

Analysis of Water Quality in New Zealand Lakes and Rivers

Prepared for Ministry for the Environment

February 2015

Prepared by:

Scott Larned
Ton Snelder*
Martin Unwin
Graham McBride
Piet Verburg
Hilary McMillan

For any information regarding this report please contact:

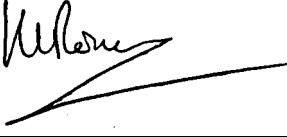

Scott Larned

+64-3-343 7834
scott.larned@niwa.co.nz

National Institute of Water & Atmospheric Research Ltd
PO Box 8602
Riccarton
Christchurch 8011
Phone +64 3 348 8987

*LWP Ltd
P.O Box 70
Lyttelton 8048

NIWA CLIENT REPORT No: CHC2015-033
Report date: February 2015
NIWA Project: MFE15503

Quality Assurance Statement		
	Reviewed by:	Helen Rouse
	Formatting checked by:	Tracy Webster
	Approved for release by:	Clive Howard-Williams

© All rights reserved. This publication may not be reproduced or copied in any form without the permission of the copyright owner(s). Such permission is only to be given in accordance with the terms of the client's contract with NIWA. This copyright extends to all forms of copying and any storage of material in any kind of information retrieval system.

Whilst NIWA has used all reasonable endeavours to ensure that the information contained in this document is accurate, NIWA does not give any express or implied warranty as to the completeness of the information contained herein, or that it will be suitable for any purpose(s) other than those specifically contemplated during the Project or agreed by NIWA and the Client.

Contents

Executive summary	7
1 Introduction	9
2 Data acquisition, organisation and processing	11
2.1 River data	11
2.2 Lake data	13
2.3 Data acquisition and organisation	14
2.4 Data processing and grooming	15
3 Analysis methods.....	20
3.1 Censored values	20
3.2 Grouping river and lake sites	20
3.3 Water quality state	21
3.4 Trend analyses	21
4 Results – river and lake state.....	29
4.1 River state – nutrients, ECOLI and CLAR	29
4.2 River state - MCI	33
4.3 Lake state	36
5 Trend results	39
5.1 River trends - nutrients, ECOLI and CLAR	39
5.2 River MCI trends	57
5.3 Lake water quality trends	61
6 Discussion	67
6.1 State and trends in water quality	67
6.2 New procedures for censored data and trend analyses.....	68
7 Acknowledgements	70
8 References.....	71
Appendix A A new approach to water quality trend assessment	74
1. Summary.....	1
1.1. Implementation	1

2.	Problems with the traditional trend testing approach	3
2.1	Unfortunate properties	4
2.2	Unfortunate inferences	4
3	A alternative approach.....	5
3.1	A refined set of questions.....	5
3.2	How may these questions be answered?	5
4	Potential issues with the proposed approach.....	8
5	Implementation.....	9
6	References.....	11
Appendix B	Nutrient concentration thresholds to achieve periphyton	1
1	Introduction	2
2	Methods.....	2
2.1	Periphyton data	2
2.2	Approximation of periphyton biomass as Chlorophyll <i>a</i>	3
	Explanatory variables	4
2.3	Derivation of concentrations thresholds	7
3	Discussion	11
4	References.....	14

Tables

Table 2-1:	River water quality variables included in this study.	11
Table 2-2:	Lake water quality variables included in this study.	13
Table 2-3:	Measurement procedures for water quality variables.	18
Table 3-1:	Parameters used in assessments of trend importance for river water quality variables.	27
Table 3-2:	Parameters used in assessments of trend importance for lake water quality variables.	28
Table 4-1:	Number of river monitoring sites by REC class and water quality variable that were included in the state analyses of nutrients, ECOLI and CLAR.	30
Table 4-2:	Number of river monitoring sites by REC Source-of-Flow class used in the state analysis of MCI scores.	33
Table 4-3:	Number of lake monitoring sites by class and water quality variable that were included in the state analyses.	36
Table 5-1:	Correspondence between pairs of estimates of RSSE with flow adjustments using TopNet synthetic flows and observed flows.	39

Table 5-2:	Number of river monitoring sites by REC class and water quality variable that were included in the 10-year trend analyses of nutrients, ECOLI and CLAR.	44
Table 5-3:	Numbers of sites in trend categories for 10-year, flow-adjusted trends across REC classes.	47
Table 5-4:	Number of river monitoring sites by REC class and water quality variable that were included in the 20-year trend analyses of nutrients, ECOLI and CLAR.	49
Table 5-5:	Numbers of sites in trend categories for 20-year, flow-adjusted trends across REC classes.	52
Table 5-6:	Numbers of sites in trend categories for the 25-year datasets for NRWQN sites.	56
Table 5-7:	Number of river monitoring sites within REC class that were included in the 10-year trend analyses of MCI scores.	58
Table 5-8:	Number of sites by lake class eligible for the 10 year trend analysis.	63
Table 5-9:	Number of lake sites in 10-year trend categories across all lake elevation × depth classes.	66

Figures

Figure 4-1:	Locations of river water quality monitoring sites used for state analyses of nutrients, ECOLI and CLAR.	31
Figure 4-2:	River water quality state in REC Source-of-Flow classes.	32
Figure 4-3:	Locations of river monitoring sites used for state analyses of MCI scores.	34
Figure 4-4:	Distribution of site-median MCI scores for hard-bottom streams, within REC Source-of-Flow class.	35
Figure 4-5:	Locations of monitoring sites used for state analyses of lake water quality.	37
Figure 4-6:	Lake water quality state in elevation × depth classes.	38
Figure 5-1:	Changes in the number of river monitoring sites that met the filtering rules for each water quality variable versus the period of site operation.	41
Figure 5-2:	Comparison of RSSSE values for the 10-year period ending 2013 for sites with flow adjustment performed using TopNet modelled flows and the observed flows.	42
Figure 5-3:	Correspondence between raw and flow adjusted trends for each water quality variable, for 10-year trends.	43
Figure 5-4:	Locations of river water quality monitoring sites used for 10-year trend analyses of nutrients, ECOLI and CLAR.	45
Figure 5-5:	Summary of 10-year flow adjusted trends.	46
Figure 5-6:	Locations of river water quality monitoring sites used for 20-year trend analyses of nutrients, ECOLI and CLAR.	50
Figure 5-7:	Summary of 20-year flow adjusted trends.	51
Figure 5-8:	Effect of flow adjustment on 20-year trends. The red line is 1 to 1.	53
Figure 5-9:	Summary of 25-year, flow-adjusted trends at NRWQN sites.	55
Figure 5-10:	Changes in the number of river macroinvertebrate monitoring sites that met the filtering rules versus the period of site operation.	57
Figure 5-11:	Locations of river macroinvertebrate monitoring sites used for 10-year trend analyses of MCI scores.	59

Figure 5-12:	Summary of 10-year MCI trends by REC class. Box-and-whisker plots show the distributions of site trends within REC classes.	60
Figure 5-13:	Changes in the number of lake monitoring sites that met the filtering rules for each water quality variable versus the period of site operation.	62
Figure 5-14:	Locations of lake water quality monitoring sites used for 10-year trend analyses of water quality variables.	64
Figure 5-15:	Summary of 10-year trends in lake water quality variables, within elevation × depth classes.	65

Executive summary

This report provides the methods and results for the first stage of a two-stage project to develop a national-scale freshwater quality model for the New Zealand Ministry for the Environment (MfE). The model is to be used to: 1) predict water quality state and trends in rivers and lakes across New Zealand; 2) assess spatial and temporal variation in the environmental pressures that affect freshwater quality; and 3) identify and quantify relationships between pressures and water quality. In the first stage of the project we updated national river and lake water quality databases and analysed water quality state and trends at river and lake sites with adequate data. In the second stage of the project we will compile agricultural pressure data and incorporate the water quality and pressure data in national-scale spatial models.

This report represents the completion of the first stage of the project. It consists of detailed methods for data processing and analysis, a concise summary of national-scale state and trends, supplementary files with site-specific information, and minimal interpretation. The analyses in the report were aligned where possible with attributes in the National Policy Statement for Freshwater Management of 2014 (NPS-FM). The supplementary files are spreadsheets with spatial data and results of the water-quality state and trends analyses for every site that met basic criteria for sampling duration and frequency. The reason for the focus on methods in this report and tabulated results in the supplementary files is that those results are intended to be used in a synthesis report by MfE, as proposed in the national Environmental Reporting Bill.

The methods used in the state and trend analyses include three advances over the methods used in previous national-scale analyses. First, we replaced censored and substituted values in the datasets with imputed values. This step alleviated problems caused by multiple tied values. Second, we developed and applied a procedure for flow-adjusting trends at all river monitoring sites, including those without flow recorders. This procedure predicts mean daily flow at each site on each sampling date using the national hydrological model TopNet. Third, we applied a new, two-step procedure for trend analysis. The new procedure first uses confidence intervals to determine trend direction where possible, then uses recognised thresholds such as NPS-FM 'bottom-lines' (the boundaries between C and D bands for ecological attributes) and critical periods to assess trend importance. A detailed rationale and methods for the censored-data and trend analysis procedures are in Appendix A.

For the assessments of trend importance in this report, we used several NPS-FM bottom lines and the MfE water clarity guideline as thresholds, and set the critical time period at 10 years. Therefore, we considered a trend important if it was projected to exceed the designated threshold within 10 years. We note that this approach is nominal and normative because the importance of any given trend is dependent on context and viewpoint. The thresholds that we used could be replaced with many other guidelines, trigger values or numeric objectives, and the critical period may be shortened or extended. Our assessments served to demonstrate the new method, not to produce definitive lists of important and unimportant water quality trends.

In addition to quantifying site-specific water quality state and trends, we summarized state and trends across river and lake monitoring sites grouped by two environmental classifications. For the river water quality summaries, monitoring sites were grouped by River Environment Classification (REC) Source- of-Flow class. For the lake water quality summaries, monitoring sites were grouped using a surface-elevation × maximum-depth classification with eight classes.

The summaries of river state indicated that variation in median nutrient and *Escherichia coli* (ECOLI) concentrations and black-disk water clarity (CLAR) was partly explained by REC-class membership. Nutrient and ECOLI concentrations were lowest in mountain Source-of-Flow classes and in the CX/Lk and CD/Lk classes. In contrast, nutrient and ECOLI concentrations were highest in the low-elevation Source-of-Flow classes, irrespective of climate. CLAR was highest and least variable at sites in the CX climate classes and the mountain, lake and hill Source-of-Flow classes. Variation in median Macroinvertebrate Community Index (MCI) scores was partly explained by REC class. In general, median MCI scores were highest in the cooler and wetter climate categories (e.g. CX, CW) and the hill Source-of-Flow categories, followed by the mountain, and then low-elevation classes.

The summaries of river trends over the 10-year period from 2004 – 2013 indicated that there were 10 times as many sites with improving trends in total phosphorus (TP) as degrading trends, three times as many sites with improving trends in dissolved reactive phosphorus (DRP) as degrading trends, and 1.5 to 1.9 times as many sites with improving trends in ECOLI, CLAR and ammoniacal nitrogen (NH₄N) as decreasing trends. In contrast, there were 1.5 times as many sites with degrading trends in oxidised nitrogen (NO₃N) as improving trends. There were nearly equal numbers of site with improving trends in total nitrogen (TN) as degrading trends. There were more sites with degrading trends in MCI scores than improving trends, but these only accounted for 17% of the sites analysed; for the remainder, the data were insufficient to confidently determine trend direction.

A comparison of 10- and 20-year trends in river water quality revealed several changes in the balance of improving and degrading trends: 1) a predominance of degrading 20-year trends in TN shifted to roughly equal proportions of degrading and improving 10-year trends; 2) the predominance of improving trends in DRP and TP increased between the 10- and 20-year periods; 3) the predominance of improving trends in NH₄N decreased between the 10- and 20-year periods; 4) the predominance of improving trends in CLAR in the 10-year trends is not apparent in the 20-year trends. The predominance of degrading trends in NO₃N has persisted between the 10- and 20-year periods, although the magnitude has shifted (twice as many degrading trends in the 20-year period versus 1.5 times as many in the 10-year period). Sites in the National River Water Quality Network (NRWQN) were used to assess 25-year trends. The predominant directions of 25-year trends in water quality variables were generally consistent with those in the 20-year trend assessment.

The analyses of lake state indicated that median nutrient and phytoplankton chlorophyll *a* (CHLA) concentrations and Trophic Level Index (TLI) levels was relatively high in most low-elevation, shallow lakes and lower in high-elevation, deep lakes. Median clarity measured as Secchi depth (SECCHI) was relatively low in low-elevation, shallow lakes and higher in high-elevation, deep lakes.

The summaries of 10-year lake trends indicated that bottom-water dissolved oxygen (DO_{bottom}) concentrations were decreasing in many of the deeper lakes (> 15 m depth). TN concentrations were decreasing and NO₃N concentrations increasing in lakes in all elevation × depth classes. No consistent differences were apparent in trends in low- and high-elevation lakes. Improving trends in CHLA, TLI, NH₄N, TN and TP occurred at three to five times as many sites as degrading trends. In contrast, degrading trends in NO₃N occurred at six sites and an improving trend at one site, and degrading trends in DO_{bottom} occurred at 13 sites, with improving trends at two sites.

Datasets and the supplementary files with state and trend statistics for monitoring sites were provided to MfE in electronic form on 17 April 2015 and 21 May 2015.

1 Introduction

As part of its national environmental reporting programme, the Ministry for the Environment (MfE) sought to develop a national freshwater-quality model (“the model”) that represents the state of, and pressures on, freshwater for the past 25 years. MfE has commissioned a NIWA-led team to develop the model as a consultancy project. The model will provide information on:

1. The current state of river and lake water quality throughout New Zealand and temporal trends in water quality in recent history;
2. The pressures that may affect water quality and changes in these pressures over time;
3. The relationships between pressures and water quality.

In particular, the model will enable relationships to be established between freshwater quality and environmental factors that explain variation in water quality (including natural factors and pressures arising from land and water resource use). The project aims to establish these relationships for time-averaged water quality conditions as a function of catchment conditions, and for temporal trends in water quality as a function of changes in pressures at and between sites.

The model will include geospatial and temporal information for the past 25 years including:

1. Water quality for river and lake monitoring sites in the national monitoring network, which is composed of state-of-environment monitoring sites operated by NIWA and the 16 regional councils and unitary authorities;
2. Data characterising aspects of the natural environment (e.g. temperature, rainfall, geology) that influences water quality;
3. Data characterising human-induced pressures that may influence water quality, including temporal and spatial data describing agricultural land use pressures, such as nitrogen leached into soil.

The national water quality model is to be developed in two stages:

- Stage 1: update national river and lake water quality databases through data compilations and processing, and analyse water quality state and trends at monitoring sites in the databases;
- Stage 2: obtain agricultural pressure data and undertake national-scale spatial modelling of water quality state and changes in state over time.

This report completes Stage 1. The report consists of detailed methods for data processing and analysis, a concise summary of national-scale state and trends, supplementary files with site-specific results, and minimal interpretation. The reason for the focus on methods and tabulated results is that the results are intended to be used in a synthesis report by MfE, as proposed in the national Environmental Reporting Bill.

The analyses in this report were aligned where possible with attributes incorporated in the National Policy Statement for Freshwater Management of 2014 (NPS-FM)¹. The NPS-FM requires regional councils, through their regional plans, to set freshwater objectives that provide for freshwater

¹ <http://www.mfe.govt.nz/publications/fresh-water/national-policy-statement-freshwater-management-2014>

values, and to set limits and develop management actions to achieve those objectives. The NPS-FM identifies multiple attributes to assist regional councils in developing numeric objectives for rivers and lakes, and policies (including limits) for achieving those objectives. The NPS-FM attributes that apply to rivers are periphyton biomass (as chlorophyll *a*), and nitrate, ammonia, and *Escherichia coli* (*E. coli*) concentrations. The attributes that apply to lakes are phytoplankton biomass (as chlorophyll *a*), and total nitrogen, total phosphorus, ammonia, *E. coli* and planktonic cyanobacteria concentrations. For attributes for which there were sufficient data, the national bottom lines (minimal acceptable state) were used as thresholds in trend analyses. For some of the remaining NPS-FM attributes, there were too few data to carry out trend analyses (e.g., lake planktonic cyanobacteria, lake *E. coli*).

The methods used in the current study include several advances on our previous national-scale water-quality analyses: 1) mean daily flows at river monitoring sites that lack flow recorders were estimated using NIWA's national hydrological model TopNet; 2) water-quality measurements that were reported by the data suppliers to be “below detection limit” (left-censored) or “above reporting limit” (right-censored) were replaced with randomised imputed values using routines based on regression-on-order statistics and survival analysis; 3) analyses and assessments of trends used confidence intervals to determine trend direction (if possible), and exceedance thresholds to assess trend importance (for those cases where trend direction was identified).

The new trend analysis procedure provides an alternative to the “traditional” procedure that tests the hypothesis that a trend has zero slope. The new procedure should help to avoid the common misinterpretation of a trend test result that fails to attain statistical significance when testing the “nil hypothesis”²—that conditions are therefore “stable” or “being maintained”. This is generally not a valid inference. The new procedure tests the *direction* of a trend. If the direction of a trend cannot be confidently inferred, the result is stated as “insufficient data to reveal the trend direction”, rather than “not statistically significant”. If a trend direction can be inferred, the method goes on to assess its environmental importance. Importance is determined by the trend magnitude and is defined by a recognised threshold toward which a trend may be heading (e.g., an NPS-FM bottom line), and a critical period to reach that threshold should the trend continue uninterrupted. A trend is categorised as important if the threshold is projected to be crossed within the critical period, or unimportant if not. A detailed rationale and methods development for the censored-data and trend analysis procedures are set out in Appendix A.

In this report, we used a selection of NPS-FM bottom lines and guideline values as trend thresholds, and 20 years (approximately one human generation) as the critical period for assessing trend importance. A caveat is required about this particular approach. The approach that we used is one of many possible approaches for assessing trend importance. Unlike objective methods for making inferences about trend direction, methods for assessing trend importance are subjective and nominal because the importance of any given trend is context-dependent. The thresholds that we used could be replaced with other guidelines, trigger values or numeric objectives, and the critical period may be shortened or extended. Our assessments served to demonstrate the new methodology, not to produce a definitive list of important water quality trends.

² The phrase “null hypothesis” can include one-sided hypotheses (e.g., trend > 0) or interval hypotheses (trend is between two limits). However traditional trend tests consider a hypothesis that the trend is exactly zero, hence we identify it as a “nil hypothesis”.

2 Data acquisition, organisation and processing

New Zealand regional and district councils carry out state-of-environment monitoring at > 1000 river and lake sites. For the monitoring sites used in this report, monthly or quarterly monitoring has been underway for 5 to 25 years, and continues to the present. A variety of physical, chemical and biological indicators of water quality are measured at these sites. In addition, water quality and biological monitoring had been carried out by NIWA since 1989 at the 77 river sites that make up the National River Water Quality Network (NRWQN).

Regional council and NRWQN river and lake monitoring data are periodically acquired and federated into databases for national-scale state-of-environment reports and investigations of monitoring performance (e.g., Sorrell et al. 2006, Ballantine et al. 2010, Larned and Unwin 2012). In the current project, existing river and lake databases were updated with data that had been collected between the last sampling dates in the most recent national compilations (Verburg et al. 2010, Unwin and Larned 2013) and the end of 2013. In this section we describe the water quality variables, data sources and organisation of the river and lake databases, and explain the data processing procedures used to derive datasets suitable for state and trend analyses.

2.1 River data

We described river water quality using eight variables that correspond to physical, chemical and microbiological conditions, and macroinvertebrate community composition (Table 2-1). In this report, we use “river water quality” as a general term to refer to some or all of the eight variables. Unless otherwise stated, we made no distinction between data collected at regional council sites and NRWQN sites, and we refer to the sites collectively as the “river monitoring network”. Data corresponding to the physical, chemical and microbiological variables came from monthly or quarterly samples; macroinvertebrate data came from annual samples.

Table 2-1: River water quality variables included in this study.

Variable type	Variable	Abbreviation	Units
Physical	Clarity	CLAR	M
	Ammoniacal nitrogen	NH4N	mg/m ³
	Nitrate nitrogen	NO3N	mg/m ³
Chemical	Total nitrogen (unfiltered)	TN	mg/m ³
	Dissolved reactive phosphorus	DRP	mg/m ³
	Total phosphorus (unfiltered)	TP	mg/m ³
Microbiological	<i>Escherichia coli</i>	ECOLI	n/100 mL
Index	Macroinvertebrate Community Index	MCI	unitless

Visual water clarity (CLAR) is a measure of light attenuation due to absorption and scattering by dissolved and particulate material in the water column. Clarity is monitored because it affects primary production, plant distributions, animal behaviour, aesthetic quality and recreational values, and because it is correlated with suspended solids, which can impede feeding in fish and cause riverbed sedimentation. Visual clarity in rivers is generally measured *in situ* as the horizontal sighting

range of a black disc (MfE 1994). At a small number of sites, clarity is measured adjacent to the river with water samples in clarity tubes.

The five nutrient species (NO₃N, NH₄N, DRP, TN and TP) were included because they influence the growth of benthic river algae (periphyton) and vascular plants (macrophytes). Nutrient enrichment from point and non-point source discharges is strongly associated with intensive land use. Nutrient enrichment can promote excessive, 'nuisance' growth of periphyton and macrophytes that can, in turn, degrade river habitat, increase daily fluctuations in dissolved oxygen and pH, impede flows, block water intakes, and cause water colour and odour problems. There are two or more methods in use to measure NO₃N, NH₄N, DRP, TN and TP concentrations, and not all methods give comparable results. Some non-comparable nutrient data were excluded from the analyses (see Section 2.4).

As discussed in Section 1, the NPS-FM identifies attributes that are intended to assist regional councils in defining freshwater objectives, and justifiable policies (including limits) for achieving these objectives. The NPS-FM attributes for rivers include two forms of nitrogen, NO₃N and NH₄N, but these attributes are based on nitrate and ammonia toxicity, not on their potential effects on periphyton and macrophytes. The reason for omitting ecologically-based nutrient attributes is that periphyton and macrophyte growth is mediated by numerous, spatially variable factors such as flood frequency and riparian shading, in addition to nutrients. However, the NPS-FM does have an attribute based on periphyton, which indirectly requires nutrient management to prevent excessive growth (Snelder et al. 2013). NO₃N and NH₄N concentrations associated with toxic effects are generally much higher than concentrations associated with proliferations of periphyton and macrophytes.

The concentration of the bacterium *Escherichia coli* (ECOLI) is used as an indicator of human or animal faecal contamination and the risk of infectious human disease from waterborne pathogens in contact-recreation and drinking water. There is an NPS-FM attribute based on ECOLI concentrations that is related to the management of human health. The national bottom line for ECOLI concentrations for secondary contact recreation is 1000/100 ml (as a median). For primary contact recreation sites the *E. coli* concentration at the "minimum acceptable state" is 540/100 ml, as a 95th percentile. Note that this report only considers median *E. coli* concentrations. State and trends in 95th percentiles of *E. coli* would need to be examined separately.

The physical, chemical and microbiological variables described above are characterized by short-term variability. For a time-integrating variable, we used the New Zealand Macroinvertebrate Community index (MCI). The MCI is a community-level biotic indicator of general river health. MCI scores are calculated using tolerance values for the macroinvertebrate taxa that are present in benthic samples. Tolerance values are weighting factors that correspond to the relative abundance of taxa along stressor gradients. We used the non-quantitative MCI in lieu of the quantitative (qMCI) or semi-quantitative (sqMCI) forms of MCI because some council datasets do not include invertebrate abundance data (Stark and Maxted 2007). Non-quantitative MCI scores are based on presence/absence data which are widely available. In contrast to the monthly monitoring frequencies for physical and chemical variables and ECOLI, which are measured monthly or quarterly, the invertebrate samples used to calculate MCI scores are generally collected once each summer. Due to the difference in sampling frequency, trend analyses of MCI scores were carried out using a different procedure that that used for the other variables (see Section 3.4).

Three additional river water quality variables were initially considered for analysis: total suspended sediment, faecal coliforms and periphyton. These variables were omitted because several regional

councils had no corresponding data and most of the remaining council datasets comprised few sites or did not meet the sampling frequency and duration criteria we applied (Sections 3.3 and 3.4.1).

2.2 Lake data

We assessed lake water quality using nine variables that correspond to physical, chemical and biological conditions (Table 2-2). In addition to dissolved inorganic and total nutrients (described in Section 2.1.1), lake water quality variables were Secchi depth, bottom-water oxygen, phytoplankton biomass as chlorophyll *a*, and the Trophic Level Index.

Table 2-2: Lake water quality variables included in this study.

Variable type	Variable	Abbreviation	Units
Physical	Secchi depth	SECCHI	M
	Bottom-water dissolved oxygen	DObottom	mg/L
Chemical	Ammoniacal nitrogen	NH4N	mg/m ³
	Oxidised nitrogen	NO3N	mg/m ³
	Total nitrogen (unfiltered)	TN	mg/m ³
	Dissolved reactive phosphorus	DRP	mg/m ³
	Total phosphorus (unfiltered)	TP	mg/m ³
Phytoplankton	Chlorophyll <i>a</i>	CHLA	mg/L
Index	Trophic Level Index	TLI	unitless

Secchi depth (referred to as SECCHI) is a measure of water clarity and gives an indication of the amount of light-scattering and light-absorbing particulate and dissolved matter in lakes. SECCHI measures the maximum depth at which a black and white Secchi disk is visible to an observer at the lake surface.

Bottom-water dissolved oxygen (DObottom) concentration is the oxygen concentration at the deepest point in lakes where oxygen-depth profiles were measured. DObottom is a measure of the trophic state of lakes that stratify during summer; low DObottom concentrations is generally associated with high primary production and ecosystem respiration. DObottom oxygen also indicates the vulnerability of lakes to internal loading of phosphorus and ammonium, as these nutrients are released from sediments under anoxic (no oxygen) or hypoxic (low oxygen concentrations) conditions.

Chlorophyll *a* concentration (CHLA) is used as a measure of lake phytoplankton biomass. High chlorophyll *a* concentrations may occur during periods of high internal and/or external nutrient loading, and are the primary indicators of eutrophication. Phytoplankton chlorophyll *a* concentrations are also used to calculate Trophic Level Index scores, as described below.

The Trophic Level Index (TLI) is an indicator variable that summarises data related to trophic state and potential primary production. The TLI is used to classify New Zealand lakes into trophic classes (e.g., oligotrophic, eutrophic); TLI scores increase with increasing eutrophication. There are two versions of TLI in use in New Zealand, one with three variables (TLI3) and one with four variables

(TLI4) (Burns et al. 2000, Verburg et al. 2010). TLI3 scores are derived from log-transformed concentrations of CHLA, TN and TP. TLI4 uses SECCHI data in addition to CHLA, TN and TP concentrations. However, SECCHI data were not available for all lake monitoring sites in the current study. Moreover, SECCHI data are strongly influenced by factors that are independent of trophic state, such as fine glacial sediment and tannins. We used the TLI3 to maximize the number of sites in the TLI dataset. To ensure consistent calculations, we calculated TLI scores for all lakes in our national dataset, (using the formulae given by Sorrell et al. 2006, Section 2.2.2), and used these scores in lieu of TLI scores provided in council datasets.

2.3 Data acquisition and organisation

River water-quality monitoring data have been acquired from regional councils and NIWA for two recent studies for MfE, a national scale state-of-environment analysis (Ballantine et al. 2010) and an assessment of statistical power in the National Environmental Monitoring and Reporting (NEMaR) programme (Unwin and Larned 2013). For the Ballantine et al. (2010) study, all available data from the beginning of systematic river water quality monitoring until the end of 2007 were federated into a common format and stored in a customised MS-Access database. The database was updated in 2012 for the NEMaR study; after that update, the ending dates for individual sites and variables in the database ranged from 2010 and 2012. We refer to this dataset as the NEMaR database.

Since 2014, regional council river water quality monitoring data for the period 2004 to 2013 (inclusive) have been federated into the Land Air Water Aotearoa (LAWA) database³. We used these data to update the NEMaR database so that all sites had ending dates in December 2013. This involved matching sites in the LAWA database to their NEMaR counterparts, and further processing to resolve inconsistencies between the two datasets, as described in Section 2.4.1.

The NEMaR dataset included NRWQN data only to the end of 2012. To update these records, and to include any changes associated with retrospective data auditing over the intervening period, we reimported the full NRWQN dataset to December 2013, overwriting all previous data.

MCI scores in the NEMaR database were calculated from raw council and NRWQN invertebrate data, provided as taxa lists and counts or coded-abundance classes for each taxon (Unwin and Larned 2013). To resolve inconsistencies in taxonomic resolution among councils, we recalculated all scores using the taxonomic levels in the original MCI user guide (Stark and Maxted 2007). All scores were calculated using tolerance values for hard-bottom streams, for the reasons discussed by Unwin and Larned (2013).

To update the MCI data for the present study, we used MCI scores to the end of 2013 as received from LAWA. These scores are based on raw taxonomic data that has yet to be federated, and are thus potentially confounded by taxonomic inconsistencies similar to those we encountered in developing the NEMaR dataset, but resolving these was beyond the scope of the current project. Of 12,462 scores in the NEMaR dataset, 10,184 (81.7%) were identical with those provided by the councils, and 290 (2.3%) differed by more than two taxa, so we assume that any resulting errors are small. The LAWA datasets contained updated scores for all councils except Environment Bay of Plenty and Waikato Regional Council.

Our lake dataset was composed of data to the end of 2009 acquired for a previous MfE project (Verburg et al. 2010), and updated data acquired from regional councils and unitary authorities. For

³ Available for download at <http://www.lawa.org.nz/explore-data/freshwater/>

most variables and councils, the pooled data were current to at least mid-2013, and often to late 2014. We also added data for 26 lakes sampled intensively from 1992 to 1996 (Burns and Rutherford 1998), which we refer to as the Burns dataset. Profile data (i.e., temperature and dissolved oxygen measured along a vertical transect from surface to either maximum depth, or to well below the hypolimnion) were available for all lakes in the Burns dataset, and for 81 lakes in the pooled council dataset, for a total of 92 lakes after merging lakes common to both datasets.

2.4 Data processing and grooming

The original river and lake data used in this study, as received from council and NRWQA records, varied widely in reporting formats, reporting conventions for variable names, site identifiers, date and time formats, units of measurement, and other data structure elements. We imported the datasets into an MS-Access database, and applied a consistent set of reporting conventions. We manually inspected the datasets and used time-series plots and other diagnostics to identify and correct errors. Common errors included mislabelled site-names, georeferencing errors, incorrect units and data transcription errors. We also developed a flagging system to attach information to individual data points; some council datasets include similar flags, but most do not. Flags include censored data, unit conversions (e.g., from mg/L to µg/L), and values that were synthesised from other data (e.g., MCI, TLI). For consistency, we used the NRWQN data when an NRWQN site coincided with a regional council river monitoring site (24 river sites in 11 regions). Twelve LAWA river sites were omitted from our final rivers dataset because they could not be reliably matched to the River Environment Classification (REC); most of these sites are on drainage channels. An additional eight LAWA sites were omitted because they mapped onto lakes rather than rivers. Our final rivers database comprised 844 sites, representing the 77 NRWQN sites plus 767 of the 1,023 state-of-environment monitoring sites listed in the LAWA dataset.

In addition to water quality data, the following spatial data were associated with each river monitoring site: site name, location and regional council identifier (if available), NZMS260 grid reference (converted from NZTM as necessary), and NZReach number (as defined in REC geodatabase). The following spatial data were associated with each lake monitoring site: lake name, regional council identification code, NZMS260 grid reference (converted from NZTM as necessary), mean and maximum depth, surface area, perimeter and water surface elevation. In addition, each lake was assigned a class based on the lake typology in Sorrell et al. (2006), and the two classes used in the NPS-FM, seasonally stratified lakes and polymictic lakes. After compiling the site data, each lake was assigned a unique identifier.

Lakes were assigned to NPS-FM classes based on maximum depth. Seasonally stratified lakes are stratified throughout summer whereas polymictic lakes do not stratify or destratify repeatedly during summer, typically as a result of their shallow depths compared with seasonally stratified lakes. Lake stratification is associated with high variability in surface-water (epilimnion) nutrient concentrations. Mean depth is a good indicator of the potential to stratify during summer. However, mean-depth data were not available for most lakes. Maximum depth data were available for most lakes, and we used 15-m maximum depth as the cut-off between polymictic and stratifying lake classes. This cut-off is consistent with the recommendations of the lake experts panel that informed the NPS-FM.

2.4.1 River data processing

River water-quality data were processed in several steps to ensure that the data were accurate and the datasets used for analyses were internally consistent.

Step 1. Comparable field and laboratory methods. The first data processing step was to assess methodological differences for all variables. For most of the variables, two or more measurement procedures were represented in the datasets. We grouped data by procedure, then pooled data for which different procedures gave comparable results, based on assessments in Davies-Colley et al. (2012) and Larned and Unwin (2012). Data measured using the less-common and non-comparable methods were eliminated. Table 2-3 lists the most common procedures used for each variable, and the procedures corresponding to data retained for analysis.

The data produced by multiple procedures used to measure ECOLI, NO₃N, CLAR and MCI were pooled, based on the assumption that the different procedures gave comparable results. In contrast, some procedures used to measure TN and TP are unlikely to give comparable results. Most councils and the NRWQN use the alkaline persulfate digestion method and unfiltered water samples. A smaller group of councils uses a sulphuric acid digestion procedure to measure total Kjeldahl nitrogen (TKN) and calculates TN as TKN + NO₃N. At least one council uses filtered samples for the data labelled TN and TP, although the filtered samples are more correctly labelled total dissolved nitrogen and phosphorus. The alternative methods could generate substantial differences in reported TN and TP concentrations. Therefore, only TN and TP measured by the persulfate digestion method with unfiltered samples were retained for analysis.

Step 2. Error correction and adjustment. The second data processing step was to manually inspect the data, correct identifiable errors, and rescale data. We used quantile plots to identify and remove gross outliers for each variable. Where necessary, values were adjusted to ensure consistent units of measurement across all datasets. Site location information was used to match sites to the correct NZReach and correct errors in site names.

Step 3. Comparison of NEMaR and LAWA data. The third data processing step was to compare the NEMaR and LAWA databases for all site × date × variable records common to both datasets (N = 255,595). To identify and characterise differences between the two datasets, we calculated the ratio $R_{\text{NEMaR:LAWA}}$ for each datum, and assigned the record to one of four categories based on the value of this ratio. The first and largest category (N = 240,346; 94.0% of the total), comprising records for which $0.99 \leq R_{\text{NEMaR:LAWA}} < 1.01$, was taken as representing equality, after allowing for minor differences ($\pm 1\%$) associated with different levels of decimal rounding.

We defined ratio bands for the second category to capture records which differed by a factor of two, reflecting the common practice of reporting left-censored data (i.e., values below detection limit) as half the detection limit. This category, defined by threshold values of $0.49 \leq R_{\text{NEMaR:LAWA}} < 0.51$ or $1.99 \leq R_{\text{NEMaR:LAWA}} < 2.01$, comprised 5,488 records, representing 2.1% of the total. We resolved such cases on a council × variable basis, generally taking the higher of the two values while taking note of any flags indicating left-censored values, and preserving these flags in the final dataset.

The third category was intended to capture records for which the difference was less than a factor of two, but more than the $\pm 1\%$ threshold assumed to indicate equality, and was defined by $0.51 \leq R_{\text{NEMaR:LAWA}} < 0.99$ or $1.01 \leq R_{\text{NEMaR:LAWA}} < 1.99$. A total of 8,354 records fell into this category, representing 3.3% of the total. On inspection, it became clear that the great majority of these records were associated with rounding errors which exceeded $\pm 1\%$, but rarely exceeded $\pm 10\%$. (e.g., 0.032 vs. 0.03; 0.168 vs. 0.17). In all cases, we resolved such records by using whichever dataset was reported to the highest number of significant digits. Other instances were associated with nutrient data which had been synthesised in the NEMaR dataset (e.g., TN = NO₃N + Total Kjeldahl Nitrogen),

but were available at source in the LAWA dataset; for these cases we gave priority to the LAWA dataset.

The final category comprised 1,400 records (0.5% of the total) for which the NEMaR and LAWA data differed by more than a factor of two. Most of these records (1,015) were associated with two datasets: 660 TN records provided by Northland Regional Council, where the NEMaR data had been incorrectly synthesised; and 355 DRP records from Gisbourne District Council (GDC). Inspection of the raw GDC data indicated that DRP was the only variable common to both NEMaR and LAWA, and also showed evidence of inconsistently recorded dates. Since we required accurate sample dates to flow-adjust each dataset, we omitted all GDC data from our final dataset. In all other cases where $R_{\text{NEMaR:LAWA}}$ differed by more than a factor of two, we gave priority to the LAWA dataset.

Step 4. Censored and substituted values. The final data-processing step concerned censored and substituted values. For several water-quality variables, some true values are too low or too high for laboratories to measure with precision. For very low values of a variable, the minimum acceptable precision corresponds to the “detection limit” for that variable; for very high values of a variable, the minimum acceptable precision corresponds to the “reporting limit” for that variable. Cases where values of variables are below the detection limit or above the reporting limit are often indicated by the data entries “<DL” and “>RL”, where DL and RL are the laboratory detection limit and reporting limit, respectively. In some cases, the censored values had been replaced (by the monitoring agency) with substituted values to facilitate statistical analyses. Common substituted values are $0.5 \times \text{detection limit}$ and $1.1 \times \text{reporting limit}$. Water-quality datasets from New Zealand rivers and lakes often include DRP, TP and NH_4N measurements that are below detection limits, and ECOLI and CLAR measurements that are above reporting limits. Although common, replacement of censored values with constant multiples of the detection and reporting limits can result in misleading results when statistical tests are subsequently applied to those data (Helsel 2012). Data that we received that were composed of censored and substituted values were replaced with imputed values using the procedures in Section 3.1.

Table 2-3: Measurement procedures for water quality variables. MCI procedures are from Stark et al. (2001). Procedures retained: data generated by the procedures in this column, and corresponding monitoring sites, were retained for analysis in this study.

Variable	Measurement procedures	Procedures retained
ECOLI	Colilert QuantiTray 2000 Membrane filtration	Both procedures (presumed to give comparable results)
NO3N	Nitrate-N, filtered, Ion chromatography Nitrate-N + nitrite-N (or “NNN”), filtered, cadmium reduction Nitrate + Nitrite-N – Nitrite-N (filtered, Azo dye colourimetry)	All procedures (nitrite presumed to be negligible in unpolluted water)
TN	Unfiltered, persulfate digestion Filtered, measured as dissolved inorganic+organic nitrogen Mixed, by Kjeldahl digestion (TKN + NNN)	Unfiltered, persulfate digestion
TP	Unfiltered, persulfate digestion Filtered, measured as dissolved inorganic+organic phosphorus	Unfiltered, persulfate digestion
DRP	Filtered, molybdenum blue colourimetry	Filtered, molybdenum blue colourimetry
CLAR	Black-disk Horizontal clarity tube	Both procedures (presumed to give comparable results)
MCI	Collection procedures C1, C2, C3, C4 Processing procedures P1, P2, P3	All procedures (presumed to give comparable presence/absence data for calculating non-quantitative MCI scores)
SECCHI	Secchi disk Horizontal clarity tube	Secchi disk
CHLA	Acetone pigment extraction, spectrofluorometric measurement. In situ and laboratory fluorometry	Acetone pigment extraction, spectrofluorometric measurement.
TLI	Calculated by the authors	
DO _{bottom}	In situ, automatic profilers Grab-samples, DO measured on boat	Both procedures (presumed to give comparable results)

2.4.2 Lake data processing

Data processing and grooming for the updated lakes database involved some of the same considerations as for the rivers database, particularly with regard to managing alternative methodologies for variables such as TN. There were also several processing issues specific to lakes. The lakes database was inherently more complex than the rivers database, due to the presence of multiple sampling sites within lakes, samples taken from multiple depths at each site; and, replicate samples taken at the same site x depth. To accommodate these factors, we partitioned the suite of water quality variables to be analysed into four groups: variables measured at or near the surface (nutrients, ECOLI); variables typically measured from an integrated sample of a range of depths (e.g.,

0-20 m; CHLA); variables measured at or near the lake bed (DO); and variables not associated with any specific depth (Secchi, TLI).

The first step in reducing multiple measurements into single values for each lake x sampling date x variable combination was excluding sites and depths that were not fully representative of each lake. At the site level, we used the available metadata for each lake, together with our own knowledge of specific lakes, to flag nonrepresentative sites. These sites excluded were generally located at or near a shoreline (rather than in the main body of the lake) or in secondary arms and outlets of larger lakes. Most lakes had at least one suitable monitoring site, but twelve (including all eight lakes in the Otago region: Lakes Dunstan, Hawea, Johnson, Onslow, Tuakitoto, Waihola, Wakatipu, and Wanaka) were left with no suitable sites and dropped out of our working dataset. For lakes with multiple suitable sampling sites, we averaged across sites within dates.

At the sample level, we relied on water depth information in the raw dataset to inform our choice as to data suitability. For records where sample depth was explicitly specified in the raw data, we used the shallowest depth (typically 0-1 m) for which data were available. Records where sample depth was indicated only by a descriptive term clearly referring to a surface or near-surface sample (e.g., 0-25m tube, composite, epilimnion, surface, top) we accepted the data as given; samples described in other ways (including anoxic, bottom, deep, hypolimnion, middle, photic zone, thermocline) were rejected. We then estimated TLI for all records with sufficient data, as described in Section 2.2, and added these to the pooled nutrient/ECOLI/SECCHI data. The final dataset, excluding DO_{bottom} (which was limited to lakes with profile data), represented 233 sites on 156 lakes, of which 106 lakes were represented by a single site; 30 by two sites; 14 by three sites; five by four sites; and one (Lake Wairarapa) by five sites.

Extracting meaningful bottom DO information from the available profile data was potentially confounded by profiles where the probe had bottomed out, so that the deepest readings may have been taken in unconsolidated sediment below the bottom of the water column. In some cases such records were readily apparent due to the presence of small negative DO readings, but other instances – where the recorded DO value was exactly zero – were more problematic. In the absence of any flags identifying values below detection limits, zero is a legitimate value for bottom DO, particularly for highly eutrophic lakes. We therefore estimated bottom DO for each lake and sampling date by identifying the maximum depth for which a non-negative DO record was available, accounting for profiles at multiple sites within some lakes, and then averaging all DO values recorded at that depth.

We conclude this section with a general caveat about data processing. Of the many issues described in this section, by far the most pervasive was inconsistent reporting of censored data, particularly for values below detection limits. In both the NEMaR and LAWA river water quality databases, and also in the newly created lakes database, we found evidence of detection limits varying across time for the same variable at the same site; data that were left-censored but not flagged as such; data that were flagged as left-censored but appeared to be well above detection limits elsewhere in the contributing data set; data reported as half or double the prevailing detection limit, and data that appeared to consist of substituted values (e.g., $0.5 \times$ detection limit), but were not flagged as such. Despite our best efforts, it is likely that some errors were undiscovered and therefore uncorrected.

3 Analysis methods

3.1 Censored values

In this study, we used a three step process to impute replacements for censored values. For comparative purposes we also performed equivalent analyses using the traditional substitution rules (i.e., left censored values substituted with values corresponding to one half the detection limit and right censored values increased by 10%).

Step 1. Left-censored data. We manipulated “less than” data using ROS (Regression on Order Statistics) to impute replacement values (Helsel 2012). The ROS procedure produces a separate replacement value for each censored datum. This procedure accommodates multiple censoring limits, which typically occurs when detection limits change over time. Briefly, the ROS method develops probability plotting positions for each data point (censored and uncensored) based on the ordering of the data. A relationship between data values and the uncensored probability plotting positions is fitted by least-squares regression, and this relationship is then used to predict the concentrations for the censored values based on their plotting positions. The ROS procedure produces estimated values for the censored data that are consistent with the distribution of the uncensored values, when distribution of these values in time is unknown. We randomised the predicted values to avoid inducing trends that would be associated with sequential plotting positions, which for the censored values is their order of appearance in time-series.

Step 2. Right-censored data. The right-censored data in our datasets were limited to ECOLI and CLAR measurements. All right-censored data were replaced with values estimated using a procedure based on “survival analysis” (Helsel 2012). These models are routinely used to estimate the survival time of samples beyond the period of observation or experiment. In this approach a parametric distribution is fitted to the uncensored values data using maximum likelihood. The values for the censored observations are then estimated by randomly sampling values larger than the censored values from this distribution.

Step 3. Striping. In some cases, laboratory results for low nutrient concentrations were reported on a semi-discrete scale (e.g., 1-2 decimal places), resulting in horizontal lines of plots of water quality variable versus time, or “striping”. These stripes correspond to tied data, which can pose problems for trend analyses, such as producing trends with slopes of exactly zero. Replacement of these tied values by imputation of randomised ROS values is inappropriate, because the striped concentrations are not the result of censoring. Instead, we “jittered” these results about their reported values to minimise the occurrence of ties. Jittering adds and subtracts random increments of < 2% of the reported values.

3.2 Grouping river and lake sites

River and lake monitoring sites were grouped into environmental classes to aid in summarizing the results of state and trend analyses, and to account for some variation in water quality associated with environmental heterogeneity. River sites were grouped by River Environment Classification (REC) Source- of-Flow class (Snelder and Biggs 2002). Lake sites were grouped by water surface elevation and maximum depth. Two elevation classes (0 – 300 m a.s.l., and > 300 m a.s.l.) and four depth classes (0-5 m, 5 – 15 m, 15 – 50 m, > 50 m) were used to define eight elevation × depth classes.

The rationale for the elevation × depth classification was: 1) elevation corresponds closely to catchment land-use and vegetation, which influence external loading to lakes; and 2) depth corresponds to lake mixing regime, which influences nutrient concentrations during summer in the surface layer, efficiency of burial of nutrients in the sediment, and DO bottom concentrations. Lakes in the < 5 m depth class are likely to be wind-mixed frequently throughout the year, lakes in the 5 – 15 m depth class are likely to be mixed occasionally during summer by surface cooling, lakes in the 15 – 50 m and > 50 m depth classes are expected to be seasonally stratified.

3.3 Water quality state

For each river and lake site, we characterised the current water quality state as percentiles (5th, 20th, 50th, 80th, 95th) of the distribution of measured values for the period 2009 to 2013 (inclusive). These percentiles were calculated using the Hazen method.⁴

The statistical robustness of determinations of water quality state depend on variability in the measurements between sampling dates and on sample size (i.e., the number of sampling dates). As a general rule, the rate at which confidence increases for estimates of population statistics levels off with increasing sample size greater than 30 (i.e., there are diminishing returns on increasing confidence with increasing sample size; McBride, 2005). Shifts in data distributions due to temporal trends can affect estimates of water quality state if they are assessed over long periods. We assessed state over a period of five years because it represented a reasonable trade-off between sample size and resistance to the effects of trends. For monthly sampling, which has been the norm since around 2010, a period of 5 years will yield at least 30 samples (allowing for some missing data). For rivers, we applied three filtering rules to ensure that site median values were reliable: 1) less than 50% of the values for a variable were censored; 2) values for at least 90% of monthly or quarterly sampling dates were available, including imputed values; 3) the 30 values were distributed over four of the five years from 2009 to 2013. Site by variable combinations that did not comply with these rules were excluded from the state analysis. For lakes, the rules for each variable were: 1) less than 50% of the values for a variable were censored; 2) values for at least 80% of sampling dates were available, including imputed values; 3) the values were distributed over four of the five years from 2009 to 2013. Relaxing the filtering rule about sampling dates from 90% to 80% greatly increased the yield of lake sites for the state analysis.

3.4 Trend analyses

3.4.1 Sampling dates and time periods for trend analyses.

Trend analysis is only meaningful for a specified time period over which the dataset being analysed has few missing values. The datasets provided by the regional councils had variable starting and ending dates, variable sampling frequencies (monthly or quarterly), and variable numbers of missing values. We therefore determined the time periods for trend analyses by examining the trade-off between the number of qualifying sites (i.e., sites that met our filtering rules concerning missing and censored values) and the time period. The trade-offs for each river and lake variables are presented graphically in Section 4. We assessed trends using monthly data preferentially, and quarterly data

⁴ (<http://www.mfe.govt.nz/publications/water/microbiological-quality-jun03/hazen-calculator.html>) Note that there are many possible ways to calculate percentiles. The Hazen method produces middle-of-the-road results, whereas the method used in Excel does not (McBride 2005, chapter 8).

when monthly data were not available, provided the filtering rules were met.⁵ We applied two filtering rules to identify the river sites to be included in trend analyses for each water quality variable: 1) 90% of the sampling dates in each of 90% of the years in a trend period had to have observations. For all variables except MCI, the rule about 90% of sampling dates applied to monthly or quarterly samples. For MCI, the 90% rule applied to annual sampling; 2) the number of censored values in a trend period had to be < 15% of the total number of observations. For lakes, the rules for each variable were: 1) 80% of the sampling dates in each of 80% of the years in a trend period had to have observations; 2) the number of censored values in a trend period had to be < 15% of the total number of observations. Relaxing the filtering rule about sampling dates from 90% to 80% greatly increased the yield of lake sites for the trend analysis but may have increased the number of sites for which there was insufficient data to establish trend direction with confidence.

In addition to the river and lake sites in the aggregated datasets, we analysed trends in water quality variables for a 25 years period at each of the 77 NRWQN sites. The variables used for the NRWQN trend analyses were DRP, NH₄, NO₃, TN, TP and CLAR. ECOLI was excluded from the analyses because it has only been monitored since 2004.

3.4.2 Statistical trend analyses

As noted in the Introduction, the method used for statistical trend analyses in this study differs in several ways from the approach used in previous analyses of national water quality data (e.g., Scarsbrook 2006, Ballantine et al. 2010). In the previous analyses, the non-parametric Seasonal Kendall Sen slope estimator was used with the Kendal trend test of a nil hypothesis, and trends that were determined to be statistically significant were then subdivided based on magnitude: trends with slopes > 1% per year were categorised as “meaningful”, and the remainder as “significant but not meaningful”. Several problems with the previous (“traditional”) approach have been identified: 1) conclusions about the significance of trends are strongly influenced by sample size in addition to trend magnitude; 2) the failure to reject the nil hypothesis is often treated as evidence that there is no trend (e.g., that water quality conditions are “stable” or “being maintained”); 3) the use of an arbitrary slope (e.g., 1%) to define meaningful trends is not a rigorous way to identify environmentally important trends. To overcome these problems, the new method uses a two-fold approach. First, confidence intervals are used to draw inferences about trend direction; if a symmetric confidence interval around the trend (estimated using the Seasonal Sen Slope Estimator SSSE) does not contain zero, then the trend direction is established with confidence.⁶ If it does contain zero it is concluded that there are insufficient data to determine the trend direction. Second, if the trend direction has been identified, the subsequent trend importance assessment is based on whether this rate of water quality change would cause a widely recognised threshold (e.g., an NPS-FM bottom line) to be exceeded within a specified critical period. A detailed critique of the traditional approach for trend analyses, and a detailed explanation of the new approach used in the current study are attached as Appendix A “A New Approach to Water Quality Trend Assessment”.

⁵ Note that in the new trend assessment procedures used herein, quarterly sampling will more commonly give rise to the finding of “insufficient data to detect trend direction”. This is a more informative statement than the equivalent result obtained using traditional trend assessment methods: “not statistically significant”.

⁶ An important technical detail arises here. In order to achieve a 95% confidence level procedure, these symmetric intervals are at the 90% level, *not* 95%. The rationale is fully explained in Appendix 1. Briefly, this arises because the new direction-testing procedure uses a two one-sided (“TOST”) methodology, rather than the traditional single “two-sided” method. TOST methods are also used in bioequivalence trials, especially in the drugs-testing domain, upon which there is a substantial literature. A notable and beneficial feature of this at-first-surprising aspect is that it makes the test uniformly more powerful than the traditional two-sided approach.

Trend assessments for all water quality variables that are measured monthly or quarterly were based on estimates from the Seasonal Sen Slope Estimator (SSSE), where seasons were either months or quarters. Trends in MCI were not estimated with SSSE because the macroinvertebrates used in MCI scores are sampled annually, which precludes accounting for seasonal variation. Instead, trends in MCI scores were estimated with the Sen Slope Estimator (SSE) (Sen 1968).

Estimated slopes were calculated with the SSSE (or SSE in the case of MCI) using a modification of the “zyp” package in R (<http://www.r-project.org>). The symmetric confidence intervals around each slope were estimated using the method of Sen (1968).⁷ The estimated slopes were then standardised by dividing by the corresponding median value and expressed as percentage changes; standardisation facilitates comparisons between groups of sites. These “relativised” trends are denoted as RSSSE (or RSSE in the case of MCI scores).

To help readers understand the implications of shifting from the traditional trend analysis procedures to the procedures used in this study, we also tested trends using the traditional Kendal test of correlation. The results of the Kendal tests for all sites and variables are included in the supplementary tables for trend analyses provided to MfE. Kendal tests and estimates of Sen slope were carried out using data for which the censored values were substituted in the traditional manner (i.e., left censored values substituted by one half the detection limit and right censored values increased by 10%).

3.4.3 River flow data

Measured or modelled flow measurements need to be paired with each river water quality measurement because many water quality variables are subject to either dilution (decreasing concentration with increasing flow, e.g., conductivity) or concentration (increasing concentration with increasing flow, e.g., total phosphorus). Data can be flow-adjusted before trend analysis, to remove the effects of variation in stream flow. Flow adjustments help to identify and quantify the effects of controlling factors other than river flow (e.g., land-use change) on water-quality trends.

Many regional council water quality sampling sites either did not have flow measurements or the councils did not provide flow measurements corresponding to the sampling occasions. Of a total of 785 sites for which we had some water quality data, 547 had no flow information provided. Where flow measurements were available, we used them. In particular, flow measurements were often available at locations with anthropogenically-modified flows, e.g. downstream of hydropower stations, where flows would be otherwise difficult to estimate. Note that Helsel and Hirsch (1992, p. 332) counsel against flow-adjustment at sites where flow is subject to human manipulation, e.g., downstream of a dam. This is because that manipulation modifies the probability distribution of water quality variables. However, we take the view that this *is* permissible, because our interest lies in what actually happened, not what might have happened had there been no dam.

Where flow measurements were not available, we estimated them using a national hydrological model (TopNet), corrected using flow duration curves (FDC) that were estimated with Random Forest models. TopNet is a spatially distributed time-stepping hydrological model that combines conceptual water balance models, including snow and plant canopy sub-models, with a kinematic wave channel routing algorithm. See McMillan et al. (2013) for a detailed description of TopNet and Clark et al. (2008) for complete model equations. For the current study, TopNet was run with daily precipitation and temperature data from the New Zealand Virtual Climate Station Network (Tait et al. 2006).

⁷ A summary of the method is provided at this website: (http://vsp.pnnl.gov/help/Vsample/Nonparametric_Estimate_of_Trend.htm).

Additional model parameters were estimated directly from GIS topography, soil and vegetation data. The resulting uncalibrated national TopNet model gave predicted flow values at hourly intervals over the period 1973–2013, for each Strahler-3 sub-catchment in the REC. The hourly data for each reach were averaged over each calendar day to obtain mean daily flow time-series. Where flow time-series were required for Strahler-1 or Strahler-2 reaches, flow data were downscaled by multiplying flows from the nearest available Strahler-3 node in the REC network by the ratio of the catchment area of the required location to the catchment area of the substitute location.

We corrected for bias in the TopNet-estimated flows using Random Forests to estimate the FDC at each site (Booker and Snelder 2012). This method uses machine-learning by combining many regression trees into an ensemble to produce more accurate regressions. The Random Forest method was used to create a regression of each of the three parameters describing a Generalised Extreme Value (GEV) distribution of the FDC as a function of available catchment characteristics. Random forest models are able to explain variation in parameters describing FDCs, and the GEV distribution can represent the range of shapes of standardised FDCs found across New Zealand (Booker and Snelder 2012). The FDC-based correction was implemented by finding the TopNet-predicted FDC at each reach, and then replacing each TopNet flow value with the equivalent percentile value from the Random Forest FDC for the same reach. The FDC correction method was selected on the basis of the analyses in Booker and Woods (2014). The exception was for locations downstream of major lakes, where the Random Forest FDC correction was not available; in those cases, uncorrected TopNet flows were used.

To characterise the accuracy of the flow estimation method, we compared flow-adjusted RSSSE values for water quality variables where the flow-adjustment was based on observed flows versus TopNet-estimated flows. This comparison was based on six water quality variables for which observed flows were available for at least 30 monitoring sites between 1994 and 2013. We characterised the accuracy of the RSSSE values derived from TopNet-estimated flows in three ways. First we assessed the overall performance using the Nash-Sutcliffe Efficiency coefficient (NSE). NSE is a measure of model performance based on the similarity of a plot of observed versus predicted values to the 1:1 line (i.e., the coincidence of two sets of values; Nash and Sutcliffe 1970). NSE values can range from one (a perfect fit) to negative infinity. Values larger than zero indicate that the model has some predictive capability, and the closer the NSE value is to one, the more accurate the model. Values less than zero suggest that the data is better predicted by the mean of the observations than the proposed model. Second, we used the root mean squared deviation (RMSD) to characterise model uncertainty (Piñeiro et al. 2008). Third, we characterised the rate of correct classification of trend direction (positive or negative) when RSSSE values were derived from the estimated flows. High correct classification rates provide confidence that the analysis of overall trends (which are based on the rate of positive or negative trends using the binomial test) was accurate.

3.4.4 Flow adjustment of river water quality variables

All river variables were flow adjusted except annual MCI scores. The flow adjustment procedure was performed by first fitting a second order generalised additive model (GAM) to the $\log_{10}(\text{variable value})$ versus $\log_{10}(\text{flow})$ relationship for each variable and site. The strength and form of these relationships varied considerably. In general, nutrient concentrations were positively related to flow (linear regression coefficients). However, at some sites, the rates of change in NO₃N and DRP concentrations with flow decreased or became negative at high flows, and the rates of change in TN and TP concentrations increased at high flows. The use of a second order GAM ensured this

curvilinear relationship between variable values and flow (in log-log space) was represented when it was evident.

The GAMs were used to adjust variable values in response to flow as outlined by Smith et al. (1996): $\text{adjusted value} = \text{raw value} - \text{value predicted by the regression model} + \text{mean value}$. Flow adjustments were made for all river monitoring sites irrespective of the strengths of the water quality-flow relationships at each site. Our rationale for this approach was that if flow significantly explains variation in concentration, however weak this relationship may be, the trends are potentially influenced by flow state at the time of sampling unless this relationship is accounted for. We note that flow-adjusted variable values converge toward the raw values as the strength of flow-concentration relationships decrease.

3.4.5 Importance of water quality trends

In the “equivalence testing” procedure advanced by McBride et al. (2014), a prior specification of the change that would be important to detect is advocated (when testing for differences between populations). This approach can be extended to trend analyses. For example, important trends can be defined a priori using threshold-values of different water quality variables and critical time spans. This is an improvement on the traditional approach for assessing trend importance based on trend magnitude alone. For example Ballantine et al. (2010) defined a “meaningful” change as a trend greater than 1% per year during the assessment period, and applied this definition to all water quality variable. However, obtaining agreement on the definition of a meaningful rate of change based on magnitude alone is impractical, particularly when multiple water quality variables are included in a study. In the present report, we used the approach recommended above; defining trend importance in relation to threshold-values of different variables and critical time periods. The thresholds corresponded to published guidelines, including attribute bands in the NPS-FM. We note that this approach is nominal because the importance of any given trend is dependent on context and viewpoint. Our assessment method demonstrates one possible approach to judging the importance of trend magnitudes for some of the water quality variables in this study.

Our assessment was based on identifying trends corresponding to a progressive change in a median value water quality variable from a reference state (i.e., unaffected by human activities) to an unacceptable state (i.e., exceed the designated guidelines) within 10 years. The 10-year period is based on the statutory review period for regional plans, as set out in Section 79 of the Resource Management Act 1991. Reference state and threshold values for rivers and lakes are given in Table 3-1 and 3-2, respectively.

Expected reference state median concentrations and water clarity levels have been characterised for rivers at the REC Source-of-Flow level by McDowell et al. (2013). We did not have estimates of reference states for lakes. In lieu of lake reference states, we used the lake A-bands for TN, TP and CHLA from the NPS-FM.

Thresholds for unacceptable states were based broadly on NPS-FM national bottom lines (i.e., the C/D band thresholds) for both lakes and rivers. For river and lake ECOLI, we used the NPS-FM bottom-line for secondary contact recreation, 1000/100 ml (as a median). For lake CHLA, we used the NPS-FM bottom line. The NPS-FM does not have bottom lines for CLAR in rivers; as an alternative, we used the MfE (1994) guideline. The NPS-FM does not include ecological bottom-lines for nutrients in rivers, and the nitrate and ammonia bottom-lines for toxicity were not appropriate for this study. As an alternative, we used nutrient concentrations that were predicted to cause river periphyton biomass to reach the NPS-FM periphyton bottom line in each REC Source-of-Flow class. The nutrients

associated with periphyton biomass were limited to DRP and TN. Therefore, the DRP and TN concentrations that correspond to the periphyton bottom lines were used as thresholds for assessing trend importance. Detailed methods for predicting DRP and TN concentrations that correspond to the periphyton bottom line are given in Appendix B.

In the absence of applicable thresholds, the importance of trends in the remaining river and lake variables (NO₃N, NH₄N, TP, MCI, SECCHI, DO_{bottom}, TLI) was not assessed. In addition, CLAR, DRP and TN thresholds could not be derived for some REC classes. For example, the predicted reference states for CLAR in the CD/L, WW/L, WW/Lk and WD/L classes are less than the MfE guideline CLAR value of 1.6 m. In each case where a reference state or threshold could not be derived, we report whether trends could be inferred with confidence, but did not assess trend importance.

Seven categories were used to organise the results of the assessments of trend importance (Table 3-1, Table 3-2):

1. Degrading and important: the trend direction is inferred with confidence, and the threshold is projected to be crossed within 10 years (applies to increasing trends in nutrients, ECOLI, and CHLA, and decreasing trends in CLAR).
2. Improving and important: the trend direction is inferred with confidence and the threshold is projected to be crossed within 10 years (applies to decreasing trends in nutrients, ECOLI and CHLA, and increasing trends in CLAR).
3. Degrading but not important: the trend direction is inferred with confidence, but the threshold is not projected to be crossed within 10 years.
4. Improving but not important: the trend direction is inferred with confidence, but the threshold is not projected to be crossed within 10 years.
5. Degrading but no threshold is available: the trend direction is inferred with confidence; no importance threshold is available.
6. Improving but no threshold is available: the trend direction is inferred with confidence; no importance threshold is available.
7. Insufficient: the data were insufficient to confidently determine trend direction (and therefore no importance assessment can be made).

Table 3-1: Parameters used in assessments of trend importance for river water quality variables. Sources for reference state values: McDowell et al. (2013). Sources for criteria/guideline values: ECOLI – NPS-FM bottom-line for secondary contact recreation, CLAR – MfE 1994, TN and DRP – Appendix B, this report. Units: CLAR (m), COND ($\mu\text{S}/\text{cm}$), ECOLI (n/100 ml), NH₄N (mg/m³), NO₃N (mg/m³), TN (mg/m³), DRP (mg/m³), TP (mg/m³). ND: not determined.

	Reference state								Threshold				Critical rate of change (units per year)			
REC Source of Flow class	Ref CLAR	Ref COND	Ref NH ₄ N	Ref NO ₃ N	Ref TN	Ref DRP	Ref TP	Ref ECOLI	Cr CLAR	Cr ECOLI	Cr TN	Cr DRP	Rate CLAR	Rate ECOLI	Rate TN	Rate DRP
CD/H	2.4	66	4	8	73	3	6	14	1.6	1000	430	20.7	-0.08	98.6	35.7	1.77
CD/L	0.9	105	7	143	568	5	9	34	1.6	1000	389	14.2	NA	96.6	NA	0.92
CD/Lk	2.1	88	5	21	111	4	9	11	1.6	1000	426	23.6	-0.05	98.9	31.5	1.96
CD/M	2.1	77	4	16	107	4	9	6	1.6	1000	470	42	-0.05	99.4	36.3	3.8
CW/H	3	83	5	44	150	5	9	9	1.6	1000	633	61.2	-0.14	99.1	48.3	5.62
CW/L	2.2	129	6	86	178	8	13	40	1.6	1000	541	33.2	-0.06	96	36.3	2.52
CW/Lk	3.1	95	4	7	86	2	10	1	1.6	1000	426	24.7	-0.15	99.9	34	2.27
CW/M	2.3	72	3	15	58	3	8	4	1.6	1000	549	71.4	-0.07	99.6	49.1	6.84
CX/H	4	76	3	35	80	3	8	5	1.6	1000	1667	NA	-0.24	99.5	158.7	NA
CX/L	2.4	88	5	52	122	7	9	42	1.6	1000	1399	350.3	-0.08	95.8	127.7	34.33
CX/Lk	3.2	74	4	32	116	2	6	4	1.6	1000	537	48.7	-0.16	99.6	42.1	4.67
CX/M	2	85	4	23	93	3	10	11	1.6	1000	1148	335.9	-0.04	98.9	105.5	33.29
WD/L	1.3	76	9	92	161	5	16	39	1.6	1000	230	12.2	NA	96.1	6.9	0.72
WW/H	2.1	81	4	21	108	5	10	15	1.6	1000	709	62.5	-0.05	98.5	60.1	5.75
WW/L	1.5	101	6	26	176	8	16	62	1.6	1000	336	11.2	NA	93.8	16	0.32
WW/Lk	1.4	111	8	87	214	14	21	17	1.6	1000	420	20.5	NA	98.3	20.6	0.65
WX/H	1.9	100	4	30	108	6	9	7	1.6	1000	606	61.5	-0.03	99.3	49.8	5.55
WX/L	2.5	103	5	35	147	3	9	16	1.6	1000	560	38.8	-0.09	98.4	41.3	3.58
All classes	ND	ND	ND	ND	ND	ND	ND	ND	1.6	1000	496	33.2	ND	ND	ND	ND

Table 3-2: Parameters used in assessments of trend importance for lake water quality variables. Sources for reference state values: Lake A-bands for TN, TP and CHLA from the NPS-FM (2014). Sources for criteria/guideline values: NPS-FM bottom line values. Units: CHLA (mg/L), TN (mg/m³), TP (mg/m³).

Lake Class <i>elevation</i> (depth)	NPS-FM Class	Reference state			Criteria / Guideline value			Critical rate of change (units per year)		
		Ref CHLA	Ref TN	Ref TP	Cr CHLA	Cr TN	Cr TP	Rate CHLA	Rate TN	Rate TP
0-300 m (0-5 m)	Polymictic	2	300	10	12	800	50	1	50	4
0-300 m (5-15 m)	Polymictic	2	300	10	12	800	50	1	50	4
0-300 m (15-50 m)	Stratified	2	160	10	12	750	50	1	59	4
0-300 m (> 50 m)	Stratified	2	160	10	12	750	50	1	59	4
> 300 m (0-5 m)	Polymictic	2	300	10	12	800	50	1	50	4
> 300 m (5-15 m)	Polymictic	2	300	10	12	800	50	1	50	4
> 300 m (15-50 m)	Stratified	2	160	10	12	750	50	1	59	4
> 300 m (> 50 m)	Stratified	2	160	10	12	750	50	1	59	4

4 Results – river and lake state

4.1 River state – nutrients, ECOLI and CLAR

Between 365 and 577 river monitoring sites met the filtering rules for the state analysis of nutrients, ECOLI and CLAR; qualifying sites varied by water quality variable and by REC class (Table 4-1). The geographic distribution of sites is shown in Figure 4-1. The sites are reasonably well-distributed, although there are gaps in the central North and central South Islands.

The distributions of site-median values of the water quality variables for the 2009-2013 period are summarized with box-and-whisker plots for the REC Source-of-Flow classes for which there were sufficient sites (Figure 4-2). Descriptions of the REC classes are given in Snelder and Biggs (2002). The plots in Figure 4-2 indicate that water quality state (i.e., site medians for nutrients, ECOLI and CLAR) were highly variable, but some of the variation is explained by the REC classes. Sites in the different REC Source-of-Flow classes had different water quality characteristics both in terms of their central tendencies (indicated by the median of the median site values) and their variation. For example, median CLAR was highest and least variable in the Cool-Extremely Wet (CX) climate and Mountain (M), Lake (Lk) and Hill (H) Source-of-Flow classes. The lowest REC class median for CLAR and the greatest variability occurred in the Warm-Wet Lake (WW/Lk) class. In general, median nutrient and ECOLI concentrations were lowest and CLAR was highest in mountain Source-of-Flow classes and in the CX/Lk and CD/Lk classes. In contrast, nutrient and ECOLI concentrations were highest in the low-elevation Source-of-Flow classes, irrespective of the climate category. The complete set of state analysis results is provided in the supplementary file “NationalRiverState_2009_2013”.

Table 4-1: Number of river monitoring sites by REC class and water quality variable that were included in the state analyses of nutrients, ECOLI and CLAR. The site numbers shown refer to sites where less than 50% of the values for a variable were censored, and ≥ 30 values were available, distributed over at least four of the five years from 2009 to 2013. NS: no sites met the filtering rules for that water quality variable.

REC Source- of-Flow class																		
Variable	Total	CX/M	CX/H	CX/L	CX/Lk	WX/L	CW/M	CW/H	CW/L	CW/Lk	WW/H	WW/L	WW/Lk	CD/M	CD/H	CD/L	CD/Lk	WD/L
DRP	519	3	15	3	3	2	12	102	87	21	2	126	5	1	25	80	NS	32
ECOLI	486	5	19	2	8	4	13	94	90	21	2	81	5	2	34	79	1	26
NH4N	365	3	10	2	3	1	8	53	57	19	1	100	5	NS	15	60	NS	28
NO3N	587	5	19	3	7	5	13	120	93	22	3	140	7	2	34	81	1	32
TP	577	5	19	3	6	5	12	114	93	22	3	139	7	2	34	80	1	32
TN	354	5	15	2	4	2	11	58	73	14	2	71	3	2	31	44	1	16
CLAR	454	3	20	3	7	5	11	107	84	20	2	117	5	NS	13	44	NS	13

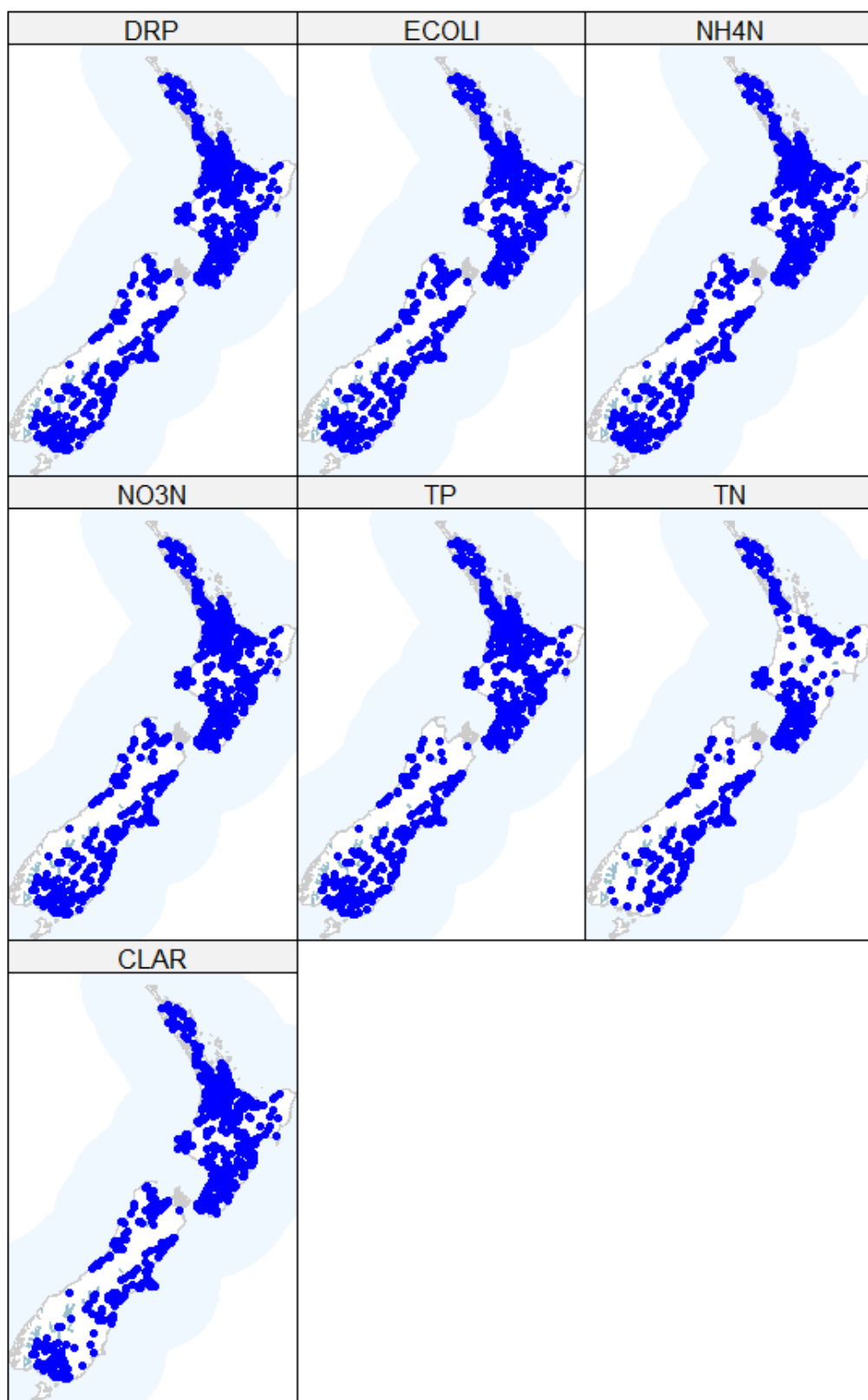


Figure 4-1: Locations of river water quality monitoring sites used for state analyses of nutrients, ECOLI and CLAR.

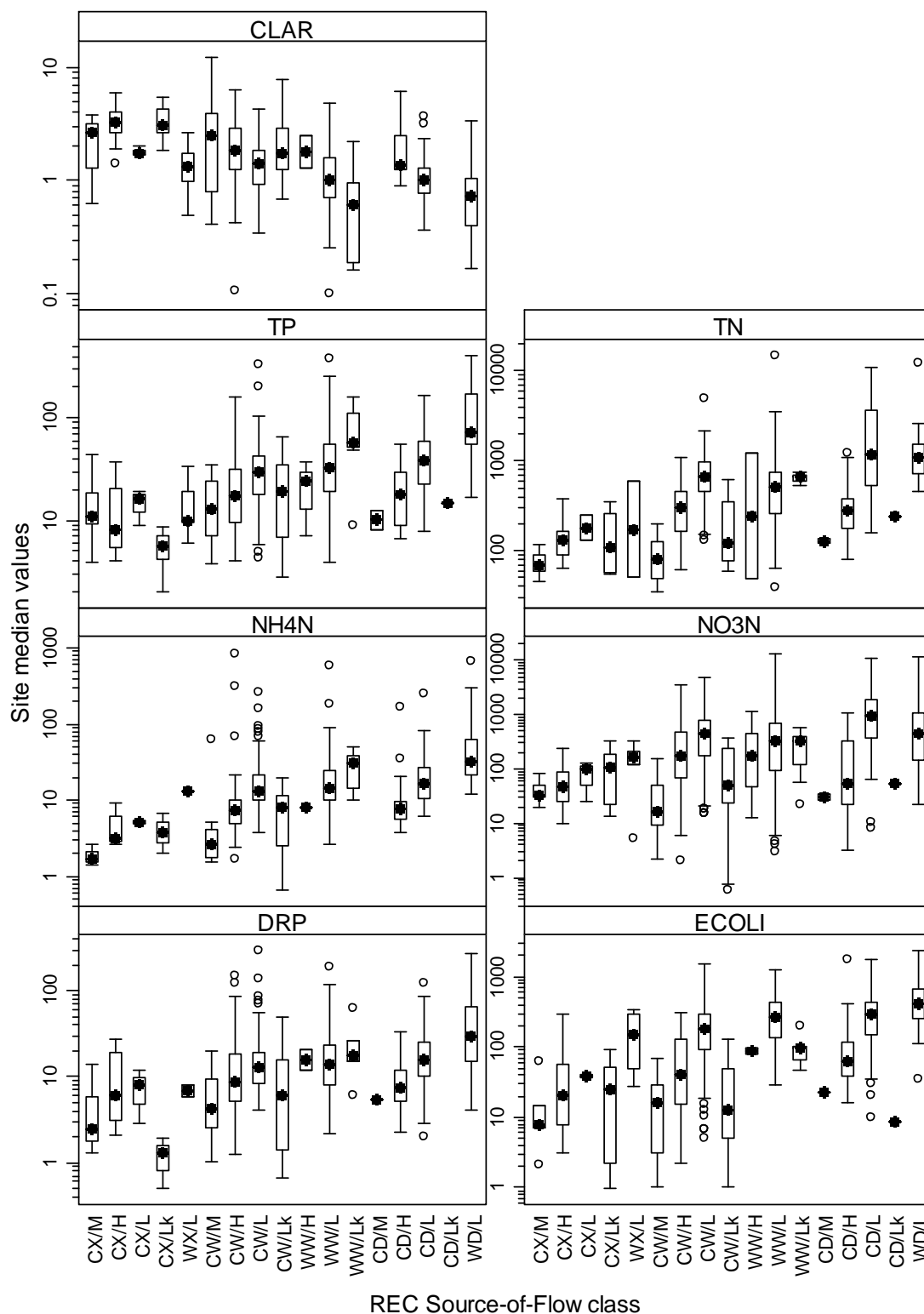


Figure 4-2: River water quality state in REC Source-of-Flow classes. Box-and-whisker plots show the distributions of monitoring site medians within REC classes. The closed circle in each box indicates the median of site medians, the box indicates the inter-quartile range and the whiskers indicate the 5th and 95th percentiles. Outliers are indicated by open circles. Note log-scale on Y-axes.

4.2 River state - MCI

A total of 510 river monitoring sites met the filtering rules for the state analysis of MCI scores (Table 4-2). The numbers of sites varied substantially by REC class, and for six classes in the WX, WW and WD climate categories, no monitoring sites qualified. The qualifying sites were reasonably well distributed geographically (Figure 4-3). However, there were gaps in the central North and South Islands and in Waikato and Bay of Plenty.

The distributions of site-median MCI scores for the 2009-2013 period are summarized with box-and-whisker plots in Figure 4-4. These plots indicate that a substantial amount of variability in median MCI scores is explained by the REC classes. In general, median MCI scores were highest in the cooler and wetter climate categories (e.g. CX, CW) and the Hill (H) Source-of-Flow categories, followed by the Mountain (M), and then Low-Elevation (L) categories. The complete state analysis results is provided in the supplementary file “NationalMCIState_2009-2013”.

Table 4-2: Number of river monitoring sites by REC Source-of-Flow class used in the state analysis of MCI scores.

REC class	Number of sites
CX/M	4
CX/H	25
CX/L	23
CX/Lk	5
WX/L	3
CW/M	14
CW/H	95
CW/L	86
CW/Lk	5
WW/H	3
WW/L	95
WW/Lk	1
CD/M	2
CD/H	35
CD/L	84
CD/Lk	1
WD/L	28
WD/Lk	1
Total	510

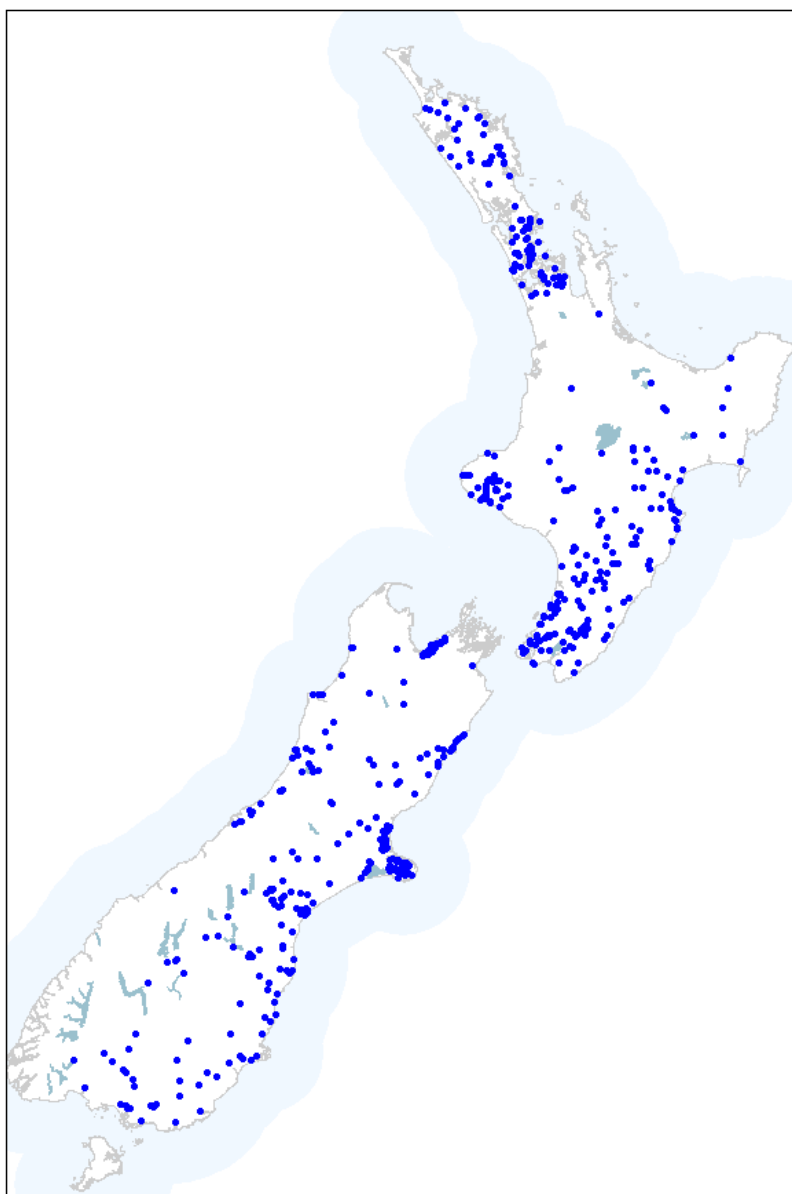


Figure 4-3: Locations of river monitoring sites used for state analyses of MCI scores.

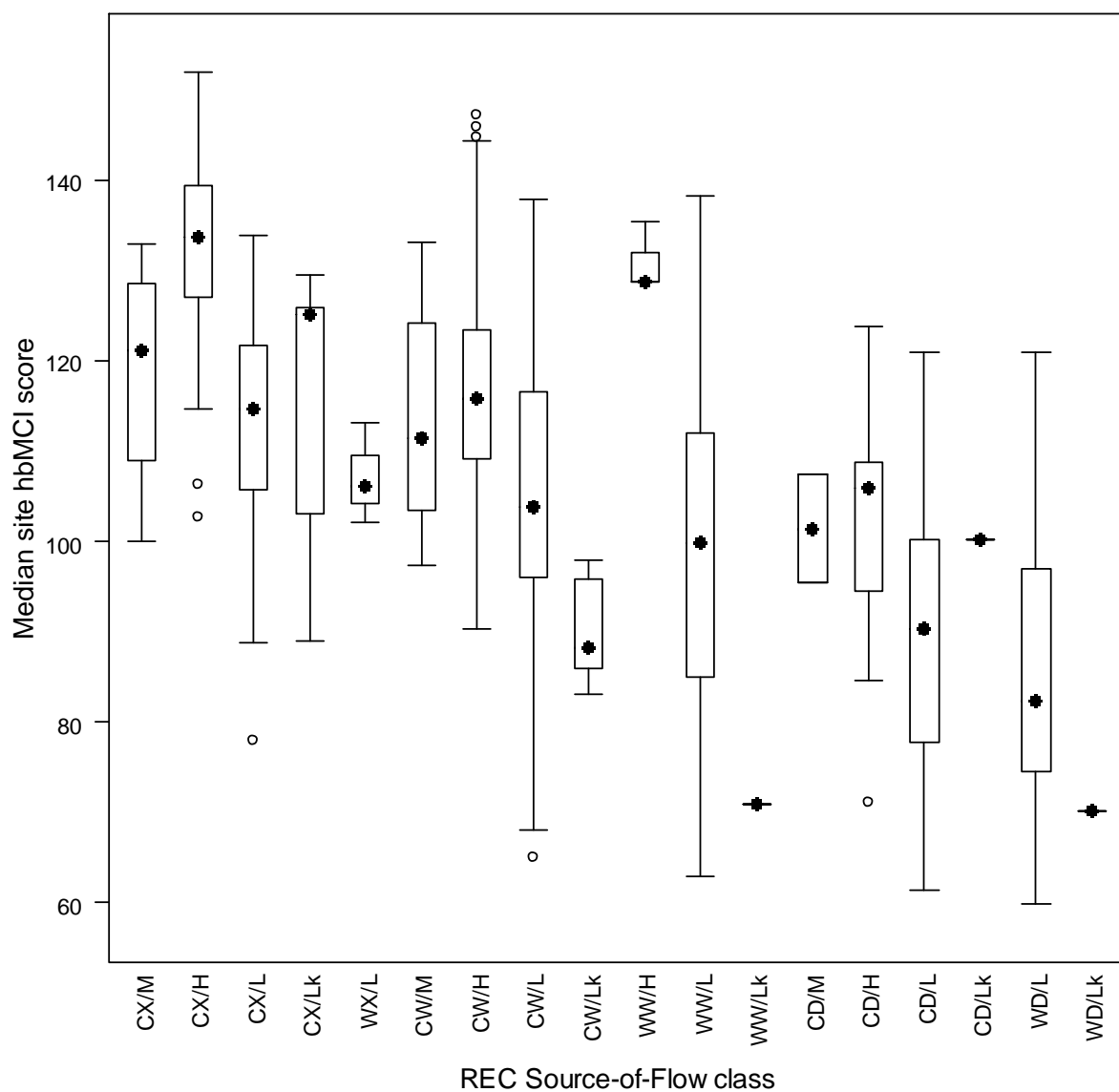


Figure 4-4: Distribution of site-median MCI scores for hard-bottom streams, within REC Source-of-Flow class. The closed circle in each box indicates the median of site medians, the box indicates the inter-quartile range and the whiskers indicate the 5th and 95th percentiles. Outliers are indicated by open circles.

4.3 Lake state

Between 20 and 84 lake monitoring sites met the filtering rules for the state analysis; as with river variables, there were different numbers of qualifying lake sites for each water-quality variable (Table 4-3). The numbers of sites also varied by lake elevation × depth classes. The qualifying lake sites are sparsely and unevenly distributed on both the North and South Islands; there are gaps in the Manawatu, Taranaki, Tasman, Marlborough, Otago and West Coast regions (Figure 4-5). Monitoring sites with NO₃N and DRP data are particularly sparse.

The distributions of site-median lake water quality for the 2009-2013 period are summarized with box-and-whisker plots in Figure 4-6. These plots indicate that a substantial amount of variability in some variables is explained by the lake elevation × depth classes. There is a clear pattern of inter-class differences for most variables: nutrient and CHLA concentrations and TLI levels are relatively high in low-elevation, shallow lakes and decrease in high-elevation, deep lakes; SECCHI is relatively low in low-elevation, shallow lakes and increases in high-elevation, deep lakes. The complete set of state analysis results is provided in the supplementary file “NationalLakeState_2009-2013”.

Table 4-3: Number of lake monitoring sites by class and water quality variable that were included in the state analyses. Elevation × depth classes are given as elevation range in top line, maximum-depth range on bottom line. The site numbers shown refer to sites where less than 50% of the values for a variable were censored, and at least 16 values were available, distributed over at least four of the five years from 2009 to 2013. NS: no sites met the filtering rules for that water quality variable.

Variable	Elevation × depth class								Total
	0-300 m 0-5 m	0-300 m 5-15 m	0-300 m 15-50 m	0-300 m > 50 m	> 300 m 0-5 m	> 300 m 5-15 m	> 300 m 15-50 m	> 300 m > 50 m	
DO _{bottom}	9	13	7	6	NS	NS	6	4	45
CHLA	18	16	7	3	2	3	15	8	72
TLI	12	14	7	3	2	3	15	9	65
NH ₄ N	12	17	6	3	1	NS	5	4	48
DRP	4	4	5	2	NS	0	5	4	24
NO ₃ N	4	1	3	3	NS	NS	4	4	19
TN	13	17	9	3	2	3	15	9	71
TP	20	19	9	3	2	3	15	5	76
SECCHI	16	19	11	6	NS	NS	5	4	61

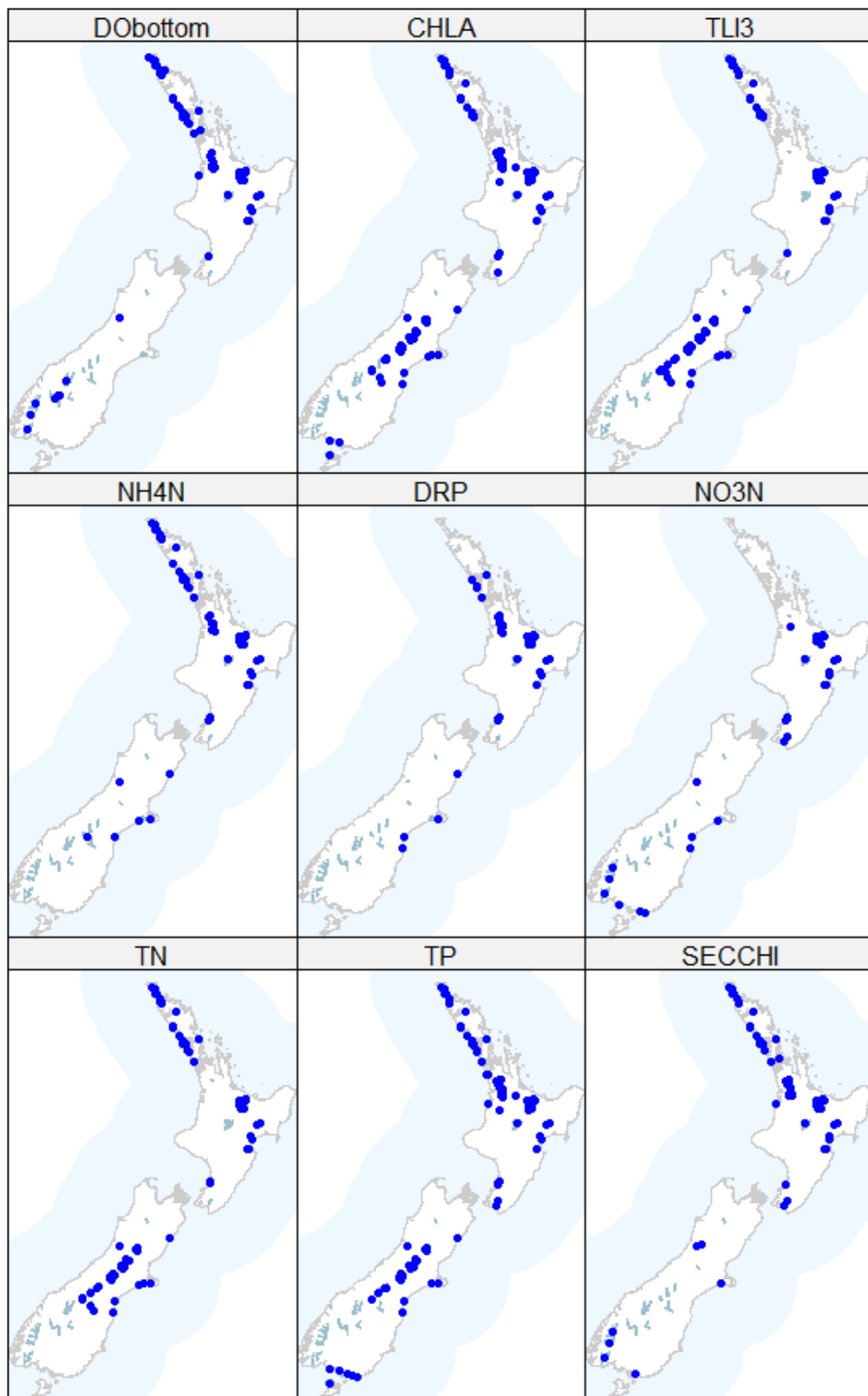


Figure 4-5: Locations of monitoring sites used for state analyses of lake water quality.

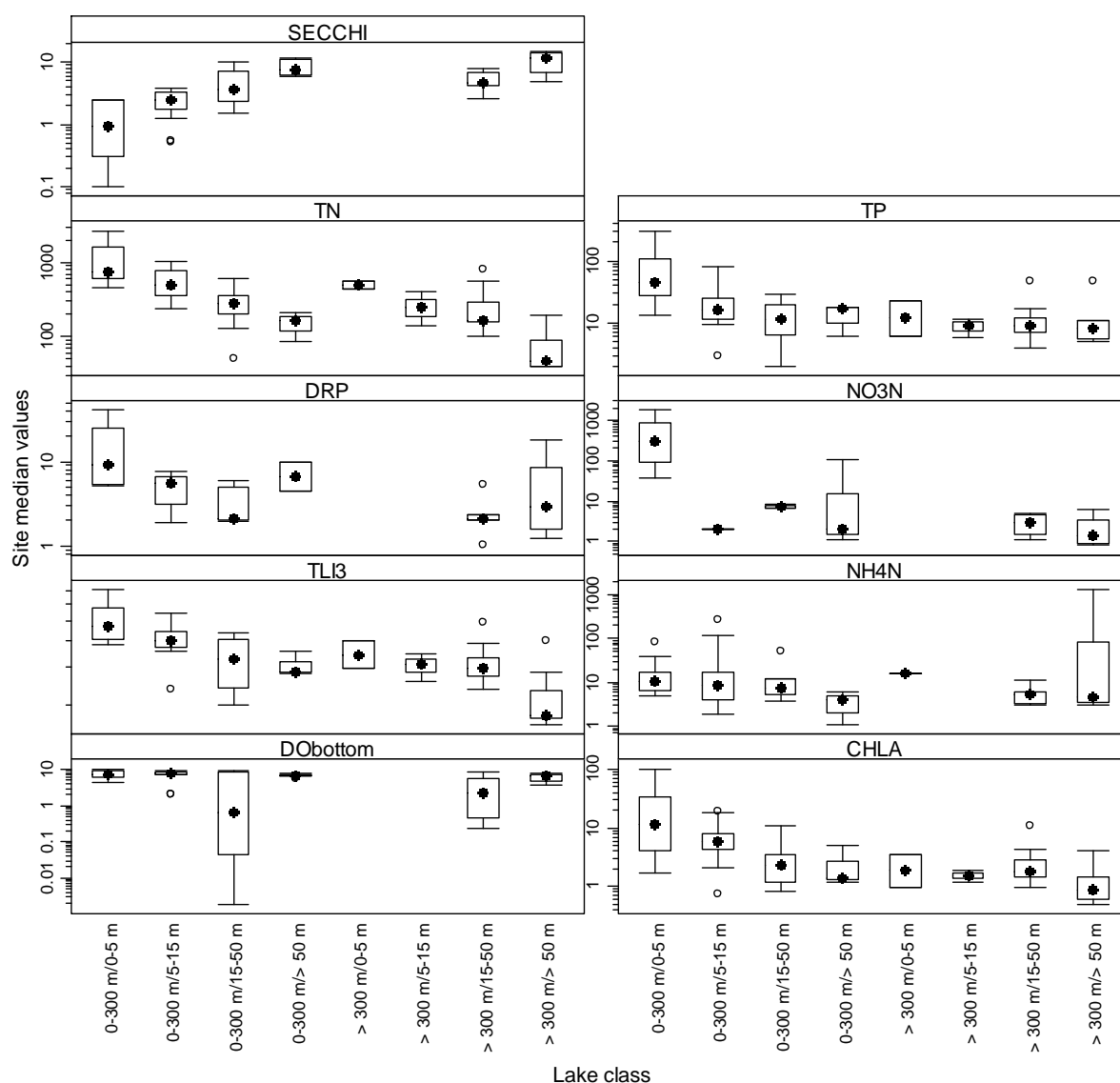


Figure 4-6: Lake water quality state in elevation x depth classes. Box-and-whisker plots show the distributions of monitoring site medians within each class. The closed circle in each box indicates the median of site medians, the box indicates the inter-quartile range and the whiskers indicate the 5th and 95th percentiles. Outliers are indicated by open circles. Note log-scale on Y-axes.

5 Trend results

5.1 River trends - nutrients, ECOLI and CLAR

5.1.1 Trade off analysis

The trade-offs between the number of qualifying river monitoring sites (i.e., sites that met our filtering rules) and the time period represented by those sites are shown for each water quality variable in Figure 5-1. Trend periods of 20 years (1994-2013) and 10 years (2004-2013) were used to make relatively robust (in terms of site number) estimates of long-term and recent monotonic trends. The 20-year period, and to a lesser degree the 10-year period, may incorporate non-monotonic trends; these were not identified or quantified in the current study, but they can be viewed in time-series plots that are supplied as supplementary files to this report. For the 77 NRWQN sites, a 25-year monitoring period was used to provide a separate set of long-term trends.

5.1.2 TopNet flows for flow adjustment

Plots of flow-adjusted RSSSE values using TopNet synthetic flows versus flow-adjusted RSSSE values using observed flows are shown in Figure 5-2. The six variables shown are those for which observed flows were available for at least 30 monitoring sites between 1994 and 2013, and robust comparisons could therefore be made between the two approaches. The results of formal tests of the correspondence between the pairs of estimates made by using TopNet and observed flows are shown in Table 5-1; these tests are based on linear regressions. ECOLI was not included in the tests in Table 5-1 due to insufficient numbers of monitoring sites with observed flow.

Table 5-1: Correspondence between pairs of estimates of RSSE with flow adjustments using TopNet synthetic flows and observed flows. R^2 : coefficient of determination for the linear regressions shown in Figure 5-2. NSE: Nash-Sutcliffe Efficiency coefficient. RMSD: root mean squared deviation. Correct classification: percent of correct classification of trend direction (positive or negative) when RSSSE values were flow-adjusted using TopNet modelled flows.

Variable	Number of sites	R^2	NSE	RMSD (%)	Correct Classification (%)
CLAR	95	0.69	0.61	0.05	80
DRP	80	0.89	0.89	0.11	91
NH4N	77	0.92	0.92	0.14	94
NO3N	88	0.93	0.92	4.39	88
TN	80	0.83	0.81	5.36	89
TP	87	0.80	0.78	0.41	72

The results indicate that the TopNet modelled flows produced flow-adjusted trend analysis results that were consistent with those produced by the observed flows. R^2 and NSE values were > 0.6 for each variable, and were generally > 0.8 (Table 5-1), which indicate good model performance (i.e., low bias and high precision). The plots in Figure 5-2 confirm the good performance, with data points and lines of best fit plotting close to the one to one line (Figure 5-2). The RMSD values were less than 0.5% for all variables except TN and NO3N. The larger RMSD values for TN and NO3N partly reflect high variation in RSSSE values for these variables. The larger RMSD values may also reflect non-linear relationships between concentration and flow for nitrogen. These relationships often increase

initially with increasing flow due to groundwater flushing, and then decreases at high flows due to dilution. Despite the higher RMSD values for TN and NO₃N, these two variables had high rates of correct classification of RSSSE values (89 and 88% respectively; Table 5-1). In general, the trend estimates produced using the TopNet modelled flows were consistent with those produced using the observed flows. While some individual site estimates of RSSSE based on the TopNet flow will be inaccurate, the overall pattern of increasing or decreasing trends is well-represented.

We also examined the correspondence between raw and flow adjusted trends for each water quality variable, for 10-year trends (Figure 5-3). The raw and flow adjusted trends were generally highly correlated, although there were substantial differences at individual sites for some variables. The plots in Figure 5-3 indicate that the effects of flow adjustment on RSSSE values were generally small, and that it is reasonable to use raw 10-year RSSSE values as indicators of trends.

5.1.3 Ten year trends (2004 – 2013)

Between 206 and 511 river monitoring sites met the filtering rules for the 10-year trend analysis of nutrients, ECOLI and CLAR (Table 5-2). The qualifying sites were reasonably well-distributed geographically, with gaps in the central North and South islands and the West Coast (Figure 5-4). All site locations, REC classes and numbers of sampling dates are included in the supplementary file “NationalRiver10YrTrend_2004-2013”.

Box and whisker plots were used to summarise the estimated trends for each of the water quality variables for the 10-year period from 2004 – 2013 (Figure 5-5). All estimated trends are included in these plots, irrespective of their importance categories (as defined in Section 3.4.5). The plots indicate that REC classes did not account for a substantial amount of the variation in trends for any variable; this is in contrast with the state analyses of river variables, where water-quality state clearly varied between REC classes (Figures 4-2 and 4-4).

The analysis of 10-year trend categories (Table 5-3) confirmed some of the patterns evident in the box and whisker plots of distributions of site RSSSE values. Summing across the improving and degrading categories, there were 10 times as many sites with improving trends in TP as degrading trends, three times as many sites with improving trends in DRP as degrading trends, and 1.5 to 1.9 times as many sites with improving trends in ECOLI, NH₄N and CLAR as decreasing trends. In contrast, there were 1.5 times as many sites with degrading trends in NO₃N as improving trends. Improving and degrading trends in TN were nearly balanced, with 13% more sites with improving trends.

Four water quality variables (DRP, TN, ECOLI, CLAR) were included in the trend importance assessments. Proportions of sites at which trends were classed as important and degrading, as defined in Section 3.4.5, ranged from <1% for ECOLI to 8% for TN, and proportions of sites at which trends were classed as important and improving ranged from 1% for ECOLI to 27% for DRP (Table 5-3). The complete 10-year trend analysis results are provided in the supplementary file “NationalRiver10YrTrend_2004-2013”.

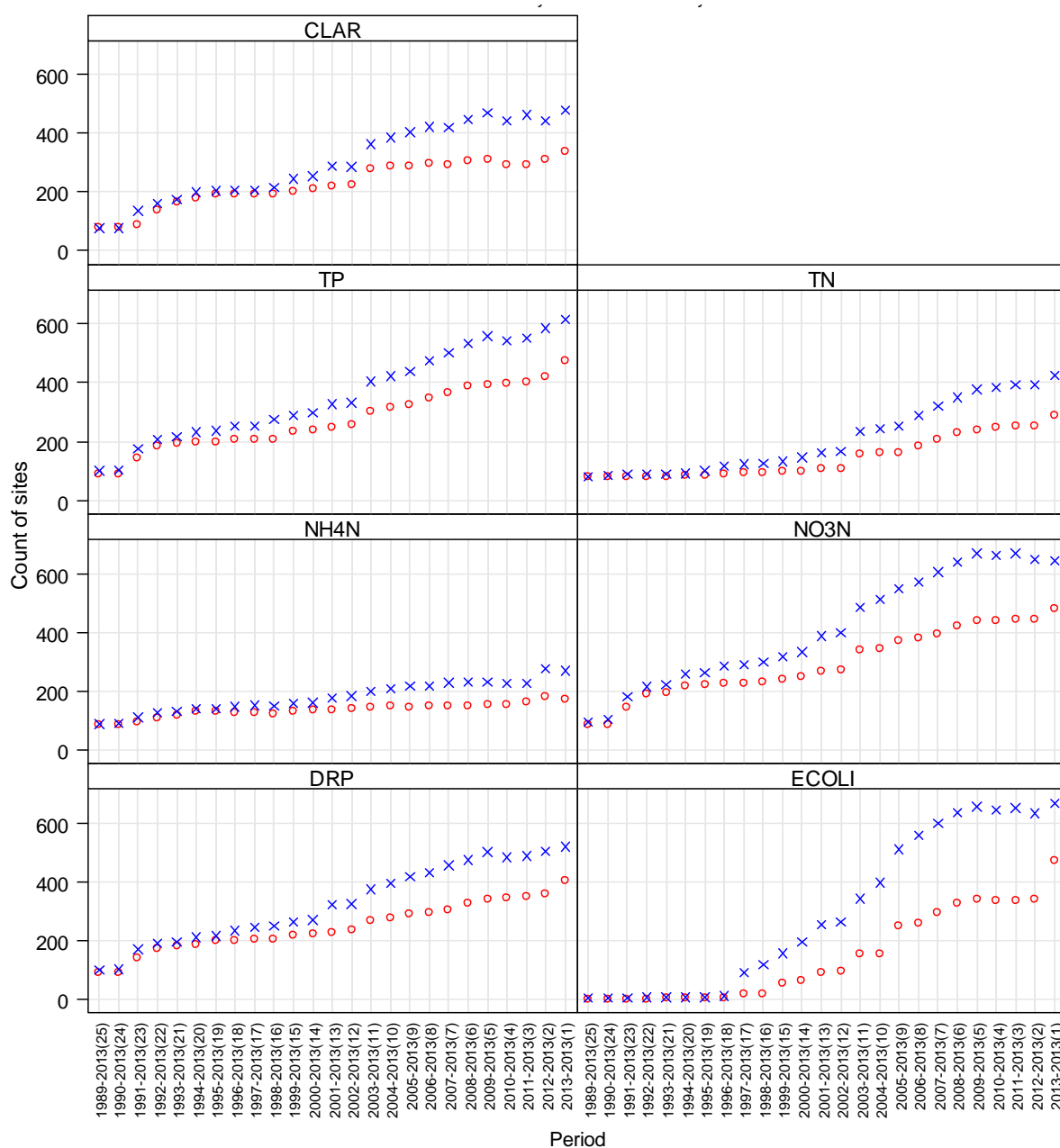


Figure 5-1: Changes in the number of river monitoring sites that met the filtering rules for each water quality variable versus the period of site operation. Durations of periods are shown in parentheses. Open circles: monthly data, crosses: quarterly data. The plots were used to select time periods for trend analyses.

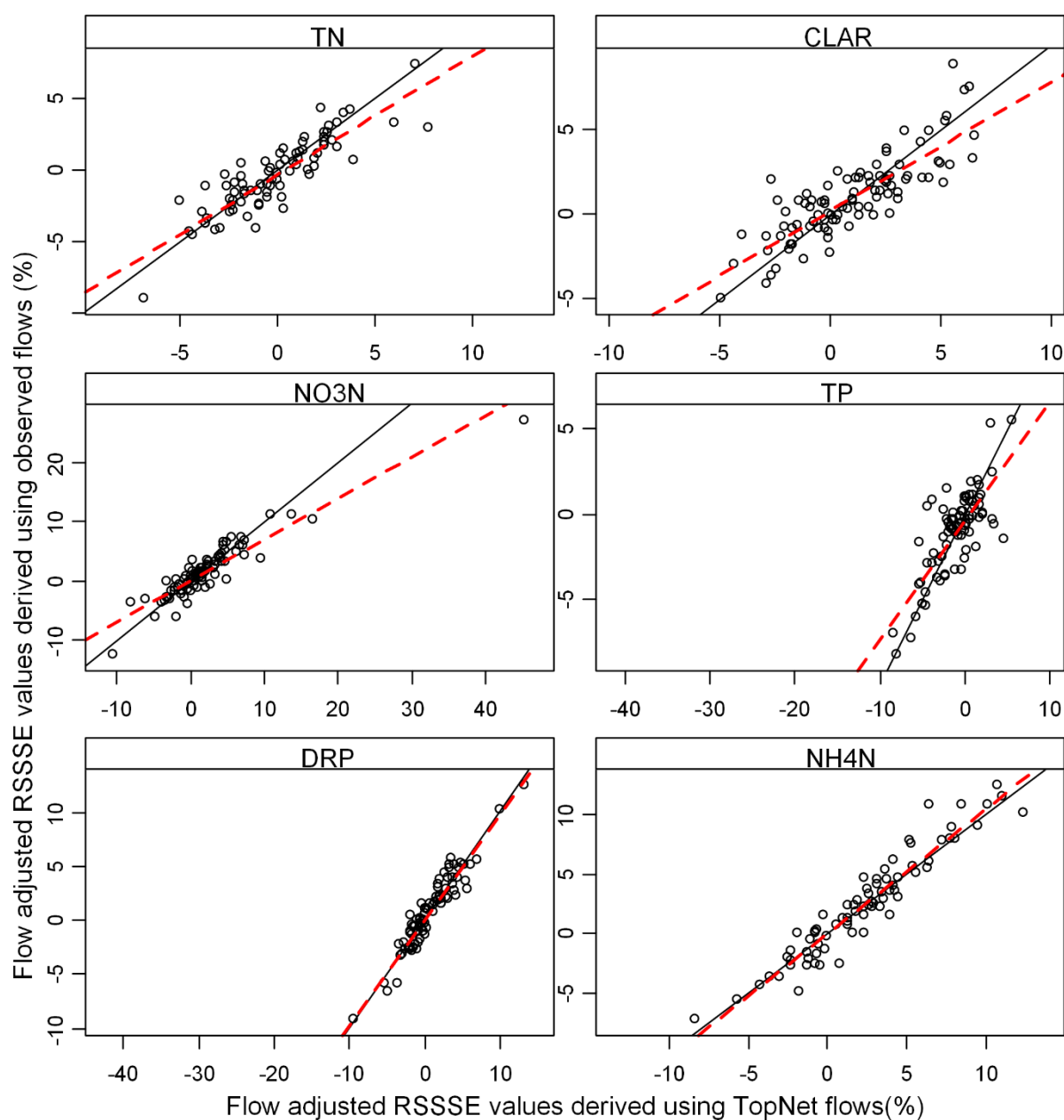


Figure 5-2: Comparison of RSSSE values for the 10-year period ending 2013 for sites with flow adjustment performed using TopNet modelled flows and the observed flows. Note that RSSSE values using observed flows are plotted on the Y-axis and RSSSE values using TopNet flows are plotted on the X-axis, following Piñeiro et al. (2008). Red dashed line: best fit linear regression of the two sets of values, black solid line: one-to-one line.

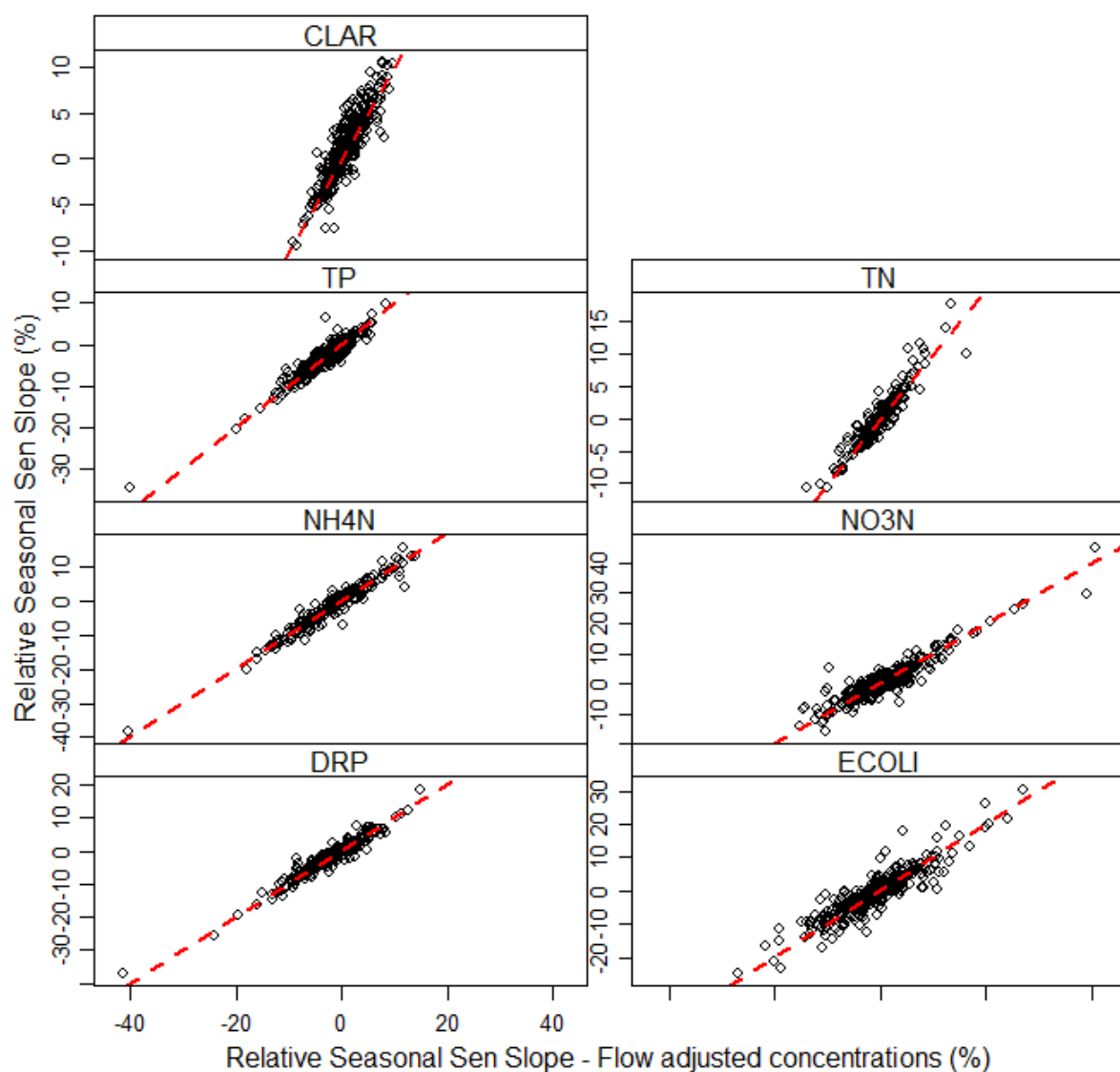


Figure 5-3: Correspondence between raw and flow adjusted trends for each water quality variable, for 10-year trends. Red line is 1 to 1.

Table 5-2: Number of river monitoring sites by REC class and water quality variable that were included in the 10-year trend analyses of nutrients, ECOLI and CLAR.
The site numbers shown refer to sites where 90% of the sampling dates nine of the years in the 2004-2013 period had observations, and less than 15% of the data for each variable consisted of the number of censored values.

Variable	Total	REC Source- of-Flow class															
		CX/M	CX/H	CX/L	CX/Lk	WX/L	CW/M	CW/H	CW/L	CW/Lk	WW/H	WW/L	WW/Lk	CD/M	CD/H	CD/L	WD/L
DRP	391	3	12	2	2	1	9	67	68	19	2	86	4	0	21	69	26
ECOLI	396	5	15	7	5	4	5	71	75	8	3	79	3	1	30	65	20
NH4N	206	3	10	6	2	1	7	31	19	13	1	36	2	0	15	41	19
NO3N	511	8	18	2	5	5	10	104	80	20	3	112	5	1	35	76	27
TP	421	3	15	2	3	4	7	68	56	21	2	109	6	1	26	70	28
TN	243	4	12	1	2	1	6	31	36	13	2	45	2	1	28	46	13
CLAR	386	3	20	5	7	5	8	91	71	20	1	95	4	0	11	35	10

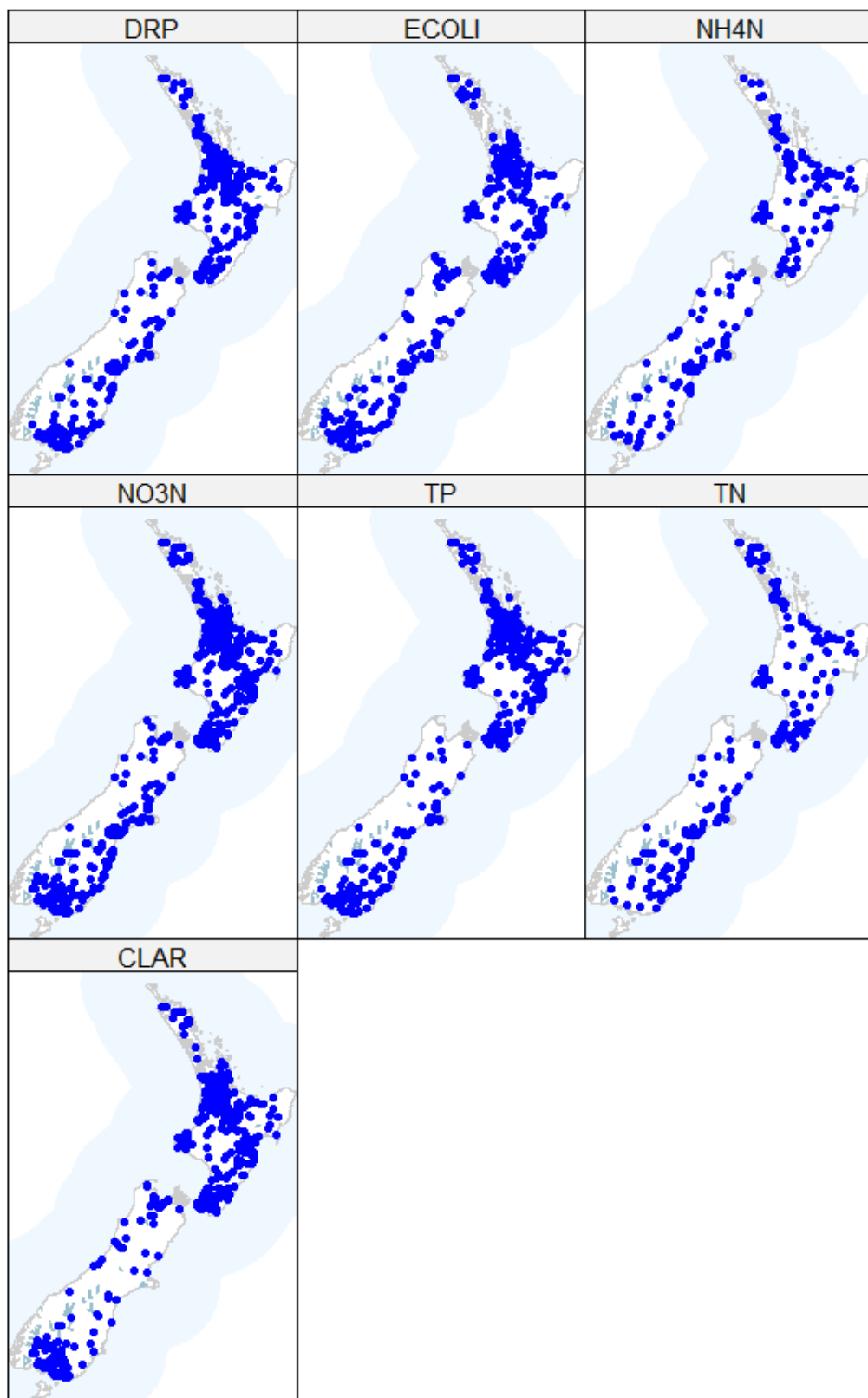


Figure 5-4: Locations of river water quality monitoring sites used for 10-year trend analyses of nutrients, ECOLI and CLAR.

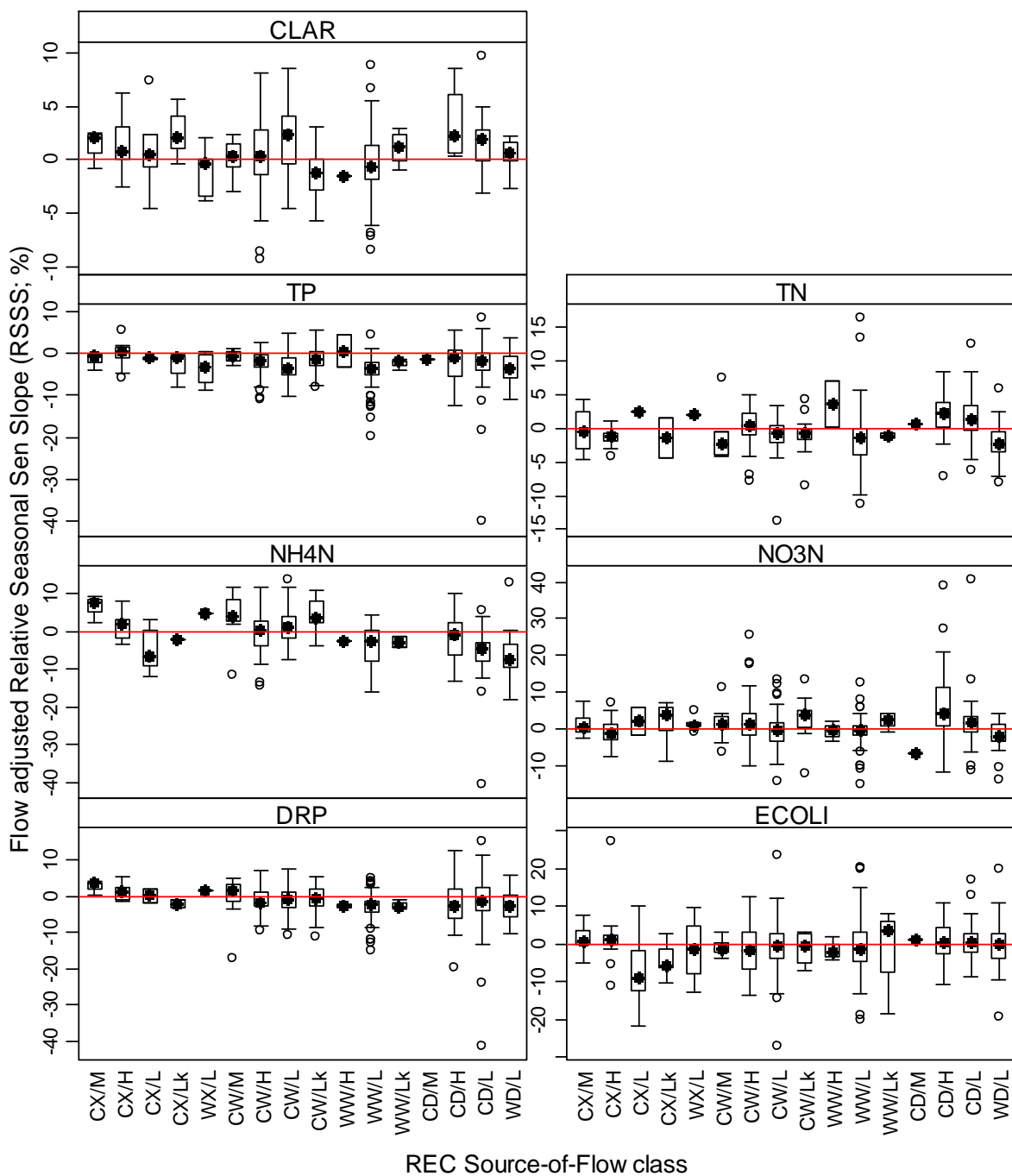


Figure 5-5: Summary of 10-year flow adjusted trends. Box-and-whisker plots show the distributions of site trends within REC classes. The closed circle in each box indicates the median of site trends, the box indicates the inter-quartile range and the whiskers indicate the 5th and 95th percentiles. Outliers are indicated by open circles.

Table 5-3: Numbers of sites in trend categories for 10-year, flow-adjusted trends across REC classes. The definitions of the importance categories are in Section 3.4.5. Degrading trends indicate increasing concentrations of nutrients and ECOLI, and decreases in CLAR. NT: trend importance could not be defined because no thresholds were available. NA: not applicable due to the availability of a threshold. The importance of CLAR trends could not be assessed for REC classes in which the estimated reference state is lower than the MfE (2004) guideline. The importance of TN and DRP trends could not be assessed for all REC classes because periphyton thresholds were not defined. In cases where the importance of trends in individual REC classes could not be assessed, the trends are included in the columns with “no threshold” in the heading.

	Importance category								
Variable	Degrading and important	Improving and important	Degrading not important	Improving not important	Degrading no threshold	Improving no threshold	Total degrading	Total improving	Insufficient data
DRP	22	107	41	92	6	2	69	201	121
ECOLI	3	5	51	76	NA	NA	54	81	261
NH4N	NT	NT	NT	NT	53	89	53	89	64
NO3N	NT	NT	NT	NT	189	122	189	122	200
TP	NT	NT	NT	NT	25	255	25	255	141
TN	19	38	29	33	20	6	68	77	98
CLAR	22	63	21	29	27	38	70	130	186

5.1.4 Twenty year trends (1994 – 2013)

Between 6 and 257 river monitoring sites met the filtering rules for the 20-year trend analysis of nutrients, ECOLI and CLAR (Table 5-4). The numbers of sites also varied substantially by REC class, and there were few or no qualifying sites for some variables in some REC classes. The qualifying sites were reasonably well-distributed geographically (apart from ECOLI, for which there were only six sites), with gaps in the central North and South islands, Marlborough and the West Coast (Figure 5-6). All site locations, REC classes and numbers of sampling dates are included in the supplementary file “NationalRiver20YrTrend_1994-2013”.

Box-and-whisker plots were used to summarise the estimated trends for each of the water quality variables for the 20-year period from 1994 – 2013 (Figure 5-7). All estimated trends are included in these plots, irrespective of their importance categories (as defined in Section 3.4.5). As with the 10-year trends reported in Section 5.1.3, the box plots for 20-year trends indicate that REC classes did not account for a substantial amount of the variation in trends for any variable.

The analysis of 20-year trend categories is shown in Table 5-5. ECOLI trends were calculated for six sites only and the categories for these trends are not considered here. For the remaining variables, there were six times more sites with improving trends in NH₄N as degrading trends, and approximately twice as many sites with improving trends in DRP and TP as degrading trends. In contrast, there were three times as many sites with degrading trends in TN as improving trends, and twice as many sites with degrading trends in NO₃N as improving trends. Improving and degrading trends in CLAR were nearly equal (Table 5-5).

There were relatively more sites with improving and important trends in DRP (18% of sites) compared with degrading and important trends (6 % of sites) (Table 5-5). In contrast, the proportions of sites with improving and important trends in TN and CLAR were roughly equal to the proportions of sites with degrading and important trends (from 2 to 5%).

A comparison of river water quality trends over 10- and 20-year time periods revealed several changes in the balance of improving and degrading trends: 1) there are roughly equal proportions of degrading and improving 10-year trends in TN, but a predominance of degrading 20-year trends; 2) the predominance of improving trends in DRP and TP increased between the 10- and 20-year periods; 3) the predominance of improving trends in NH₄N decreased between the 10- and 20-year periods; 4) the predominance of improving trends in CLAR in the 10-year trends is not apparent in the 20-year trends; and 5) the predominance of degrading trends in NO₃N persisted between the 10- and 20-year periods, but the magnitude shifted (twice as many degrading trends in the 20-year period versus 1.5 times as many in the 10-year period). The complete 20-year trend analysis results are provided in the supplementary file “NationalRiver20YrTrend_1994-2013”.

Table 5-4: Number of river monitoring sites by REC class and water quality variable that were included in the 20-year trend analyses of nutrients, ECOLI and CLAR.
The site numbers shown refer to sites where 90% of the sampling dates in 18 of the years in the 1994-2013 period had observations, and less than 15% of the data for each variable consisted of the number of censored values.

Variable	Total	REC Source- of-Flow class														
		CX/M	CX/H	CX/L	CX/Lk	WX/L	CW/M	CW/ H	CW/L	CW/Lk	WW/H	WW/L	WW/Lk	CD/H	CD/L	WD/L
DRP	209	3	6	2	2	1	6	43	16	18	1	66	4	8	19	14
ECOLI	6	0	0	0	0	0	0	4	0	2	0	0	0	0	0	0
NH4N	138	3	6	1	2	1	6	27	10	12	0	29	2	7	19	13
NO3N	257	3	7	2	3	4	6	58	17	18	1	82	5	12	23	16
TP	231	3	7	2	2	4	6	49	14	18	1	80	6	10	13	16
TN	91	3	6	1	2	1	6	24	8	12	0	12	1	9	3	3
CLAR	200	3	7	2	2	3	7	47	14	18	0	67	4	8	12	6

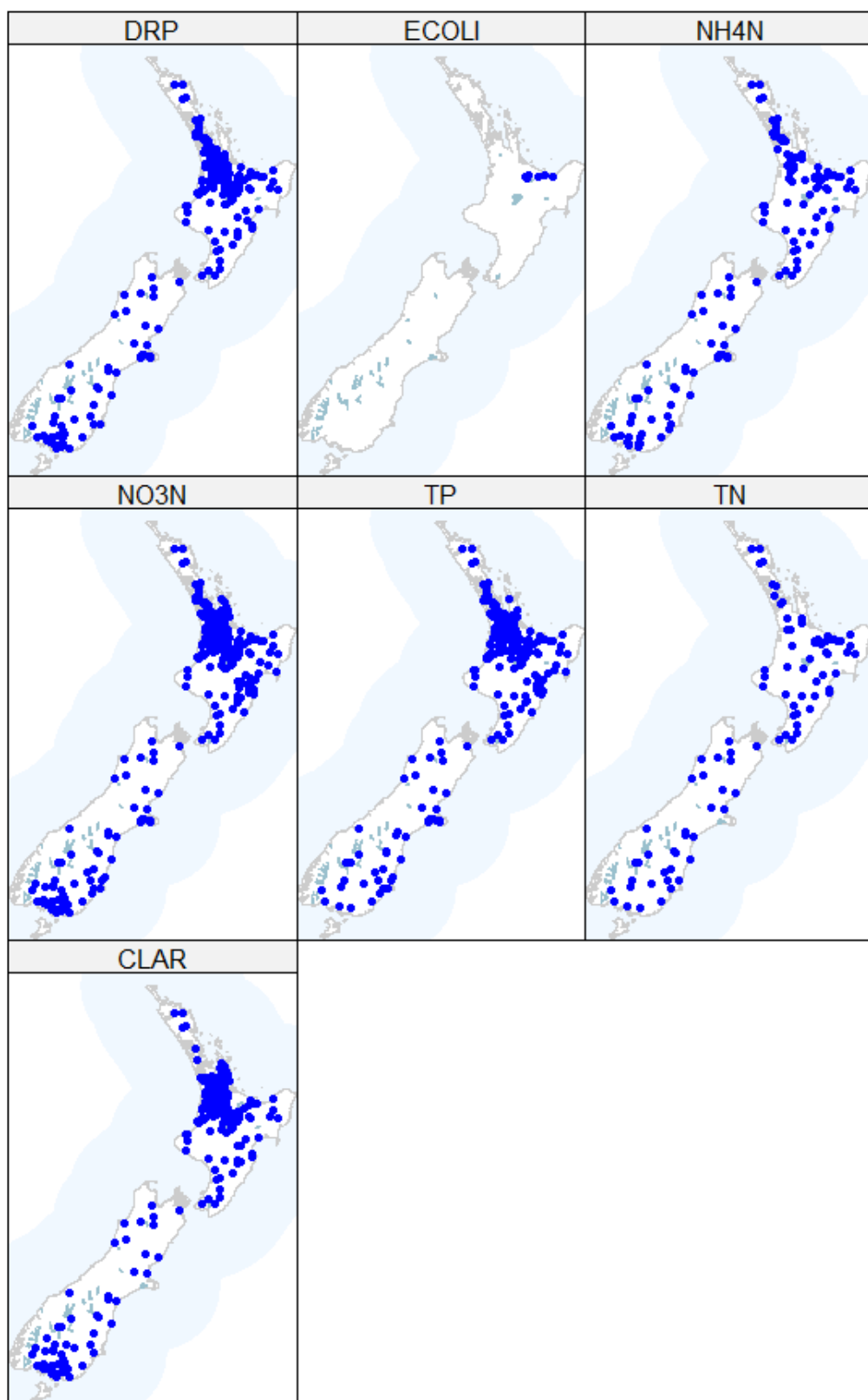


Figure 5-6: Locations of river water quality monitoring sites used for 20-year trend analyses of nutrients, ECOLI and CLAR.

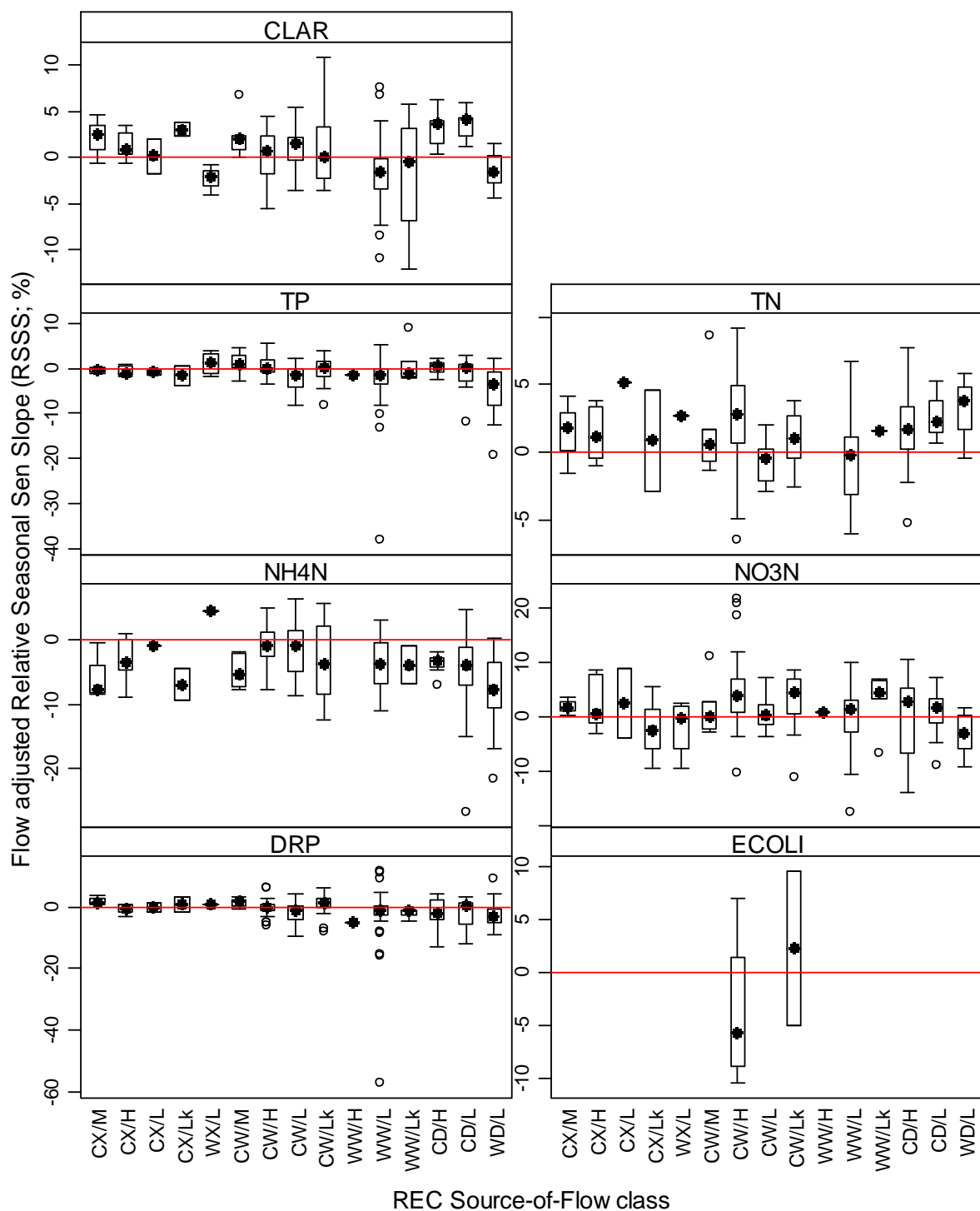


Figure 5-7: Summary of 20-year flow adjusted trends. Box-and-whisker plots show the distributions of site trends within REC classes. The closed circle in each box indicates the median of site trends, the box indicates the inter-quartile range and the whiskers indicate the 5th and 95th percentiles. Outliers are indicated by open circles.

Table 5-5: Numbers of sites in trend categories for 20-year, flow-adjusted trends across REC classes. Degrading trends indicate increasing concentrations of nutrients and ECOLI, and decreases in CLAR. NT: trend importance could not be defined because no thresholds were available. NA: not applicable due to the availability of a threshold. The importance of CLAR trends could not be assessed for REC classes in which the estimated reference state is lower than the MfE (2004) guideline. The importance of TN and DRP trends could not be assessed for all REC classes because periphyton thresholds were not defined. In cases where the importance of trends in individual REC classes could not be assessed, the trends are included in the columns with “no threshold” in the heading.

	Importance category								
Variable	Degrading and important	Improving and important	Degrading not important	Improving not important	Degrading no threshold	Improving no threshold	Total degrading	Total improving	Insufficient data
DRP	12	38	39	48	1	2	52	88	69
ECOLI	0	0	2	4	NA	NA	2	4	0
NH4N	12	76	2	13	NA	NA	14	89	35
NO3N	NT	NT	NT	NT	142	71	142	71	44
TP	NT	NT	NT	NT	46	99	46	99	86
TN	5	3	42	14	2	0	49	17	25
CLAR	4	9	26	45	45	23	75	77	48

We also examined the correspondence between raw and flow adjusted-trends for each water quality variable, for 20-year trends (Figure 5-8). The correspondence for 20-year trends is even closer than the correspondence for 10-year trends shown in Figure 5-3, because sampling dates are more random with respect to flow as the monitoring period increases. The plots in Figure 5-8 indicate that flow adjustment had minimal effects on RSSSE values.

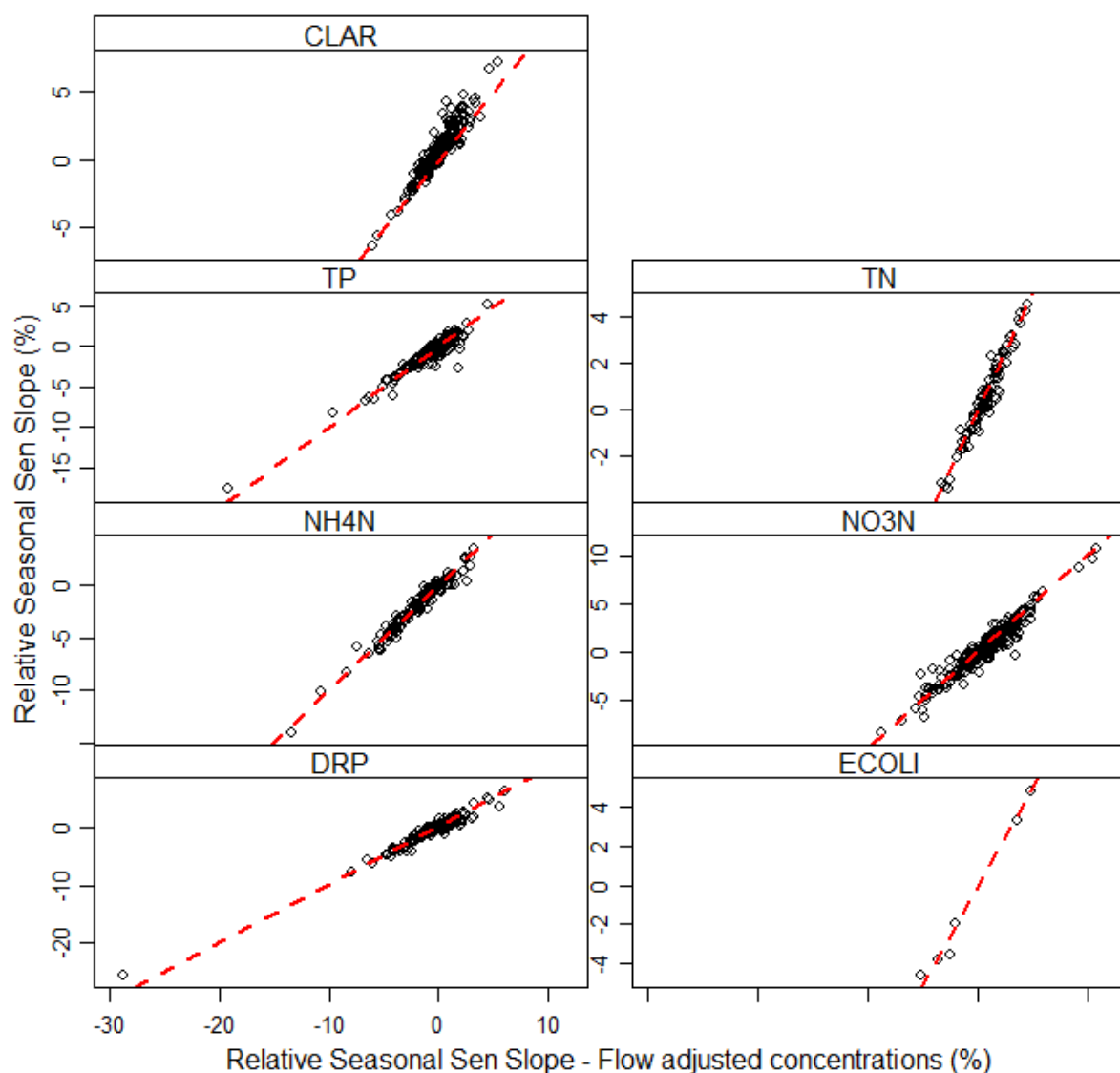


Figure 5-8: Effect of flow adjustment on 20-year trends. The red line is 1 to 1.

5.1.5 Twenty-five year trends at NRWQN sites

All 77 NRWQN sites were included in the 25-year trend analyses of six water quality variables, so no comparison of site numbers versus years of data was necessary. ECOLI was excluded from the trend analyses because ECOLI monitoring at NRWQN sites commenced in 2004. NRWQN site locations, REC classes and site-specific trends are in the supplementary file “NationalRiver25YrTrend_NRWQN2004-2013”.

Box-and-whisker plots for the 25-year trends indicate that REC classes did not account for a substantial amount of the variation in trends for any variable (Figure 5-9). Note that NRWQN sites are unevenly distributed across REC classes; 13 Source-of-Flow classes are represented in the NRWQN, but the number of sites per class ranges from 20 in the CW/H class to one in each of three classes, CX/L, WW/Lk and WX/L.

Summing across the improving and degrading categories, there were approximately four times more NRWQN sites with degrading trends in DRP and TN as improving trends, and twice as many sites with degrading trends in NO₃N as improving trends (Table 5-6). In contrast, there were 20 times more sites with improving trends in NH₄N as degrading trends, and seven times more sites with improving trends in CLAR as degrading trends. The number of sites with improving and degrading trends in TP was nearly equal. Sites with important and degrading or important and improving trends in DRP, TN and CLAR were very rare (range: 0-4% of sites) (Table 5-6).

For most water quality variables, the predominant direction of 25-year trends at the NRWQN sites is consistent with the 20-year trends (1989-2009) at the same sites reported by Ballantine and Davies-Colley (2014). In both analyses, the majority of trends for which direction could be confidently inferred were degrading for NO₃N, DRP and TN and improving for CLAR. The predominant directions of the 25-year trends were also roughly consistent with the 20-year trends in the larger national dataset (compare Tables 5-5 and 5-6), with degrading trends in NO₃N and TN and improving trends in NH₄N. The principal differences were for DRP and CLAR; most DRP trends were degrading and most CLAR trends were improving in the 25-year NRWQN analysis, whereas most DRP trends were improving and CLAR trends were evenly divided into improving and degrading in the larger 20-year analysis.

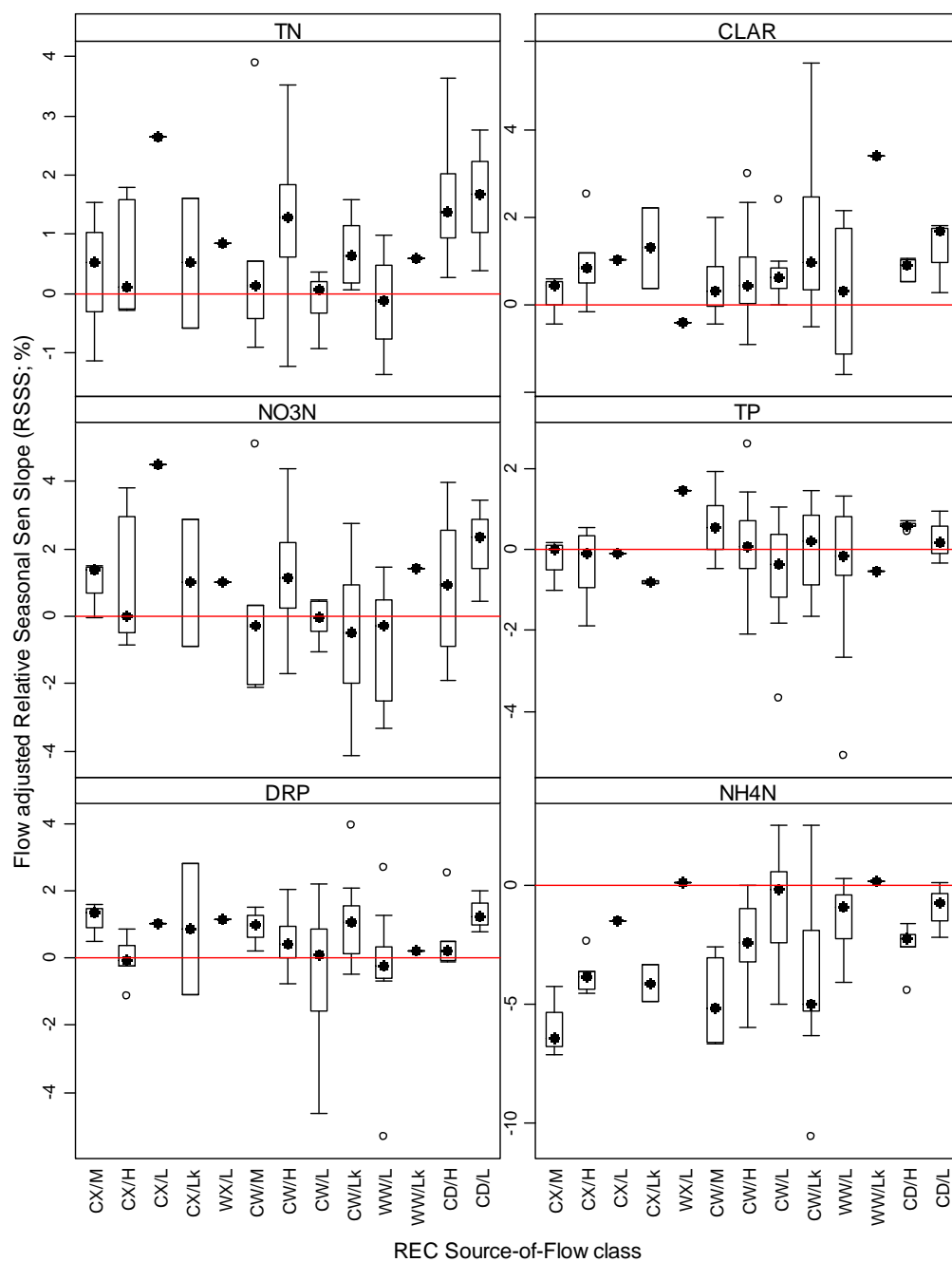


Figure 5-9: Summary of 25-year, flow-adjusted trends at NRWQN sites. Box-and-whisker plots show the distributions of site trends within REC classes. The closed circle in each box indicates the median of site trends, the box indicates the inter-quartile range and the whiskers indicate the 5th and 95th percentiles. Outliers are indicated by open circles.

Table 5-6: Numbers of sites in trend categories for the 25-year datasets for NRWQN sites. ECOLI was excluded from the analysis because no data were available prior to 2004. The definitions of the importance categories are in Section 3.4.5. Degrading trends indicate increasing concentrations of nutrients and decreases in CLAR. NT: trend importance could not be defined because no thresholds were available. NA: not applicable due to the availability of a threshold. The importance of CLAR trends could not be assessed for REC classes in which the estimated reference state is lower than the MfE (2004) guideline. The importance of TN and DRP trends could not be assessed for all REC classes because periphyton thresholds were not defined. In cases where the importance of trends in individual REC classes could not be assessed, the trends are included in the columns with “no threshold” in the heading.

	Importance category								
Variable	Degrading and important	Improving and important	Degrading not important	Improving not important	Degrading no threshold	Improving no threshold	Total degrading	Total improving	Insufficient data
DRP	0	2	37	8	2	1	39	11	27
NH4N	3	57	0	3	NA	NA	3	60	14
NO3N	NT	NT	NT	NT	40	21	40	21	16
TP	NT	NT	NT	NT	29	23	29	23	25
TN	2	1	42	10	2	0	46	11	20
CLAR	0	3	4	39	3	7	7	49	21

5.2 River MCI trends

5.2.1 Trade off analysis

The trade-offs between the number of qualifying invertebrate monitoring sites (i.e., sites that met our filtering rules for trend analyses) and the monitoring periods for those sites are shown in Figure 5-10. There were fewer than 100 sites with ≥ 20 years of MCI data. However, there were 461 sites with 10 years of data, and 10 years represented a good site number versus duration trade-off.

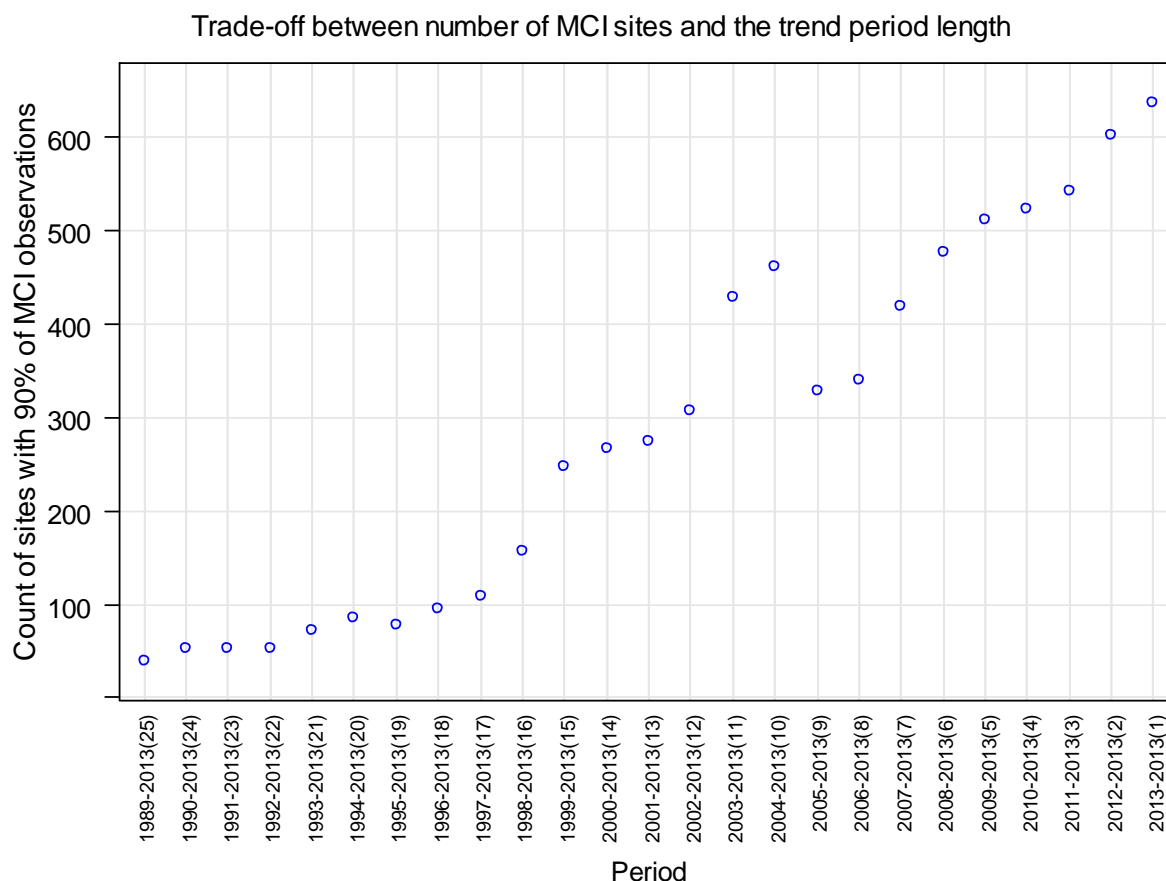


Figure 5-10: Changes in the number of river macroinvertebrate monitoring sites that met the filtering rules versus the period of site operation. Durations of periods are shown in parentheses.

5.2.2 Ten year trends in MCI scores

The numbers of sites that qualified for trend analyses of MCI scores varied widely by REC class (Table 5-7). The sites were reasonably well distributed geographically, with gaps in the central areas of the both the North and South islands and in the Waikato and Bay of Plenty regions (Figure 5-11). All site locations, REC classes and numbers of sampling dates are included in the supplementary file “NationalMCI10YrTrend2004-2013”.

Table 5-7: Number of river monitoring sites within REC class that were included in the 10-year trend analyses of MCI scores. The site numbers shown refer to sites where MCI scores were determined annually in nine out of ten years between 2004 and 2013.

REC class	Number of sites
CX/M	4
CX/H	30
CX/L	24
CX/Lk	4
WX/L	6
CW/M	14
CW/H	85
CW/L	81
CW/Lk	5
WW/H	3
WW/L	78
CD/M	2
CD/H	26
CD/L	78
WD/L	21
Total	461

Box-and-whisker plots were used to summarise the estimated trends in MCI scores for the 10-year period from 2004 – 2013 (Figure 5-12). All estimated trends are included in these plots, irrespective of their importance categories (as defined in Section 3.4.5). The plots indicate that REC classes explained little variation in MCI trend. For most REC classes, median RSSSE values were close to zero. There is no threshold for evaluating the importance of MCI trends. However, there were more sites with decreasing (i.e., degrading) trends (59 sites) than increasing (i.e., improving) trends (20 sites). For the remaining 382 sites, the data were insufficient to confidently determine trend direction. The complete 10-year MCI trend analysis results are provided in the supplementary file are included in the supplementary file “NationalMCI10YrTrend2004-2013”.

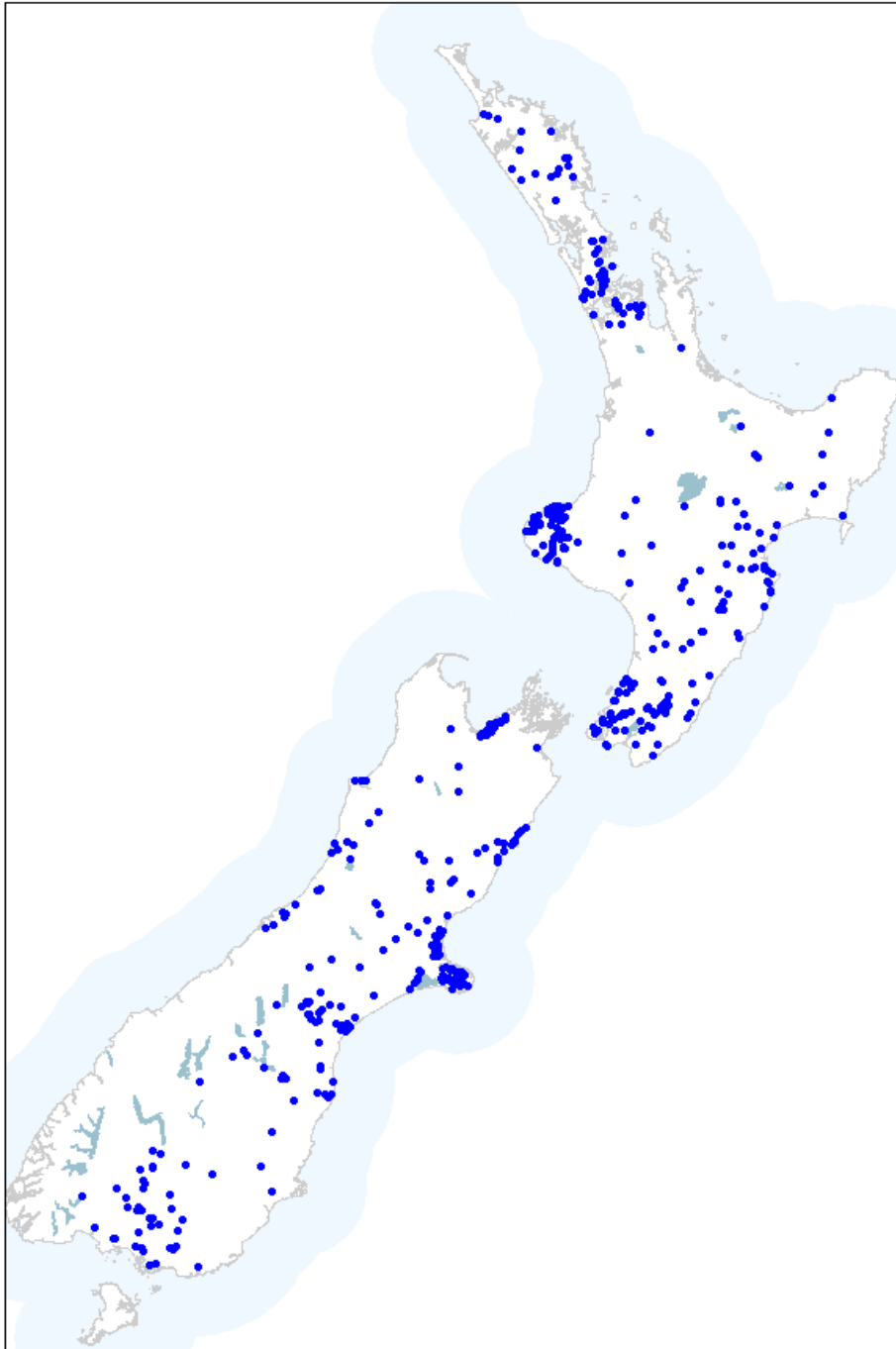


Figure 5-11: Locations of river macroinvertebrate monitoring sites used for 10-year trend analyses of MCI scores.

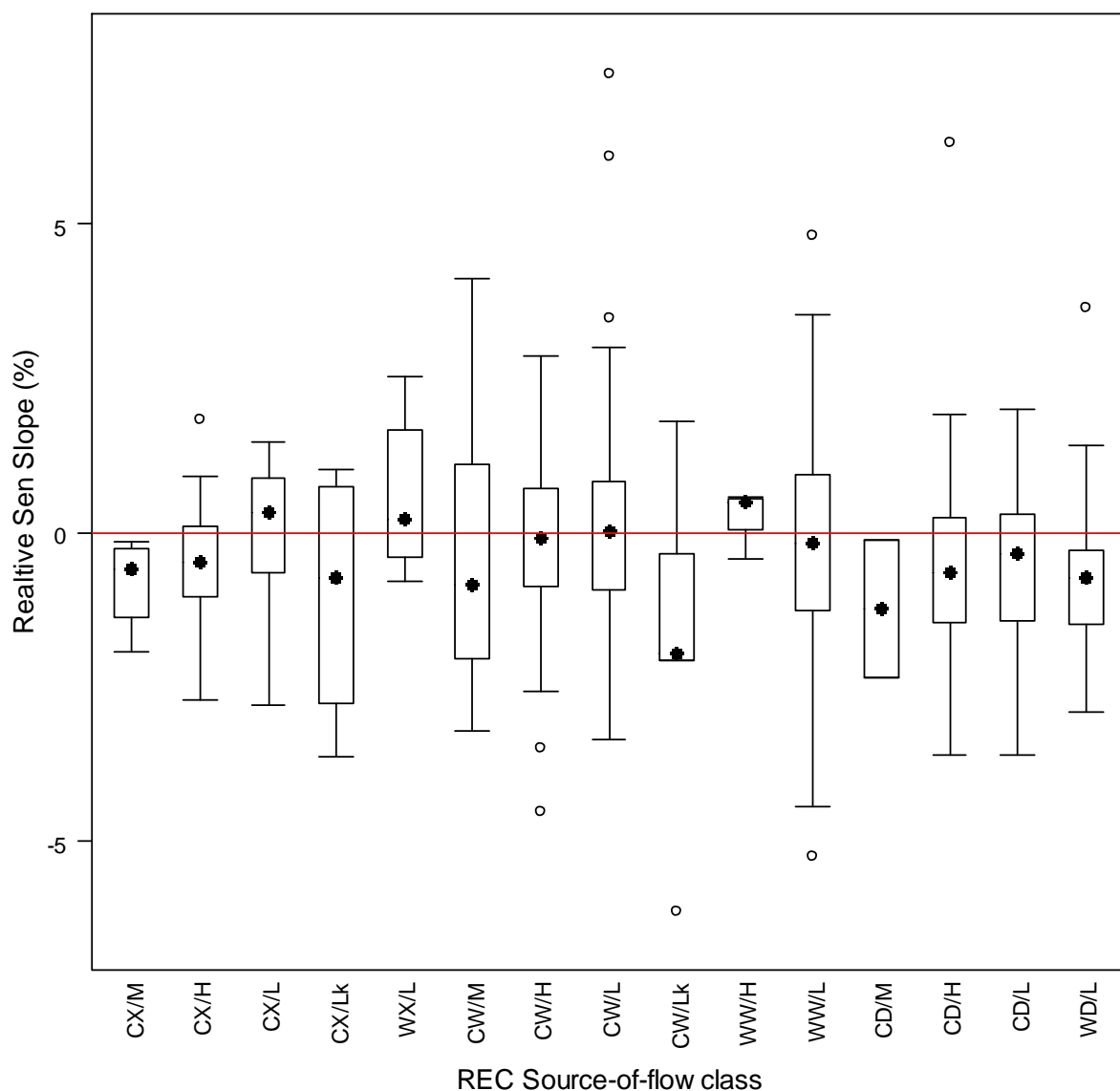


Figure 5-12: Summary of 10-year MCI trends by REC class. Box-and-whisker plots show the distributions of site trends within REC classes. The closed circle in each box indicates the median of site trends, the box indicates the inter-quartile range and the whiskers indicate the 5th and 95th percentiles. Outliers are indicated by open circles.

5.3 Lake water quality trends

5.3.1 Trade-off analysis

The trade-off analysis (Figure 5-13) indicated that the 10-year period (2004-2013) represents a good trade-off between the duration of monitoring and number of lake sites. Site yields were considerably less than for rivers at any duration, but using a ten year period kept the lake trend analysis consistent with the river water quality analysis. Trend analyses over longer periods were not carried out due to the shortage of eligible monitoring sites.

5.3.2 Ten-year trends (2004 – 2013)

Between 19 and 44 lake monitoring sites met the filtering rules for the 10-year trend analysis of water quality variables. The number of sites varied widely by lake class (Table 5-8). The qualifying lake sites are sparsely and unevenly distributed on both the North and South Islands (Figure 5-14). All site locations, lake classes and numbers of sampling dates are included in the supplementary file “NationalLakeTrend_2004-2013”.

Box-and-whisker plots were used to summarise the estimated trends for each of the lake water quality variables for the 10-year period from 2004 – 2013 (Figure 5-15). All estimated trends are included in these plots, irrespective of their importance categories (as defined in Section 3.4.5). The box plots indicate that declining trends in DO_{bottom} were concentrated in the deeper lakes, and declining trends in TN and increasing trends in NO₃N occurred in lakes in all elevation × depth classes. No consistent differences were apparent in trends in low- and high-elevation lakes.

Summing across the improving and degrading categories, improving trends in CHLA, TLI, NH₄N, TN and TP occurred at three to five times as many sites as degrading trends (Table 5-9). In contrast degrading trends in NO₃N occurred at six sites and an improving trend at one site, and degrading trends in DO_{bottom} occurred at 13 sites, with improving trends at two sites. For DRP and SECCHI, there was not a clear predominance of sites with improving trends versus degrading trends. Three water quality variables associated with thresholds were included in the importance assessments, CHLA, TN and TP. Important 10-year trends in these variables were identified at 19 - 25% of the qualifying lake sites (Table 5-9). Most of the important trends indicated improvement; no sites had degrading and important trends in CHLA or TP and two sites had degrading and important trends in TN. The complete 10-year lake trend analysis results are provided in the supplementary file “NationalLakeTrend_2004-2013”.

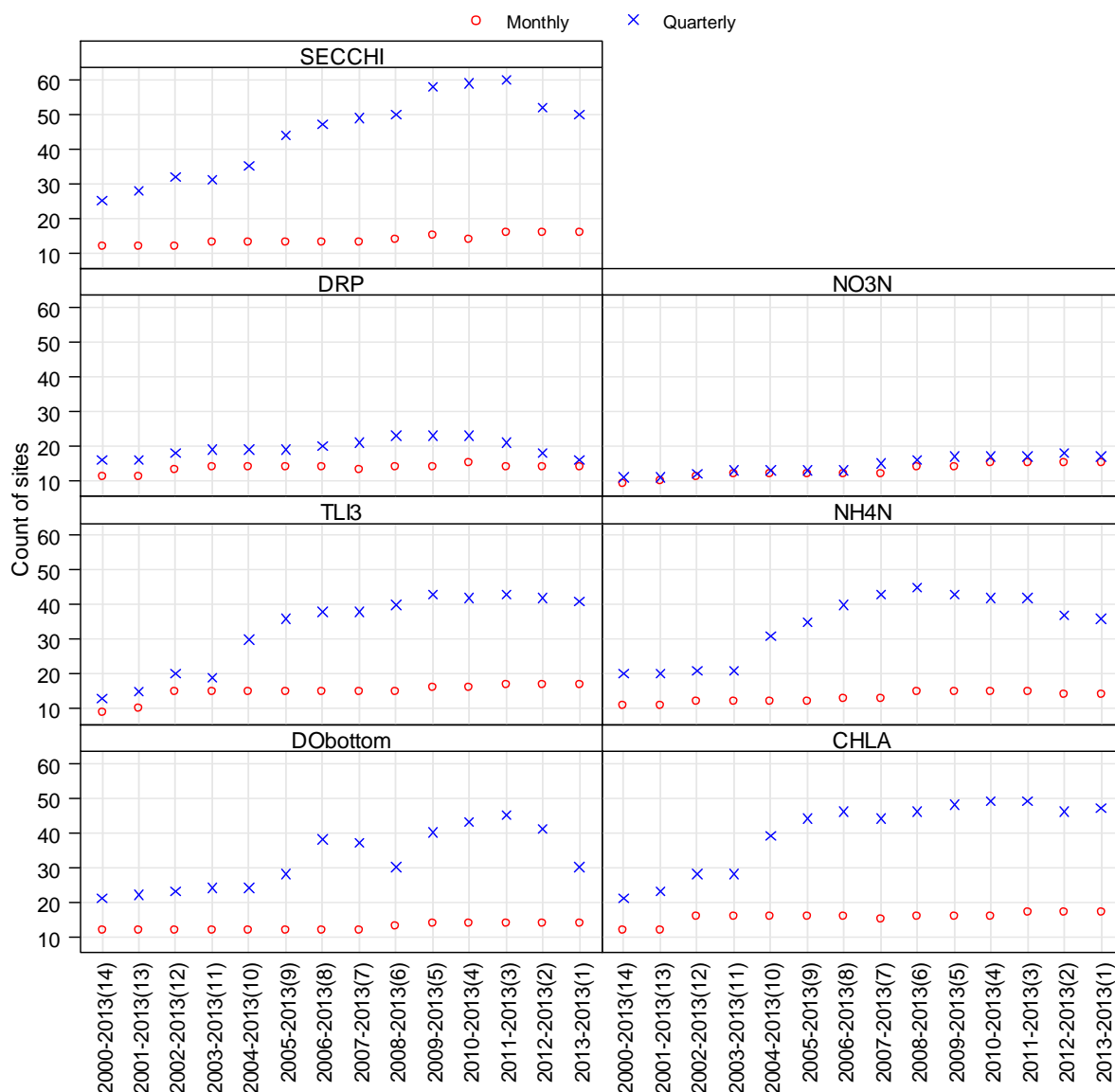


Figure 5-13: Changes in the number of lake monitoring sites that met the filtering rules for each water quality variable versus the period of site operation. Durations of periods are shown in parentheses. Open circles: monthly data, crosses: quarterly data. The plots were used to select time periods for trend analyses.

Table 5-8: Number of sites by lake class eligible for the 10 year trend analysis. Number of lake monitoring sites by elevation × depth class and water quality variable that were included in the 10-year trend analyses. The site numbers shown refer to sites where 80% of the sampling dates in 8 of the years in the 2004-2013 period had observations, and less than 15% of the data for each variable consisted of the number of censored values.

Variable	Elevation × depth class								Total
	0-300 m 0-5 m	0-300 m 5-15 m	0-300 m 15-50 m	0-300 m > 50 m	> 300 m 0-5 m	> 300 m 5-15 m	> 300 m 15-50 m	> 300 m > 50 m	
DObottom	4	5	4	3	0	0	4	4	24
CHLA	15	9	4	3	0	0	4	4	39
TLI3	10	7	3	3	0	0	4	3	30
NH4N	7	9	4	3	1	0	4	3	31
DRP	3	3	3	2	0	0	4	4	19
NO3N	1	1	1	3	0	0	3	4	13
TN	10	7	4	3	0	0	4	3	31
TP	16	11	6	3	0	0	4	4	44
SECCHI	7	8	6	6	0	0	4	4	35

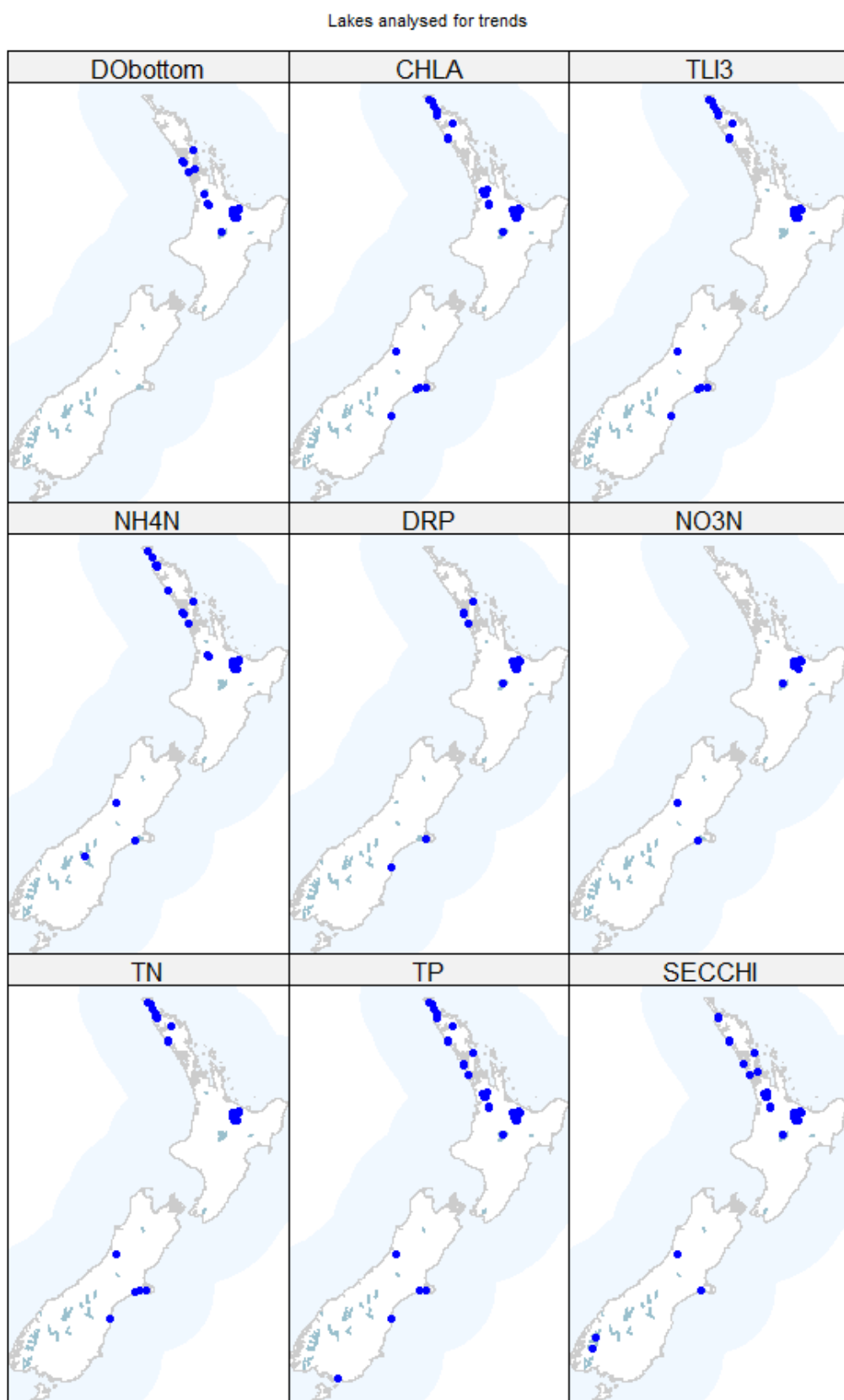


Figure 5-14: Locations of lake water quality monitoring sites used for 10-year trend analyses of water quality variables.

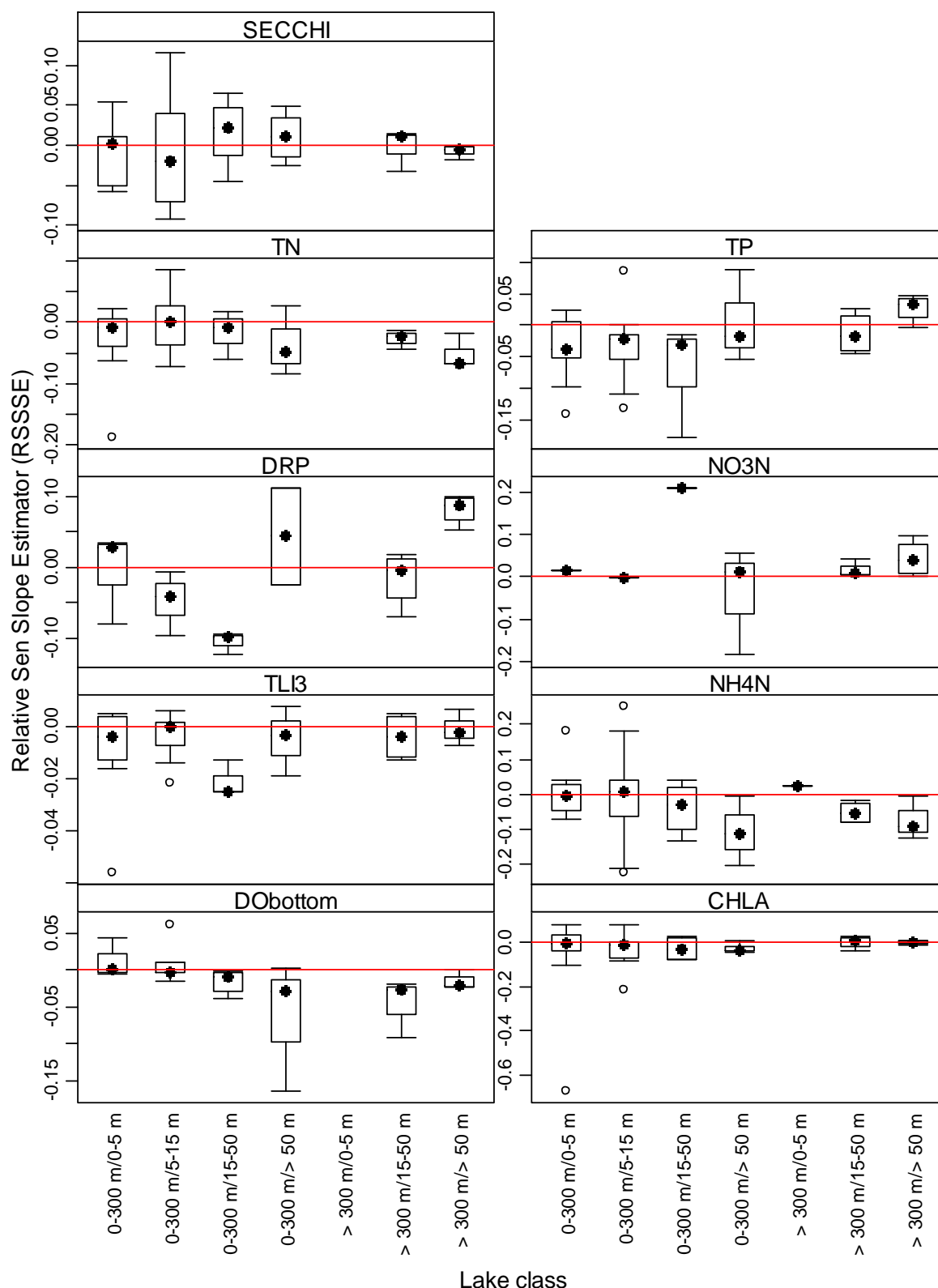


Figure 5-15: Summary of 10-year trends in lake water quality variables, within elevation x depth classes. Box-and-whisker plots show the distributions of site trends in each class. The closed circle in each box indicates the median of site medians, the box indicates the inter-quartile range and the whiskers indicate the 5th and 95th percentiles. Outliers are indicated by open circles.

Table 5-9: Number of lake sites in 10-year trend categories across all lake elevation × depth classes. The definitions of the importance categories are in Section 3.4.5. Degrading trends indicate increasing concentrations of nutrients, TLI and CHLA, and decreases in DO_{bottom} and SECCHI. NT: trend importance could not be defined because no thresholds were available. NA: not applicable due to the availability of a threshold.

	Importance category								
Variable	Degrading and important	Improving and important	Degrading not important	Improving not important	Degrading no threshold	Improving no threshold	Total degrading	Total improving	Insufficient data
DO _{bottom}	NT	NT	NT	NT	13	2	13	2	9
CHLA	NT	8	3	7	NA	NA	3	15	21
TLI	NT	NT	NT	NT	11	4	11	4	15
NH ₄ N	NT	NT	NT	NT	14	3	14	3	14
DRP	NT	NT	NT	NT	9	6	9	6	4
NO ₃ N	NT	NT	NT	NT	1	6	1	6	6
TN	2	4	3	14	NA	NA	5	18	8
TP	NT	11	6	9	NA	NA	6	20	18
SECCHI	NT	NT	NT	NT	11	12	11	12	12

6 Discussion

6.1 State and trends in water quality

The primary purpose of the state and trend analyses reported here is to provide MfE with information needed for a synthesis report on freshwater. The detailed information for each river and lake monitoring sites is contained in the supplementary files that accompany this report. The sites and corresponding water quality conditions can be aggregated in different ways to suit the synthesis report (e.g., by region, environmental class, nation-wide). Therefore, we have limited our interpretation of the summary analyses in this report to a concise overview.

6.1.1 River state and trends

In general, nutrient and ECOLI concentrations were elevated and CLAR low in the low-elevation REC Source-of-Flow classes. Nutrient and ECOLI concentrations were lowest and CLAR highest in REC mountain Source-of-Flow classes and in the CX/Lk and CD/Lk classes. These results are consistent with previous empirical and modelling studies based on the national monitoring-site network (e.g., Larned et al. 2003, Ballantine et al. 2010, Unwin et al. 2010). In addition to lowland sites, sites in the WW/Lk class had elevated nutrient concentrations and low CLAR; these sites were limited to seven North Island rivers. MCI scores were lowest in low-elevation Source-of-Flow classes, and this is also consistent with previous studies (e.g., Clapcott et al. 2013). Poor water-quality at low-elevation river monitoring sites on a national scale is generally attributed to the relatively high proportions of agricultural and urban land in low-elevation areas of New Zealand (Larned et al. 2004).

The 10- and 20-year trend analyses indicated that more monitoring sites have improving trends in DRP, TP and NH₄ than degrading trends. These patterns suggest that there have been consistent reductions in phosphorus enrichment for at least 20 years. However, the predominance of sites with degrading trends in NO₃N has persisted over the 10- and 20-year periods. While the improving trend in NH₄N is encouraging, NH₄N concentrations in most New Zealand rivers are less than 10% of the NO₃N concentrations. The predominance of sites with upward trends in NO₃N concentrations suggests that dissolved inorganic nitrogen enrichment has generally worsened.

The results of our assessment of trend importance suggested that relatively few of the sites for which directions of trends in DRP, TN and CLAR could be confidently inferred were degrading and important (i.e., few sites were predicted to exceed the NPS-FM periphyton bottom line and the MfE clarity guideline within 10 years). Relatively more sites had important and improving trends, and are predicted to reach the reference condition threshold for their respective REC class within 10 years, if improving trends continue uninterrupted. We reiterate that the our use of a 10-year critical period and NPS-FM bottom lines, reference conditions and other thresholds for assessing water quality trends in rivers was one of many possible approaches. Our primary aim was to demonstrate the method set out in Appendix A.

6.1.2 Lake state and trends

Lake water quality was generally poorer in low-elevation and shallow lakes compared with high-elevation and deep lakes. This pattern reflects the combined influence of catchment land use and lake stratification and mixing, and is consistent with recent studies (Schallenberg and Sorrell 2009, Verburg et al. 2010,). As with rivers, poor water-quality in low-elevation lakes across New Zealand has been attributed to agricultural and urban land use in the lake catchments (Drake et al. 2011).

Sites with sufficient data to compute 10-year trends were almost entirely restricted to Northland, Auckland, Waikato and Bay of Plenty regions. The lake elevation × depth classification did not explain a substantial amount of spatial variation in water quality trends, and there were few trends for which the directions could be inferred with confidence. However, some variable-specific patterns were evident, including declining trends in bottom DO in lakes >50 m deep. Reduced DO near the beds of deep lakes may increase the risk to these lakes of internal nutrient loading from sediment efflux.

The 10-year trend analyses for lake water quality variables indicated that improving trends in CHLA, TLI, NH₄N, TN and TP outnumbered degrading trends, and that degrading trends in NO₃N and DO_{bottom} outnumbered improving trends. However, the number of lakes that met the filtering rules for trend analyses was limited, ranging from 20 to 84 depending on the water quality variable, and we could only infer trend directions with confidence in 7 to 24 of those lakes. These numbers of within-site trends are too small to extrapolate across New Zealand lakes in general.

6.2 New procedures for censored data and trend analyses

Most river and lake water quality datasets contain censored data that represent laboratory results reported as less than a detection limit. In these cases, the laboratory analyst lacks confidence in the low number generated by a measurement instrument and so censors it. This is particularly common for nutrient concentrations in samples from unpolluted water. In another, smaller number of cases, laboratory measurements of *E. coli* concentrations and field measurements of water clarity are reported as greater than an upper reporting limit. Traditionally, the most common procedures for processing censored data were to substitute the censored values with a fraction of the detection limit (typically ½) or a multiple of the reporting limit (typically 1.1). These substitutions can have untoward effects on water quality trend analyses, because the nonparametric methods used for those analyses estimate the trend slope as the median of all within-season slopes. When that calculation is applied to datasets with even small proportions of censored data, overall trends of zero can result. This is because the censored data can have identical values (e.g., half of a constant reporting limit). To avoid these problems, and to encourage best practices, we have used modern imputation methods (Helsel 2012), which ensure that each censored value is replaced by a unique number. The new procedures never predict slopes of exactly zero, which is consistent with the fundamental axiom of our trend assessment methodology.

The trend analyses in this report were carried out using a new procedure that we believe provides more useful information than traditional procedure. We replaced the traditional hypothesis test (which posits that the trend slope is exactly zero) with hypotheses concerning the direction of the trend. With this approach it is axiomatic that true trend slopes cannot be zero, and the traditional hypothesis is *a priori* false; our new procedure is based on the fundamental premise that there is always a trend (i.e., a non-zero slope). The two questions that water quality managers then face are: 1) can we confidently infer the direction of a trend?; and 2) if we can infer the trend direction, is the trend environmentally important? Our procedure addresses both questions. Note that if the answer to question 1 is “no”, then the valid inference is simply “insufficient data to detect trend direction”, which is a more informative statement than the ambiguous “not statistically significant” statement in the traditional methodology. Our procedure also corrects the frequent interpretation of the failure to reject the null hypothesis as evidence that there is no trend (e.g., that water quality conditions are “stable” or “being maintained”).

The new trend analysis procedure that concerns trend direction is strictly objective. If different analysts apply the same procedure to a given dataset they will reach the same conclusions regarding trend direction: “upward”, “downward”, “insufficient data”. However, the second part of our new procedure — assessing the importance of trends for which the direction has been identified—is not strictly objective. Indeed, it cannot be. That is because the assessment requires criteria about the importance of trends in terms of magnitude and the time period during which trends persist. Obtaining agreement on such criteria will be difficult, due to the range of opinions among water resources managers. In this project we considered a water quality trend to be important if it was predicted to indicate a substantial change in water quality within 10 years. We defined a substantial change as one in which a recognised threshold would be exceeded. For thresholds associated with degrading trends, we used the NPS-FM bottom-lines for some variables, and the MFE (1994) clarity guideline. Other variables have no widely recognised ecological thresholds, and we did not assess trend importance in those variables. We recognise that there are as yet no standardised methods for assessing the importance of water quality trends in New Zealand. Our choices of thresholds and time periods were nominal and other thresholds and time periods may be justified. Therefore, our assessments should be viewed as a demonstration of the new method rather than definitive statements about the importance of different trends. Developing standardised thresholds would be a useful step toward consistent assessment of water-quality trends.

7 Acknowledgements

We thank Sheree De Malmanche (Ministry for the Environment) for project support, feedback, enthusiasm, and patience. We thank Maree Clark (Horizons Regional Council) and the LAWA team for data updates. We thank Graham Bryers (NIWA) and Julian Sykes (NIWA) for assistance with NRWQN data and GIS analyses. And we thank the many regional council staff who provided additional data and information about monitoring programmes. Helen Rouse and Clive Howard-Williams reviewed the report.

8 References

- Ballantine, D.J., Booker, D., Unwin, M., Snelder, T. (2010) Analysis of national river water quality data for the period 1998-2007. New Zealand Ministry for the Environment. *NIWA Client Report* CHC2010-038.
- Booker, D. J., Snelder, T.H. (2012) Comparing methods for estimating flow duration curves at ungauged sites *Journal of Hydrology*, 434: 78-94.
- Booker, D.J., Woods, R.A. (2014) Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments. *Journal of Hydrology*, 508: 227-239.
- Burns NM, Rutherford JC 1998. Results of monitoring New Zealand lakes 1992-1996. Vol. 1. General findings. New Zealand Ministry for the Environment. *NIWA Client Report* MfE80216/1.
- Burns, N., Bryers, G., Bowman, E. (2000) Protocol for monitoring trophic levels of New Zealand lakes and reservoirs. Ministry for the Environment.
- Clapcott, J., Goodwin, E., Snelder, T. (2013) Predictive models of benthic macroinvertebrate metrics. New Zealand Ministry for the Environment. *Cawthron Report* 2301.
- Clark, M.P., Rupp, D.E., Woods, R.A., Zheng, X., Ibbitt, R.P., Slater, A.G., Schmidt, J., Uddstrom, M.J. (2008) Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Advances in Water Resources*, 31: 1309-1324.
- Davies-Colley, R., Hughes, A.O., Verburg, P., Storey, R. (2012) Freshwater monitoring protocols and quality assurance. New Zealand Ministry for the Environment. *NIWA Client Report* HAM2012-92.
- Drake, D.C., Kelly, D., Schallenberg, M. (2011) Shallow coastal lakes in New Zealand: current conditions, catchment-scale human disturbance, and determination of ecological integrity. *Hydrobiologia*, 658: 87-101.
- Helsel, D.R. (2012) Statistics for Censored Environmental Data Using Minitab® and R. Wiley, New York.
- Helsel, D.R.; Hirsch, R.M. (1992). *Statistical Methods in Water Resources*. USGS Techniques of Water Investigations Book 4, Chapter A3, 510 p.
- Hirsch, R.M., Slack, J.R., Smith, R.A. (1982) Techniques of trend analysis for monthly water quality data. *Water Resources Research*, 18: 107-121.
- Larned, S.T., Scarsbrook, M., Snelder, T., Norton, N. (2003) Nation-wide and regional state and trends in river water quality 1996-2002. New Zealand Ministry for the Environment. *NIWA Client Report* CHC2003-051.
- Larned, S.T., Scarsbrook, M., Snelder, T., Norton, N., Biggs, B.J.F. (2004) Water quality in low-elevation streams and rivers of New Zealand: recent state and trends in contrasting land-cover classes. *New Zealand Journal of Marine and Freshwater Research*, 38: 347-366.

- Larned, S.T., Unwin, M. (2012) Representativeness and statistical power of the New Zealand river monitoring network. New Zealand Ministry for the Environment. *NIWA Client Report* CHC2012-079.
- McBride, G.B. (2005) Using Statistical Methods for Water Quality Management: Issues, Problems and Solutions. John Wiley & Sons, New York.
- McBride, G.B., Cole, R., Westbrooke, I., Jowett, I.G. (2014) Assessing environmentally significant effects: a better strength-of-evidence than a single *P* value? *Environmental Monitoring and Assessment* 186: 2729–2740.
- McDowell, R.W., Snelder, T.H., Cox, N., Booker, D.J., Wilcock, R.J. (2013) Establishment of reference or baseline conditions of chemical indicators in New Zealand streams and rivers relative to present conditions. *Marine and Freshwater Research*, 64: 387–400.
- McMillan, H.K., Hreinsson, E.O., Clark, M.P., Singh, S.K., Zammit, C., and Uddstrom, M.J. (2013) Operational hydrological data assimilation with the recursive ensemble Kalman filter. *Hydrology and Earth System Science*, 17: 21–38.
- Ministry for the Environment. (1994) Guidelines for the management of the colour and clarity of water. Water quality guidelines 2.
- Nash, J., Sutcliffe, J.V. (1970) River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, 10: 282–290.
- Piñeiro, G., Perelman, S., Guerschman, J.P., Paruelo, J.M. (2008) How to evaluate models: observed vs. predicted or predicted vs. observed? *Ecological Modelling*, 216: 316–322.
- Quinn, J.M., Raaphorst, E. (2009) Trends in nuisance periphyton cover at New Zealand National River Water Quality Network sites 1990–2006. New Zealand Ministry for the Environment. *NIWA Client Report* HAM2008-194.
- Scarsbrook, M.R. (2006). State and trends in the National River Water Quality Network (1989–2005). New Zealand Ministry for the Environment. *NIWA Client Report* HAM2006-131.
- Schallenberg, M., Sorrell, B. (2009) Factors related to clear water vs. turbid water regime shifts in New Zealand lakes and implications for management and restoration. *New Zealand Journal of Marine and Freshwater Research*, 43: 701–712
- Sen, P.K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association* 63: 1379–1389.
- Smith, D.G., McBride, G.B., Bryers, G.G., Wisse, J., Mink, D.F. (1996) Trends in New Zealand's national river water quality network. *New Zealand Journal of Marine and Freshwater Research*, 30: 485–500.
- Snelder, T., Biggs, B.J.F. (2002) Multiscale river environment classification for water resources management. *Journal of the American Water Resources Association*, 38: 1225–1239.
- Snelder, T., Biggs, B., Kilroy, C., Booker, D. (2013) National Objective Framework for periphyton. New Zealand Ministry for the Environment. *NIWA Client Report* CHC2012-122.

- Sorrell, B., Unwin, M., Dey, K., Hurren, H. (2006) A snapshot of lake water quality. New Zealand Ministry for the Environment. *NIWA Client Report* CHC2006–145.
- Stark, J D., Boothroyd I., Harding J., Maxted J., Scarsbrook M. (2001). Protocols for sampling macroinvertebrates in wadeable streams. New Zealand Ministry for the Environment.
- Stark, J D., Maxted, J. (2007) A user guide for the MCI. New Zealand Ministry for the Environment. Cawthron Report 1166.
- Tait, A., Henderson, R., Turner, R., Zheng, X. (2006) Thin plate smoothing spline interpolation of daily rainfall for New Zealand using a climatological rainfall surface. *International Journal of Climatology*, 26: 2097-2115.
- Unwin, M., Snelder, T., Booker, D., Ballantine, D., Lessard, J. (2010) Modelling water quality in New Zealand rivers from catchment-scale physical, hydrological and land-cover descriptors using random forest models *NIWA Client Report* CHC2010-037.
- Unwin, M. Larned, S.T. (2013) Statistical models, indicators and trend analyses for reporting national-scale river water quality. New Zealand Ministry for the Environment. *NIWA Client Report* CHC2013-033.
- Verburg, P., Hamill, K., Unwin, M., J. Abell, J. (2010) Lake water quality in New Zealand 2010: status and trends. New Zealand Ministry for the Environment. *NIWA Client Report* HAM2010-107.

Appendix A A new approach to water quality trend assessment

Authors:

Graham McBride (NIWA, Hamilton)
Ton Snelder (LWP, Christchurch)
Martin Unwin (NIWA, Christchurch)
Doug Booker (NIWA, Christchurch)
Piet Verburg (NIWA, Hamilton)
Scott Larned (NIWA, Christchurch)

8 May 2015

Contents

1.	Summary	1
1.1.	Implementation	1
2.	Problems with the traditional trend testing approach	3
2.1	Unfortunate properties	4
2.2	Unfortunate inferences	4
3	An alternative approach.....	5
3.1	A refined set of questions.....	5
3.2	How may these questions be answered?	5
3.2.1	Infer trend direction	5
3.2.2	Insufficient data?	6
3.2.3	What estimator?.....	7
3.2.4	Trend importance?	7
4	Potential issues with the proposed approach.....	8
5	Implementation.....	9
6	References.....	11

1. Summary

We propose a new trend assessment procedure, different from those commonly undertaken, yet based on peer-reviewed science. A particular benefit is that it avoids the use of ambiguous or undefined inferences often made in past trend analyses as typically seen in phrases such as: “no trend”, “not significant”, “significant trend”, or “trend detected”. The problem with these statements is that the words “significant” and “trend” are not defined: they are only applied *post hoc* to the result of a performed test and their invocation can be too-heavily influenced by the size of the sampling dataset from which inferences are being drawn. Consequently, apparent information about the trend’s magnitude may be an illusion.

Instead we provide information we consider to be more useful for water management. In the first instance this concerns inferences about the trend direction (“upward”, “downward”, or “insufficient data to infer trend direction”). Secondly, if the direction can be inferred, an assessment can be made about its importance. That can be judged by calculating where such an uninterrupted trend is heading *were it to continue*. By “heading” we mean calculating the time to cross a threshold, such as a bottom line as defined in the National Policy Statement for Fresh Water Management (NPS-FM).⁸

We will use the words “confidence” (regarding the inferred trend direction) and “best estimate” (regarding the time until a threshold will be crossed). In doing so, we avoid using the ambiguous term “significant”.

To achieve these outcomes the word “trend” is defined as any change in a key feature of a water quality variable’s statistical distribution over time—for example, a change in a distribution’s median. It is a broadly accepted reality that the biophysical world always changes over time, even if only gradually, so it follows fundamentally that we should observe the following:

There is always a trend, even though it may be judged to “gradual” or “small”.

As a consequence of this definition, one can never infer “no trend”; in its place we may infer that there are insufficient data to confidently establish the trend direction. This contrasts with some current practice in which failure to reject a null hypothesis (positing no trend whatsoever) is interpreted to mean that the hypothesis should be accepted, or that the situation is “stable”. To do so is formally illogical, as explained in the following technical sections.

In some cases these new methods could render obsolete methods that have previously been advocated to water scientists and managers (e.g., Ballantine & Booker 2011). In particular, Ballantine (2012) suggested halving detection limits, but only as an interim measure until better methods are implemented in relevant software. More dramatically, the change from testing nil hypotheses to testing directional hypotheses is wholly new and not foreshadowed⁹, simply reflecting the nature of progress in research efforts honing our tools.

1.1. Implementation

All calculations are being conducted in the R environment.

We continue with the traditional Seasonal Sen Slope estimator to make inferences about trends, accounting for seasonality (e.g., as advocated by Gilbert 1987), but with modifications to accommodate our new procedure. Seasonal Sen slopes (the median of all possible within-season

⁸ <http://www.mfe.govt.nz/publications/fresh-water/national-policy-statement-freshwater-management-2014>

⁹ A brief speculation about adopting the “three-valued logic” is given in McBride (2005, p.81).

slopes in the dataset¹⁰) are widely used in water quality, particularly because they are robust—minimising compliance with assumptions. This procedure also calls for a suitable confidence interval about the estimated trend line—provided we can infer the trend direction, as explained above.

Note that the approach to date has used *two* trend assessment procedures: the Seasonal Kendall test to make inferences about the trend magnitude and the Seasonal Sen Slope Estimator for the trend magnitude (and direction). Difficulties can arise precisely because these are independent procedures, i.e., a zero estimated slope that is statistically significant from zero.¹¹ Our new approach draws all its conclusions about *both* trend and direction using the (suitably modified) Sen Slope Estimator. This guarantees that such paradoxical conclusions cannot be reached.

Four special categories of data require attention in trend analysis studies, prior to conducting the trend assessments, all having to do with censored and other “tied” values which can be troublesome for trend assessments.

1. *Less-than data.* Trend analysis procedures for water quality data usually encounter left-censored data in the form of values that are “less than detection limit”. This occurs when laboratories lack confidence in their actual measurement. Most previous New Zealand (and international) studies have used a simple replacement rule, typically substituting one half the limit for each censored value. The above-mentioned paradoxical result can then arise because of the resulting ties in the dataset, i.e., confidence that the trend is in a particular direction yet its estimated trend slope is zero. We have addressed this issue using randomised imputation of results from the ROS (Regression on Order Statistics) procedure, so that tied “less-than” replacement values are all distinct. The method allows for detection limits that change over time (i.e. multiple detection limits).
2. *Greater-than data.* Laboratories and field teams may report “greater than” censored values, typically confined to microbiological assays and water clarity. Usually (and in all cases for microbial contaminant *E.coli*) they comprise no more than 4% of current records (i.e., no more than two “greater-thans” in a period of at least 50 observations). In that case simple substitution with tied values somewhat larger than the upper limit of detection is appropriate (e.g., 10%) as these will not affect the estimated median slope value. However, for water clarity observations “greater-than” values can exceed 5% of the observations. When this is the case we will replace the censored values based on a random sample drawn from a regression survival model (a distribution fitted to the data using maximum likelihood), which are used routinely in “survival analysis” for right-censored data.
3. *Lost flag.* In some datasets the “less-than” or “greater-than” qualifier has been detached from the appropriate detection limit. In that case expert judgement seems to be the best approach to determining if indeed the original data were censored (in which case Categories 1 or 2 would apply) or are actually striping (see Category 4 below).
4. *Striping.* Laboratory results for low levels of nutrient species (e.g., dissolved reactive phosphorus) are sometimes reported on a rather discrete scale (two significant digits), resulting in what may be called “striping”. These stripes (manifest as one or more

¹⁰ When accounting for seasonality only within-month slopes are computed. For example, all possible slopes between each January’s data are computed (and the median taken), but a January datum is not compared to a February datum, etc.

¹¹ This has been recognised for some time (McBride 2005, section 11.1.3). It has also been observed by other researchers in the field (e.g., pers. comm. Dr Mike Joy, Massey University, Palmerston North to Graham McBride, NIWA). It occurs when there are many “ties” in the dataset being analysed. Much more rarely, one can infer absolute certainty that there is no trend (McBride 2005, op. cit.).

horizontal data-lines on a graph of water quality variable versus time) can also result in numerous ties. However replacement of these values by imputation of randomised ROS values is inappropriate in this case, because the striped concentrations are not the result of censoring. Instead we have “jittered” these results (i.e., added and subtracted small random increments to their reported value) to avoid ties.

2. Problems with the traditional trend testing approach

Herein, without loss of generality, we discuss issues concerning hypothesis tests as they are applied to parametric trend assessments, e.g., using linear regressions (as elaborated in applied statistics texts, such as Zar (1984)). We do not delve strongly into particular problems with procedures for estimating the magnitude of a trend—because that topic is relatively straightforward. Issues to do with data censoring are discussed as a separate sets of issues, in section 0.

The following hypothesis test procedure is adopted in standard software.¹² The hypothesis tested is that the true trend slope is *exactly* zero.¹³ A trend is declared to have been detected when this hypothesis is rejected, a “statistically significant” result—in which case the test has returned a “*P*-value” less than a prescribed “significance level”. This level is denoted by α , usually taken as $\alpha = 0.05$.¹⁴ (In some implementations statistical significance is attained if a “test statistic” is greater than a “critical value”: this is entirely equivalent to finding that $P < \alpha$.) The *P*-value is obtained from the sample data, using the results of mathematical statistical theory concerning the distribution of the test statistic when the tested hypothesis is true.

So what are these terms: P-value and significance level?

The *P*-value is the probability of obtaining data at least as extreme as have been obtained, *assuming the tested hypothesis is true*.¹⁵ So it is retrospective. It contains three data items: (i) the assessed trend slope, (ii) the variability about that trend, and (iii) the number of data (the “sample size”).

The significance level is the probability of committing a “Type I error”—which is to reject a true hypothesis.¹⁶ It covers all future possibilities of reaching such a conclusion, so it is prospective. No information on the risk of committing a Type II error (failing to reject a false hypothesis) is required for, or supplied by, the test procedure. Nor is any information required or supplied concerning the magnitude of a trend that would be important to detect.

¹² For simplicity “significance test” and “hypothesis test” herein are treated as the same, although there are distinctions between them (to do with whether their apparatus concerns an alternative hypothesis).

¹³ So not only is it a “null” hypothesis—*sensu* “nullify” (Berkson 1942)—it is “nil” (Cohen 1994).

¹⁴ The value of α should be set before data are collected. Otherwise there is a risk of adjusting α to get a more favourable result (e.g., re-setting α from 0.05 to 0.10); a “fishing expedition”.

¹⁵ More correctly, substitute “test statistic” for “data” in this sentence.

¹⁶ For any hypothesis (nil, interval or one-sided) α is a number. For nil hypotheses it is defined at the “effect size” at which the hypothesis is true (most usually taken as zero effect size). For one-sided or interval hypotheses the Type I error risk is a function defined over the region where the hypothesis is true. In that case α is the *maximum* Type I error probability (Conover 1980), defined at the boundary of the region where the hypothesis is true. The Type II error risk (denoted by β) is always a function, defined over the region where the hypothesis is false. So it makes no sense to talk of Type II error risks in isolation from the effect magnitude.

2.1 Unfortunate properties

The P -value generally decreases with the number of available data.

Consequently, with many samples nil hypothesis trend tests will routinely return a statistically significant result when in fact the trends that are present would be considered minor by water resource managers and decision-makers—reflecting a precautionary approach to the burden-of-proof. Conversely, with only a few data, environmentally important trends may often be declared to be ‘not significant’—reflecting a permissive approach.

If the trend is environmentally important, this approach could be considered appropriate—a large sampling effort from a given population over a given trend assessment period can lead to rejection of a false hypothesis that with a lesser effort would not have been rejected. But this approach is not appropriate if the trend is considered to be trivial.

2.2 Unfortunate inferences

Two common inferences arise:

- Declaring a trend to be *statistically* significant, often taken to mean that something important has been found, does not necessarily imply that the trend is *environmentally* significant. It may merely reflect a large number of samples or small variability, leading to a small P -value.
- Declaring a trend to be not significant may merely reflect an inadequate sampling effort.

Note that in some cases that the effects of these two items can (roughly) cancel each other out, such that statistical significance may correctly imply environmental significance. Indeed, one writer has noted that this should be a goal: “Assuming an investigator desires to detect only differences that are of practical importance, and not merely differences of any magnitude, he should impose the added safeguard of not employing sample sizes that are larger than he needs to guard against the second type of error.” (Fleiss 1981, p. 33). This is a logically correct inference for hypothesis tests, and may be appropriate for the design of experiments, with a defined start and finish. But it is in complete disharmony with the needs of an accumulating data collection programme, as is common in monitoring systems—any coherence between statistical significance and environmental significance would be fleeting.

A third common inference, related to the above, is also of concern (as has been noted by many authors¹⁷):

- Declaring non-significant results as “no trend” or “stable” is formally illogical.

The detailed explanation of this logic (and lack thereof) appears in a footnote.¹⁸

¹⁷ Misinterpretation of P -values has been discussed in various literatures for decades (e.g., Berkson 1942, Carver 1978, McBride *et al.* 1993, Schervish 1996, Johnson 1999; Greenland & Poole 2013, Burnham & Anderson (2014). See also <http://www.indiana.edu/~stigtsts/>. It lead one prominent statistician, toward the end of a distinguished career (Nelder 1999), to opine that: “The most important task before us in developing statistical science is to demolish the P -value culture, which has taken root to a frightening extent in many areas of both pure and applied science, and technology....”.

¹⁸ The P -value can be characterised as: $P = \text{Prob}(D \mid H \text{ is true})$, where “ D ” denotes all data at least as extreme as was obtained, “ \mid ” means “given that” and “ H ” denotes the tested hypothesis. If this P -value is large, i.e., there is a good chance of obtaining the data if the hypothesis is true, it does not follow that the hypothesis should be accepted. That’s because the probability $\text{Prob}(D \mid H \text{ is false})$ may also be large. And to claim “no trend” on the basis of a large P -value is to ‘invert the conditionals’ in the probability statement, by claiming that $\text{Prob}(D \mid H \text{ is true}) = \text{Prob}(H \text{ is true} \mid D)$. This inverted probability [$\text{Prob}(D \mid H \text{ is true})$] is actually a Bayesian probability (Oakes 1986, p. 16) and to get it we must use Bayes’ rule. Details of that calculation need not concern us here, except to note that the Bayesian probability

3 A alternative approach

In considering the unfortunate inferences above, it is more than tempting to conclude that the standard trend assessment apparatus is not asking the right question: it asks “Is the trend slope zero?”

3.1 A refined set of questions

We suggest that a more appropriate set of questions is four-fold:

1. Can we be confident about the *direction* of the trend?
2. Is our sampling effort sufficient to reach such a conclusion?
3. If so, what can we infer about the likely *magnitude* of the trend?
4. Does the trend appear to be important?

3.2 How may these questions be answered?

Based on a new published procedure (McBride *et al.* 2014) we propose the following responses to items 1–4:

1. Use a re-interpretation of the test of a nil hypothesis so that the inference made concerns trend *direction*, rather than its *magnitude* (positing no trend).
2. Sampling effort efficacy comes automatically from the outcome of step 1.
3. Use standard estimators.
4. State *a priori* what an environmentally important rates of change would be, as judged by time-until-threshold-crossing, such as those provided by the NPS-FM.

We now elaborate on these responses.

3.2.1 Infer trend direction

This section draws heavily on material in McBride *et al.* (2014).

A number of authors have proposed recasting of nil hypothesis testing procedures, in what may be called a “three-valued logic” (Bohrer 1979; Harris 1997&2001; Bansal *et al.* 2012). Many years ago Kaiser (1960) pointed out that accepting the alternative hypothesis (that a trend is not zero) doesn't of itself enable us *logically* to conclude anything about the direction-of-change. Researchers employing nil hypothesis testing *alone* cannot therefore come to any logical conclusion about the sign of an effect. Kaiser (1960), following a lead by Hodges & Lehmann (1954), pointed out that the only way to use hypothesis tests to come to a conclusion about the direction-of-change is to employ “two one-sided” (TOST) procedures, separately testing the twin hypotheses: trend > 0; trend < 0.¹⁹ Three outcomes are possible): (i) Confidence that the change is in the positive direction; (ii) Confidence that the change is in the negative direction; (iii) There are insufficient data to be

would define “*D*” as only the data obtained, not all values at least as extreme. (Bayesian probabilities conform to the “likelihood principle”: only the actual observations should enter the probability calculations, not other possible outcomes that were not observed—Lee 1997, p. 193.)

¹⁹ Kaiser posed his discussion in terms of differences in the means (denoted by μ) between two populations, i.e., $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$. We can simply extend this to contemplate trends within a single population, as above (e.g., Dixon & Pechman 2005).

confident about the direction-of-change.²⁰ Conclusion (iii) is valid in that it is a statement that can be accepted, even though it flows from the apparatus for testing a nil hypothesis. More recently, this approach has been endorsed by others (Jones & Tukey 2000; Goudey 2007), though there is inertia resisting its adoption (judging by the lack of its availability in statistical software).

Note that Jones & Tukey (2000) cast these outcomes (i)–(iii) in terms of a decision theoretic approach. For example, (i) becomes *act as if* trend > 0. We assume herein that we can instead interpret the results as *confidence* that trend > 0.

Hypothesis splitting?

The validity of splitting interval hypotheses into two separate directional hypotheses, each of size α , has been demonstrated by Berger (1982, see also Berger & Hsu 1996, p. 288). This result relies on the properties of the IUT (Intersection-Union Tests). Berger & Hsu emphasised that “An important feature...is that each of the individual test is performed at level- α , but the overall test also has the same level α ”. Jones & Tukey (2000) showed that this also applies to nil hypothesis tests when assessing the direction of change. There is great benefit to be had in this hypothesis splitting.²¹

So what is α in this directional approach?

Under the directional approach advocated here the word “significance” need not be associated with α . Instead, α becomes the maximum risk of falsely inferring a positive trend when in fact the trend is negative (and vice versa). In other words, the “maximum probability of erroneous confidence” (Williams *et al.* 1999). Given the historical baggage associated with the word “significant” it seems wise to banish it from this proposal.

An alternative method

This procedure can be interpreted using a symmetric $100(1-2\alpha)\%$ confidence interval on the Sen slope (see <http://www.webapps.cce.vt.edu/ewr/environmental/teach/smprimer/sen/sen.html>).²² If that interval does not contain the nil value (zero) then the trend direction is known with confidence.

Note that “... the proposed procedure is uniformly more powerful than the conventional procedure” (Jones & Tukey 2000). That is because the conventional procedure’s nil hypothesis is rejected when $P \geq \alpha$, whereas the threshold for the new procedure to detect direction is equivalent to $P \geq 2\alpha$.

3.2.2 Insufficient data?

If the symmetric $100(1-2\alpha)\%$ confidence interval about the trend slope contains the nil value zero (i.e., $P_{(1)} \geq \alpha$ or $P_{(2)} \geq 2\alpha$), then we conclude that there are insufficient data to discern the trend direction.

²⁰ Kaiser’s approach—adopted by Bohrer (1979), Harris (1997&2001) and Bansal *et al.* (2012)—includes the nil hypothesis as well as two directional alternatives (which is why it has been called a “three-valued logic”—Harris 1997). However, that inclusion of this unsatisfactory hypothesis isn’t necessary has been demonstrated by Jones & Tukey (2000), following proposals by Tukey (1991) and Williams *et al.* (1999). One needs to consider only two hypotheses (e.g., trend > 0; trend < 0) in order to reach one of the *three* conclusions (i)–(iii).

²¹ Consider an ongoing trend of constant slope and variance. Tests of one-sided hypotheses, for example positing that the trend slopes are greater than zero, can be shown to cause P -values to fall or rise as the sample size is increased (McBride *et al.* 2014), depending on the truth of the hypothesis tested. In contrast P -values from two-sided tests of nil hypotheses will always fall with increasing sample size—leading to the noted unfortunate properties.

²² These are $100(1-2\alpha)\%$ confidence intervals, not $100(1-\alpha)\%$ intervals, because each of the two one-sided tests is performed at level α , as shown in Jones & Tukey (2000). Berger & Hsu (1996), discussing interval tests, noted “The fact that the TOST seemingly corresponds to a $100(1-2\alpha)\%$, not $100(1-\alpha)\%$, confidence interval initially caused some concern”, and go on to note that many authors now accept this feature.

3.2.3 What estimator?

Standard Sen slope estimators are ideal for this task. They effectively mimic linear trends and are our “best guesses”. A Sen Slope is the median of all slopes between separate data in a record. Note that even though our data can contain many censored values (see section 0) we deliberately do not propose using the trend estimate procedure of Akritas (1995), advocated for censored data by an influential writer (Helsel 2012). That is because: (a) application to seasonal models is not included in the software we will use (the R freeware package), and (b) its results do not appear to be satisfactory (as judged by inspection of Figure 12.7 in Helsel 2012).

3.2.4 Trend importance?

In the procedure advanced by McBride *et al.* (2014), a prior specification of the change that would be important to detect is advocated (when testing for differences between populations). This can be extended to trends within a single population. For example, Ballantine *et al.* (2001) defined a “meaningful” change as a trend greater than 1% per annum of the trend assessment period median value. However obtaining agreement on such a figure among environmental scientists and managers for a large nation-wide study is impractical.²³ Instead we appeal to thresholds provided by guidelines or other types of water quality targets, such as the “bands” defined by the National Objectives Framework, and a “critical time period”. There are two ways that this could be implemented, as follows:

1. A critical time period is specified for which the crossing of the threshold is *a priori* judged to be sufficiently imminent to be either of concern (in the case of a degrading trend) or an indication of effective management.
2. Where the threshold is specified in terms of the sample median we can use the Sen slope to assess whether the threshold will be crossed within the critical time period *were the trend to continue uninterrupted*.
3. The trend is categorised as important (plus either improving or degrading) if the threshold is projected to be crossed *within* the critical period, or unimportant if not. Any degrading trend for a site-variable that is already below the National Bottom Line could be flagged as an important trend.
4. An alternative to trend-importance-assessment would define a trend that, were it to continue uninterrupted, would produce an appreciable change within a generation. We define this rate of change as the rate at which water quality that was initially in a reference (i.e. natural un-impacted) state would need to change to reach a broadly unacceptable state in a nominal period of 20 years. Expected reference state median concentrations, or distance in the case of water clarity, for rivers have been characterised for classes at the Source-of-Flow level of the River Environment Classification (REC) (Snelder and Biggs 2002) by McDowell *et al.* (2013). Unacceptable states could be based broadly on the NPS-FM bottom lines for both lakes and rivers. Ecological nutrient criteria for rivers, which are not provided by the NPS-FM (although toxicity are), could be based on concentration criteria derived for meeting periphyton bottom lines (S. Elliot personal communication). The NPS-FM does not have bottom line criteria for water clarity but an alternative criteria such as the MfE (1994)

²³ The fact that such specification (for a “power analysis”) has to be reported when seeking funding for medical studies is of little help, since such a study is generally much more specific in scope and focus.

guideline could be used. The rates would be derived in terms of rates of degradation but when the sign is reversed, the rate would define an important improving trend.

Both of these methods (Items 3 and 4) will involve making some pragmatic choices such as selecting the relevant criteria and choosing the length of time periods. In addition, assessment of trend importance at some sites may not be possible because the appropriate criteria may not have been established. For example, for a small number of REC classes, the reference water clarity has been assessed by McDowell *et al.* (2013) to be less than the MfE (1994) clarity guideline of 1.6m.

4 Potential issues with the proposed approach

Four issues have been identified with the proposed trend assessment approach.

- 1 “Trends always exist” may conflict with common usage—causing confusion, or at least, discomfort. However we regard that as a minor issue compared with the unfortunate inferences often drawn from nil hypothesis tests when the word “trend” is not defined. This is especially the case when the failure to reject a nil hypothesis (because $P > \alpha$) is interpreted to mean “no trend”. In such a case our proposed procedure make the logical and much-more-helpful statement: “Insufficient data to discern the trend direction”.
- 2 Bands will have to be defined for water quality variables not included in the NPS-FM.
- 3 The performance of a large number of test procedures raises the prospect of whether there should be an adjustment to cater for the problem of “overall error rate”. That is, some of the individual inferences of finding confidence for particular sites and variables in the discernment of trend direction may have been in error, and so the effective value of α should be reduced in some way to guard against finding too many “false positives” in the overall study. We park that issue for the time-being. However, we note that often-used adjustments that guard against the possibility of making *at least one error* (e.g., using Bonferroni adjustments) will be far too restrictive. Instead, if α -adjustments are to be done, the FDR (False Discovery Rate) approach should be used (McBride 2005, sec. 4.6) as it controls the *proportion* of such errors.
- 4 A related issue concerns the interpretation of the $100(1-2\alpha)\%$ confidence interval for trend direction-testing. That is, the common interpretation of a confidence interval has it that one can say with confidence that the tested value lies between the interval’s numeric limits. Actually that is a Bayesian statement (Reckhow and Chapra 1983, McBride 2005, p. 59)—because it concerns the probability of a hypothesis given the data at hand. That being so, the statement’s calculation has invoked a flat (“non-informative”) prior distribution, which may well be appropriate. In that case, issues identified in point 3 above would seem to evaporate.

5 Implementation

We propose to focus on monotonic trend analysis, using non-parametric methods. This is the international norm for water quality variables, because assumptions required by parametric methods can be difficult to meet, and the sheer effort required to examine assumption-compliance for many variables at hundreds of sites is not practicable. Also, a long-term record showing minimal monotonic trend may be comprised of a larger upward trend in the first part of the record followed by a similarly sized trend in the opposite direction for the second part of the record. In our reporting we will make efforts to alert readers to such patterns for particular site-variable combinations.

All trend calculations are being conducted in the R statistical computing environment.

We continue with the traditional Seasonal Sen Slope estimator to make inferences about trends, accounting for seasonality²⁴, but with estimation of the confidence interval about the estimated trend line, as explained above.

However, before trend assessments are conducted some we note four special categories of data treatment that must first be addressed. All of these data treatments are associated with dealing with censored and other “tied” values which can be troublesome for trend assessments.

1. *Less-than data.* Trend analysis procedures for water quality data usually encounter left-censored data in the form of values that are “less than detection limit”. This occurs when laboratories lack confidence in their actual measurement. Most previous New Zealand (and international) studies have used a simple replacement rule, typically substituting one half the limit for each censored value. The above-mentioned paradoxical result can then arise because of the resulting ties in the dataset, i.e., confidence that the trend is in a particular direction yet its estimated trend slope is zero. We have addressed this issue using randomised imputation of results from the ROS (Regression on Order Statistics) procedure (Helsel 2012), so that tied “less-than” replacement values are all distinct.²⁵ It includes consideration of datasets in which detection limits change over time.
2. *Greater-than data.* Laboratories and field teams may report “greater than” censored values, typically confined to microbiological assays and water clarity. Usually (and in all cases for microbial contaminant *E.coli*) they comprise no more than 4% of current records (i.e., no more than 2 “greater-thans” in a period of at least 50 observations). In that case simple substitution with tied values somewhat larger than the upper limit of detection is appropriate (e.g., 10%) as these will not affect the estimated median slope value. However, for water clarity observations “greater-than” values can exceed 5% of the observations. When this is the case we will replace the censored values based on a random sample drawn from a regression survival model (a distribution fitted to the data using maximum likelihood), which are used routinely in “survival analysis” for right-censored data.
3. *Lost flag.* In some datasets the “less-than” or “greater-than” qualifier has been detached from the appropriate detection limit. In that case expert judgement seems to

²⁴ When accounting for seasonality only within-month slopes are computed. For example, all possible slopes between each January’s data are computed (and the median taken), but a January datum is not compared to a February datum, etc.

²⁵ The ROS method develops probability plotting positions for each data point (censored and uncensored) based on the ordering of the data. A least squares line is then fit by regressing the concentrations to the uncensored probability plotting positions. ROS can accommodate multiple censoring limits.

be the best approach to determining if indeed the original data were censored (in which case Categories 1 or 2 would apply) or are actually striping (in which case Category 4 would apply).

4. *Striping*. Laboratory results for low levels of nutrient species (e.g., dissolved reactive phosphorus) are sometimes reported on a rather discrete scale (two significant digits), resulting in what may be called “striping”. These stripes (manifest as one or more horizontal data-lines on a graph of water quality variable versus time) can also result in numerous ties. However replacement of these values by imputation of randomised ROS values is inappropriate in this case, because the striped concentrations are not the result of censoring. Instead we have simply jittered these results about their reported value to avoid ties.

Finally, as in previous assessments, we deliberately avoid accounting for autocorrelation issues, for reasons given in McBride (2005, section 2.11)—see also Ellis (1989, Appendix 2C), Loftis *et al.* (1999) and de Gruijter *et al.* (2006, page 42). In essence, autocorrelation (of residuals, once trend and seasonality have been accounted for) only inflates the residual’s standard error if the trend assessment period is no longer than the data record.

6 References

- Akritas, M.G.; Murphy, S.A.; LaValley, M.P. (1995). The Thiel-Sen estimator with doubly censored data and applications to astronomy. *Journal of the American Statistical Association* 90(429): 170–177.
- Ballantine, D.; Booker, D.; Unwin, M.; Snelder, T. (2010). Analysis of national river water quality data for the period 1998–2007. NIWA Client Report: CHC2010-038 for Ministry for the Environment (Project MFE10502): sec. 2.4.3.
- Ballantine, D.; Booker, D. (2011). Greater Wellington state of environment river water quality data assessment and trend analysis (2003–2010). NIWA Client Report: HAM2011-126 (Project WRC11203).
- Ballantine, D. (2012). Water quality trend analysis for the Land and Water New Zealand website (LAWNZ). NIWA Client Report 2012-080 (Project ELF12218), prepared for Horizons Regional Council, Palmerston North.
- Bansal, N.K.; Hamedani, G.G.; Sheng, R. (2012). Bayesian analysis of hypothesis testing problems for general population: a Kullback-Leibler alternative. *Journal of Statistical Planning and Inference* 142(7): 1991–1998.
- Berger, R.L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics* 24: 295–300.
- Berger, R.L.; Hsu, J.C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* 11(4): 283–319 (with discussion).
- Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association* 37: 325–335.
- Bohrer, R. (1979). Multiple three-decision rules for parametric signs. *Journal of the American Statistical Association* 74: 432–437.
- Burnham, K.P.; Anderson, D.R. (2014). *P* values are only an index to evidence: 20th- vs. 21st-century statistical science. *Ecology* 95(3): 627–630.
- Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review* 48, 378–399.
- Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist* 49(12): 997–1003.
- Conover, W.J. (1980). *Practical Nonparametric Statistics*. 2nd ed. Wiley New York.
- de Grijter, J.; Brus, D.; Bierkens, M.; Kotters, M. (2006). *Sampling for Natural Resource Monitoring*. Springer, Berlin.
- Dixon, P.M.; Pechman, J.K. (2005). A statistical test to show negligible trend. *Ecology* 86(7): 1751–1756. Comments by R.J. Camp *et al.* (2008) 89(5): 1469–1472. Authors' response 89(5): 1473.
- Ellis, J.C. (1989). Handbook on the design and interpretation of monitoring programmes. Report NS 29, Water Research Centre, Medmenham, England (first published April 1990).
- Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd ed. Wiley, New York.
- Gilbert, R.O. (1987). *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold, New York, New York.
- Goudey, R. (2007). Do statistical inferences allowing three alternative decisions give better feedback for environmentally precautionary decision-making? *Journal of Environmental Management* 85: 338–344.

- Greenland, S.; Poole, C. (2013). Living with P values: Resurrecting a Bayesian perspective on frequentist statistics." *Epidemiology* 24(1): 62–78 [includes discussion by A. Gelman (pp. 69–72) and the authors' rejoinder (pp. 73–78)].
- Harris, R.J. (1997). Reforming significance testing via three-valued logic. In L.L. Harlow, S.A. Mulaik, J. H. Steiger (eds.), *What if There Were No Significance Tests?* (pp. 145–174). Mahwah, NJ: Lawrence Erlbaum.
- Harris, R.J. (2001). *A Primer of Multivariate Statistics*. 3rd ed. Lawrence Erlbaum, Mahwah, NJ.
- Helsel, D.R. (2012). *Statistics for Censored Environmental Data Using Minitab® and R*. Wiley, New York.
- Helsel, D.R.; Hirsch, R.M. (1992). *Statistical Methods in Water Resources*. USGS Techniques of Water Investigations Book 4, Chapter A3, 510 p.²⁶
- Hodges, J.L.; Lehmann, E.L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society, Series B* 16: 261–268.
- Johnson, D.H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management* 63(3): 763–772.
- Jones, L.V.; Tukey, J.W. (2000). A sensible formulation of the significance test. *Psychological Methods* 5(4): 411–414.
- Kaiser, H.F. (1960). Directional statistical decisions. *Psychological Review* 67(3): 160–167.
- Lee, P.M. (1989). *Bayesian Statistics: An Introduction*, 2nd ed. Arnold, London.
- Loftis, J.C.; McBride, G.B.; Ellis, J.C. (1991). Considerations of scale in water quality monitoring and data analysis. *Water Resources Bulletin* 27(2): 255–264.
- McBride, G.B., Loftis, J.C.; Adkins, N.C. (1993). What do significance tests really tell us about the environment? *Environmental Management* 17(4): 423–432 (errata: 18: 317).
- McBride, G.B. (2005). *Using Statistical Methods for Water Quality Management: Issues, Problems and Solutions*. John Wiley & Sons, New York.
- McBride, G.B.; Cole, R.; Westbrooke, I.; Jowett, I.G. (2014). Assessing environmentally significant effects: A better strength-of-evidence than a single P value? *Environmental Monitoring and Assessment* 186(5): 2729–2740. doi: 10.1007/s10661-013-3574-8.
- McDowell, R.W.; Snelder, T.H.; Cox, N.; Booker, D.J.; Wilcock, R.J. (2013). Establishment of reference or baseline conditions of chemical indicators in New Zealand streams and rivers relative to present conditions. *Marine and Freshwater Research* 64(5): 387–400.
- MfE (1994). Water quality guidelines No. 2: Guidelines for the management of water colour and clarity. Ministry for the Environment, Wellington. 77 p.
- Nelder, J.A. (1999). Statistics for the millennium: from statistics to statistical science. *The Statistician* 48(2): 257–269.
- Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. Wiley, New York.
- Reckhow, K.H.; Chapra, S.C. (1983). *Engineering Approaches for Lake Management. Volume 1: Data Analysis and Empirical Modelling*. Butterworth, Boston.
- Schervish, M.J. (1996). P values: What they are and what they are not. *American Statistician* 50(3): 203–206.

²⁶ Free copy at <http://water.usgs.gov/pubs/twri/twri4a3/>

- Sen, P.K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association* 63: 1379–1389.
- Singh, A.; Singh, K. (2013). ProUCL Version 5.0.11 Technical Guide: Statistical software for environmental applications for data sets with and without nondetect observations. U.S. Environmental Protection Agency Report EPA/600/R-07/041.
- Snelder, T.H.; Biggs, B.J.F. (2002). Multi-scale river environment classification for water resources management. *Journal of the American Water Resources Association* 38(5): 1225–1240.
- Tukey, J.W. (1991). The philosophy of multiple comparisons. *Statistical Science* 6(1): 100–116.
- Williams, V.S.L.; Jones, L.V.; Tukey, J.W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioural Statistics* 24: 42–69.
- Zar, J.H. (1984). *Biostatistical analysis* 2nd ed. Englewood Cliffs, NJ: Prentice-Hall.

Appendix B Nutrient concentration thresholds to achieve periphyton objectives across climate and source of flow REC classes

Authors:

Ton Snelder (LWP, Christchurch)
Michelle Greenwood (NIWA, Christchurch)
John Quinn (NIWA, Hamilton)
Sandy Elliott (NIWA, Hamilton)
May 2015

1 Introduction

The periphyton attribute of the National Objectives Framework (NOF) is based on restricting the biomass of periphyton (measured as chlorophyll *a* on the beds of streams and rivers). Nutrient concentrations typically increase the growth rate of periphyton, which increases the likelihood that periphyton biomass thresholds will be exceeded. Achieving the NOF periphyton objectives therefore requires defining nutrient concentration criteria that ensure that periphyton biomass is restricted to the required levels.

The MfE periphyton guidelines (MFE, 2000) defines nutrient concentration thresholds for the management of periphyton biomass. However, nitrogen and phosphorus thresholds derived using these guidelines appear to be overly restrictive. For example, some nitrogen and phosphorus thresholds derived using the MfE periphyton guidelines for some streams are less than the reference state water quality predicted using McDowell *et al.* (2013). The reasons for this appear to be that the original research that developed these criteria (Biggs, 2000) used data from sites at which only nutrients were limited the growth of periphyton biomass. In reality sites are likely to be limited by other factors such as adequate substrate and light. This study attempted to define nutrient concentration criteria for management of periphyton biomass using periphyton and nutrient data associated with the National River Water Quality Network (NRWQN). These sites are probably more heterogeneous than the sites that (Biggs, 2000) used and, therefore, we expect the criteria we derive to be more lenient.

2 Methods

2.1 Periphyton data

The NRWQN comprises 77 sites located on 48 of New Zealand's rivers and was specifically designed to broadly represent variation in main-stem rivers across New Zealand (Davies-Colley *et al.*, 2011). Since 1989 a range of water quality variables and a visual assessment of the cover of filamentous and mat forming algae has been carried out monthly at NRWQN sites and flows are monitored continuously (Davies-Colley *et al.*, 2011; Smith and McBride, 1990).

In this study we analysed data for the time period 1989 to 2010 (22 years) but excluded several NRWQN sites for various reasons. The sites AK1, AK2 and GS1 (see (Snelder *et al.*, 2014) for site codes) were excluded because they are located on deep rivers with silty beds that lack periphyton. RO2 and RO6 were excluded due to a large number of missing periphyton observations. Sites WH3 and WH4 were excluded because they are dominated by macrophytes. Three sites on large rivers including the Waikato (HM2, HM4 and HM5) and the Clutha (DN4) rivers were excluded due to logistical difficulties in sampling periphyton and artificially fluctuating water levels.

We split the records for some sites into two portions to account for significant changes that had occurred at the site through the period of operation of the NRWQN. The two portions of the record were treated as separate sites in the analyses that follow. Four sites (HM1, RO4, WA6, WN3, and DN2) were split due to changes in site locations that were required for operational reasons. Five sites (including RO3, HV5, WN2, TK2 and DN1) were split due to very obvious changes in mean water quality. Thirteen sites in the South Island (NN3, NN5, GY1, CH1, TK3, TK4, TK6, AX1, AX2, AX3, AX4, DN4 and DN9) were split because they were colonized by the invasive alga *Didymosphenia geminata* (Kilroy *et al.*, 2009). The abundance of *D. geminata* responds to very different factors to the other bloom forming taxa in New Zealand rivers (Kilroy *et al.*, 2009). For this reason, we retained the

portion of the record prior to the establishment of *D. geminata* but did not include the second portion. Site TK1 was split due to the pre-commissioning failure of the Opuha dam and the subsequent establishment of *D. geminata*. After excluding some sites and splitting others we had a total of 78 sites comprising either the entire record of the NRWQN site or part thereof.

Monthly observations of periphyton cover in two categories (mats and filaments) have been made by visual assessment at most of the NRWQN's 77 sites since 1989. The cover of two categories; filaments (> 2cm long) and mats (> 2mm thick) were measured as continuous variables. In the field, mats were distinguished from thin films when the texture of the underlying substrate could not be seen and the layer could be scraped or peeled. Where possible, ten replicate observations of 0.5 m radius patches of riverbed were made at equally spaced points across a wadeable cross-section of the river using an underwater viewer. However, because the NRWQN rivers are medium to large some observations were confined to the wadeable margin from one bank or, at 10% of sites, to observations from bridges or cableways (see (Quinn *et al.*, 2009) for details).

For each site and sample date we calculated the mean of the ten replicate observations to produce a time-series of monthly observations of the proportion of the bed covered by filaments and mats.

2.2 Approximation of periphyton biomass as Chlorophyll *a*

The NOF periphyton attribute state bands (A-D) are defined in terms of biomass thresholds and an exceedance frequency. Biomass is defined by Chlorophyll *a* per unit area of the stream bed (mg Chl *a* m⁻²). The periphyton threshold between the C and D band is a maximum chlorophyll *a* of 200 mg m⁻². The value of 200 mg Chl *a* m⁻² is considered to be a point at which the health of streams is significantly compromised. The thresholds for the A/B and B/C bands are 50 and 120 mg Chl *a* m⁻² respectively. Given that exceedances do occasionally occur naturally (i.e. even without human influence), the NOF periphyton attribute specifies an exceedance frequency of once a year on average, based on monthly measurements of periphyton, or approximately 8% of the time. The NOF makes an exception to this exceedance frequency (of twice per year on average, or approximately 17% of the time) for stream types that are naturally productive due to geological enrichment and particularly long accrual periods. For the 78 NRWQN site by date combinations we computed the 92nd percentile of WCC (i.e. the value of WCC that was exceeded 8% of the time) from the full 22 year dataset.

Chlorophyll *a* is not routinely monitored at NRWQN sites. However, the cover observations can be aggregated into a single metric called Weighted Composite Cover (WCC), which is defined as 0.5 x average cover by filaments + average cover by mats (Matheson *et al.*, 2012). The weighting of the two periphyton categories is based on the guideline that they are problematic if they exceed 30% and 60% of the visible stream bed (generally < 0.75 m deep) respectively (MFE, 2000). This WCC was used as a proxy measure in the analysis.

The relationship between WCC and chlorophyll *a* was investigated as part of the development of the NOF periphyton attribute using a combined Horizons Regional Council and Environment Canterbury dataset (n = 1084) (Cathy Kilroy, NIWA, *pers comm*). A linear relationship was fitted to square-root transformed mean %WCC and log₁₀ transformed mean chlorophyll *a* from each site (n = 66, r² = 0.59):

$$\log_{10}(\text{chlorophyll } a) = 0.291 + 0.307(\sqrt{\text{WCC}})$$

We used this regression equation to back calculate the WCC equivalents of chlorophyll *a* for the NOF periphyton attribute band thresholds. The NOF chlorophyll *a* thresholds of 50, 120 and 200 mg/m²

equate to 21, 34 and 43% WCC based on this equation. These thresholds are similar to the levels of 20, 45 and 55% WCC that (Matheson *et al.*, 2012) proposed correspond to excellent, good and poor ecological condition as indicated by the MCI.

Explanatory variables

We derived a number of explanatory variables for each site from the observed water quality and hydrological data. The relevant set of explanatory variables for the regression models was selected based on a conceptual model that represented periphyton abundance as a consequence of counteracting processes of biomass accrual and loss (Biggs, 1996). The details of these variables are provided by (Snelder *et al.*, 2014). For the analyses that follow, the explanatory variables needed to be available for every segment of the national digital river network represented by the River Environment Classification (REC; (Snelder and Biggs, 2002). We used a combination of available digital river network variables and new variables that were based on modelled values to provide values of the explanatory variables for each segment of the river network.

We assumed that differences in accrual rates between our sites were determined by the rate of growth and that this is controlled primarily by nutrient supply, light and temperature. We assumed that biomass loss is determined primarily by hydrological disturbance (i.e., high flows, and changes of flows; (Biggs, 1996). We expected biomass loss processes due to hydrological disturbance to be mediated by channel morphology with loss rates being lower where stream substrates are stable and consist of larger substrata (i.e., gravels, cobbles and boulders (Doyle and Stanley, 2006; Uehlinger, 1991). We also anticipated that biomass accrual would be related to hydrological conditions, particularly the magnitude of base or low flows. Sites with small low flows are likely to have periods of low water velocity that favours the development of some filamentous taxa (Flinders and Hart, 2009; Suren *et al.*, 2003).

Differences in loss rates between sites may also arise due to differences in invertebrate grazer abundance and grazing rates (Dodds and Welch, 2000; Rutherford *et al.*, 2000). Our environmental data did not include invertebrate grazers explicitly. However, we assumed that some differences in grazer density between sites may be accounted for by the combination of substrate size and hydrological indices (Quinn and Hickey, 1990).

We represented nutrient concentration using the median of the observed values. We evaluated the median concentration of Dissolved Reactive Phosphorus (DRP, mg m^{-3}), Total Nitrogen (TN, mg m^{-3}) and Dissolved Inorganic Nitrogen (DIN, mg m^{-3}). We calculated the ratio of DIN to DRP to define NPratio. The nutrient variables were log (base 10) transformed to make their distributions approximately normal and their relationship with site mean periphyton cover more linear. The values of these variables for every segment of the digital river network were represented by modelled values produced by (Unwin *et al.*, 2010).

We used several hydrological indices derived from the mean daily flow time-series for each site to characterize the flow regime components (see (Snelder *et al.*, 2014). The frequency of large floods was represented by the number of events per year that exceeded three times the long-term median flow (FRE3). If the time interval between an event dropping below the threshold and the next event rising above the threshold was less than 5 days, only a single event was counted. The frequency of changes of flows was represented by hydrological reversals (Reversals). Reversals are occasions on which the direction of daily change in flows reverses (i.e., the number of occasions on which increasing flows (rising hydrograph limbs) changed to falling limbs and vice versa (Olden and Poff,

2003). Sites with frequent reversals have many hydrograph peaks. Rates of increase of flow were represented by the number of days on which flow was less than that of the previous day (nNeg). We estimated nNeg for each site by first counting the number of days in each year for which the flow reduced on the subsequent day. nNeg for each site was the mean of these values over years. Sites with steep rising limbs have large values of nNeg. We used the mean annual 7-day low flow divided by the mean flow to represent the low flow magnitude (7DayFlowMins). We derived this explanatory variable for each site by first estimating the minimum of a 7-day moving average flow in each year of record. 7DayFlowMins was the mean of these annual values divided by the mean daily flow for the entire record. Predicted Values of these four hydrological indices were available for the REC network based on (Snelder and Booker, 2013).

The observed proportion of bed covered by the substrate size classes in each year were transformed into a single substrate index, as described by (Jowett and Richardson, 1990). We used the mean of the annual values as our explanatory variable (Substrate). Predicted values of Substrate for the REC network were available from the Freshwater Environments of New Zealand (FENZ) database based on modelled values produced by (Leathwick *et al.*, 2011).

We calculated the 95th percentile of the monthly spot measurements of water temperature at each site and used this value to represent summer maximum water temperature (T95). We used a method similar to that of (Julian *et al.*, 2008) to estimate Photosynthetically Active Radiation reaching the riverbed (PAR, $\mu\text{mol m}^{-2} \text{s}^{-1}$). The details of the method are provided by (Snelder *et al.*, 2014). We fitted regression models to the NRWQN values of T95 and PAR using explanatory variables that were available from the FWENZ database and made predictions of these variables for all segments in the REC network.

Regression relationships

Data from the 78 NRWQN sites/date combinations identified by (Snelder *et al.*, 2014) were used to develop regression models that explained between-site variation in the 92nd percentile of WCC as a function of explanatory variables. The 92nd percentile values of WCC were square root transformed prior to analyses to make its distribution approximately normal.

Two separate regression models of the 92nd percentile values of WCC as a function of the explanatory variables were fitted using stepwise linear regression (see (Snelder *et al.*, 2014) for details). Model 1 was intended for evaluating TN concentration thresholds and included TN and the ratio of DIN to DRP as predictors but not DRP. Model 2 was intended for evaluating DRP concentration thresholds and included only DRP as the nutrient predictor.

A summary of the fitted models is provided in Table 1 and Figure 1 below. Both models were unbiased but had large uncertainties, which was consistent with the relatively low proportion of variation explained (Table 1).

Table 10 Summary of the regression models fitted to the 92nd percentile values of WCC at the NRWQN sites.

	Model 1	Model 2
r^2	0.40	0.34
Adjusted r^2	0.38	0.30
Significant variables retained after stepwise eliminations	FRE3, 7DayFlowMins, nNeg, T95, PAR, log10TN, log10NPratio	FRE3, 7DayFlowMins, nNeg, T95, PAR, log10DRP

Nash-Sutcliffe Efficiency ²⁷ (NSE)	0.4	0.36
Model uncertainty (RMSD*)	1.96	2.03
Model bias	0	0

*Uncertainty is expressed in terms of the square root transformed response.

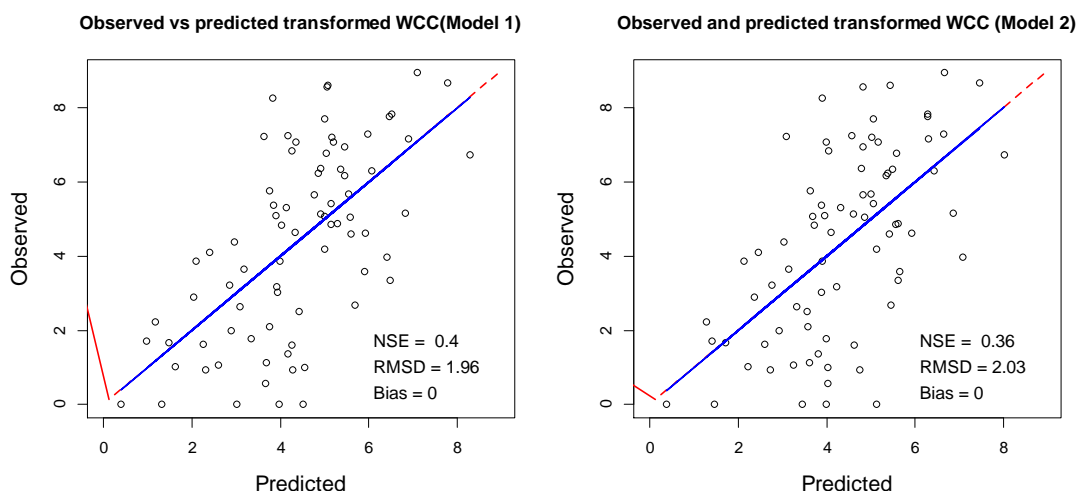


Figure 1: Comparison of observed 92nd percentile values of WCC (square root transformed) with that predicted by the regression models 1 and 2. Model 1 includes TN and DIN:DRP and Model 2 includes DRP. The red line is 1:1 and the blue is a line of best fit. NSE is the Nash-Sutcliffe Efficiency.

We tested these regression models on an independent dataset comprising sites in Canterbury and the Whanganui-Manawatu regions. The NRWQN models were not able to predict the pattern of the observed 92nd percentile values of WCC at the Canterbury and the Whanganui-Manawatu sites (Nash Sutcliffe Efficiencies < 0) (Figure 2). The model predictions were also low relative to the observations (under prediction). For example the highest observed value was 131% whereas the largest prediction was 62%. This may indicate that the NRWQN data is systematically low, and/or that models generated for the NRWQN are not good at predicting WCC at new sites. However, without a more complete dataset we were unable to investigate this further.

²⁷ The Nash–Sutcliffe model efficiency coefficient is used to assess the predictive power of models. Nash–Sutcliffe efficiencies can range from $-\infty$ to 1. Efficiencies of greater than zero indicate the model has predictive power. An efficiency of 1 corresponds to a perfect match of modelled to the observed data. An efficiency of 0 ($E = 0$) indicates that the model predictions are as accurate as the mean of the observed data and negative values indicate the model is a poorer predictor than simply assuming the mean value.

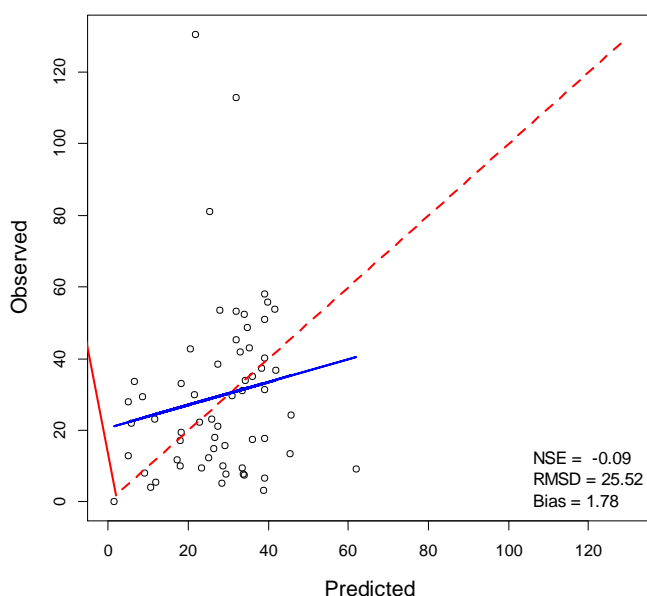


Figure 2: Comparison of observed 92nd percentile values of WCC at sites in the Whanganui-Manawatu and Canterbury regions with values predicted using the regression model fitted to the NRWQN sites. The red line is 1:1 and the blue is a line of best fit.

2.3 Derivation of concentrations thresholds

We used the two NRWQN regression models to make multiple sets of predictions of the 92nd percentile values of WCC for all segments in the REC network with stream order > 3. For each set of predictions, we held the nutrient concentrations TN (Model 1) and DRP (Model 2) constant over all segments. We incrementally varied the range of TN and DRP values from 1 to 10,000 mg m⁻³ and 0.1 to 500 mg m⁻³ respectively and made a new set of predictions to all segments for each increment. Because the TN model included the DIN to DRP ratio as a predictor (Table 1), we set a uniform DIN:DRP ratio of 13.5, which is the median N:P ratio across all segments in the national (REC) network. This corresponds to approximately co-limiting conditions. The model predictions were back-transformed by squaring the model output, and also corrected for bias using a smearing coefficient (Duan, 1983).

We stratified the results by REC Source-of-flow classes (Snelder and Biggs, 2002). REC Source-of-flow classes subdivide New Zealand's rivers on the basis of differences in catchment climate and topography. This is an appropriate classification to summarise the results because the classes can be expected to discriminate differences in the drivers of periphyton (e.g. as represented by the variables in our regression models; Table 1). Therefore we expected our analysis to produce reasonably similar nutrient concentrations within a REC Source-of-flow class and for there to be large differences in concentrations between classes.

For each REC Source-of-flow class, the proportion of segments that exceeded each of the three proxy chlorophyll *a* thresholds (92nd percentile values of WCC of 21%, 34% and 45%) was calculated for each value of TN and DRP. The concentrations of TN and DRP for which 5%, 10% and 20% of the segments exceeded the thresholds was linearly interpolated from these data.

We nominated the TN and DRP criteria for each REC Source-of-flow class to be the concentration at which 20% of segments were predicted to exceed the each of the three WCC thresholds. The proportion exceedance approach was taken because the high model uncertainty at the site scale (i.e., large RMSD values; Table 1) indicated the model predictions for individual segments was low. This could lead to overly restrictive criteria for some segments and it was assumed that a 20% exceedance level was an appropriate trade-off between defining nutrient criteria that are overly restrictive and the reverse; criteria that are not sufficiently protective. The results for 20% of segments exceeding the thresholds are presented in Table 2 and 3 below. Results for the TN and DRP concentrations at which 5% and 10% of segments were predicted to exceed the each of the three WCC thresholds are included in Appended Table 1 and 2.

The analysis produced a DRP criteria to achieve the periphyton attribute state in the WDL class that appeared unrealistically low (0.5 mg m^{-3}), which would result in a very high TN:DRP ratio (460, c.f. average across all classes of 14.3). This may be because there were no WDL sites in the NRWQN dataset from which the model was defined. We replaced the DRP criteria for WDL class with the average of two values; (1) the product of the predicted TN for the class divided by the average TN:DRP ratio for all classes (14.3) and (2) the product of the predicted TN for the class divided by the average DIN:DRP ratio of the most closely related class to WDL (27.4 for CDL). This produced a DRP level of 12.2 mg m^{-3} that was used to replace the analysed concentration for class WDL.

To provide an indication of the sensitivity of the thresholds to the model uncertainty we repeated the analysis with the model uncertainty (RMSD) subtracted and added from/to the predictions. This provides an estimate of the standard error for the TN and DRP thresholds. The interval was very large reflecting the large model uncertainty and the effect of log transformation of the nutrient variables (Table 2 and 3).

Table 2: The TN concentration (mg m⁻³) for which 20% of all segments (order > 3) belonging to each REC class exceeded the WCC cover thresholds of 21, 34 and 43% for 8% of the time. These WCC thresholds correspond to the NOF chlorophyll *a* thresholds of 50, 120 and 200 mg m⁻². For each WCC threshold, the three sets of TN concentration thresholds correspond to model predictions (middle value) and the lower (left value) and upper (right value) confidence limits corresponding to the uncertainty (RMSD) of the regression model. See Appended Table 1 for results corresponding to 5% and 10% of segments exceeding the threshold.

	21%			34%			43%		
	Lower CI		Upper CI	Lower CI		Upper CI	Lower CI		Upper CI
CWL	3	6	494	4	269	1655	7	541	3267
WWL	3	4	299	3	58	1076	4	336	2104
WDL	5	3	144	3	6	664	3	230	1311
WDLk	5	4	128	3	219	1196	5	293	2407
WWLk	3	3	186	3	213	1276	3	420	2514
WXL	3	5	490	4	261	1597	6	560	3156
WWH	4	11	598	6	339	1995	12	709	3864
WXH	3	7	545	4	293	1795	8	606	3587
CWH	4	7	565	5	299	1893	8	633	3720
CWLk	4	4	381	4	217	1293	4	426	2555
CXL	5	233	1307	22	731	4217	253	1399	8270
CXH	8	249	1507	44	846	5004	279	1667	9861
CDLk	5	3	180	3	219	1268	4	426	2518
CXLk	3	5	469	4	250	1633	5	537	3222
CXM	7	92	1035	17	591	3434	161	1148	6774
CWM	3	5	498	4	267	1701	5	549	3367
CDH	3	3	368	3	219	1284	4	430	2531
CDL	3	4	345	3	162	1178	5	389	2335
CDM	3	3	452	3	245	1473	3	470	2904
CXGM	4	10	554	6	322	1760	11	598	3449
CWGM	5	3	497	3	269	1791	3	575	3455
ALL	4	6	455	4	252	1503	6	496	2961

Table 3: The DRP concentrations (mg m^{-3}) for which 20% of all segments (order > 3) belonging to each REC class exceeded the 92nd percentile WCC cover thresholds of 21, 34 and 43%. These WCC thresholds correspond to the NOF chlorophyll *a* thresholds of 50, 120 and 200 mg m^{-2} . NA values occur where predicted values are higher than the maximum value or lower than the minimum value in the current dataset. For each WCC threshold, the three sets of DRP concentration thresholds correspond to model predictions (middle value) and the lower (left value) and upper (right value) confidence limits corresponding to the uncertainty (RMSD) of the regression model. See Appended Table 2 for results corresponding to 5% and 10% of segments exceeding the threshold.

	21%			34%			43%		
	Lower CI		Upper CI	Lower CI		Upper CI	Lower CI		Upper CI
CWL	NA	0.4	28.1	0.2	4.4	300.1	0.4	33.2	NA
WWL	NA	0.3	7.2	NA	0.6	101.1	0.3	11.2	381.2
WDL	NA	NA	0.5	NA	0.3	32.8	0.1	12.2*	124.1
WDLk	NA	0.3	4.1	NA	0.6	103.1	0.3	11.4	377.3
WWLk	NA	0.3	11.8	0.1	0.4	185.1	0.3	20.5	NA
WXL	NA	0.4	32.7	0.3	8.9	330.6	0.4	38.8	NA
WWH	NA	0.6	52.2	0.3	17	NA	0.6	62.5	NA
WXH	NA	0.4	51.3	0.2	13.6	NA	0.5	61.5	NA
CWH	NA	0.5	52	0.3	15.9	NA	0.5	61.2	NA
CWLk	NA	0.3	21.2	0.3	1.1	227.2	0.3	24.7	NA
CXL	0.3	5.4	295	1	91.9	NA	10.6	350.3	NA
CXH	0.4	13.5	452.7	2.1	140.9	NA	15.9	NA	NA
CDLk	NA	0.3	20.5	0.2	0.5	214.1	0.3	23.6	NA
CXLk	NA	0.3	41.2	0.2	12.2	432.1	0.3	48.7	NA
CXM	0.3	6.6	285.5	1.2	88.8	NA	9.5	335.9	NA
CWM	NA	0.4	60	0.3	17.8	NA	0.4	71.4	NA
CDH	NA	0.3	17.2	0.2	0.8	184.3	0.3	20.7	NA
CDL	NA	0.3	12.3	0.1	1	127.8	0.3	14.2	486.1
CDM	NA	0.3	35.1	0.3	11.3	381.4	0.3	42	NA
CXGM	NA	0.8	84.4	0.5	27.2	NA	0.9	100.7	NA
CWGM	NA	0.3	90	0.3	27.9	NA	0.3	106.1	NA
ALL	NA	0.4	28.1	0.2	5.8	296	0.4	33.2	NA

* This value was adjusted as described above from an analytical result of 0.5 mg m^{-3}

3 Discussion

The TN and DRP concentration criteria we derived in this analysis are variable between REC Source-of-flow classes (Table 2 and 3). This variability is a result of the strong influence of the hydrological indices in the regression models, which is also consistent with the conceptual model of periphyton abundance (Biggs, 1996). The variation is also consistent with our knowledge of variability in flow regimes between REC classes (Snelder *et al.*, 2005).

REC classes representing wet and extremely wet environments (i.e., CX, CW, WW) have frequent high flows and relatively high base flows as a result of climate driven frequent rainfall (Snelder *et al.*, 2005). Our nutrient concentration criteria for these classes were higher than classes with dry climates due to the more frequent flushing by high flows and relative lack of low flow events. The criteria also varied in expected ways with variation in the topography (second level) of the REC. For example, within the CW climate class the nutrient criteria are lowest for the lake (Lk) topography class. This reflects the buffering of flow variation in catchments with lakes and consequently lower flow variability than would be expected for this climate class were lakes not present. There is also a tendency for the Hill topography class to have higher nutrient criteria than the Mountain class. This is due to hill dominated catchments having a relatively non-seasonal and consistent response to precipitation, whereas mountain dominated catchments tend to have stable winter flows due to precipitation falling as snow and high low flows due to snow-melt during summer.

There was a tendency for the warmer classes to have lower nutrient concentration criteria than the cool classes. This is also consistent with the influence of temperature (T95) and light (PAR) in the regression models. Both of these variables had positive regression coefficients in our models indicating the periphyton abundance increases with increasing values of these predictors. It is also consistent with the expectation that summer water temperatures and solar radiation are generally higher in the warm (generally northern) regions of New Zealand. Thus, after accounting for the other variables in the models, nutrient concentrations must be lower to achieve a given periphyton abundance threshold in a Warm REC class compared to a Cool class.

The nutrient guidelines provided by Biggs (2000) are shown in Table 4 and the equivalent guideline values for each REC class are shown in Table 5. We estimated the equivalent guideline values in each REC class by first finding the mean value of FRE3 in each class based on the predicted values of FRE3 that are available for all REC segments (Snelder and Booker, 2013). We converted the REC class mean values to mean days of accrual ($Da = 1/FRE3 \times 365$) and interpolated the DIN and DRP criteria from the Biggs guideline values (Table 4). The Biggs' guideline values for DIN (Table 4) were converted to equivalent TN values based on a comparison of the site medians for these two species observed for the NRWQN. A regression ($DIN \text{ (site median)} = 0.75TN \text{ (site median)}$) had an r^2 of 96% and was used to convert the interpolated DIN criteria for each REC class to an equivalent TN criteria (Table 5).

Table 5 indicates the present study's thresholds (as equivalent DIN thresholds) are significantly higher than those of Biggs (2000). This is an expected outcome because Biggs (2000) thresholds were developed for sites in which nutrients were the only limitation on periphyton growth, whereas the present study uses sites that have a degree of light limitation and probably are also affected by other factors such as substrate suitability and hydrological variability.

Table 4: MFE periphyton guidelines (Biggs 2000) criteria for DIN and DRP for chlorophyll *a* <50 and < 200 mg m⁻² chl *a* (this approximates 21% and 43% WCC). Days of accrual is based on the mean value of days of accrual (1/FRE3 x 365).

Days of accrual	<50 mg m ⁻² chl <i>a</i>		<200 mg m ⁻² chl <i>a</i>	
	DIN (mg m ⁻³)	DRP (mg m ⁻³)	DIN (mg m ⁻³)	DRP (mg m ⁻³)
20	<20	<1	<295	<26
30	<10	<1	<75	<6
40	<10	<1	<34	<2.8
50	<10	<1	<19	<1.7
75	<10	<1	<10	<1
100	<10	<1	<10	<1

Table 5: DIN and DRP concentration thresholds for each REC Source-of-flow class as suggested by the MFE periphyton guidelines (Biggs 2000) based on the mean value of days of accrual (1/FRE3 x 365) for segments in class. The TN values have been derived by dividing the DIN values derived from the periphyton guidelines by 0.75 (the mean ration of DIN to TN for sites in the NRWQN).

REC Class	Days of accrual	<50 mg m ⁻² chl <i>a</i>		<200 mg m ⁻² chl <i>a</i>	
		TN (mg m ⁻³)	DRP (mg m ⁻³)	TN (mg m ⁻³)	DRP (mg m ⁻³)
CWL	42	13	1	41	3
WWL	42	13	1	42	3
WDL	54	13	1	24	2
WDLk	45	13	1	36	2
WWLk	51	13	1	25	2
WXL	35	13	1	72	4
WWH	40	13	1	47	3
WXH	34	13	1	79	5
CWH	49	13	1	27	2
CWLk	75	13	1	13	1
CXL	26	18	1	203	13
CXH	31	13	1	92	6
CDLk	78	13	1	13	1
CXLk	48	13	1	29	2
CXM	38	13	1	54	3
CWM	57	13	1	22	2
CDH	59	13	1	21	1
CDL	58	13	1	22	1
CDM	63	13	1	19	1
CXGM	34	13	1	77	5
CWGM	77	13	1	13	1

Our results are sensitive to assumptions and thresholds. For example, the derived nutrient criteria were sensitive to the exact value of periphyton WCC that was assumed to be equivalent to the NOF chlorophyll *a* thresholds of 50 and 200 mg m⁻². In addition, when the uncertainty of the regression model (i.e., RMSD) was carried through into the derivation of nutrient thresholds the range between the upper and lower confidence intervals was very large. These sensitivities arise because of the form of the regression model (i.e., the log10 and square root transformations) and the large model uncertainty. We note that many studies have found that relationships between nutrients and periphyton abundance are highly uncertain (see (Matheson *et al.*, 2012) and (Snelder *et al.*, 2014)).

There were some significant limitations associated with this analysis and the derived nutrient criteria should be treated with caution and only be used as a provisional guide. The NRWQN data was not collected with the intention of deriving nutrient concentration criteria and this limitation probably contributes to the large uncertainties in the derived criteria. The model fitted to the NRWQN data performed poorly when tested with independent data. It is likely that there are differences in the protocols that are being used to observe periphyton cover and this means the NRWQN dataset is not comparable with data being collected by regional councils. There is clearly a need to improve the data collection methods and consistency and to collect data for a wider range of sites. Finally our regression models that underlie the method did not explain a lot of variation in periphyton abundance and predictions made using these models have high uncertainty. The reasons for the poor performance of these models are unknown but may reflect processes associated with periphyton growth and loss that were not represented in the regression models.

4 References

- Biggs, B.J.F., 1996. Hydraulic Habitat of Plants in Streams. *Regulated Rivers: Research and Management* 12:131–144.
- Biggs, B.J.F., 2000. Eutrophication of Streams and Rivers: Dissolved Nutrient-Chlorophyll Relationships. *Journal of the North American Benthological Society*. 19:17–31.
- Davies-Colley, R.J., D.G. Smith, R.C. Ward, G.G. Bryers, G.B. McBride, J.M. Quinn, and M.R. Scarsbrook, 2011. Twenty Years of New Zealand's National Rivers Water Quality Network: Benefits of Careful Design and Consistent Operation. Wiley Online Library. <http://onlinelibrary.wiley.com/doi/10.1111/j.1752-1688.2011.00554.x/full>. Accessed 16 Oct 2014.
- Dodds, W.K.K. and E.B. Welch, 2000. Establishing Nutrient Criteria in Streams. *Journal of the North American Benthological Society* 19:186–196.
- Doyle, M.W. and E.H. Stanley, 2006. Exploring Potential Spatial-Temporal Links between Fluvial Geomorphology and Nutrient-Periphyton Dynamics in Streams Using Simulation Models. *Annals of the Association of American Geographers* 96:687–698.
- Duan, N., 1983. Smearing Estimate: A Nonparametric Retransformation Method. *Journal of the American Statistical Association* 78:605–610.
- Flinders, C.A. and D.D. Hart, 2009. Effects of Pulsed Flows on Nuisance Periphyton Growths in Rivers: A Mesocosm Study. *River Research and Applications* 25:1320–1330.
- Jowett, I.G. and J. Richardson, 1990. Microhabitat Preferences of Benthic Invertebrates in a New Zealand River and the Development of in-Stream Flow-Habitat Models for Deleatidium Spp. *New Zealand Journal of Marine and Freshwater Research* 24:19–30.
- Julian, J.P., M.W. Doyle, and E.H. Stanley, 2008. Empirical Modeling of Light Availability in Rivers. *Journal of Geophysical Research* 113:G03022.
- Kilroy, C., S.T. Larned, and B.J.F. Biggs, 2009. The Non-Indigenous Diatom *Didymosphenia Geminata* Alters Benthic Communities in New Zealand Rivers. *Freshwater Biology* 54:1990–2002.
- Leathwick, J., T. Snelder, W. Chadderton, J. Elith, K. Julian, and S. Ferrier, 2011. Use of Generalised Dissimilarity Modelling to Improve the Biological Discrimination of River and Stream Classifications. *Freshwater Biology* 56:21–38.
- Matheson, F., J. Quinn, and C. Hickey, 2012. Review of the New Zealand Instream Plant and Nutrient Guidelines and Development of an Extended Decision Making Framework: Phases 1 and 2 Final Report. Prepared for the Ministry of Science and Innovation, Envirolink fund., NIWA, Hamilton, New Zealand.
- MFE, 2000. New Zealand Periphyton Guideline: Detecting, Monitoring and Managing Enrichment of Streams. Wellington.
- Olden, J.D. and N.L. Poff, 2003. Redundancy and the Choice of Hydrologic Indices for Characterizing Streamflow Regimes. *River Research & Applications* 19:101–121.

- Quinn, J.M. and C.W. Hickey, 1990. Characterisation and Classification of Benthic Invertebrate Communities in 88 New Zealand Rivers in Relation to Environmental Factors. *New Zealand Journal of Marine and Freshwater Research* 24:387–410.
- Quinn, J.M., E. Raaphorst, N.I. of Water, and A.R. (NZ), 2009. Trends in Nuisance Periphyton Cover at New Zealand National River Water Quality Network Sites, 1990-2006. National Institute of Water & Atmospheric Research Limited.
- Rutherford, J.C., M.R. Scarsbrook, and N. Broekhuizen, 2000. Grazer Control of Stream Algae: Modeling Temperature and Flood Effects. *Journal of Environmental Engineering* 126:331–339.
- Smith, D.G. and G.B. McBride, 1990. New Zealand's National Water Quality Monitoring Network - Design and First Year's Operation. *Water Resources Bulletin* 26:767–775.
- Snelder, T.H. and B.J.F. Biggs, 2002. Multi-Scale River Environment Classification for Water Resources Management. *Journal of the American Water Resources Association* 38:1225–1240.
- Snelder, T.H. and D.J. Booker, 2013. Natural Flow Regime Classifications Are Sensitive to Definition Procedures. *River Research and Applications* 29:822–838.
- Snelder, T.H., D.J. Booker, J.M. Quinn, and C. Kilroy, 2014. Predicting Periphyton Cover Frequency Distributions across New Zealand's Rivers. *JAWRA Journal of the American Water Resources Association* 50:111–127.
- Snelder, T.H., R. Woods, and B.J.F. Biggs, 2005. Improved Eco-Hydrological Classification of Rivers. *River Research and Applications* 21:609–628.
- Suren, A.M., B.J.F. Biggs, M.J. Duncan, L. Bergey, and P. Lambert, 2003. Benthic Community Dynamics during Summer Low-flows in Two Rivers of Contrasting Enrichment 2. Invertebrates. *New Zealand Journal of Marine and Freshwater Research* 37:71–83.
- Uehlinger, U.E., 1991. Spatial and Temporal Variability of the Periphyton Biomass in a Prealpine River (Necker, Switzerland). *Arch. Hydrobiol* 123:219–237.
- 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.

Appended Table 1. The TN concentrations (mg m⁻³) for which 5% and 10% of all segments (order > 3) belonging to each REC class exceeded the 92nd percentile WCC cover thresholds of 21, 34 and 43%.

	5% exceed								
	21%			34%			43%		
	Lower CI		Upper CI	Lower CI		Upper CI	Lower CI		Upper CI
CWL	1	2	283	1	39	1055	2	321	2093
WWL	1	1	212	1	3	680	1	217	1438
WDL	2	1	3	1	2	485	1	10	1041
WDLk	2	1	25	1	5	833	2	30	1097
WWLk	1	1	27	1	5	871	1	39	1757
WXL	1	2	262	1	28	1041	2	97	2098
WWH	1	2	172	2	77	959	2	220	1867
WXH	1	2	84	1	35	565	2	135	1669
CWH	1	2	345	1	209	1251	2	417	2477
CWLk	1	1	222	1	5	887	1	238	1852
CXL	1	5	771	2	437	2583	7	873	5031
CXH	2	6	722	3	423	2488	7	832	4889
CDLk	2	1	27	1	7	1070	1	46	2086
CXLk	1	1	54	1	10	543	1	54	1042
CXM	2	3	526	2	283	1683	4	576	3260
CWM	1	1	293	1	59	1106	2	339	2257
CDH	1	1	234	1	6	898	1	252	1781
CDL	1	1	221	1	4	789	1	235	1532
CDM	1	1	114	1	209	1243	1	413	2476
CXGM	1	2	275	2	146	1151	2	364	2291
CWGM	2	1	247	1	216	1386	1	436	2767
ALL	1	2	245	1	11	903	2	267	1817

	10% exceed								
	21%			34%			43%		
	Lower CI		Upper CI	Lower CI		Upper CI	Lower CI		Upper CI
CWL	2	3	399	2	217	1321	3	432	2621
WWL	2	2	238	2	8	848	2	251	1694
WDL	2	2	12	2	2	576	2	93	1153
WDLk	2	2	43	2	24	1072	2	65	2162
WWLk	2	2	52	2	26	973	2	106	1961
WXL	2	2	338	2	214	1260	3	424	2493
WWH	2	3	449	3	236	1408	3	491	2807
WXH	2	3	415	2	220	1268	3	443	2685
CWH	2	3	444	2	235	1490	3	482	2942
CWLk	2	2	266	2	28	1074	2	304	2103
CXL	2	29	960	5	536	3216	55	1082	6292
CXH	3	31	937	7	529	3066	55	1025	6010
CDLk	2	2	51	2	46	1145	2	146	2241
CXLk	2	2	269	2	108	1085	2	323	2267
CXM	3	10	705	4	403	2372	13	800	4656
CWM	2	2	413	2	220	1371	2	439	2704
CDH	2	2	272	2	34	1068	2	315	2127
CDL	2	2	257	2	13	925	2	278	1846
CDM	2	2	416	2	221	1336	2	432	2696
CXGM	2	3	378	2	237	1337	3	456	2593
CWGM	2	2	433	2	232	1665	2	478	3103
ALL	2	2	316	2	130	1157	3	375	2289

Appended Table 2. The DRP concentrations (mg m⁻³) for which 5% and 10% of all segments (order > 3) belonging to each REC class exceeded the 92nd percentile WCC cover thresholds of 21, 34 and 43%.

	5% exceed								
	21%			34%			43%		
	Lower CI		Upper CI	Lower CI		Upper CI	Lower CI		Upper CI
CWL	NA	0.1	7.7	NA	0.3	102.9	0.1	11.4	387.3
WWL	NA	0.1	0.3	NA	0.2	45	0.1	0.4	172.9
WDL	NA	NA	0.1	NA	0.1	19.2	NA	0.2	73.3
WDLk	NA	0.1	0.3	NA	0.2	83.3	0.1	0.4	309
WWLk	NA	0.1	0.5	NA	0.1	71.9	0.1	0.9	272.1
WXL	NA	0.1	11	0.1	0.3	146.8	0.1	13.5	NA
WWH	NA	0.2	6.6	NA	0.6	89.5	0.2	10.8	336.6
WXH	NA	0.1	2.8	NA	0.4	67.7	0.1	4.9	245.4
CWH	NA	0.1	20.9	0.1	1	218	0.1	23.6	NA
CWLk	NA	0.1	10.9	0.1	0.2	125.2	0.1	11.6	480.7
CXL	0.1	0.3	113.3	0.2	34.1	NA	0.3	133	NA
CXH	NA	0.4	110.7	0.2	33.3	NA	0.4	130.4	NA
CDLk	NA	0.1	11.2	0.1	0.1	142.2	0.1	12.9	NA
CXLk	NA	0.1	2.9	NA	0.4	62.1	0.1	3	233.5
CXM	NA	0.3	75.1	0.2	22.9	NA	0.3	88.1	NA
CWM	NA	0.1	29.3	0.1	4.4	308.6	0.1	33.8	NA
CDH	NA	0.1	2	NA	0.2	87.6	0.1	5.3	331.6
CDL	NA	0.1	0.4	NA	0.2	52.2	0.1	0.6	197.3
CDM	NA	0.1	22.8	0.1	0.5	251.9	0.1	28.4	NA
CXGM	NA	0.2	32.9	0.1	10.6	371.1	0.2	41.2	NA
CWGM	NA	0.1	16.4	0.1	3.2	468.9	0.1	53.7	NA
ALL	NA	0.1	2.8	NA	0.3	82.3	0.1	5.4	311.9

	10% exceed								
	21%			34%			43%		
	Lower CI		Upper CI	Lower CI		Upper CI	Lower CI		Upper CI
CWL	NA	0.2	15.7	0.1	0.7	178.7	0.2	19.9	NA
WWL	NA	0.1	0.8	NA	0.2	63.9	0.2	1.3	243.2
WDL	NA	NA	0.2	NA	0.2	24.2	NA	0.2	92.9
WDLk	NA	0.2	0.6	NA	0.2	86.9	0.2	1.6	330.5
WWLk	NA	0.2	2.4	NA	0.2	89.2	0.2	7.3	335.2
WXL	NA	0.2	18.9	0.1	0.9	208.9	0.2	24.3	NA
WWH	NA	0.2	32.4	0.1	3.7	321.9	0.2	31.1	NA
WXH	NA	0.2	20.6	0.1	2	276.6	0.2	30.8	NA
CWH	NA	0.2	32.2	0.2	9.1	340.5	0.2	37	NA
CWLk	NA	0.2	13.8	0.1	0.3	155.4	0.2	16.3	NA
CXL	0.1	0.7	166.8	0.3	52.4	NA	1.1	197.5	NA
CXH	0.2	1.4	180.9	0.5	55.9	NA	1.8	213.3	NA
CDLk	NA	0.2	13.7	0.1	0.2	167.3	0.2	18.4	NA
CXLk	NA	0.2	17.4	0.1	1.4	207	0.2	21.8	NA
CXM	0.1	0.8	147.8	0.3	44.4	NA	1	173.1	NA
CWM	NA	0.2	39.5	0.2	12	419.6	0.2	46.6	NA
CDH	NA	0.2	11.3	0.1	0.3	126	0.2	12.7	478.1
CDL	NA	0.2	1.7	NA	0.3	75.7	0.2	3.4	285.1
CDM	NA	0.2	28.2	0.2	2.7	304.3	0.2	33.9	NA
CXGM	NA	0.3	46.8	0.2	14.4	484.8	0.3	55.1	NA
CWGM	NA	0.2	35	0.2	9.8	493	0.2	58.7	NA
ALL	NA	0.2	13.5	0.1	0.8	148.2	0.2	16	NA