropogenic effects on water quality, it is necessary to control for the confounding effects of climate.

Our study suggests two relatively simple statistical treatments of trend data may improve the quality of inferences drawn from trend analyses. First, climate indices such as the SOI could be treated as a covariate in trend analysis of individual sites in much the same way that flow adjustment is performed. Second, large scale studies seeking to explain variation in trends between many sites could include the influence of the SOI, as defined by the correlation of the monthly observations with the monthly value of the SOI, as an explanatory variable. However, the SOI measures one aspect of climate variation and it may be that more appropriate or additional measures should be used. We therefore recommend that further research on the role of climate on water quality and practical methods to account for climate variation's effect on water quality measurements is undertaken.

1 Introduction

This report is part of a larger project that analysed state and trends in the water quality of New Zealand's lakes and rivers (Larned *et al*., 2018a; Larned *et al*., 2018b). In the larger project, river water quality trends were assessed for sites that are monitored as part of the State of Environment (SOE) programmes operated by Regional Councils and unitary authorities and the National River Water Quality Network (NRWQN) operated by NIWA. Trends were analysed for time periods of differing lengths, all of which finished in 2017 (Larned *et al.*, 2018). The present study aimed to provide supplementary information to assist with the interpretation of the river water quality trend assessments. In particular, the study aimed to assist interpreting differences in trend assessments pertaining to different time-period lengths (e.g., a ten-year trend period from 2007 to 2016) and time-period windows (e.g., a ten-year trend period window from 2007 to 2016 compared to the window from 2008 to 2017).

There were three key objectives set out in the study scope. First, the study was to characterize the effects of data variability on confidence in assessment of temporal trends in water quality. Second, the study was to evaluate the effects of trend period window on trend direction, magnitude and confidence. The third objective was also to provide guidance on the use of the 'indeterminate trend' category when reporting and modelling water quality trends.

The third objective was intended to provide a rationale for accounting for statistical confidence in trend evaluations both on a case by case basis and when aggregating trends over groups of sites to represent the general changes on water quality over a domain of interest. However, subsequent to the commencement of the present study, McBride (2019) showed how a statistic describing the level of confidence in a trend could be produced and provided guidance on describing this confidence. In addition, Snelder and Fraser (2018) showed how the statistics describing the level of confidence in individual trends could be aggregated to produce a statistic describing the proportion of improving trends (PIT) over a domain of interest. The techniques described by these two studies have provided a basis for maximising the information available from indeterminate trends and obviated the need to undertake work to achieve the third objective in the present study. This report does not address the third objective set out in the original scope and refers the reader to McBride (2019) and Snelder and Fraser (2018) for guidance on methods for accounting for statistical confidence in trend evaluations. It is noted that the analyses of lake and river water quality trends Larned *et al.* (2018) and Larned *et al.* (2018) and the present study have made extensive use of the methods set out by McBride (2019) and Snelder and Fraser (2018).

This report aimed to address the first and second objectives of the original scope. To do this we undertook analyses in three steps:

1. assessment of the factors that are associated with variation in the confidence of trend evaluations,
2. assessment of variation in trends with time-period length, and
3. assessment of variation in trends with time-period window.

The statistical confidence of trend analyses is affected by both the inherent variability in water quality at monitoring sites and sample size. Because water quality monitoring is conducted at a fixed time interval (e.g., monthly), sample sizes for sites in state-of-environment monitoring networks are a function of the time-period length. Therefore, the ability to confidently detect a trend of a given magnitude is a function of the data variability and the time-period length. At step one we therefore investigated which aspects of water quality data were most strongly associated with the level of confidence achieved by trend analyses. We used the results of this analyses to guide an evaluation of the minimum trend magnitude that can be detected with a specified level of confidence.

Trends are always specific to a time-period length and a time period window. It is generally observed that as time-period length increases the absolute magnitude of the trends decreases. Further, for any time-period length, there are large changes in the direction and magnitude of individual site trends pertaining to different trend period windows. The aggregate behaviour of the individual site trends across many sites can result in significantly different proportions of degrading or improving sites between time-period windows.

Although a trend assessment provides no information about the causes of the detected trends, there is always an interest in attributing the observed water quality changes to drivers such as land use. Attributing trends to causes is of particular interest in the context of managing to limits because this helps to understand the efficacy of management actions and contributes to the feedback part of the policy cycle. Differences in site and aggregate trends between time-period windows (for example, 10-year trends ending in 2013 compared to ending in 2017) invites speculative suggestions of what has caused these changes. However, water quality trends are frequently detected with high levels of confidence at sites that are unimpacted by human activities. This suggests that differences in results obtained for different trend period windows are at least partly associated with natural processes. A previous study by Scarsbrook *et al.* (2003) showed that variability in trends between time periods was associated with the El Niño Southern Oscillation (commonly called ENSO). Therefore, another important consideration in attributing trends to causes is the degree to which water quality variation is associated with these natural cycles. At step three we quantified the variability in trends of the same time-period length, but for different time-period windows and examined the association between this variability and ENSO.

2 Data

2.1 2017 trends dataset

We used the river water quality input dataset and analysis outputs of Larned *et al.* (2018), including both the raw water quality observation data and the flow adjusted water quality observations. The water quality observation data set was sourced from regional councils, LAWA and NIWA, as described by Larned *et al.* (2018) and represents the most up to date available national water quality dataset.

Larned *et al.* (2018) evaluated site trends for three time periods, 10, 20 and 28 year, all of which ended in December 2017. Trends were evaluated for all combinations of site, variable and time period for which the minimum data requirements were met (see Larned *et al.* (2018) for details). In this report, we refer to this set of trends as the “2017 trends dataset”.

Table 1 provides a summary of the water quality variables used in this report, as well as the number of trends that were calculated for each time period by Larned *et al.*, (2018). Figure 1 shows a map of the sites used in this study, by variable, and distinguished by the maximum time-period length for which trend analysis was performed.

Table 1. River water quality variables included in this study.

				Number of site trends		
Variable type	**Variable**	**Abbreviation**	**Units**	**10 year**	**20 year**	**28 year**
Physical	Visual clarity	CLAR	m	457	230	78
	Turbidity	TURB	NTU	718	79	77
Chemical	Ammoniacal nitrogen	NH4N	mg/m^3	731	298	106
	Nitrate nitrogen	NO3N	mg/m^3	749	309	112
	Total nitrogen (unfiltered)	TN	mg/m^3	660	162	83
	Dissolved reactive phosphorus	DRP	mg/m^3	750	328	122
	Total phosphorus (unfiltered)	TP	mg/m^3	663	307	110
Microbiological	*Escherichia coli*	ECOLI	cfu/100 mL	748	152	16
Macroinvertebrate	Macroinvertebrate Community Index	MCI	unitless	575	334	72

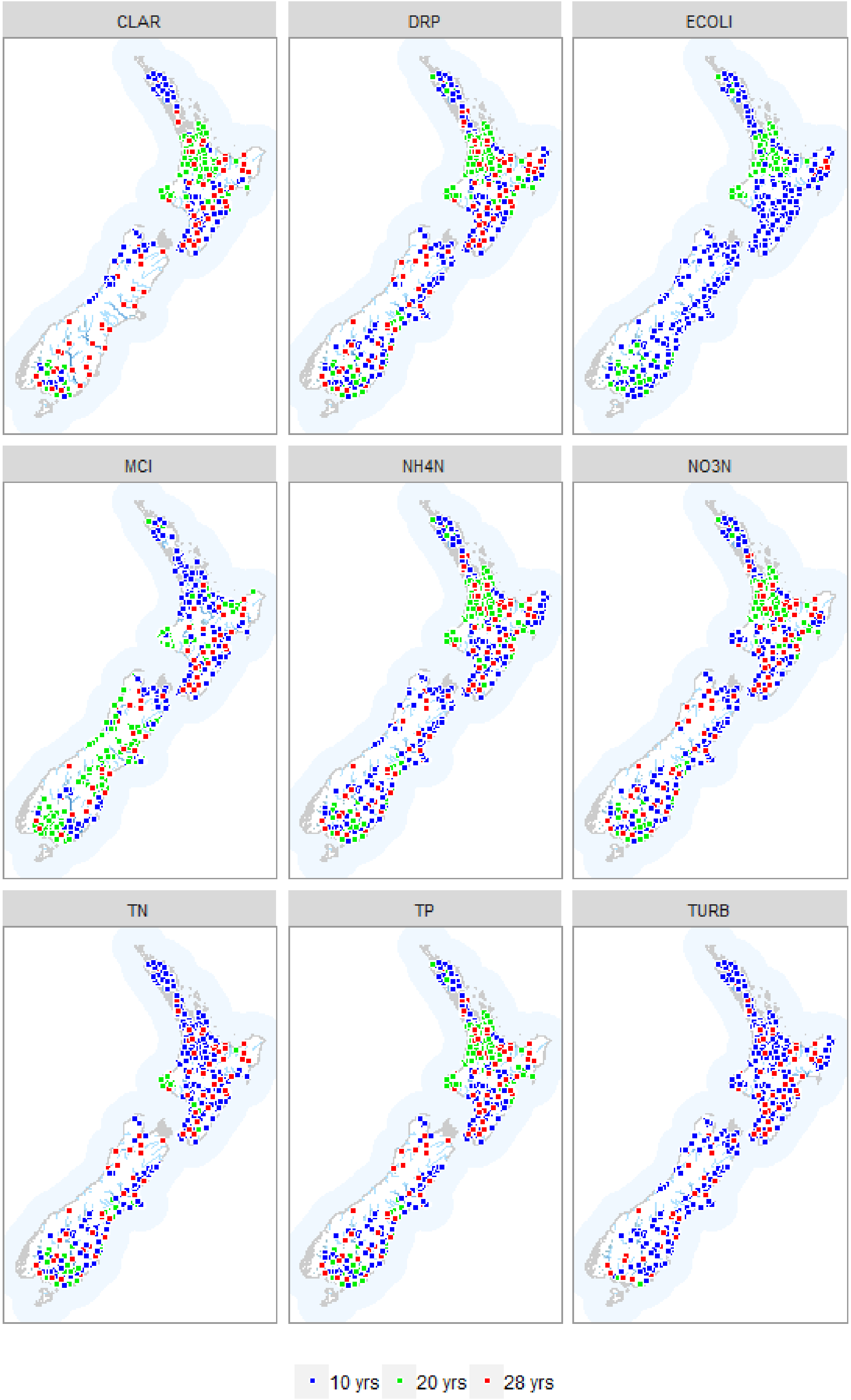

Figure 1. River water quality monitoring sites included in the 2017 trends dataset analyses. Colours indicate the longest trend period analysed at that site.

2.2 NRWQN rolling trends dataset

The 2017 trends dataset maximises the number of sites nationally for which there are ‘up to date’ trend assessments (i.e., trends ending in 2017). However, the 2017 trends dataset had some inconsistencies that may have influenced our analyses. In particular:

(1) the total number of sites represented in each time period decreases as the time-period length increases (due to increasing numbers of monitoring sites over time), resulting in differing sample sizes for the different time periods; and

(2) the time periods analysed represent specific time-period windows, but one of our study aims was to examine the general effect of time-period window on trends.

To overcome these inconsistencies, we conducted some analyses associated with step 2 of this study and all of the analyses associated with step 3 using only the NRWQN sites (Figure 2). The NRWQN sites have consistent record of monthly sampling for 28 years 1990 – 2017 for most of the water quality variables shown in Table 1. The NRWQN data have few missing values except for TN and NH4N in 1994. In addition, monitoring of ECOLI at NRWQN sites only started in 2004. NRWQN sites are categorised as “baseline” and “impact” (Smith and Maasdam, 1994; Figure 2).

For the NRWQN sites, we evaluated trends using rolling analysis periods of 5, 10 and 20 years. (i.e., trend period lengths of 5, 10 and 20 year with trend period windows starting in 1990 and incrementing by one year to a final period ending in 2017). This resulted in 24, 19 and 9 time-period windows of length 5, 10, and 20 years respectively. MCI was excluded from the rolling trend analysis because in the analyses associated with step 3, monthly values were required but MCI is based on annual observations. From here on we refer to these trends as the “NRWQN rolling trends dataset”.

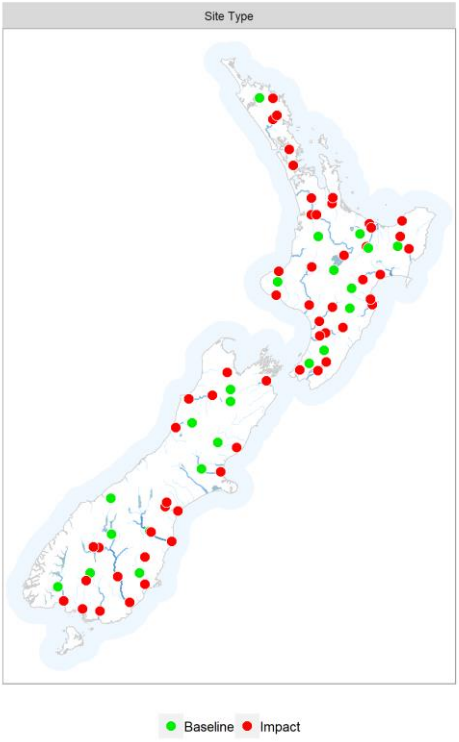

Figure 2. Location of NRWQN sites showing their classification into impact and baseline sites.

3 Methods

3.1 Site trends

Site trends were calculated using the methodology outlined in Larned *et al.* (2018). Briefly trends were analysed for all site and variable combinations with observations in >90% of years and >90% of seasons (Figure 1). Trend analyses were conducted using flow adjusted data following the procedure described in Larned *et al.* (2018) Where water quality observations were strongly associated with the flow on the sampling occasion, they are statistically adjusted to account for the influence of flow on the variable; at sites where the flow-concentration relationships are weak, raw observations are used in the assessment.

For assessments of trends in water quality variables other than MCI, we used seasons defined by months preferentially, and quarters when there were insufficient monthly observations. For some sites and variables, there was more than one sample within some seasons. In these cases, we used the median of the within-season values for (or the year for the invertebrate samples) and conducted the trend analyses with these data.

Trend analyses for every site and variable were undertaken using the LWP-Trends library and produced two key statistics: the Sen slope and the probability that the true trend was decreasing. The Sen slope describes trend magnitude and is expressed as the rate of change of the variable (year^{-1}). In this report, trends across sites and variables are made comparable by expressing them as a relative Sen slope (RSS), which is calculated by dividing the Sen slope by the median of all the observations within the time period.

Traditionally trends are declared to be detected with confidence when statistical confidence exceeds a nominal level. The probability that the true trend was decreasing provides a continuous measure of confidence in the trend direction. Larned *et al.* (2018) evaluated the probability that the true trend was decreasing from the Kendall *p*-value as follows:

$$P(S < 0) = 1 - 0.5 \ \times pvalue$$

$$P(S > 0) = 0.5 \ \times pvalue,$$

where $pvalue$ is the *p*-value returned by Kendall test (either seasonal or non-seasonal), S is the S statistic returned by Kendall test (either seasonal or non-seasonal) and P is the probability that the trend was decreasing. The trend direction is interpreted as decreasing when P > 0.5 and increasing when P < 0.5. For some of the analyses in this study that were not dependent on trend direction, we used the raw *p*-value to represent confidence in the evaluated trend. A *p*-value of <0.1 is equivalent to a >95% confidence that a trend is decreasing, or a <5% confidence that the trend is decreasing (i.e., 95% confidence that the trend is increasing).

The precision in the Sen slope is evaluated as part of the Sen Slope calculation. Briefly, the Sen slope is calculated by determining all possibly inter-observation slopes and ordering them from highest to lowest. The inter-observation slope is converted to a Z-score, allowing the evaluation of the probability of exceedance of any given slope. The Sen slope is the median (50th percentile) of all intersample slopes, and the 90% confidence interval of the Sen slope is defined by the 5th to 95th percentile.

3.2 Trend aggregation

We summarised the results over many sites using two measures of ‘aggregate trend’. First, we used the median of site RSS values as an aggregate measure of trend magnitude and examined the distribution of the RSS values using box plots. Sites were grouped in a variety of ways including nationally (i.e., all sites) and by sites belonging to designated classes (see section 3.5.2).

The second measure of aggregate trend was the proportion of improving trends (PIT) statistic (Snelder and Fraser, 2018). The PIT statistic was also calculated for groups of sites including nationally (i.e., all sites) and by sites belonging to designated classes (see Section 3.5.2). The PIT statistic is derived from the probability that the true trend was improving of the individual site trends. 'Improving trends' corresponded to decreasing trends in nutrient and ECOLI concentrations and TURB and increasing trends in CLAR and MCI. Conversely, 'degrading trends' corresponded to increasing trends in nutrient and ECOLI concentrations and TURB, and decreasing trends in CLAR and MCI.

The PIT statistic for a given water-quality variable assumes that the trends at multiple monitoring sites distributed across a domain of interest (e.g., a spatial domain such as the whole of New Zealand or a class of rivers) represent independent samples of the population of trends, for all sites within that domain. Let the sampled sites within this domain be indexed by s, so that $s \in \{1, \dots, S\}$ and let I be a random Bernoulli distributed variable which takes the value 1 with probability p and the value 0 with probability $q = 1 - p$. Therefore, $I_s = 1$ denotes an improving trend at site $s \in \{1, \dots, S\}$ when the estimated $p_s \geq 0.5$ and a degrading trend as 0 when $p_s < 0.5$. Then, the estimated proportion of sites with improving trends in the domain is:

$$PIT = \sum_{s=1}^{s=S} I_s / S$$

Because the variance of a random Bernoulli distributed variable is $Var(I) = p(1 - p)$, and assuming the site trends are independent, the estimated variance of PIT is:

$$Var(PIT) = \frac{1}{S^2} \sum_{s=1}^{s=S} Var(I_s) = \frac{1}{S^2} \sum_{s=1}^{s=S} p_s(1 - p_s)$$

PIT and its variance represent an estimate of the population proportion of improving trends, within a spatial or environmental domain, and the uncertainty of that estimate. It is noted that the proportion of degrading trends is the complement of the result (i.e., 1 - PIT). The estimated variance of PIT can be used to construct 95% confidence intervals[1] around the PIT statistics as follows:

$$CI_{95} = PIT \pm 1.96 \times \sqrt{Var(PIT)}$$

We calculated PIT and its confidence interval for all water quality variables over only the NRWQN rolling trends dataset; PIT statistics for the 2017 trends dataset were presented in Larned *et al.* (2018).

3.3 Variation in the confidence of trend evaluations

There were two tasks in this analysis: (1) to investigate differences in the confidence of trend evaluations, and (2) to evaluate the minimum trend magnitude that can be detected with a specified level of confidence. We characterised the confidence of trend evaluations in two ways: the confidence in the estimated trend magnitudes and the confidence in the estimated trend direction. The confidence (or precision) in the estimated trend magnitude was quantified as the difference between upper and lower 90% confidence intervals of the trend Sen slope.

[1] Note that +/- 1.96 are approximately the 2.5th and 97.5th percentile of a standard normal distribution.

The confidence in the estimated trend direction was quantified by the Kendall test p-value; the closer this value is to zero, the greater the confidence in trend direction. Note that the trend probability statistic quantifies the confidence that the trend was decreasing and is calculated from the Kendall test p-value (Section 3.1). We did not use the trend probability in these analyses because we were interested in the confidence in direction and not the direction itself.

We used both the 2017 trends and the NRWQN rolling trends datasets to explore whether characteristics of the observation datasets were associated with variation in the confidence in evaluations of trend magnitude and direction. First, we calculated several statistics from the observation datasets to characterise the site/variable observations. Characteristics that we hypothesised might be related to variability in trend evaluation confidence included: the variability of the observations, number of observations, the time-period length, number of unique values, degree of censoring, median observation value and trend magnitude. The complete list of characteristics that we used are in Table 2. For sites and variables with more than one observation within seasons, we used the median of within-season values to calculate the statistics listed in Table 2.

We also hypothesised that there would be differences in the confidence of trend evaluations between water quality variables, due to between-variable differences (e.g., differences in physical properties, chemical behaviour, analytical techniques and precision etc), and therefore all analyses were performed separately for each variable. We calculated the strength of the association between the characteristics of the observations listed in Table 2 and the confidence in the estimated trend magnitudes and direction using the non-parametric Spearman rank correlation coefficient which is a suitable test when the distributional assumptions of a parametric correlation test (e.g., absence of outliers, normality of variables, linearity, and homoscedasticity) are not met. The Spearman rank correlation coefficient is similar to the Pearson correlation coefficient, in that it represents the proportion of variability that is common to two variables. However, the Spearman correlation coefficient is computed from ranks, and therefore evaluates the monotonic relationship between two variables rather than the linear relationship that is evaluated by the Pearson correlation coefficient.

Table 2: Observation characteristic statistics

Statistic name	Description
nObs	Number of observations
PeriodLength	Time-period length of the trend analysis (years)
prop.censored	Proportion of observations that are censored
prop.unique	Number of unique observations as a proportion of total number of observations. This provides an indication of the analytical precision.
nuniqueObs	Total number of unique observations (prop.unique x nObs)
absRSS	The absolute Sen Slope divided by the median (% year^{-1}; RSS)
absSenSlope	The absolute Sen Slope (units of the original variable year^{-1})
Median	The median of the observations
sdlog10	The standard deviation of the log of the observations

For task two we used the combined trends dataset (both the 2017 trends and the NRWQN rolling trends) to explore minimum detectable trends at three nominated levels of confidence: 80%, 90%, 95%. For each variable, we ranked the trends in descending order of absolute magnitude (both as RSS and Sen Slopes) and calculated the cumulative count of statistically significant trends (at each of the three confidence levels) and divided by the cumulative count of the total number of trends, thus obtaining a relationship between the percentage of significant trends for all trends above a certain trend magnitude. We then identified the lowest RSS and Sen slopes at which 90% of the trends above this value were detectable at the three confidence levels.

3.4 Variation in trends with time-period length

We investigated how trend magnitude (RSS) and proportion of improving trends (PIT) vary with time-period length using both the 2017 trends dataset and the NRWQN rolling trends dataset. We plotted the distributions of RSS values associated with differing time-period lengths for both datasets (i.e., 10, 20 and 28 years for the 2017 trends dataset and 5, 10 and 20 years for the NRWQN rolling trends dataset). The boxplots of RSS values provided a graphical representation of how the distribution of RSS values varied with time-period length.

We also plotted PIT statistics for all variables derived from the 2017 trends dataset and the NRWQN rolling trends dataset. For the 2017 trends dataset we calculated the PIT statistic using all sites (i.e., representing the proportion of improving trends at the national scale) for the three time-period windows ending in 2017 of length 10, 20 and 28 years. For the NRWQN rolling trends dataset, we calculated the PIT statistic using all sites (i.e., representing the proportion of improving trends at the national scale) for the 24, 19 and 9 time-period windows of length 5, 10, and 20 years and ending in yearly increments up to 2017. The NRWQN rolling trends dataset therefore represented the distribution of PIT statistics for each of the time-period lengths but for the smaller NRWQN dataset.

3.5 Variation in trends with time-period window

3.5.1 Temporal variability in aggregate trends with time-period window

To characterise the temporal variation in trends across time-period windows, we graphed the NRWQN rolling trends dataset for each variable to demonstrate changes in the distribution of RSS values for each window. Similarly, for each variable we calculated and plotted the PIT statistic derived from the NRWQN rolling trends dataset for each time-period window and time-period length.

Variability in evaluated trends between different time-period windows can be related to sampling error, trends in water quality due to anthropogenic actions and trends associated with other forcing (e.g., ENSO) or to a combination of these three factors. Scarsbrook *et al.* (2003) showed that temporal patterns in water quality observations and water quality trends were correlated with the El Niño Southern Oscillation (commonly called ENSO) climate pattern. ENSO constitutes the single largest source of inter-annual variability in the global climate system (Diaz *et al.*, 1992). While this pattern is best known for the extremes of the oscillation (i.e., El Niño and La Niña) the phenomenon is in fact part of a continuum reflecting changes in sea level atmospheric pressure in the tropical Pacific Ocean (Allan *et al.*, 1996).

The SOI is calculated as the normalized anomalies of the monthly mean sea level pressure difference between Tahiti and Darwin. The SOI typically ranges from -30 to 30 and is quasi-periodic with a typical period of 3 to 7 years. An El Niño phase is defined when the SOI < 0 and a la Niña phase when the SOI > 0. Examples of the impact of ENSO on New Zealand climate can be seen at https://www.niwa.co.nz/our-science/climate/information-and-resources/clivar/elnino. Several other indices have been developed, but the SOI is most frequently used (Allan *et al.*, 1996; Mosley, 2000). Monthly values of the SOI for the period from 1989 to 2017 were obtained from the Australian Bureau of Meteorology (http://www.bom.gov.au/climate/current/soi2.shtml) and we use the Troup convention, whereby normalized index values are multiplied by 10.

Following Scarsbrook *et al.* (2003), we examined the association between water quality trends and climate forcing in two steps. First, for each site and variable, we characterised the association between individual water quality observations and the SOI. Second, for each trend period length, we examined the aggregate behaviour of site trends in each trend period window in relation to the linear trend in the SOI for the corresponding window. These two steps are described in more detail in the subsequent sections.

3.5.2 Association between water quality observations and the SOI

The Pearson correlation coefficient was used to indicate the strength of linear relationships between SOI and individual water quality variables at each site (Figure 3). We used all the observations in the entire period of record at for each site to calculate the correlation coefficients. To reduce the variability associated with regular seasonal changes, all water quality observations were deseasonalised by employing 12-month centred moving averages and the SOI was treated similarly (Figure 3). For each variable, we classified each site as either a negative or positive responder to the SOI depending on whether the Pearson correlation coefficient was positive or negative (Figure 3). We mapped these SOI response classes to examine their spatial distribution.

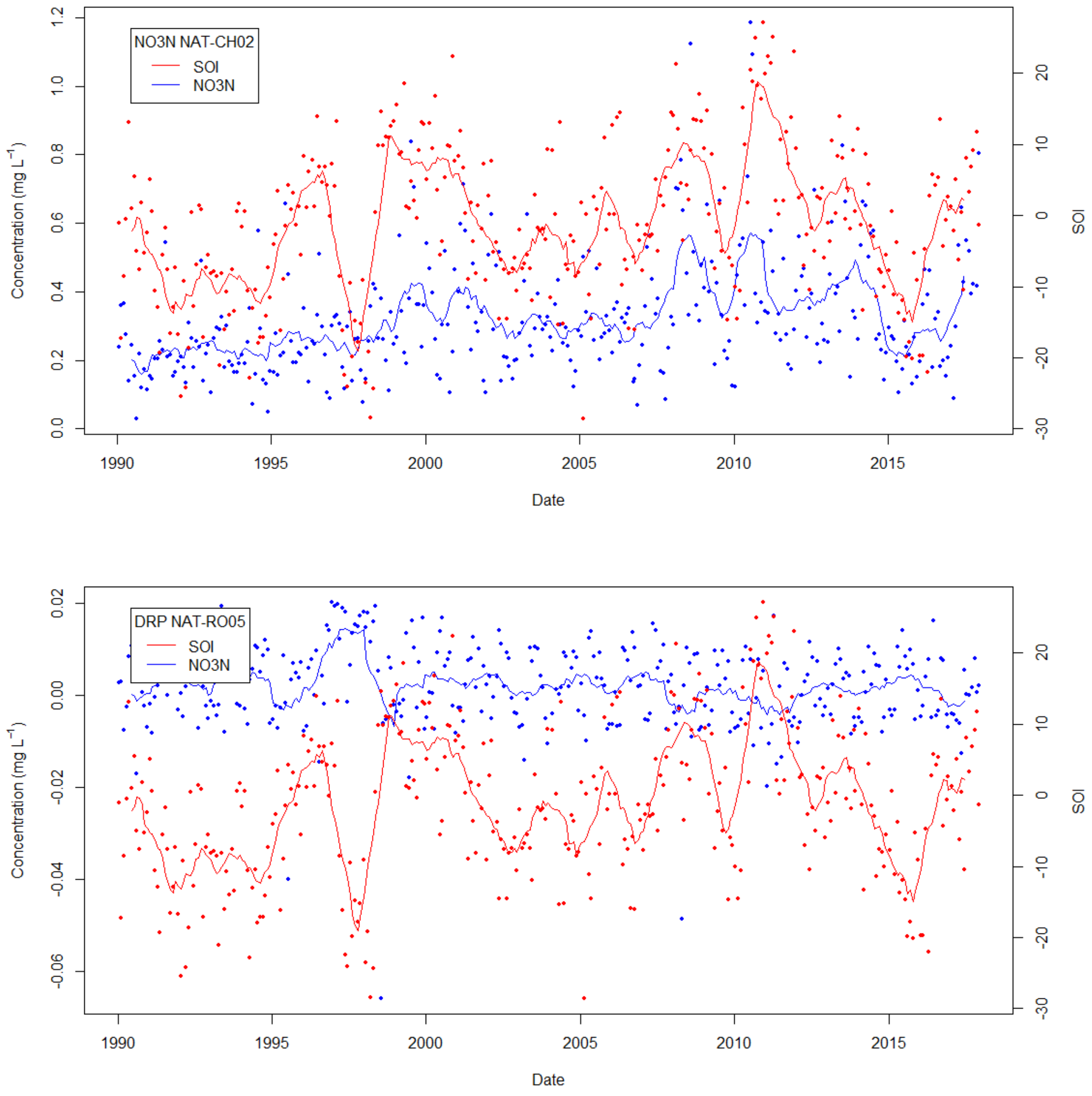

Figure 3. Examples of correlations between water quality observations and SOI. The top plot is nitrate at CHO2 (Hurunui @ SH1 Bridge), which has a correlation of 0.72. The lower plot is DRP at RO05 (Rangitaiki @ Te Teko) which has a correlation of -0.55. The lines are the deseasonalised data (I.e., 12-month centred moving average values) and the points are the individual monthly observations.

To provide further insights into the reasons for the spatial distribution of SOI response classes, we used statistical modelling. A statistical classification model was used to explain the geographic distribution of SOI response classes using geographic coordinates and catchment characteristics as explanatory variables. The statistical modelling methods and detailed results are provided in Appendix 2.

3.5.3 Association between water quality trends and SOI trends

We used the SOI response classes to examine the association between trends in each water quality variable and trends in the SOI in three steps. First, we quantified the monotonic trend in the SOI for every time-period window associated with the rolling trends (i.e., 5, 10 and 20-year time-period lengths starting from 1990 with time-period windows incrementing by one year to a final time-period window ending in 2017). We quantified the trend in the SOI by regressing the monthly values of SOI against their respective dates. We used simple linear regression because the SOI data are approximately normally distributed.

At the second step we plotted the RSS and PIT statistics against the end year of each time-period window for all eight water quality variables and the three time-period lengths. On these plots we grouped sites by their SOI response classes and superimposed the SOI trends for the corresponding time-period windows. We expected that there would be relationships between the "aggregate trend" (i.e., the mean RSS values and PIT statistics) and the SOI trends. In addition, we expected that there would be systematic differences in those relationships between SOI response classes. In particular, we expected that when the SOI trend was negative, the positive SOI response class would tend toward negative RSS values and, for variables where decreasing RSS indicates improvement, this would produce high PIT statistics, and vice versa. For variables where increasing trends suggest improvement (e.g. CLAR), in this example, the reverse relationship with PIT would be expected. There are four combinations of SOI trend and SOI response class, each associated with a differing expected aggregate trend response, which are summarised on Table 3.

Table 3. Expected aggregate water quality trend outcome.Each of the four combined SOI response and SOI trend classes are represented by the four cells in the table.

	SOI trend	
SOI response class	**Decreasing**	**Increasing**
Positive	High PIT^1 & Negative RSS	Low PIT^2 & Positive RSS
Negative	High PIT^1 & Positive RSS	Low PIT^2 & Negative RSS

Notes:

1. Low PIT for CLAR
2. High PIT for CLAR

The third step used the site correlation with the SOI and the SOI trend in each trend period window to explain the between site differences in RSS values and PIT statistics for each time-period window. For each water quality variable and time-period length (i.e., taking all the time-period windows) we fitted a linear regression model of the form:

$$RSS_{SW} \sim SOICor_S + SOITrend_W \qquad \textit{Equation 1}$$

where RSS_{SW} is the trend magnitudes of all sites and all windows (i.e., the individual RSS values), $SOICor_S$ is the observation - SOI correlation of all sites and $SOITrend_W$ is the linear trend in the SOI for the time-period window.

For each water quality variable and time-period length (i.e., taking all the time-period windows) we fitted a second linear regression model of the form:

$$PIT_{WC} \sim SOIClass \times SOITrend_w \qquad \textit{Equation 2}$$

where PIT_{WC} is the PIT statistic for each SOI response class and all time-period windows, SOIClass is the SOI response class of the sites from which PIT is calculated and $SOITrend_W$ is the linear trend in the SOI for the time-period window. This second model specifies an interaction between $SOIClass_C$ and $SOITrend_W$ (indicated by the multiplication operator in the above model). The interaction allows the effect of the SOI trend to depend on the SOI class of each site, which allows the model to represent the expectations set out in Table 3.

4 Results

4.1 Variation in confidence of trend evaluations

Correlations between measures of confidence in the trend direction and various distributional characteristics of the observation datasets are shown in Figure 4 and Figure 5. Confidence in trend direction is quantified by the p-value and the confidence trend magnitude (or precision of) is quantified by the 90% confidence intervals of the Sen slope. A full set of correlation matrices pertaining to each individual water quality variable is provided in Appendix A.

For all variables, the p-value is strongly negatively correlated to the magnitude to the trend and moderately negatively correlated to the trend period length, i.e. as the trend magnitude and trend periods increase, confidence in the trend direction increases. Other characteristics of the water quality observations were also associated with the confidence in trend direction. For DRP, TP and NH4N, the p-value was moderately positively correlated with the proportion of unique observations (i.e., the p-values are influenced by measurement precision). These variables have low measurement precision leading to many observations having the same numeric value, and therefore low proportions of unique observations. There were also moderate negative correlations between confidence in trend direction and Sen slope precision.

Variation in the confidence in trend magnitude was most strongly correlated with the absolute RSS (i.e., the trend magnitude). The correlation coefficients were positive indicating that trend precision is lower for higher magnitude trends. There were also strong negative correlations with time-period length and number of observations, indicating that estimates of trend magnitudes become more precise with increasing dataset size. It is also noted however, that trend magnitude and period are moderately negatively correlated hence these observations are not independent.

The variability in the observations (as characterised by the standard deviation of log10 of the observations) only had moderate to low correlations with the p-value. The Sen slope confidence intervals for TURB, MCI and NH4N increase as the observation variability increased; the relationship for other water quality variables were in the same direction, but the correlations were weaker.

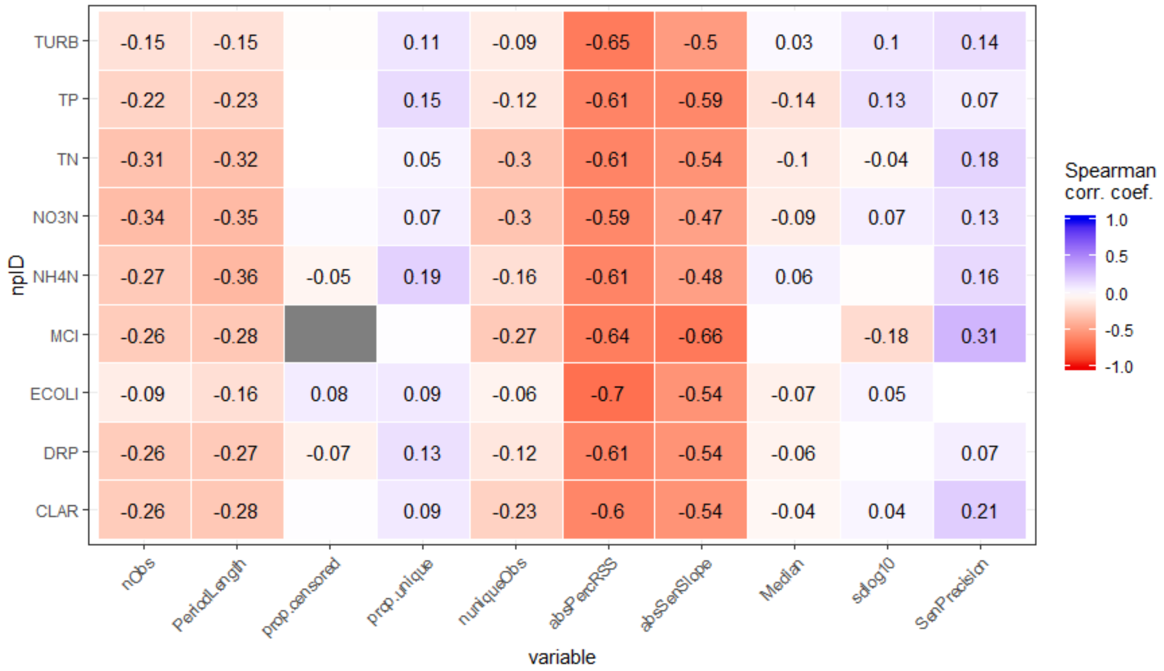

Figure 4: Correlations between the trend analysis Kendall test p-value and characteristics of the observation datasets and the evaluated trends, by water quality variable.Values in the cells are the Spearman rank correlation coefficient. Cells with no values had Spearman rank correlation coefficient p-values > 0.05.

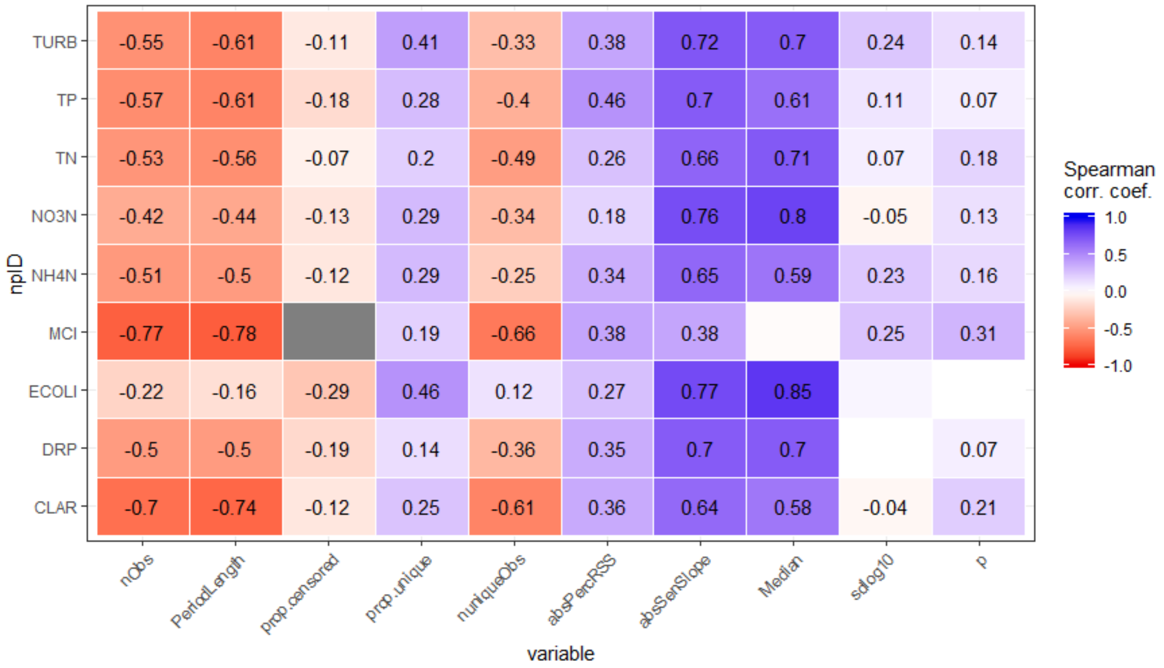

Figure 5: Correlations between width of the Sen slope 90% CI (precision) and characteristics of the observation datasets and the evaluated trends, by water quality variable. Values in cells are the Spearman rank correlation coefficient. Cells with no values had correlations with p-values >0.05.

Generally, the minimum detectable trend magnitudes (RSS and Sen slopes) decreased as time-period increased for all confidence levels (Table 4). For example, for a 5-year time-period length, 90% of trends with RSS >12.6% were detected with a confidence of 95% but for a 10-year time-period length 90% of trends with RSS >6.6% were detected with this level of confidence. There were some deviations from this pattern between the 5 and 10-year values, which is likely due to the differences in the sampled sites in each group (a maximum of 77 NRWQN sites for 5 years, compared with approximately 300 sites for the 10-year samples).

Table 4. Minimum detectable trend magnitudes as RSS and absolute Sen Slopes, by time-period length, confidence level and water quality variable. The values in the cells are the trend magnitudes at which 90% of trends have the indicated confidence levels. The cells are coloured on one colour scale (white to red) from lowest to highest RSS across all variables. Sen Slope cells are coloured from white to red, with scales individualised to each water quality variable.

Variable	Time-period length	Absolute RSS (%)			Absolute Sen Slope[1]		
		Confidence level					
		80th	90th	95th	80th	90th	95th
CLAR	5	4.4	9.7	12.6	1.2E-01	2.7E-01	4.4E-01
	10	1.3	2.5	3.6	2.4E-02	8.1E-02	1.3E-01
	20	0.3	0.5	0.9	3.8E-03	9.5E-03	2.5E-02
	28	0.2	0.3	0.3	2.4E-03	3.6E-03	4.0E-03
DRP	5	2.7	5.0	7.9	2.9E-04	9.1E-04	3.7E-03
	10	0.7	1.5	2.6	3.7E-03	1.8E-03	1.7E-03
	20	0.1	0.4	0.6	5.6E-06	2.3E-05	2.1E-03
	28	0.0	0.3	0.4	1.6E-06	3.0E-05	4.6E-05
ECOLI	5	6.9	12.6	22.9	1.5E+01	4.3E+01	4.7E+01
	10	2.9	7.0	11.8	4.7E+01	5.7E+01	9.2E+01
	20	0.6	1.9	2.4	3.4E+00	8.3E+00	1.3E+01
	28	0.4	0.4	1.2	2.1E+00	2.1E+00	2.1E+00
MCI	10	1.4	3.0	4.6	1.3E+00	2.7E+00	1.0E+00
	20	0.4	0.7	1.6	3.8E-01	6.9E-01	1.7E+00
	28	0.2	0.3	0.4	1.7E-01	3.1E-01	4.2E-01
NH4N	5	3.4	6.9	10.6	8.5E-04	2.5E-03	6.8E-03
	10	0.9	2.2	3.4	1.3E-04	3.5E-03	1.1E-01
	20	0.1	0.4	0.7	1.1E-05	4.6E-05	7.9E-05
	28	0.0	0.2	0.5	0.0E+00	1.7E-05	4.6E-04
NO3N	5	3.6	6.8	10.6	1.5E-02	2.5E-02	6.0E-02
	10	0.9	1.9	3.2	3.2E-03	1.1E-02	6.7E-02
	20	0.1	0.4	0.6	1.3E-04	4.1E-04	1.6E-03
	28	0.1	0.1	0.4	5.0E-05	2.5E-04	7.4E-04
TN	5	2.4	4.2	5.7	1.6E-02	3.5E-02	4.7E-02
	10	12.7	12.7	7.3	4.1E-03	2.4E-02	1.0E-01
	20	0.1	0.4	0.5	1.7E-04	8.7E-04	2.2E-03
	28	0.1	0.1	0.2	1.2E-04	4.9E-04	6.7E-04
TP	5	4.3	9.1	12.8	1.2E-03	3.6E-03	6.7E-03
	10	1.4	2.9	4.9	4.0E-04	1.1E-03	8.0E-03
	20	0.3	0.6	0.9	6.9E-05	1.5E-04	2.4E-04
	28	0.1	0.2	0.4	1.8E-05	3.1E-05	1.1E-04
TURB	5	5.8	10.6	16.4	2.6E+00	6.9E+00	1.3E+01
	10	1.7	3.3	4.8	5.1E-01	1.5E+00	2.4E+00
	20	0.5	0.9	1.4	1.5E-02	5.1E-02	1.9E-01
	28	0.2	0.5	0.6	1.1E-02	2.3E-02	6.7E+00

4.2 Variation in trends with time-period length

RSS values decreased with increasing time-period length for both the 2017 trends dataset and the NRWQN rolling trends dataset (Figure 6). For a given time-period length, the RSS values were lower for the NRWQN rolling trends data than the 2017 trends data. This is likely related to differences in the sites represented in the two datasets. The NRWQN sites are generally located on larger rivers, which may damp the impact of drivers of water quality change (e.g., land use changes or climate variation effects) compared to smaller catchments. The differences in the RSS values between datasets were generally smaller than those between time periods (Figure 6).

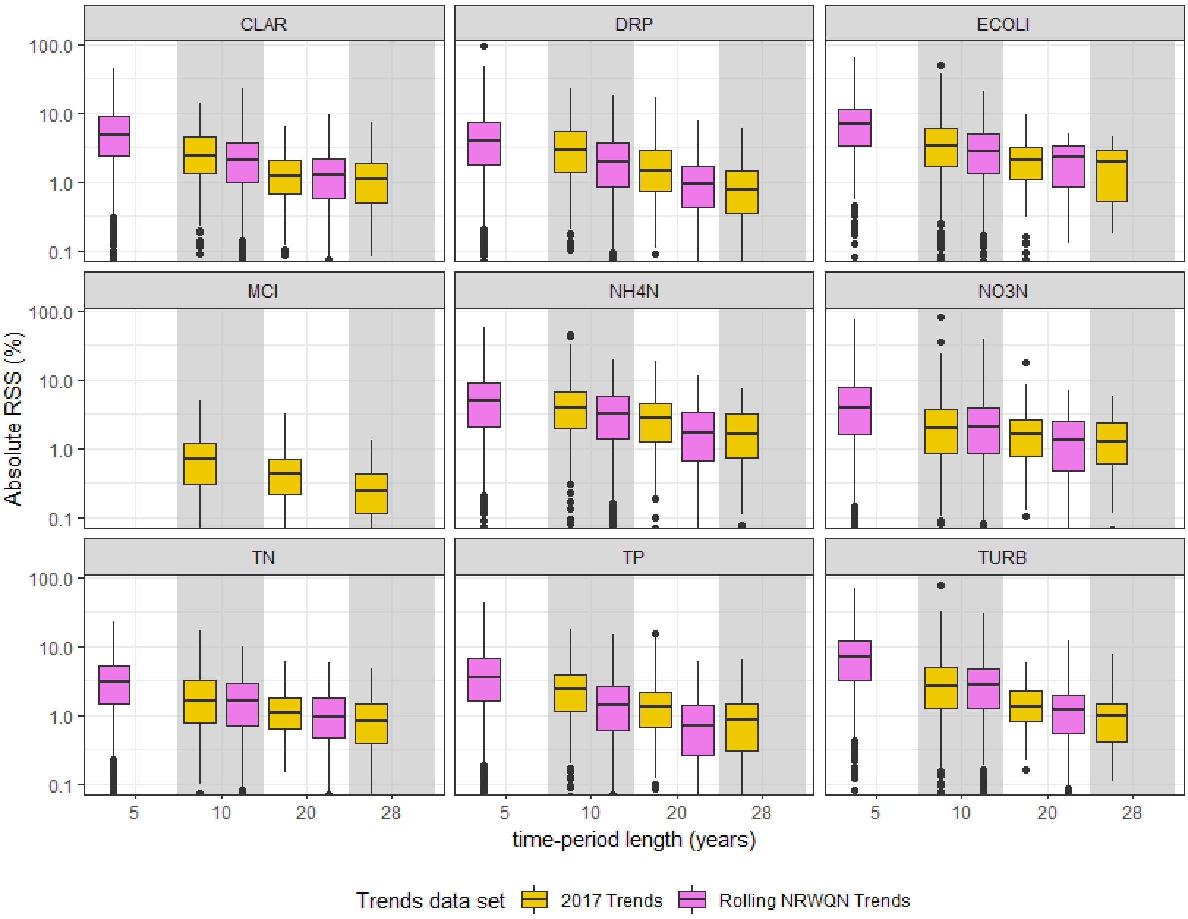

Figure 6. Distributions of RSS values associated with differing time-period lengths for the 2017 trends dataset (10, 20 and 28 years) and the NRWQN rolling trends dataset (5, 10 and 20 years).

PIT statistics were variable with respect to time-period length for both the 2017 trends dataset and the NRWQN rolling trends dataset (Figure 7). For the 2017 trends dataset, PIT was >50% (i.e., a majority of sites were improving) for most time-period windows except for TN and TURB. For each variable, there was at least one significant difference in PIT statistics (i.e., non-overlapping 95% confidence intervals) between time-period lengths (e.g., DRP for the 20-year period compared to the 28-year period).

The PIT statistics were highly variable for the NRWQN rolling trends dataset (Figure 7). The distribution of PIT across time-period windows was generally evenly split between the majority of sites improving and degrading (i.e. the distributions lay either side of the 50% line on Figure 7). Exceptions to this were CLAR and TN for time-period lengths of 10 and 20-years, for which PIT was >50% and <50% for all time-period windows (i.e., the entire distributions were above and below the 50% line on Figure 7).

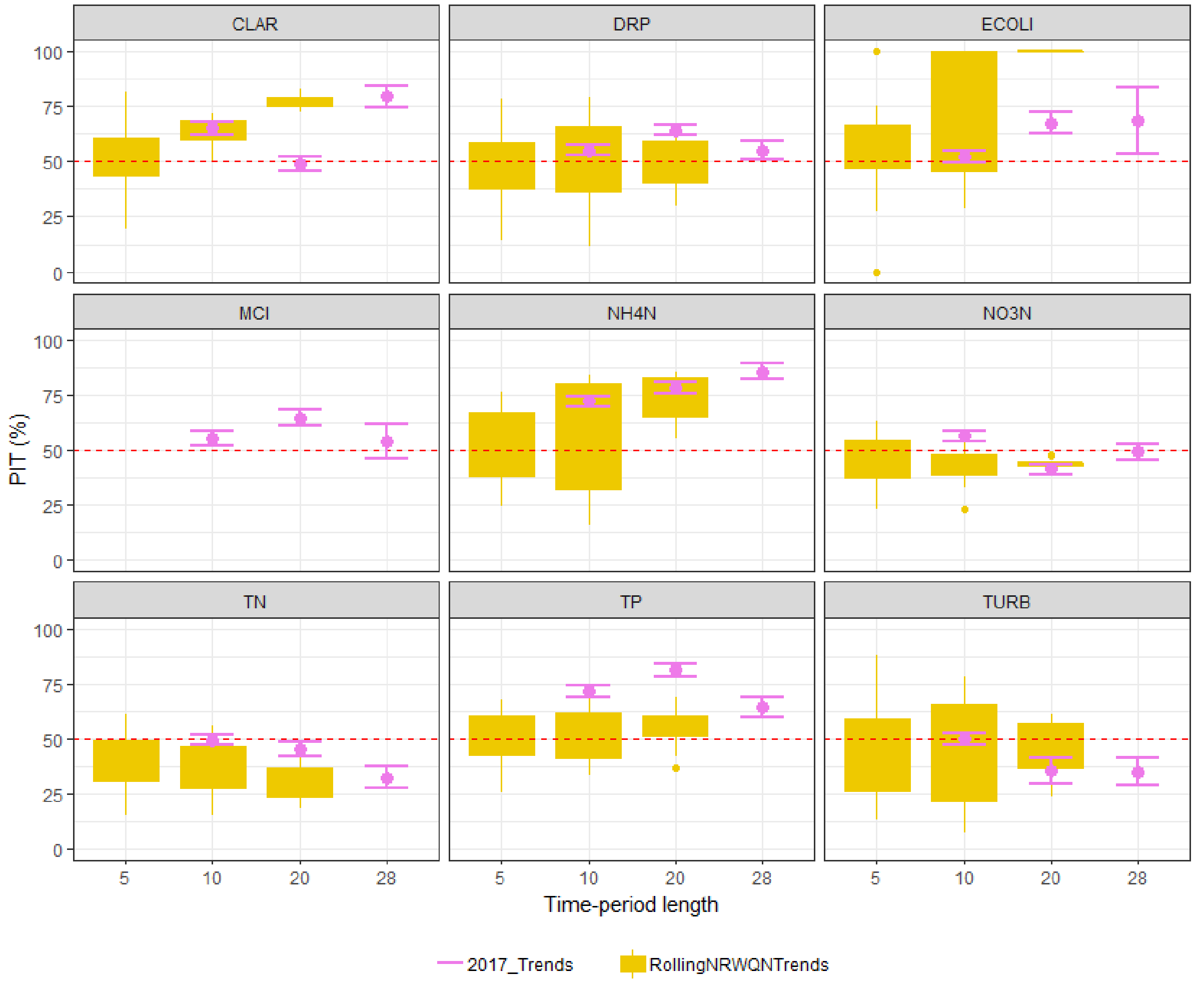

Figure 7: PIT statistics derived from the 2017 trends dataset (10, 20 and 28 years) and the NRWQN rolling trends dataset (5, 10 and 20 years) for differing trend period lengths. A single PIT statistic is shown for each variable and time-period length for the 2017 trends dataset, which pertains to the period ending 2017. The error bars show the 95% confidence intervals for this PIT statistic. The boxplots show the distribution of PIT statistics for each variable for all the analysed time-period windows for the NRWQN rolling trends dataset. No confidence intervals are shown for the PIT statistics associated with the NRWQN rolling trends dataset.

4.3 Variation in trends with time-period window

4.3.1 Temporal variation in aggregate trends

For all variables, the distribution of NRWQN site trend magnitudes was sensitive to time-period of analysis (Figure 8). For all variables and time-period windows, the site median RSS fluctuated between positive and negative values, but the variability of the site median RSS values decreased with increasing time-period duration (Figure 8). For example, for the 5-year time-period length, there were windows for which CLAR had absolute site median RSS values greater than 5%, whereas for the 10 and 20-year time periods the highest absolute site median RSS values were <3% and <0.5% respectively (Figure 8).

The changes in the RSS values were quasi periodic for most variables and trend durations. For example, for the 5-year time period there were up to 3 peaks and troughs (e.g., CLAR, DRP; Figure 8). The temporal patterns in the median RSS values differed between variables and were inversely related for some pairs of variables (e.g., NH4N compared to DRP and NO3N for the 10- and 20-year time-period windows; Figure 8).

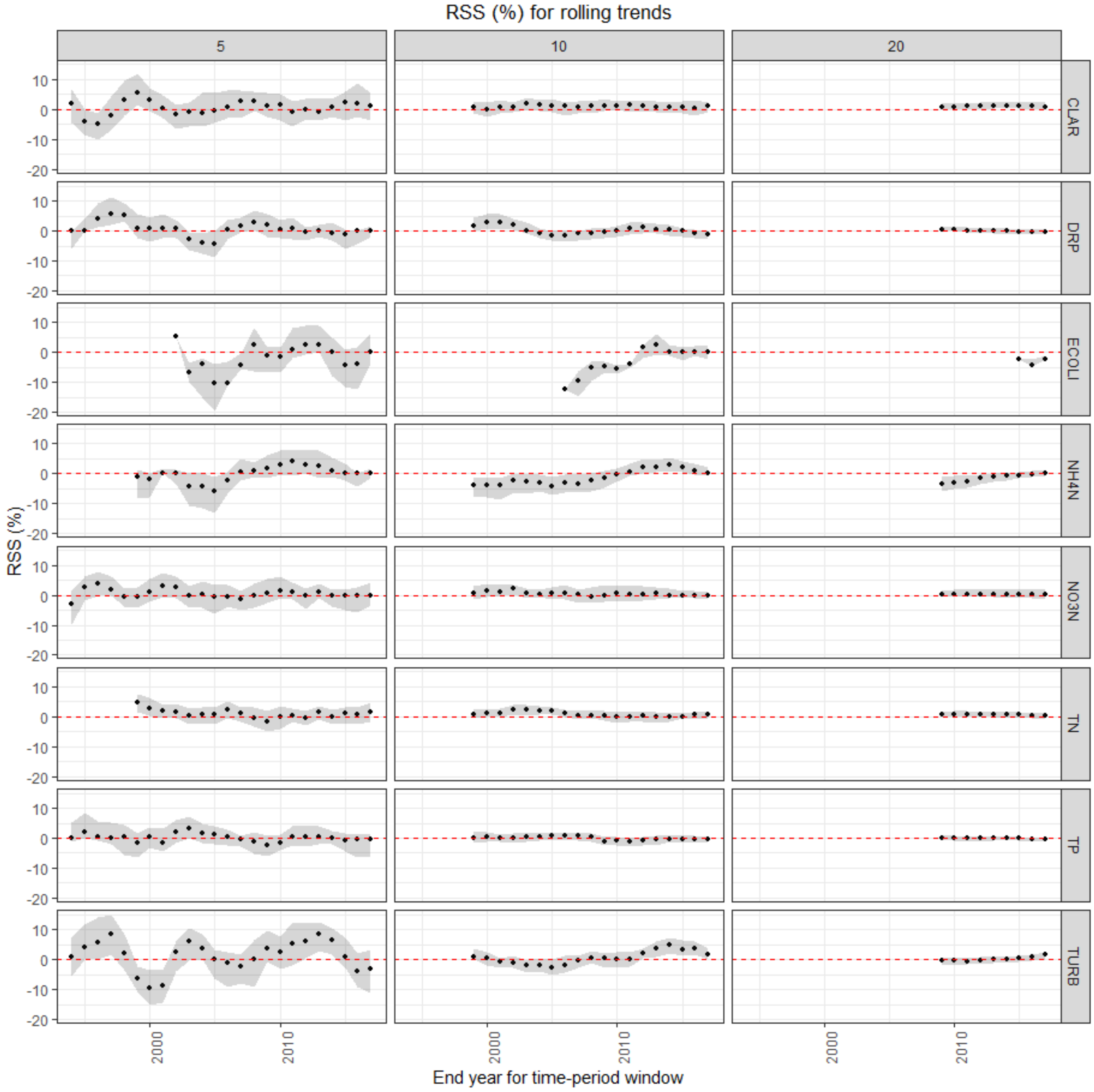

Figure 8. Temporal variation in trend magnitude with time-period window. The points represent the median RSS value (over all NRWQN sites) and the grey ribbon indicates the interquartile range.

For all variables, the proportion of sites with improving trends (PIT statistic) was sensitive to time period of analysis (Figure 9). For all variables and time-period windows, the PIT statistic fluctuated between values greater than and less than 50% (Figure 9). The variability of the

PIT statistics decreased with increasing time-period duration (Figure 9). For example, for the 5-year time period there were examples of greater than 95% confidence that the majority improving of sites improved (i.e., the lower 95% confidence interval for PIT was greater than 50%) to majority degrading trends (i.e., the upper 95% confidence interval for PIT was less than 50%) within a one or two year change in the trend period, whereas this did not occur for the 20-year period.

The temporal variation of the PIT statistics exhibited the same pattern to the temporal variation in the trend magnitudes and were quasi-periodic for most variables and trend durations. For example, for the 5-year time period there were up to 3 peaks and troughs for both the trend magnitude and PIT statistics for CLAR; Figure 8 and Figure 9). The temporal patterns in the PIT statistics differed between variables and were inversely related for some pairs of variables (e.g., NH4N compared to TN, NO3N and TP for the 10- and 20-year time-period windows; Figure 9).

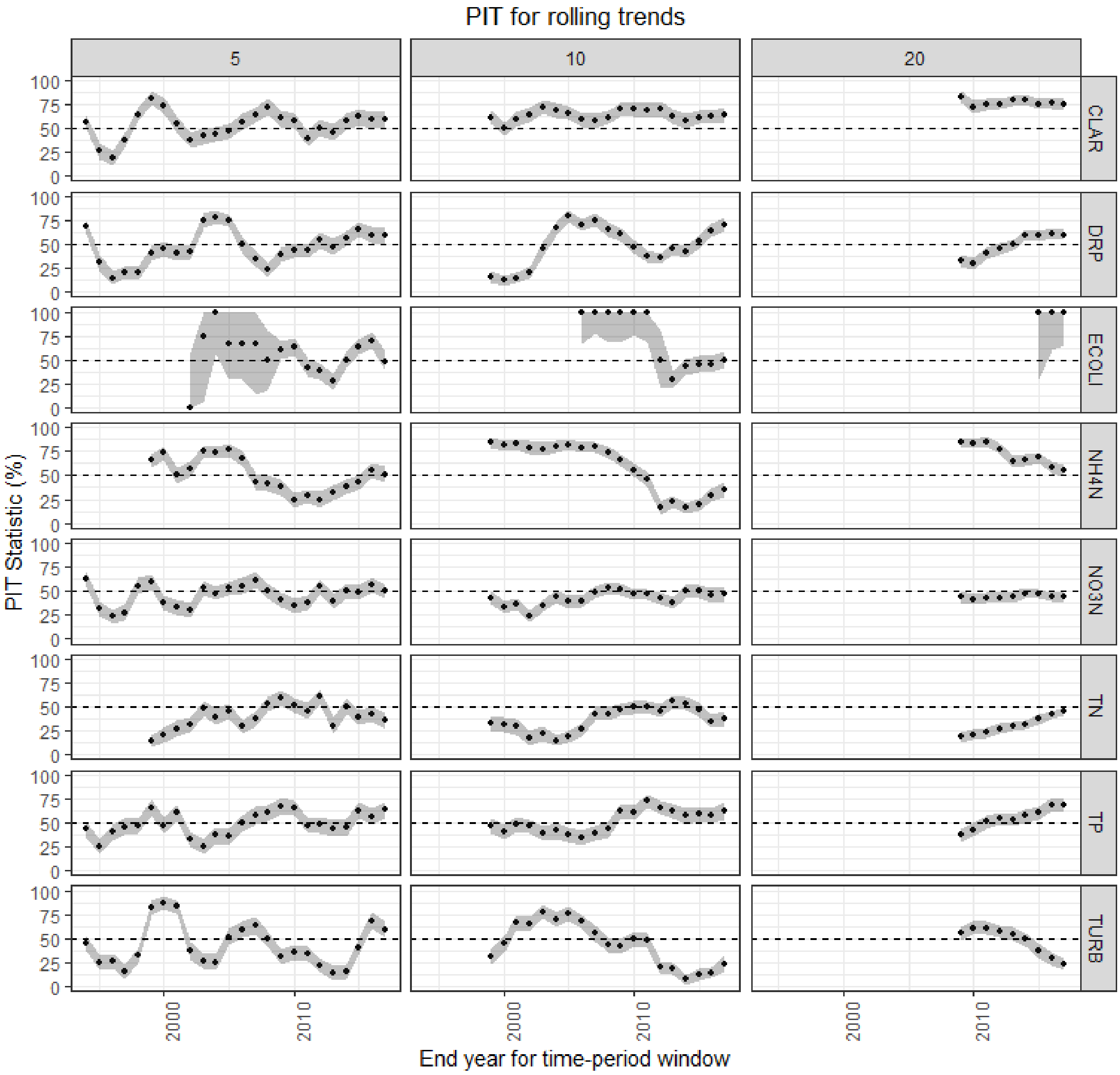

Figure 9. Variation in the characteristic trend magnitude with time-period window. The points represent the PIT statistic (proportion of improving NRWQN sites) and the grey ribbon indicates the 95% confidence interval of PIT.

4.3.2 Association between water quality observations and SOI

There was significant variation in the strength and direction of the relationship between the monthly deseasonalised water quality observations and the monthly deseasonalised SOI at individual NRWQN sites (Figure 10). Pearson correlation coefficient values ranged from-0.5 to 0.7. The mean of the absolute values of correlation differed by water quality variable and was lowest for TURB and highest for DRP (Table 5). Site type (i.e., impact or baseline) explained a maximum of 12% of the variation in the correlation coefficients and was only significant for DRP, NH4N and TN (note the *p*-value was 0.06 for NO3N) (Table 5).

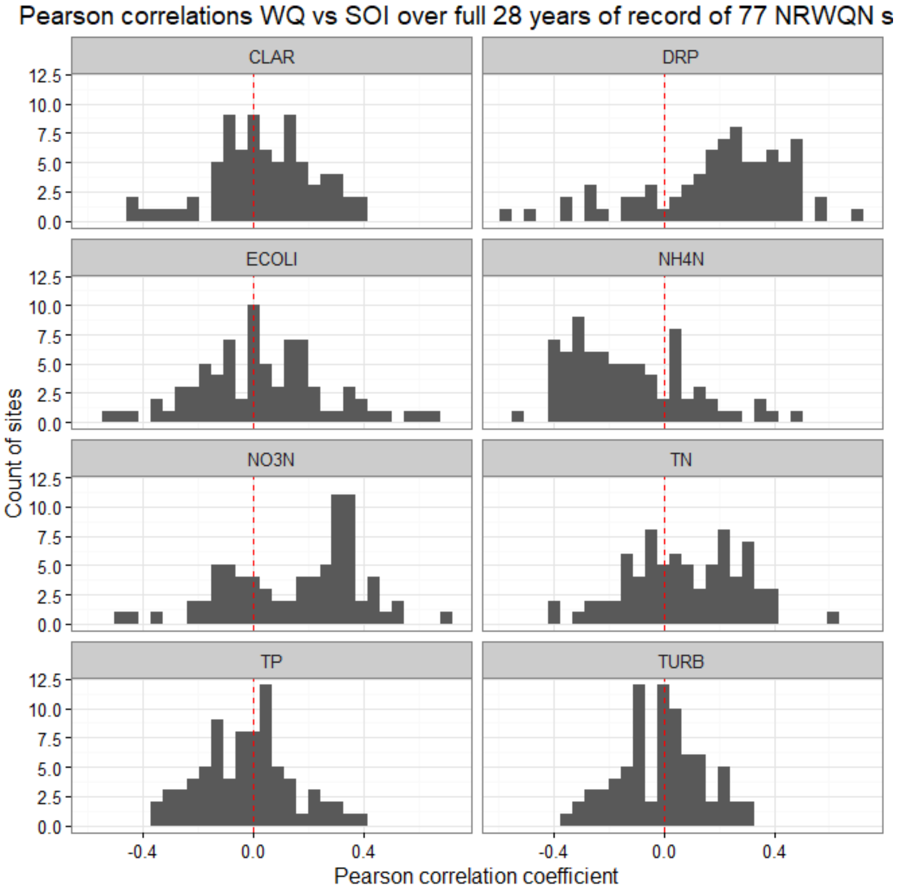

Figure 10. Distributions of Pearson correlation coefficient measuring the linear relationships between water quality observations and the SOI at the NRWQN sites. The red line indicates a correlation coefficient of zero.

Table 5. Pearson correlations of monthly deseasonalised water quality observation and the monthly deseasonalised SOI by water quality variable. The mean absolute correlations are shown for all sites and sites grouped by site type (impact or baseline). The ANOVA statistics test whether site type (i.e., impact or baseline) explains the value of the site correlations.

Variable	**Mean absolute correlation**			**ANOVA statistics**	
	All sites	**Impact sites**	**Baseline sites**	**R^2 (%)**	**P value**
CLAR	0.15	0.16	0.13	2	0.19
DRP	0.29	0.26	0.35	7	**0.02**
ECOLI	0.19	0.19	0.20	0	0.75
NH4N	0.22	0.20	0.26	12	**0.002**
NO3N	0.24	0.26	0.20	5	0.06
TN	0.18	0.19	0.15	9	**0.01**
TP	0.13	0.13	0.15	2	0.25
TURB	0.12	0.12	0.13	0	0.83

Sites were predominantly assigned to the positive SOI response class (i.e. concentrations are on average higher when SOI > 0) for all variables except NH4N (Table 6).

Table 6. Proportion of the 77 NRWQN sites in the positive and negative SOI response classes.

Variable	**Positive**	**Negative**
CLAR	60	40
DRP	81	19
ECOLI	61	39
NH4N	27	73
NO3N	70	30
TN	62	38
TP	45	55
TURB	55	45

The spatial patterns associated with the SOI response classes indicated that the direction of response of water quality to the SOI for a given site differs by water quality variable (Figure 11). The mapped patterns suggest that the causes of SOI response class membership are not purely spatial and are complex (e.g., for a given water quality site, adjacent sites could belong to different classes (Figure 11). See Appendix A2 for further analysis of spatial variation in SOI response classes.

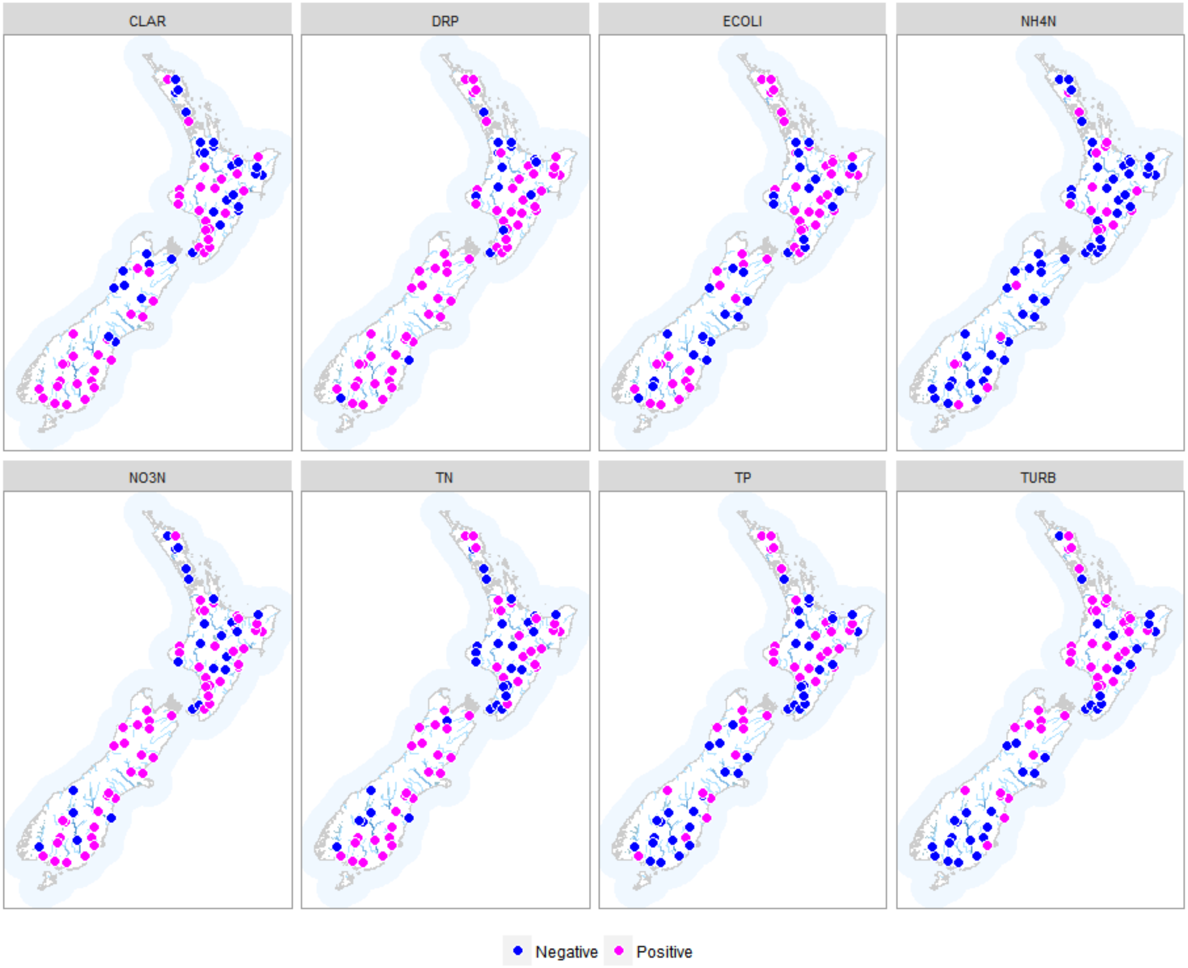

Figure 11. Geographic patterns in SOI response classes for NRWQN sites and the six water quality variables.

4.3.3 Association between water quality and SOI trends

The linear trend in the SOI varied between time-period windows for all three trend period lengths (i.e., 5, 10 and 20 years; Figure 12). All three trend periods lengths were represented by trend period windows with both positive and negative SOI trends. However, the magnitudes of the SOI trends generally decreased with increasing time-period length (Figure 12).

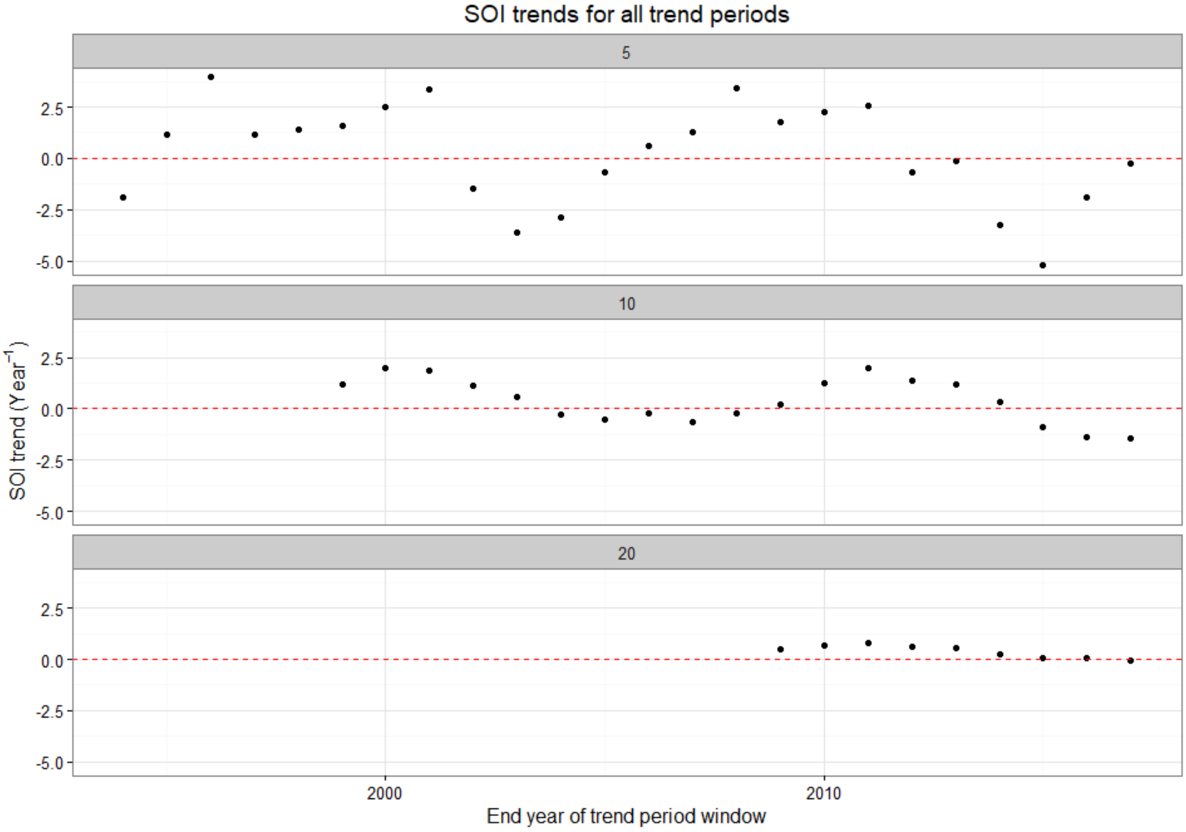

Figure 12. Linear trends in the SOI for three trend period lengths of 5, 10 and 20 years. Each point represents the linear trend in the SOI for the time-period window (panels) with the indicated end year (x-axis).

For all variables, the distribution of NRWQN site trend magnitudes for sites grouped by SOI response class were sensitive to the time-period of analysis (Figure 13). There was a degree of correspondence between the linear trend in the SOI for the time-period window and the site trend magnitudes (Figure 13). For example, for DRP and the 5-year time-period lengths, the median RSS values for the positive SOI response group tended to be positive for time-period windows when the SOI trend was positive and vice versa. There was also a degree of correspondence between the expected aggregate water quality trend outcome (Table 3) and the variation in median RSS values of the sites grouped by positive and negative SOI response classes. For example, when the SOI trend was positive, the RSS values for DRP for sites in the positive SOI response class tended to be positive and the RSS values for DRP for sites in the negative SOI response class tended to be negative (Figure 13).

The degree to which median RSS values for the individual variables followed the expected outcomes (Table 3) differed (Figure 13). In addition, within a variable the correspondence of median RSS values and expected outcomes (Table 3) differed between trend period lengths (Figure 13). For example, for NO3N and the five-year trend period windows, there were trend periods where both positive and negative SOI response classes had median RSS values that were positive (e.g., periods ending 2000, 2001, 2002 and 2003; Figure 13). This was contrary to the expected outcomes described in Table 3. However, for the 10-year time-period length, the median RSS values for the positive and negative SOI response classes were in close agreement with expectations for most time-period windows (i.e., the negative class had negative median RSS values when the linear trend in the SOI was positive and vice versa; Figure 13).

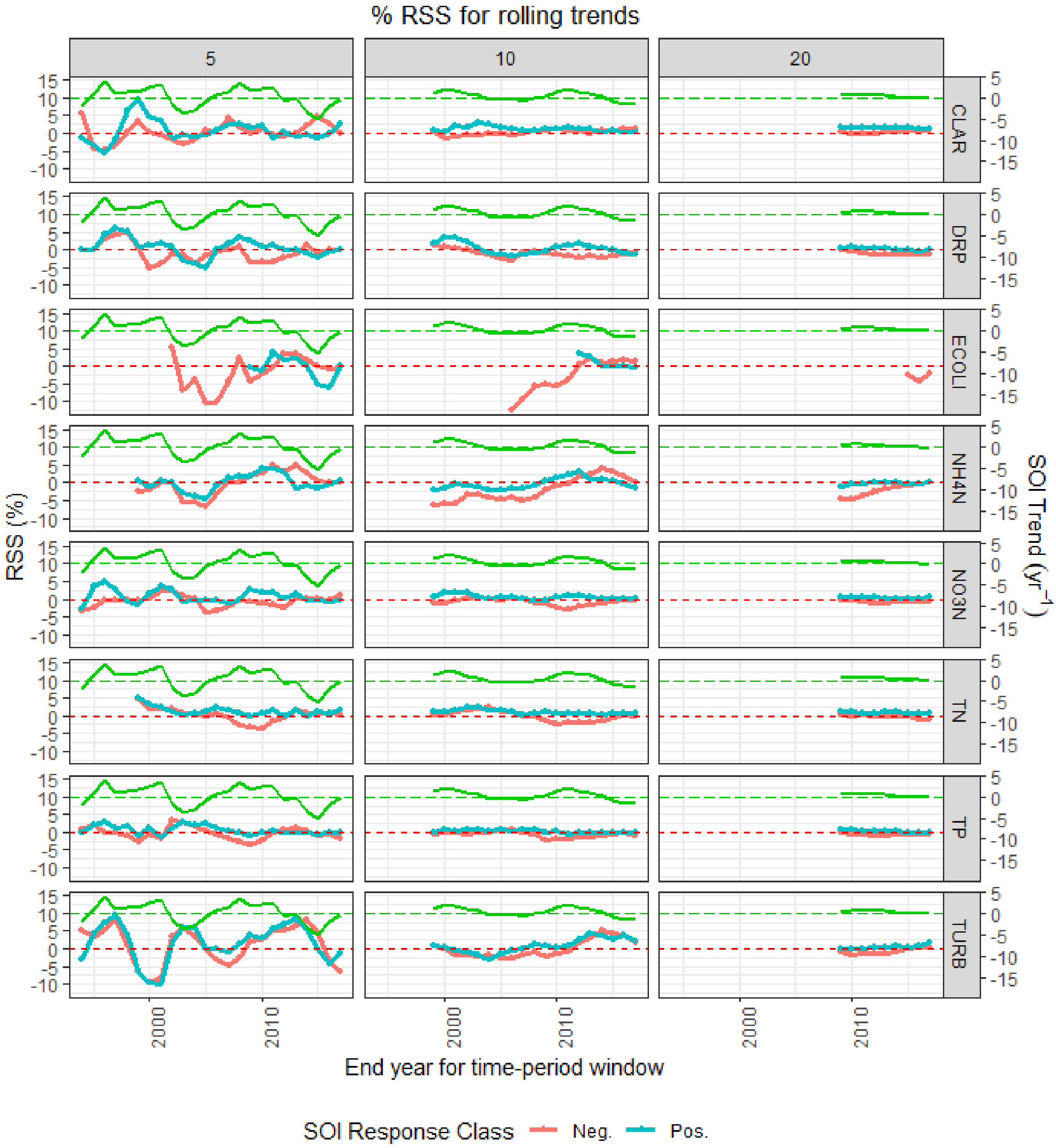

Figure 13. SOI trend and aggregate water quality trends (represented by median RSS values) for NRWQN sites grouped by SOI response class. The columns represent the three trend period lengths and the rows represent the water quality variables. The linear trend in the SOI is the green line on each panel. The aggregate water quality trends are shown by the red and blue lines which indicate the median of the site RSS values for the positive and negative SOI response classes respectively.

For all variables, the PIT statistics for sites grouped by SOI response class were sensitive to time-period of analysis (Figure 14). There was a degree of correspondence between the linear trend in the SOI for the time-period window and the site trend magnitudes (Figure 14). For example, for NO3N and the 10- and 20-year time-period lengths, the PIT statistics for the positive SOI response group tended to be below 50% (i.e., a majority of sites were degrading) for time-period windows when the SOI trend was positive and vice versa. Note that the reverse pattern is evident for CLAR because increasing clarity indicates improvement; hence the PIT statistics for the negative SOI response group tended to be below 50% when the SOI trend was positive and vice versa.

There was a degree of correspondence between the expected aggregate water quality trend outcome (Table 3) and the variation in PIT statistics of the sites grouped by positive and negative SOI response classes. For example, when the SOI trend was positive, the PIT statistics for DRP for sites in the positive SOI response class tended to be below 50% (i.e., a majority of sites were degrading) and the PIT statistics for DRP for sites in the negative SOI response class tended to be above 50% (Figure 14).

The degree to which PIT statistics for the individual water quality variables followed the expected outcomes (Table 3) differed (Figure 14). In addition, within a variable the correspondence of PIT statistics and expected outcomes (Table 3) differed between trend period lengths (Figure 13). For example, for NO3N and the five-year trend period windows, there were trend periods where both positive and negative SOI response classes had PIT statistics above or below 50% (i.e., a majority of sites were improving or degrading) (e.g., periods ending 2000 - 2003; and 2005 – 2008; Figure 14). This was contrary to the expected outcomes described in Table 3. However, for the 10-year time-period length, the PIT statistics for the positive and negative SOI response classes were in close agreement with expectations for most time-period windows (i.e., the negative class had PIT>50% when the linear trend in the SOI was positive and vice versa; Figure 14).

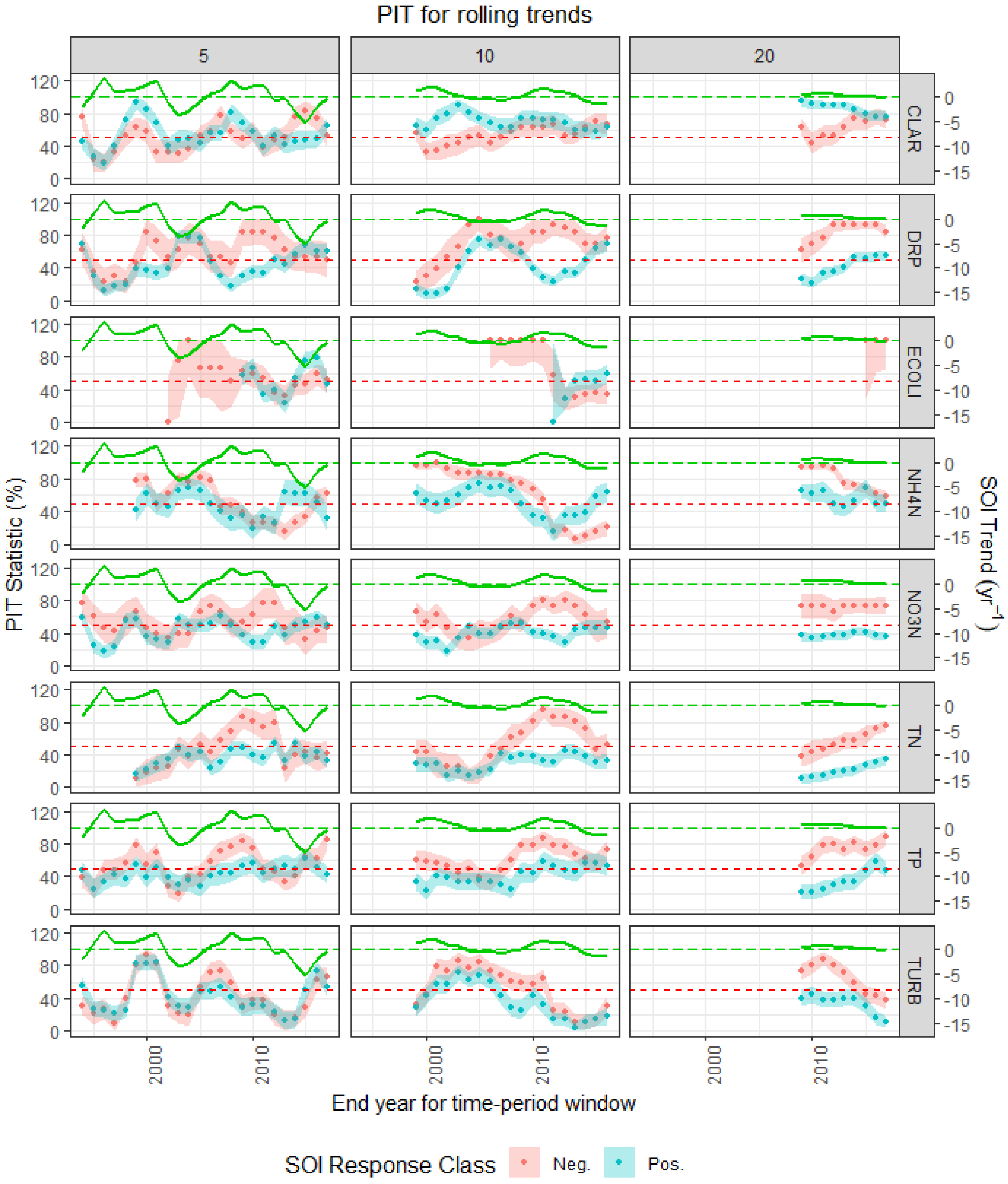

Figure 14. SOI trend and aggregate water quality trends (represented by PIT statistics) for NRWQN sites grouped by SOI response class.The columns represent the three trend period lengths and the rows represent the water quality variables. The SOI trend is the green line on each panel. The PIT statistics for each SOI response class are shown by the red and blue points and lines and the ribbon represents the 95% confidence interval for PIT.

Figure 15 illustrates the relationship between the SOI trend and trend magnitude for DRP in each of the 15 time-period windows of 10-year duration. The relationship between the RSS values and SOI correlation (shown by the red regression line in Figure 15) corresponds to the direction of the SOI trend in each time-period window. For example, for the time-period window ending 2000 the regression slope was strongly positive, which corresponds to the positive SOI trend for that window (see Figure 12). For the time-period window ending 2017 the regression slope was strongly negative, which corresponds to the negative SOI trend for that window (see Figure 12).

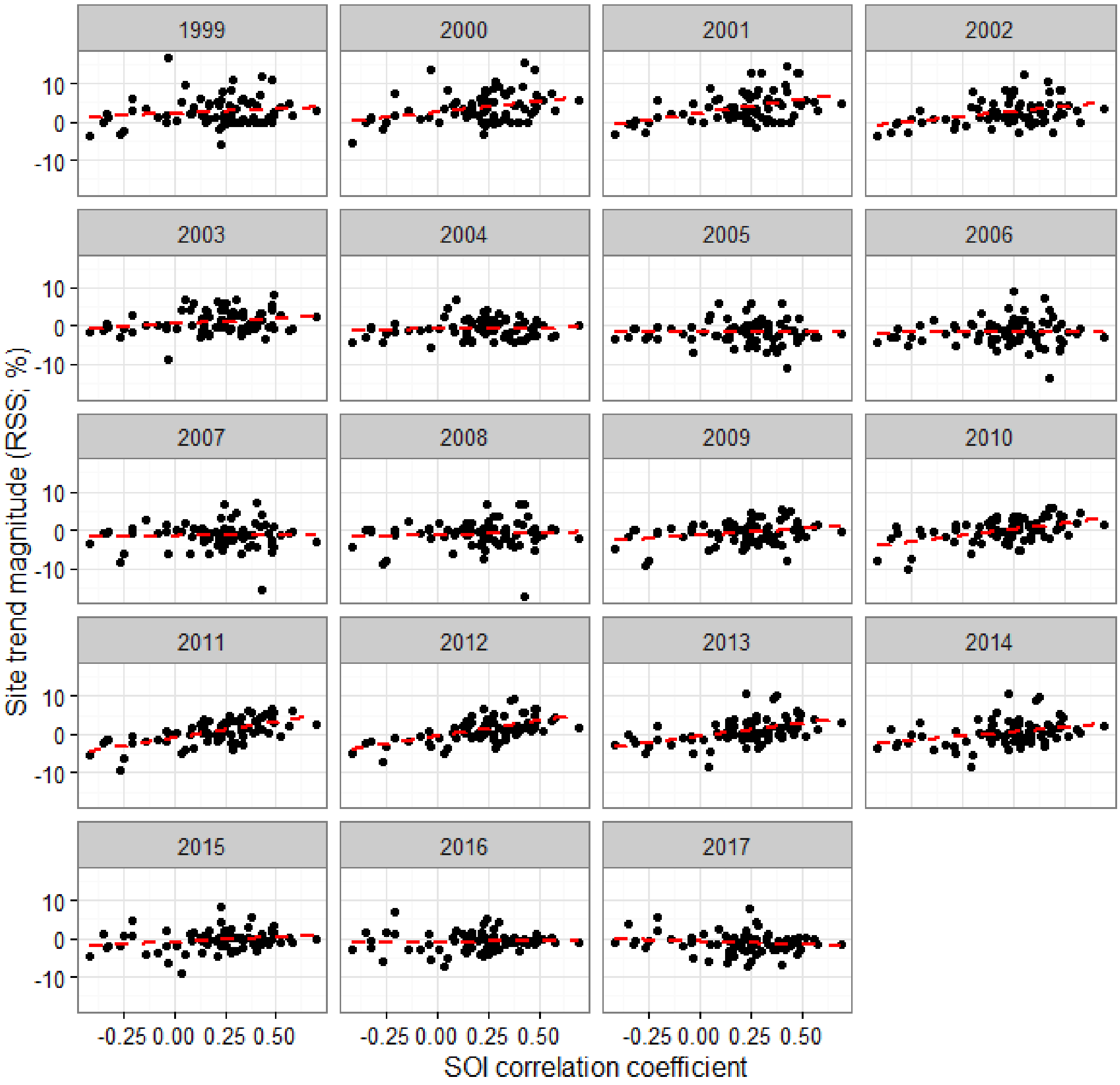

Figure 15. Example of the relationship between the SOI trend and the magnitude of DRP trends for the 10-year time-period windows. Each panel represents a time-period window, labelled according to the end year of the window, each of which has a different linear trend in the SOI (see Figure 12 for SOI trend directions and magnitudes for each window). The x-axis represents the SOI correlation coefficient for the site.

The details of fitting Equation 1 to the site RSS values for each water quality variable and all time-period durations are shown in Table 7. Most models were statistically significant indicating that variation in between-site trend magnitude is partly explained by the combination of observation - SOI correlation and the linear trend in the SOI for the time-period window.

The variation in site RSS values explained by the models differed between water quality variables and was highest for DRP, NH4N and TN and lowest for CLAR, ECOLI and TP (Table 7). In general, the variation in RSS explained by the models increased with time-period length. The models explained in the order of 1%, 4% and 20% of the RSS values, depending on the variable for time-period lengths of 5, 10 and 20-years (Table 7).

Table 7. Explanation of between site variation in RSS by observation - SOI correlation and the linear trend in the SOI for the time-period window. Each model reported in the table represents the application of Equation 1 to the site RSS and SOI correlations for each time period window within each of the three time period lengths (5, 10 and 20 years).

Variable	**Time-period length (Years)**	**Number of cases**	**Variation explained (R^2 %)**	**P-value**
CLAR	5	1837	0.3	0.080
	10	1460	4	**0.000**
	20	693	16	**0.000**
DRP	5	1837	9	**0.000**
	10	1460	20	**0.000**
	20	693	27	**0.000**
ECOLI	5	705	2	**0.000**
	10	405	2	**0.009**
	20	7	16	0.705
NH4N	5	1452	1	**0.000**
	10	1460	4	**0.000**
	20	693	29	**0.000**
NO3N	5	1837	3	**0.000**
	10	1460	7	**0.000**
	20	693	24	**0.000**
TN	5	1452	1	**0.000**
	10	1460	6	**0.000**
	20	693	27	**0.000**
TP	5	1837	1	**0.000**
	10	1460	5	**0.000**
	20	693	18	**0.000**
TURB	5	1837	0.8	**0.001**
	10	1460	1	**0.001**
	20	693	19	**0.000**

The details of fitting Equation 2 to the time-period window PIT statistics for each water quality variable and time-period duration are shown in Table 8. Most models were statistically significant indicating that variation in PIT between time-period windows is partly explained by the combination of SOI response class and the linear trend in the SOI for the time-period window.

The variation in PIT statistics explained by the models was reasonably uniform among the water quality variables and increased with time-period length (Table 8). For most variables, the models explained in the order of 10%, 30% and 80% of the variation in PIT statistics, depending on the water quality variable, for time-period lengths of 5, 10 and 20-years respectively.

Table 8. Explanation of variation in PIT statistics by SOI response class and the linear trend in the SOI for the time-period window.

Variable	**Time-period length (Years)**	**Number of cases**	**Variation explained (R^2 %)**	**P-value**
CLAR	5	48	11	0.166
	10	38	45	**0.000**
	20	18	87	**0.000**
DRP	5	48	43	**0.000**
	10	38	57	**0.000**
	20	18	83	**0.000**
ECOLI	5	26	3	0.894
	10	18	37	0.087
NH4N	5	38	11	0.256
	10	38	19	0.064
	20	18	85	**0.000**
NO3N	5	48	29	**0.002**
	10	38	54	**0.000**
	20	18	99	**0.000**
TN	5	38	11	0.267
	10	38	31	**0.005**
	20	18	94	**0.000**
TP	5	48	23	**0.010**
	10	38	44	**0.000**
	20	18	87	**0.000**
TURB	5	48	11	0.170
	10	38	12	0.233
	20	18	93	**0.000**

5 Discussion

The primary purposes of the analyses reported here were to investigate three aspects of variability in water quality trends;

1. variation in the confidence of trend evaluations,
2. variation in trends with time-period length, and
3. variation in trends with time-period window.

The first set of analyses (section 4.1) established that variation in the confidence of trend assessments was linked to a variety of factors but most strongly linked to trend magnitude itself. The range of the 90th confidence intervals (a measure of precision of in trend magnitude) was positively related to trend magnitude (RSS) and was positively related to confidence in trend direction (*p*-value). The trend period length was also moderately negatively correlated the with confidence in predicted trend direction, while the precision of the data (proportion of unique observations) for DRP, TP and TN were moderately positively correlated to the confidence in predicted trend direction. Based on these observations, we were able to approximately define the minimum detectable trend magnitude at a given level of confidence for each water quality variable, for a given time-period length (Table 4). In general, as the time-period length increases, smaller trends can be detected with confidence.

This information provided by Table 4 can inform the design of monitoring and trend evaluation strategies. If a trend magnitude of interest is defined, the information in Table 4 can be used to determine the minimum trend-period length required to detect trends of interest at a given level of confidence. Or, alternatively, if trend-period length and a trend magnitude of interest are both defined, but it is found that this magnitude/time period combination is unlikely to be detected with sufficient confidence, consideration can be given to increasing the frequency of monitoring.

The second set of analyses (section 4.2) established that trend magnitudes tend to decrease with increasing time-period length. This indicates that comparing trends evaluated over different time-period lengths is inappropriate. Apparent decreases in trend magnitude may be due to the dampening effect of lengthening time period. Further, as was demonstrated in section 4.3.3, other confounding factors, such as climate, can influence trend magnitude and direction, and different period lengths will be subject to different climate conditions (even if some of the periods overlap).

The third set of analyses (section 4.3) established that trends are sensitive to the time-period of analysis (i.e., time-period window; Figure 8). This means that there can be large variation in the magnitude of trends at individual sites when the trend period window is shifted (e.g., changing a ten-year trend period from 2007 – 2016 to the period 2008-2017). The analyses also established that the changes in trend magnitudes over many sites between trend periods are, to some extent, consistent because there can be large differences in the proportion of sites with improving (and conversely, degrading) trends between trend periods (Figure 9). This indicates that the trend directions over many sites are to some extent synchronous (i.e., exhibiting the same change from positive to negative and vice versa) between trend periods.

The sensitivity of trends to the time-period window has implications for the reporting of trends as part of state of environment reporting. Regional and national state of environment reports are produced at regular intervals, often of between two and five years, and generally, water quality trends are presented in these reports. Our results indicate that there will often be large fluctuations in the proportions of trends indicating improving or degrading conditions in adjacent reporting periods.

Based on earlier work by Scarsbrook *et al.* (2003), we hypothesised that the observed synchronisation of site trends is partly driven by climatic forcing. We represented climatic forcing by the SOI which measures the El Niño Southern Oscillation (ENSO) climate pattern. Part of the evidence for the involvement of the ENSO climate pattern in the synchronisation of site trends is our observation that the volatility of site trend magnitudes and directions reduces with increasing time-period duration (Figure 8, Figure 9). This observation is consistent with the pattern of reducing volatility in the monotonic trend in the SOI as time-period increases (Figure 12).

We showed that there are associations between monthly water quality observations and the monthly value of the SOI at many sites and that some of these were relatively strong i.e., > 0.25; Figure 10). Our analysis was simplistic; we did not test for the effect of lags on the correlation between water quality and SOI. However, because our results were based on flow adjusted water quality data, we assume the observed patterns are not simply due to the influence of climate on river flows (Scarsbrook *et al.,* 2003). Furthermore, we showed that water quality observations at baseline sites were similarly associated with the monthly value of the SOI as at impact sites (Table 5). This suggests that the observed variation in water quality is associated with aspects of natural climate variability (e.g., temperature and hydrological regimes) rather than with the influence of changing land management practices in association with climate variation.

We also showed that site water quality observations could be either positively or negatively associated with the SOI (Figure 10). Scarsbrook *et al.* (2003) had previously shown that a significant proportion of the variation in the direction of the association could be explained by six climate regions that were a simplification of the eight rainfall and three temperature regions of New Zealand (Salinger and Mullan, 1999). However, in this study, the mapped patterns of SOI response class membership did not strongly suggest that the relationship between SOI and water quality was regional (Figure 11). We therefore used statistical classification models to attempt to explain spatial variation in the SOI response classes. The models explained the SOI response class membership as a function of both geographic coordinates (i.e., similar to the geographically defined climate regions used by Scarsbrook *et al.*, 2003) and several characteristics of each site's catchment (e.g., elevation, geology, topography). Our hypothesis was that the geographic region may differentiate differences in the regional climate's response to the SOI and that catchment characteristics may differentiate the between site water quality response to regional climate. Although our statistical models did statistically discriminate SOI response classes, the models were generally weak (i.e., no better than satisfactory; Appendix B, Table 10). In addition, the relationships fitted by the spatial model offered little insight into the mechanisms determining SOI response class membership. CLAR, TURB and TP were related to both North and East indicating geographic variation in the water quality – SOI association (Figure 11. This may be because variation in measurements of these variables are associated with the degree of surface runoff and this is more strongly associated with the immediate influence of climate (i.e., rainfall) than other catchment characteristics. We therefore conclude that this study has merely indicated that SOI response classes vary in geographic and environmental space and that the mechanisms underlying these responses are yet to be determined.

Finally, our analyses showed that aggregate trends (i.e., the mean trend magnitude (RSS) and the proportion of improving trends (PIT)) were strongly associated with the combination of the SOI response class of sites and linear trend in the SOI for the time-period window under consideration (Figure 13, Figure 14, Table 7, Table 8). The outcomes were not precisely as hypothesised on Table 3, but there was a general tendency for the expected pattern to be followed. We consider that deviations from the hypothesised pattern can be expected because climate is not the only influence on trends and anthropogenic drivers such as increasing land use intensity will also be influencing the trends. We showed that the expected pattern was generally followed for all trend period durations that we analysed (i.e., 5, 10 and 20). However, we found that as time-period length increased, the SOI explained a greater proportion of the between site trend magnitudes (RSS; Table 7) and between time-period window proportion of improving trends (PIT; Table 8). This may indicate that site trends become more synchronous at longer time scales because differences in the short-term responses of individual catchments (e.g., lags and sudden responses to large changes in anthropogenic drivers) are dampened.

Our findings are consistent with those of a limited number of studies in other countries. In international studies of rivers draining agricultural areas, long-term records generally show a clear upward trend in nitrate concentrations since the 1960s (e.g., Betton *et al.*, 1991; Burt *et al.*, 1988; Burt and Worrall, 2009; Van Herpe and Troch, 2000). However, alongside the longer-term trends, other temporal variation of nitrate fluxes and concentrations in rivers, have been observed at seasonal, interannual (i.e., over 2–6 years) and decadal time scales (Burt and Worrall, 2009; Gascuel-Odoux *et al.*, 2010; Van Herpe and Troch, 2000). Inter-annual cycling in the norther hemisphere has been explained by climatic drivers and associated with the North Atlantic Oscillation (NAO). This has been demonstrated for nitrogen in rivers (Mitchell *et al.*, 1996; Monteith *et al.*, 2000; Straile *et al.*, 2003) but also for a range of water quality variables in both rivers and lakes (Weyhenmeyer, 2004).

A study of long-term records in 30 coastal rivers of western France by Gascuel-Odoux *et al.* (2010) indicated inter-annual cyclic behaviour in the fluxes and concentrations of nitrate. The study used deterministic hydrological modelling to highlight that the behaviour results from the interaction of climate, hydrology and land management practices. The study showed that the causes of the observed behaviours was complex. They found that the timing and amount of nitrate leached at the scale of soil profiles is strongly controlled by the balance of the fertilisation and by the rate of the biological processes in the soil. Inter-annual variations in leaching was therefore related to climate variables (temperature, rainfall) and to management practices. However, at the catchment scale, the behaviour of nitrate concentrations and fluxes was strongly influenced by the buffering effect of the groundwater system, which delayed delivery to the streams and dampened the variation in nitrate leaching.

Gascuel-Odoux *et al.* (2010) showed that although the inter-annual climatic variation can strongly influence nitrate concentrations and fluxes at the catchment scale, long term agricultural changes were the main drivers of the long-term trends. However, catchment hydrology induced large variations in the dynamics of the response to climatic and anthropogenic drivers. Gascuel-Odoux *et al.* (2010) concluded this variability presents difficulties to assessing the effects of pollution mitigation measures.

Our study indicates that attributing trends to causes is complex and that the role of climate is significant and needs to be accounted for. The results of our study do not mean that climate is responsible for long term changes in water quality. However, our results indicate that climate exerts considerable influence over water quality at inter-annual time scales. Because our trend analyses always involve periods in which climate is variable, attributing trends to causes will need to control for climate variation. Our study suggests two relatively simple statistical treatments of trend data may improve the quality of inferences drawn from trend analyses.

First, climate indices such as the SOI could be treated as a covariate in trend analysis of individual sites in much the same way that flow adjustment is performed. Second, large scale studies seeking to explain variation in trends over many sites could include the influence of the SOI, as defined by the correlation of the monthly observations with the monthly value of the SOI, as an explanatory variable. This type of analysis would benefit from the inclusion of additional explanatory variables such as of changes in land use or land management (e.g., (Snelder, 2018). This type of approach could statistically control for the influence of the SOI and evaluate the component of the trend that was attributable to land use or land management. Finally, the SOI is one index that measures an aspect of climate variation. It may be that more appropriate or additional measures of climate variability are appropriate. We therefore recommend that further research on the role of climate on water quality and practical methods to account for the effect of climate variation on water quality measurements is undertaken.

Acknowledgements

We thank Dan Elder, Annabelle Ellis, Sheree De Malmanche and Chris Daughney (Ministry for the Environment) for project support and feedback. We thank Abi Loughnan and the LAWA team for data updates. We thank the many regional council staff who provided additional data and information about monitoring programmes. We thank Graham Bryers (NIWA) for assistance with NRWQN data. We thank Scott Larned, Maree Patterson of Horizons Regional Council and Brydon Hughes for reviewing early drafts of this report.

References

Allan, R., J. Lindesay, and D. Parker, 1996. El Niño Southern Oscillation & Climatic Variability. CSIRO publishing.

Betton, C., B.W. Webb, and D.E. Walling, 1991. Recent Trends in NO3-N Concentration and Loads in British Rivers. Vienna Symposium, August., pp. 169–180.

Breiman, L., 2001. Random Forests. Machine Learning 45:5–32.

Breiman, L., J.H. Friedman, R. Olshen, and C.J. Stone, 1984. Classification and Regression Trees. Wadsworth, Belmont, California.

Burt, T.P., B.P. Arkell, S.T. Trudgill, and D.E. Walling, 1988. Stream Nitrate Levels in a Small Catchment in South West England over a Period of 15 Years (1970-1985). Hydrological Processes 2:267–284.

Burt, T.P. and F. Worrall, 2009. Stream Nitrate Levels in a Small Catchment in South West England over a Period of 35 Years (1970–2005). Hydrological Processes: An International Journal 23:2056–2068.

Cutler, D.R., J.T.C. Edwards, K.H. Beard, A. Cutler, K.T. Hess, J. Gibson, and J.J. Lawler, 2007. Random Forests for Classification in Ecology. Ecology 88:2783–2792.

Diaz, H.F., H.F. Diaz, and V. Markgraf, 1992. El Niño: Historical and Paleoclimatic Aspects of the Southern Oscillation. Cambridge University Press.

Gascuel-Odoux, C., P. Aurousseau, P. Durand, L. Ruiz, and J. Molenat, 2010. The Role of Climate on Inter-Annual Variation in Stream Nitrate Fluxes and Concentrations. Science of the Total Environment 408:5657–5666.

Hanley, J.A. and B.J. McNeil, 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology 143:29–36.

Hastie, T. and R. Tibshirani, 1990. Generalized Additive Models. Chapman and Hall.

Larned, S., T. Snelder, and M. Unwin, 2016. Water Quality in New Zealand Rivers; Modelled Water Quality State. NIWA CLIENT REPORT, NIWA, Christchurch, New Zealand.

Larned, S., A. Whitehead, C. Fraser, T. Snelder, and J. Yang, 2018a. Water Quality State and Trends in New Zealand Rivers. Analyses of National-Scale Data Ending in 2017. NIWA, NIWA, Christchurch, New Zealand.

Larned, S., T. Snelder, A. Whitehead, and C. Fraser, 2018b. Water Quality State and Trends in New Zealand Lakes. NIWA Client Report, NIWA, Christchurch, New Zealand.

McBride, G.B., 2019. Has Water Quality Improved or Been Maintained? A Quantitative Assessment Procedure. Journal of Environmental Quality.

Mitchell, M.J., C.T. Driscoll, J.S. Kahl, G.E. Likens, P.S. Murdoch, and L.H. Pardo, 1996. Climatic Control of Nitrate Loss from Forested Watersheds in the Northeast United States. Environmental Science & Technology 30:2609–2612.

Monteith, D.T., C.D. Evans, and B. Reynolds, 2000. Are Temporal Variations in the Nitrate Content of UK Upland Freshwaters Linked to the North Atlantic Oscillation? Hydrological Processes 14:1745–1749.

Mosley, M.P., 2000. Regional Differences in the Effects of El Niño and La Niña on Low Flows and Floods. Hydrological Sciences Journal 45:249–267.

Salinger, M.J. and A.B. Mullan, 1999. New Zealand Climate: Temperature and Precipitation Variations and Their Links with Atmospheric Circulation 1930–1994. International Journal of Climatology: A Journal of the Royal Meteorological Society 19:1049–1071.

Scarsbrook, M.R., C.G. McBride, G.B. McBride, and G.G. Bryers, 2003. Effects of Climate Variability on Rivers: Consequences for Long Term Water Quality Analysis1. JAWRA Journal of the American Water Resources Association 39:1435–1447.

Smith, D.G. and R. Maasdam, 1994. New Zealand's National Water Quality Network. 1. Design and Physio-Chemical Characterisation. New Zealand Journal of Marine & Freshwater Research 28:19–35.

Snelder, T., 2018. Assessment of Recent Reductions in E. Coli and Sediment in Rivers of the Manawatū-Whanganui Region: Including Associations between Water Quality Trends and Management Interventions. LWP Client Report, LWP Ltd, Christchurch, New Zealand.

Snelder, T. and C. Fraser, 2018. Aggregating Trend Data for Environmental Reporting. WP Client Report 2018-01, LWP Ltd, Christchurch, New Zealand.

Snelder, T.H., S.T. Larned, and R.W. McDowell, 2018. Anthropogenic Increases of Catchment Nitrogen and Phosphorus Loads in New Zealand. New Zealand Journal of Marine and Freshwater Research 52:336–361.

Straile, D., D.M. Livingstone, G.A. Weyhenmeyer, and D.G. George, 2003. The Response of Freshwater Ecosystems to Climate Variability Associated with the North Atlantic Oscillation.

Svetnik, V., A. Liaw, C. Tong, and T. Wang, 2004. Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. P. Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Vardi, and G. Weikum (Editors). Springer, Cagliari, Italy, pp. 334–343.

Unwin, M., T. Snelder, D. Booker, D. Ballantine, and J. Lessard, 2010. Predicting Water Quality in New Zealand Rivers from Catchment-Scale Physical, Hydrological and Land Cover Descriptors Using Random Forest Models. NIWA Client Report: CHC2010-0.

Van Herpe, Y. and P.A. Troch, 2000. Spatial and Temporal Variations in Surface Water Nitrate Concentrations in a Mixed Land Use Catchment under Humid Temperate Climatic Conditions. Hydrological Processes 14:2439–2455.

Weyhenmeyer, G.A., 2004. Synchrony in Relationships between the North Atlantic Oscillation and Water Chemistry among Sweden's Largest Lakes. Limnology and Oceanography 49:1191–1201.

Whitehead, A., 2018. Spatial Modelling of River Water-Quality State. Incorporating Monitoring Data from 2013 to 2017. NIWA Client Report, NIWA, Christchurch, New Zealand.

Wild, M., T. Snelder, J. Leathwick, U. Shankar, and H. Hurren, 2005. Environmental Variables for the Freshwater Environments of New Zealand River Classification. Christchurch.

A1 Correlation matrices for characteristics of the trends and observation datasets

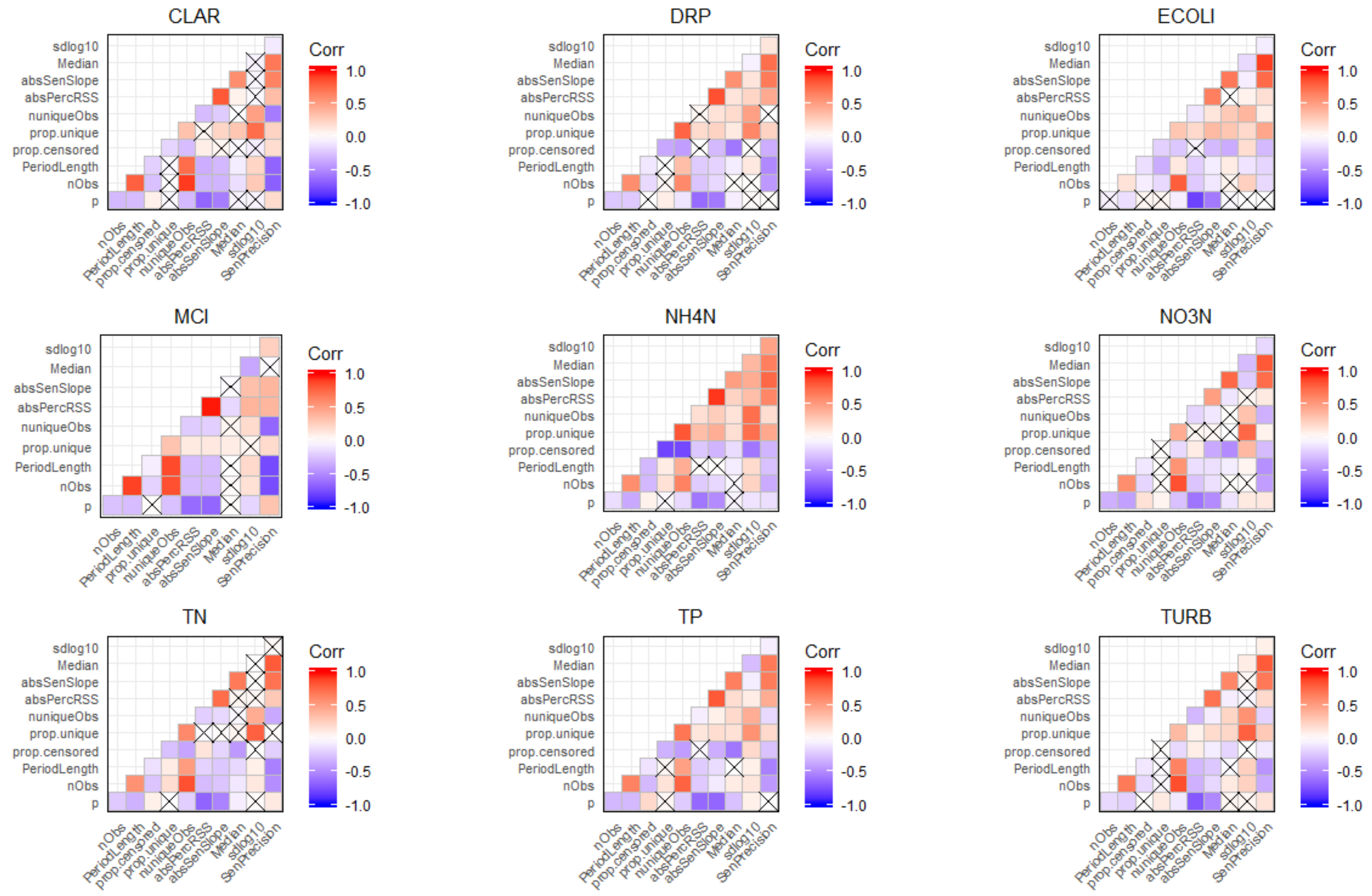

Figure 16. Matrices showing the correlation between all trend analysis outcomes and the characteristics of the observation datasets and the evaluated trends, by water quality variable. This includes all the variables in Table 2 as well as the p-value (p) and Sen slope 90% confidence intervals (SenPrecision). Values in the cells are the Spearman rank correlation coefficient. Cells with no values had Spearman rank correlation coefficient p-values > 0.05

A2 Statistical modelling of SOI response classes

A2.1 Methods

The strength of the relationship between individual water quality observations at a site and the SOI was characterised by the Pearson correlation between deseasonalised SOI and the deseasonalised water quality observations at each site as described in Section 3.5.2**Error! Reference source not found.**. Sites were assigned to SOI response classes based on the sign of the Pearson correlation coefficient and irrespective of the level of confidence (i.e., the *p*-value) of the coefficient. This decision is justifiable on the basis that the *p*-values indicate confidence in the sign of the correlation coefficient at individual sites, and hence their assignment to SOI response classes. The acceptable risk of making incorrect assignments at individual sites is arbitrary (i.e., alpha value of 0.05 is arbitrary but is generally accepted). The confidence in the SOI response class assignment for individual sites can be disregarded when considering all sites globally because it is assumed that incorrect classifications will cancel each other (i.e., as many sites will be misclassified as positive as sites misclassified as negative). Therefore, the "face value" of each site's correlation coefficient (i.e., the direction indicated the Pearson correlation coefficient) was used to assign each site to an SOI response class, irrespective of the *p*-value.

Statistical classification models were used to discriminate between sites assigned to the negative and positive SOI response classes based on many potential predictor variables including geographic coordinates and catchment characteristics. The statistical modelling used the same approach as those used to generate predications of river and lake water quality state (Whitehead, 2018, Fraser et al. 2019) and other predictions of water quality at regional to national scales (e.g., Larned et al., 2016; Unwin *et al.*, 2010). The approach combines the monitoring site locations with a spatial framework provided by a database representing the national river network. The database contains a range of variables that represent the characteristics of the catchments upstream of every segment of the river network (Wild *et al.*, 2005). The statistical spatial models used the same catchment characteristics as Larned et al. (2016) and Snelder *et al.* (2018) as predictor variables in the models (Table 9).

Table 9. Predictor variables used in spatial models of SOI response class.

Predictor	Abbreviation	Description	Unit
Geography and topography	North	Site location coordinate	m
	East	Site location coordinate	m
	usArea	Catchment area	m^2
	usLake	Proportion of upstream catchment occupied by lakes	%
	usCatElev	Catchment mean elevation	m ASL
	usAveSlope	Catchment mean slope	degrees
	segAveElev	Segment mean elevation	degrees
Climate and flow	usAvTWarm	Catchment averaged summer air temperature	degrees C x 10
	usAvTCold	Catchment averaged winter air temperature	degrees C x 10
	usAnRainVar	Catchment average coefficient of variation of annual rainfall	mm y^{-1}r
	usRainDays10	Catchment average frequency of rainfall > 10 mm	days $month^{-1}$
	usRainDays20	Catchment average frequency of rainfall > 20 mm	days $month^{-1}$
	usRainDays100	Catchment average frequency of rainfall > 100 mm	days $month^{-1}$
	segAveTCold	Segment mean minimum winter air temperature	degrees C x 10
	usFlow	Estimated mean flow	m^3 s^{-1}
Geology*	usHard	Catchment average induration or hardness value	Ordinal*
	usPhos	Catchment average phosphorous	Ordinal*
	usParticleSize	Catchment average particle size	Ordinal*
Land cover	usPastoral	Proportion of catchment occupied by combination of high producing exotic grassland, short-rotation cropland, orchard, vineyard and other perennial crops (LCDB3 classes 40, 30, 31, 33)	Proportion
	usIndigForest	Proportion of catchment occupied by indigenous forest (LCDB3 class 69)	Proportion
	usUrban	Proportion of catchment occupied by built-up area, urban parkland, surface mine, dump and transport infrastructure (LCDB3 classes 1,2,6,5)	Proportion
	usScrub	Proportion of catchment occupied by scrub and shrub land cover (LCDB3 classes 50, 51, 52, 54, 55, 56, 58)	Proportion
	usWetland	Proportion of catchment occupied by lake and pond, river and estuarine open water (LCDB3 classes 20, 21, 22)	Proportion
	usBare	Proportion of catchment occupied by bare ground (LCDB3 classes 10, 11, 12,13,14, 15)	Proportion
	usExoticForest	Proportion of catchment occupied by exotic forest (LCDB3 class 71)	Proportion
	usGlacial	Proportion of catchment occupied by ice (LCDB3 classes 14)	Proportion

We fitted the classification model using Random Forest (RF) modes; the same type of statistical model that Whitehead (2018) and Fraser *et al.* (2019) used to model water quality state but in a classification mode. RF models are a machine learning method that automatically detects and fits non-linear relationships and high order interactions, both of which we expected may be involved in discriminating SOI response class membership due to the sites being located over large environmental gradients (Unwin *et al.*, 2010). Determining and specifying non-linearities and interactions in more traditional statistical models such as linear models requires significant skill and insight by the modeller into the relationships being modelled. Because RF models automatically detect and fit these complex relationships, it is more likely that results generated by different modellers will be comparable.

A RF model is an ensemble of individual classification and regression trees (CART). In a regression context, CART partitions the observations (the site SOI response classes) into groups that minimise the misclassification of sites based on a series of binary rules or splits that are constructed from the predictor variables. CART models require no distributional assumptions and automatically fit non-linear relationships and high order interactions. However, single regression trees have the limitations of not searching for optimal tree structures, and of being sensitive to small changes in input data (Hastie and Tibshirani, 1990). RF models reduce these limitations by using an ensemble of trees (a forest) and making predictions based on the average of all trees (Breiman 2001). Detailed descriptions of RF models and their diagnostic tools are described in detail in Breiman (2001) and Cutler *et al.* (2007).

Misclassification rates and receiver operating curves (ROCs) were used to evaluate the performance of the classification models. ROC plots show the true positive rate (sensitivity) against the false positive rate (1−specificity) as the probability threshold used to classify a case varies from 0 to 1 (Hanley and McNeil, 1982). Good models have high true positive rates and relatively small false positive rates and, therefore, have ROC plots that rise steeply at the origin, and level off near the maximum value of 1. The ROC plot for a poor model lies near the diagonal, where the true positive rate equals the false positive rate for all thresholds. The model performance was quantified using the area under the ROC curve (AUC). AUC is a measure of the performance of a binary classifier, with a good model having an AUC near 1, while a poor model will have an AUC near 0.5 (Hanley and McNeil, 1982). The following rules of thumb were used to express the quality of the model indicated by AUC in narrative terms: very good (0.9 – 0.8); good (0.8 - 0.7); satisfactory (0.7 - 0.6); poor (0.6 - 0.5).

The relationships between predictor and response variables represented by RF models were represented by importance measures and partial dependence plots (Breiman 2001; Cutler et al. 2007). The importance of each predictor variable is indicated by the degree to which prediction accuracy decreases when the response variable is randomly permuted. Importance is defined in this study as the total decrease in node impurities (a measure of misclassification rate) associated with splits based on the predictor variable, averaged over all trees.

A partial dependence plot is a graphical representation of the marginal effect of a predictor variable on the response variable, when the values of all other predictor variables are held constant. The benefit of holding the other predictors constant (generally at their respective mean values) is that the partial dependence plot effectively ignores their influence on the response variables. Partial dependence plots do not perfectly represent the effects of each predictor variable, particularly if predictor variables are highly correlated or strongly interacting, but they do provide an approximation of the modelled predictor-response relationships that are useful for model interpretation (Cutler et al. 2007).

RF models can include any of the original set of predictor variables that are chosen during the model fitting process. Inclusion of marginally important and correlated predictor variables does not degrade the performance of the RF models. However, these predictor variables may be redundant (i.e. their removal does not affect model performance) and their inclusion can complicate model interpretation. We used the predictor elimination procedure (Svetnik *et al.*, 2004) to remove redundant predictor variables from the models. The procedure first assesses the model error (miss-classification rate) using a 10-fold cross validation process. The predictions made to the hold out observations during cross validation are used to estimate the miss-classification rate and its standard error. The model's least important predictor variables are then removed in order, with the miss-classification rate and its standard error being assessed for each for each successive model. The final, 'reduced' model is defined as the model with the fewest predictor variables whose error is within one standard error of the best model (i.e. the model with the lowest cross validated miss-classification rate). This is equivalent to the "one standard error rule" used for cross validation of classification trees (Breiman *et al.*, 1984).

An alternative approach is to choose the model with the smallest error. We used the former procedure as it retains fewer predictor variables than the latter procedure, while achieving an error rate that is not different, within sampling error, from the "best solution". Importance levels for predictor variables were not recalculated at each reduction step to avoid over-fitting (Svetnik *et al.*, 2004).

All calculations were performed in the R statistical computing environment (R Development Core Team 2009) using the randomForest package and other specialised packages.

A2.2 Results

All RF models were able to significantly discriminate positive and negative response classes on the basis of the site geographic coordinates and catchment characteristics of the upstream catchment (Table 10). Model performance as indicated by AUC were at least satisfactory (0.7 - 0.6) for CLAR, DRP. NO3N, TN and TURB but were poor (0.6 - 0.5) for ECOLI, NH4N, and TP (Table 10).

Table 10. Performance of the RF models of SOI response class.

Variable	Misclassification rate (%)	AUC
CLAR	36	0.64
DRP	21	0.62
ECOLI	39	0.56
NH4N	35	0.59
NO3N	29	0.70
TN	35	0.71
TP	55	0.53
TURB	39	0.63

Model simplification reduced the number of predictor variables in all models considerably. The CLAR model included three predictors (North, usPhos and East; Figure 17). The TN model included two predictors (usRain and usArea) and the NO3N model included two predictors (segTmin and usTmin) (Figure 17). All other models included two predictors (Figure 17). The models of SOI response class for CLAR, TURB and TP included only North and East as predictors indicating geographic variation in the direction of water quality – SOI association. This may be because variation in measurements of these predictors are associated with the degree of surface runoff and this is more strongly associated with climate (i.e., rainfall) than other catchment characteristics. For the other water quality variables, the predictors and partial plots characterising the fitted relationships between the predictor and response (probability the site belonged to the Positive class) were difficult to interpret. A mix of predictors were included in these models that suggest SOI response class is associated with catchment geology (i.e., usCalc, usPhos), land cover (i.e., usIntensiveAg, usWetland, usUrban), climate (i.e., usRain, usTmin, segTwarm) and as well as river size (i.e., usArea, MeanFlow) and catchment topography (usElev). It is difficult to understand why these variables are associated with SOI response class and we conclude that the results indicate that the causes of water quality response to variation in climate are complex.

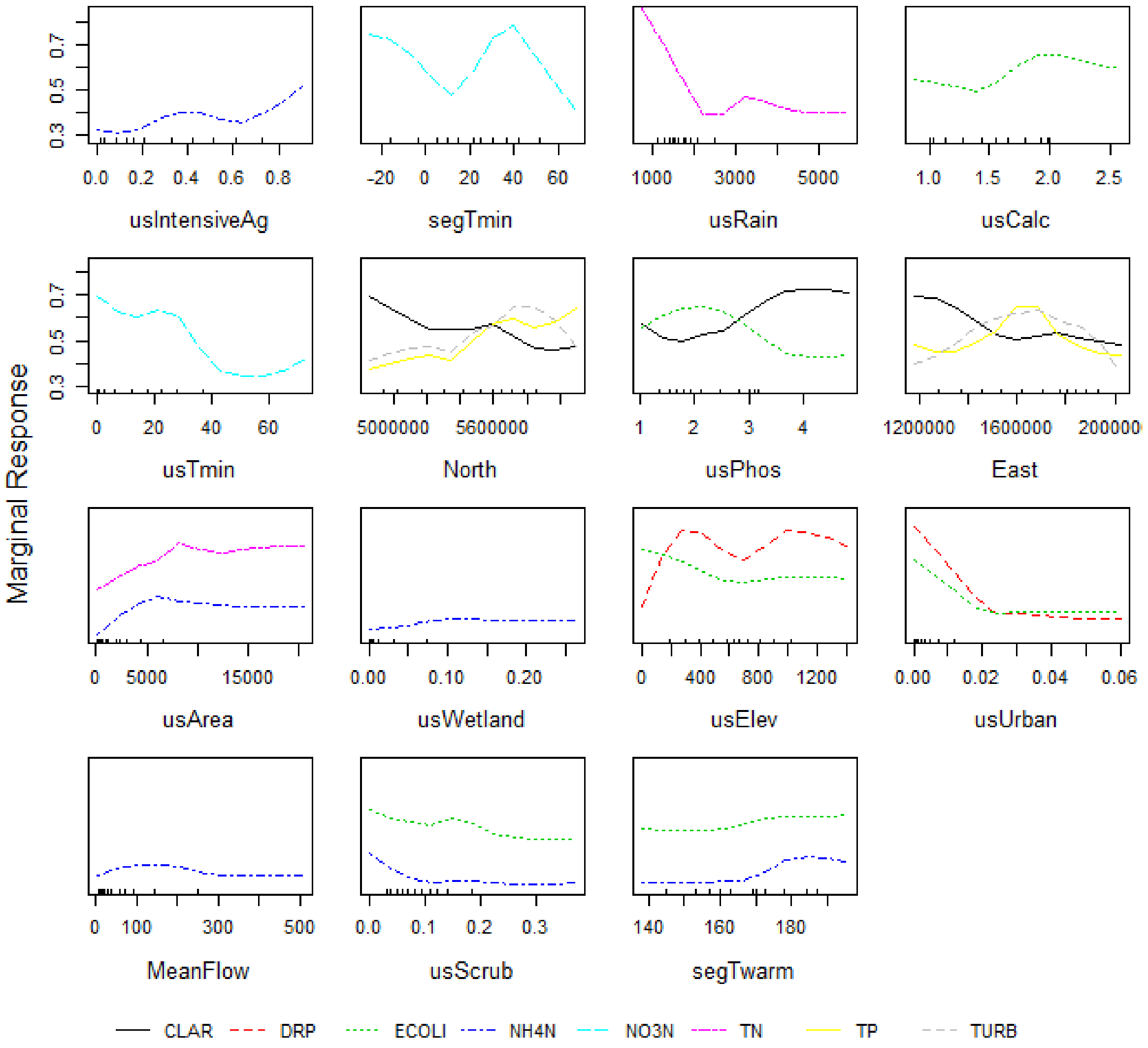

Figure 17. Partial plots characterising the relationships between the 15 predictors included in at least one of the RF models and the response (probability the site belonged to the Positive class) fitted by the simplified RF models.