Strategic assessment of New Zealand’s freshwaters for recreational use: a human health perspective

Escherichia coli in rivers and planktonic cyanobacteria in lakes

September 2016

Prepared By:

Ton Snelder

Susie Wood*

Javier Atalah*

* Cawthron Institute

For any information regarding this report please contact:

Ton Snelder

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

LWP Ltd
PO Box 70
Lyttelton 8092
New Zealand

LWP Client Report Number: LWP Client Report 2016-011

Report Date: September 2016

LWP Project: 2016-011

Quality Assurance Statement

Version	Date	Reviewed by	
1	15 September 2016	Roger Young and Joanne Clapcott[1]	Cawthron Institute

[1] Copies of the review comments and responses and changes made are available on request from MFE.

Table of Contents

Table of figures

Table of tables

Executive Summary

The concentration of *Escherichia coli* (*E.coli*) is an indicator of human or animal faecal contamination and the risk of infectious human disease from waterborne pathogens. Many planktonic cyanobacteria produce toxins (cyanotoxins) that are a health risk to humans during recreational activities. Total cyanobacterial biovolume is used to assess the degree of risk they pose.

In this study, three datasets were obtained from regional council state of environment monitoring (SOE) and other water quality monitoring programmes; *E.coli* concentrations in rivers, water quality parameters in lakes, and four water quality variables and their concurrent planktonic cyanobacterial biovolumes in lakes. *E coli* also indicate risks to human health in lakes, however, current datasets are temporally and spatially limited preventing their inclusion in the current study. A number of analyses were undertaken that aimed to provide a strategic assessment of New Zealand's freshwaters for recreational use from a human health perspective.

Monthly *E.coli* measurements from 753 sites on rivers with durations from between 5 and 24 years were analysed to extract several types of summary information. First, the relationship between flow and *E.coli* concentration was quantified using linear regression models. Second, the timing of *E.coli* exceedances was assessed by counting the proportion of samples in each month that exceed concentration thresholds of 260 and 540 *E.coli* 100 mL^{-1}. These thresholds define attribute bands in the National Policy Statement for Freshwater Management (NPS-FM). Third, the median and 95th percentile *E.coli* concentrations for each site were calculated over all observations. The NPS-FM uses these summary statistics and the attribute band thresholds to define the *E. coli* attribute state for a water body. Fourth, the percentage of samples that exceeded 260 and 540 *E.coli* 100 mL^{-1} (PercGT260 and PercGT540) were calculated for each site over all observations. The four site summary statistics (median, 95th percentile, PercGT260 and PercGT540) were combined with environmental data describing river catchments to make national scale spatial predictions of these values.

Quarterly SOE measurements of lake water quality variables comprising chlorophyll *a* (total phytoplankton biomass), total nitrogen (TN), total phosphorus (TP), Secchi depth, and the trophic level index (TLI3) were obtained for up to 99 lakes (differed by variable) over the period 2009 to 2013. The median values of these variables were extracted for each lake. In addition, concurrent observations of cyanobacterial biovolume, chlorophyll *a*, TN, TP, and Secchi depth were obtained for 37 lakes from various regional councils and research programme datasets. The site median values of the water quality variables pertaining to the SOE data were combined with environmental data describing lakes and their catchments to make spatial predictions of these values for all of New Zealand's 3,821 lakes larger than 1 hectare. Multiple linear regression was used to establish the relationship between cyanobacterial biovolume (80th percentile) and median chlorophyll *a*, TN, TP, and Secchi depth using datasets from 37 lakes. Predicted values of chlorophyll *a*, TN, TP, and Secchi depth from the spatial models were included as 'new data' in the cyanobacterial biovolume multiple linear regression equation and used to estimate cyanobacterial biovolume for all New Zealand lakes.

The analysis indicated that relationships between *E.coli* concentrations and flow are generally weak. While relationships were stronger at some sites, the broad conclusion was that exceedance of *E.coli* threshold values is not strongly determined by flow state. The

analysis of timing of exceedances of *E.coli* thresholds indicated that exceedance of *E.coli* thresholds can occur at any time of the year. Exceedances of thresholds tend to be more frequent in autumn and spring in lowland rivers. However, the frequency of *E.coli* exceedances was not strongly predicted by month or season. These analyses indicate that broad scale patterns of the four *E.coli* statistics (site median, 95th percentile, PercGT260 and PercGT540) are relevant descriptions for assessing human health risk and that accounting for flow state or time of the year would not strongly influence conclusions drawn from these statistics.

Good performance was achieved for the four spatial models of the *E.coli* statistics (median, 95th percentile, PercGT260 and PercGT540). Predicted values of all statistics described similar spatial patterns when mapped. Values were high in low-elevation parts of rivers on the east coasts of the North and South Island, and in the inland Waikato, Ruamahunga Valley, Rangitikei-Manawatu coastal plain, Taranaki Ring Plain, and Auckland and Northland regions. Values were generally low in rivers rising in major mountain ranges (e.g. Southern Alps, Kahurangi, Kaimanawa, and Tararua Ranges), in large areas of the Department of Conservation estate (e.g. Fiordland, Westland, Te Urewera, Egmont, Whanganui and Tongariro National Parks), and in smaller, native forest-dominated areas of Northland and the Coromandel Peninsula. Rivers with catchments in low elevation areas that were characterised by high values of the *E.coli* variables coincided with land used for intensive agriculture and areas associated with urban centres.

Satisfactory to good performance was achieved for the five lake spatial models (chlorophyll *a*, TN, TP, Secchi depth and TLI3) and predictions were made for all lakes with an area greater than 1 hectare. Because of the limited number of lakes represented in the dataset used to develop the models there is a likelihood that values are over predicted in lowland areas with native catchments e.g., West Coast South Island. The mapped predictions for all five variables had similar spatial patterns, with high values of chlorophyll *a*, TN, TP and TLI3 and low values of Secchi depth, in low-elevation areas on the coasts of the North and South Island. Values of chlorophyll *a*, TN, TP and TLI3 were also high in inland areas of both islands that are dominated by agricultural land use such as Southland, parts of Otago, Hawkes Bay, Bay of Plenty, Waikato and Northland. Values of chlorophyll *a*, TN, TP and TLI3 were generally low and Secchi depth was high in the inland areas of the South Island.

Strong statistical relationships were established between the 80th percentile cyanobacterial biovolume measured in 37 lakes and chlorophyll *a*, TN, TP and Secchi depth. The most parsimonious multiple linear regression model included chlorophyll *a*, TP, and Secchi depth and had a R^2 of 0.64. When the data from spatial models was transformed into cyanobacterial biovolume, spatial patterns were similar to those observed for the other lake water parameters described above.

The predictions made using the spatial models provide a description of regional to national scale patterns in *E.coli* in rivers and cyanobacterial biovolume in lakes. The predictions are uncertain at the site-scale and actual data should be used in preference to the modelled predictions. However, the broader-scale predictions will be useful for strategic purposes such as quantifying the proportion of New Zealand’s rivers and lakes that have high and low human health risks and identifying areas of most concern.

1 Introduction

River and lake water quality across New Zealand was characterised by a recent national analysis of state and trends at over 1000 monitoring sites (Larned *et al.*, 2015). The sites are monitored as part of the State of Environment (SOE) programmes operated by regional councils, unitary authorities and the national river water quality network operated by NIWA. Larned *et al.* (2015) provided information on several monitored water quality variables that are measured at 844 river sites and 156 lakes. The datasets underlying these analyses contained quarterly or monthly measurements of physical, chemical, and biological variables over time periods from as early as 1990 to 2013.

In the current report, the river *E.coli* and lake water quality monitoring data underlying the Larned *et al.* (2015) study were utilised. The temporal variability in the key indicator of human health in rivers, the faecal indicator bacterium *Escherichia coli (E.coli)* was investigated. The data were then used to develop spatial models that predicted *E.coli* in all New Zealand's rivers and water quality in all lakes with an area greater than 1 hectare. The benefit of spatial modelling is that it provides a large-scale assessment of water quality that is more representative than assessments based on aggregating raw monitoring site data. The latter approach can lead to conclusions about water quality patterns that are biased by the non-random locations of monitoring sites (Larned and Unwin, 2012).

The water quality variables that are routinely measured by SOE monitoring of lakes are focussed on understanding the impacts of nutrients. Although *E.coli* is a relevant indicator of human health risks in lakes it has often not been included SOE monitoring programmes. However, nutrients in lakes can stimulate growth of planktonic cyanobacteria, which can reach high densities, form blooms, and some species produce toxins that are harmful to humans. The presence of cyanobacteria in lakes is known to be related to water quality variables that characterise lake trophic status. In this study, an additional dataset was collated consisting of cyanobacterial biovolumes (a measure of total cyanobacterial biomass) and four water quality variables. We investigated the relationship between cyanobacterial biovolume and the water quality variables. These relationships were then used to transform water quality predictions, made using the lake spatial models, into predictions of cyanobacterial biovolume for all New Zealand lakes. The *E. coli* in rivers and cyanobacterial biovolume in lakes spatial models provide a broad scale assessment of New Zealand's freshwaters for recreational use, from a human health perspective.

This report provides a detailed description of the methods used to extract variables from available data and to produce spatial predictions at unmonitored locations. The methods used to prepare the water quality variables data, make assessments of the representativeness of the monitoring sites, and to undertake the spatial modelling are described. The results provide national maps of river *E.coli*, lake water quality and cyanobacterial biovolume. Measures of model performance and the important relationships between water quality variables and predictors are described and discussed.

2 Data

2.1 River water quality data

We used the SOE data for rivers analysed by Larned *et al.* (2015) for the current study. Detailed methods for obtaining and grooming these data are provided by Larned *et al.* (2015). In this study we were only interested in the microbiological measure *E. coli*. The concentration of the bacterium *E.coli* 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 a National Policy Statement for Freshwater Management (NPS-FM; Ministry for the Environment, 2014) attribute based on *E. coli* concentrations that is related to the management of human health for recreation. The national bottom line for *E. coli* concentrations for secondary contact recreation is 1000 *E. coli* 100 mL^{-1} (as an annual median). The NPS-FM also stipulates that the "minimum acceptable state" for waters managed for primary contact recreation is an *E. coli* concentration of 540 *E.coli* 100 mL^{-1}, as a 95th percentile. The minimum acceptable state is consistent with an exposure to a moderate risk of infection (less than 5% risk) when undertaking activities likely to involve full immersion (Ministry for the Environment, 2014). The NPS-FM stipulates the "A-band" as a concentration of 260 *E.coli* 100 mL^{-1} as a 95th percentile. The A-band is consistent with an exposure to a low risk of infection (up to 1% risk) when undertaking activities likely to involve full immersion (Ministry for the Environment, 2014).

The *E.coli* data in the Larned *et al.* (2015) dataset comprised monthly or quarterly samples that extended from 1990 at some sites to the end of 2013. In this study we used these data to assess relationships between *E.coli* and flow, the timing of *E.coli* exceedances of the thresholds described above, and the proportion of samples for which *E.coli* concentrations exceeded thresholds. The robustness of these assessments increases with the number of samples. We therefore maximised the number of samples by retaining the entire time series for each site and applied the single filtering rule that sites be associated with at least 30 samples to ensure the data were representative of each site. The potential issue with this approach is that the data can be influenced by trends. This can mean that analyses carried out on the entire time series is not representative of the more recent past. This study assumed that broad scale (i.e. aggregated analyses) were not strongly affected by trends based on the results of Larned *et al.* (2015). They found that very few sites in the national dataset had important (i.e. meaningful) trends in *E.coli* over the last 10 or 20 years.

E.coli measurements can be too high to be measured with precision. These are recorded in the data as greater than the "reporting limit" and are referred to as right-censored data. All right-censored data were replaced with values estimated using a procedure based on "survival analysis" (Helsel, 2012). 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 (see Larned *et al.*, 2015 for details). A final criterion was that fewer than 50% of the observations were censored (i.e. below the detection limit). This resulted in a dataset comprising 753 sites (Figure 1).

Approximately 25% of samples in our dataset were associated with flow measurements that were either obtained from a water level (flow) recorded or were provided by the monitoring organisation. There was no flow data provided for 65% of the sites and 75% of all samples. For these samples, flow estimates were obtained from the TopNet national hydrological model (see Larned *et al.*, 2015 for details).

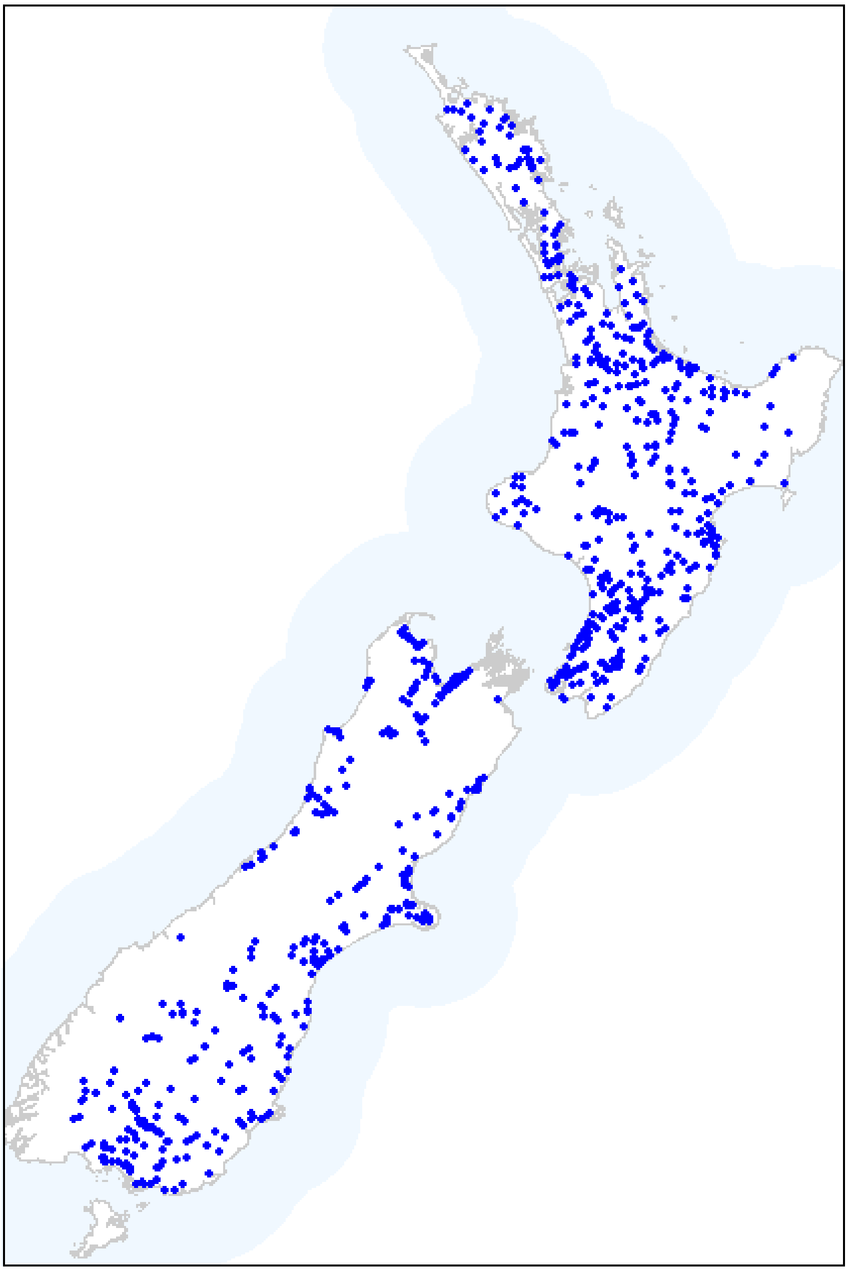

Figure 1. Location of the river state of environment sites that were the source of the* Escherichia coli *data used in this study.

2.2 Lake water quality data

Lake water quality data were derived from two sources. First, the SOE data for lakes analysed by Larned *et al.* (2015) were used in the current study for spatial modelling. Detailed methods for obtaining and grooming these data are provided by Larned *et al.* (2015). The lake SOE data analysed by Larned *et al.* (2015) included five water quality variables that correspond to physical, chemical and biological conditions (Table 1). The variables included total nitrogen (TN) and total phosphorus (TP), the visual clarity indicator

Secchi depth, phytoplankton biomass as chlorophyll *a*, and the trophic level index (TLI3, comprising TN and TP and chlorophyll *a*; Burns *et al*., 2000).

This study used only lake SOE data for the five-year period from 2009 to 2013. Two filtering rules were applied to ensure that the SOE data were representative of each lake and variable. First, at least eight samples were available for the five-year period. Second, less than 50% of the observations of each variable were censored (i.e. below the detection limit). This was a relaxation of the inclusion rule of Larned *et al.* (2015) who required 18 samples in the five-year period. This relaxation of the inclusion rules increased the number of sites used in this study compared to Larned *et al.* (2015) (Figure 2).

Measurements of some of the lake variables can be too low to be measured with precision. These are recorded as less than the "detection limit" and are referred to as left-censored data. We imputed replacement values for the left censored data using Regression on Order Statistics (ROS) (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 (see Larned *et al.*, 2015 for details).

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

Variable type	Variable	Abbreviation	Units	Number of lakes
Physical	Secchi depth	SECCHI	m	64
Chemical	Total nitrogen	TN	mg/m^3	82
	Total phosphorus	TP	mg/m^3	99
Phytoplankton	Chlorophyll *a*	CHLA	mg L^{-1}	92
Water quality index	Trophic Level Index	TLI3	unitless	76

Although *E.coli* has been sampled in 37 lakes nationally as part of SOE programmes (Martin Unwin, *pers comm*), the samples are sporadic and insufficient for robust spatial modelling. We therefore focussed on a national assessment of planktonic cyanobacteria in lakes. An increasing number of cyanobacterial species are known to include toxin-producing strains. These natural toxins, known as cyanotoxins, are a threat to humans during recreational activities and the risk to humans is estimated by measuring planktonic cyanobacterial biovolumes. The NPS-FM defines an attribute for cyanobacteria in lakes based on biovolumes that is related to the management of human health for recreation. Cell size varies among cyanobacterial species and the use of biovolumes as an NPS-FM attribute allows these differences to be accounted for. Additionally, toxin concentration per cell is thought to be more closely related to cyanobacterial biovolume than to total cell number. The national bottom line for total cyanobacterial biovolume is >1.8 mm^3 L^{-1} of potentially toxic cyanobacteria, or >10 mm^3 L^{-1} when no known toxin producing species or no toxins are detected. The suggested statistic to make this assessment is an 80th percentile, using a minimum of 12 samples collected over three years. Concentrations of >0.5-1.8 mm^3 L^{-1} and >0.5 mm^3 L^{-1} of potentially toxic cyanobacteria as 80th percentile defines the C and A attribute bands respectively (note: there is no B band for planktonic cyanobacteria). The same values are used to trigger the action, alert and surveillance mode levels in the New Zealand Guidelines for Managing Cyanobacteria in Recreational Fresh Waters (Wood *et al.*, 2009).

Cyanobacterial biovolumes are not routinely measured as part of SOE monitoring and were not included in the dataset analysed by Larned *et al.* (2015). However, a dataset comprising cyanobacterial biovolume data and all, or a subset of chlorophyll *a*, TN, TP, and Secchi depth was obtained for 66 lakes. Our approach was therefore based on the assumption that cyanobacterial biovolume would be related to trophic state as represented by one or more of the variables in the SOE dataset (i.e. chlorophyll a, TN, TP, TLI3, Secchi depth; Wood *et al.*, 2016, Smith *et al.*, 2016). Our strategy was to first develop a relationship between these water quality variables and cyanobacterial biovolume and then to use the spatial predictions developed from the SOE data to extrapolate the relationship nationally.

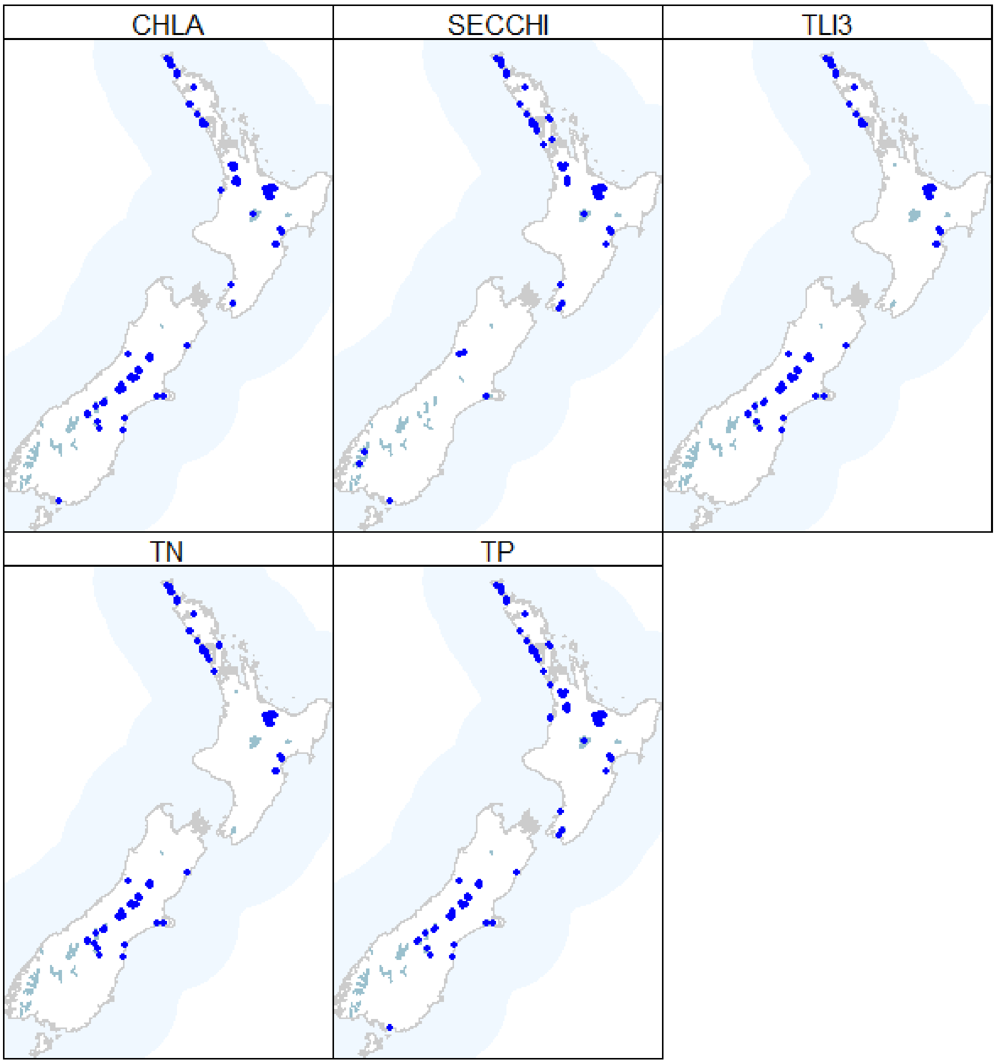

Figure 2. Locations of lake state of environment monitoring sites associated with the five water quality variables. *The locations shown on each panel correspond to the sites that were included in this study for each variable listed in Table 1.*

We obtained the dataset comprising cyanobacterial biovolume data and chlorophyll *a* TN, TP, and Secchi depth from various sources including regional council monitoring programmes and research projects (Appendix 1 and 2). The per lake sample numbers in our cyanobacterial biovolume dataset ranged from 1 to 475 (Appendix 1 and 2). When data were provided as cell concentrations these were converted to biovolumes using values given in Appendix 4 of Wood *et al.* (2009). When genera or species data was not available in this document, values were obtained from the 'biovolume calculator' at http://www.depi.vic.gov.au/water/rivers-estuaries-and-wetlands/blue-green-algae/blue-green-algae-resources**.**

Lakes where sample size was less than seven were excluded from further analysis. This value was chosen based on our expert judgement to maintain as many datasets as possible while avoiding biases caused by low or only targeted (e.g. during a bloom event) sampling effort (Appendix 1 and 2). This left a total of 37 lakes in the dataset (Figure 3; Appendix 1). These datasets were biased geographically, for example, useable data were only available from three lakes in the South Island (Figure 3). However, because these datasets were only used to establish relationships between cyanobacterial biovolumes and water quality variables (see Section 3.1.2) this was not considered problematic. Although some geographic (altitudinal, latitudinal and longitudinal) patterns have been observed in planktonic cyanobacteria at a national scale, trophic state is considerably more important in structuring cyanobacterial communities (Wood *et al.*, In review). Additionally, in general the dominant and bloom forming taxa are found nationwide (Wood *et al.*, In review) and were all represented in the dataset used in this study.

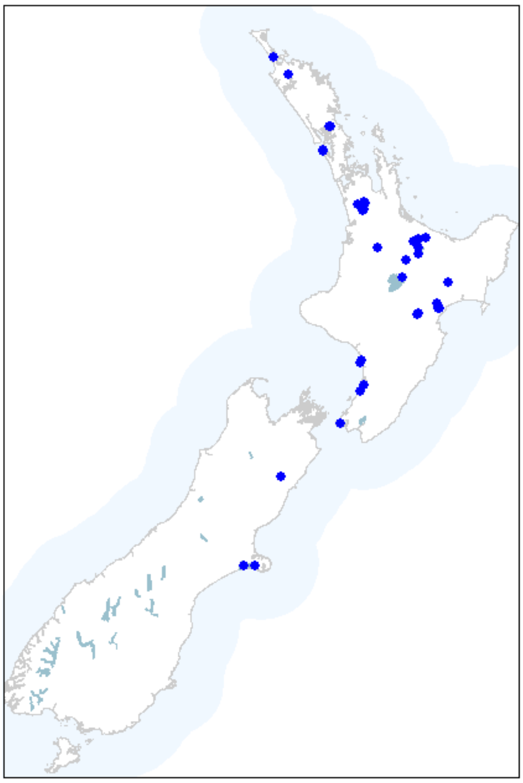

Figure 3. Locations of lake sites with corresponding cyanobacterial biovolume and water quality variables where n ≥ 7.

2.3 River predictor data

The Freshwater Ecosystems of New Zealand database (FENZ) provides a database of characteristics of the 550,000 segments of the digital river network that represents the streams and rivers of New Zealand (Wild *et al.*, 2005). The characteristics of the upstream catchment of all segments have been successfully used as predictors in spatial models of various water quality variables (Unwin *et al.*, 2010). For each of the river SOE sites we obtained the same suite of predictors used by Unwin *et al.* (2010) for use in the spatial models (Table 2).

Table 2. Predictor variables used in spatial models of river **Escherichia coli*****.***

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

*Geological variables are based on regolith, using averages of ordinal values assigned to LRI top-rock categories by (Leathwick *et al.*, 2003). The variables usHard and usPsize characterise physical regolith conditions; and usPhos characterises regolith fertility.

The digital river network is also the spatial framework for the River Environment Classification (REC; Snelder and Biggs, 2002). River monitoring sites were grouped into environmental classes to aid in summarising the results of analyses, and to account for some variation in water quality associated with environmental heterogeneity. River sites were grouped by REC Source-of-flow classes which are defined by the combination of climate and topography categories shown in Table 3.

Table 3. *Levels, categories, and notation for the river environment classification.*

Classification level	Defining characteristics	Categories	Notation	Category membership criteria
1	Climate (10^3 – 10^4 km^2)	Warm-extremely-wet	WX	Warm: mean annual temperature ≥ 12°C Cool: mean annual temperature < 12°C
		Warm-wet	WW	Extremely Wet: mean annual effective precipitation[1] ≥1500 mm
		Warm-dry	WD	
		Cool-extremely-wet	CX	Wet: mean annual effective precipitation > 500 and < 1500 mm
		Cool-wet	CW	
		Cool-dry	CD	Dry: mean annual effective precipitation ≤ 500 mm
2	Topography (10^2 – 10^3 km^2)	Glacial-mountain	GM	GM: M and % permanent ice > 1.5%
		Mountain	M	M: > 50% annual rainfall volume above 1000 m ASL
		Hill	H	H: 50% rainfall volume between 400 and 1000 m ASL
		Low-elevation	L	L: 50% rainfall below 400 m ASL
		Lake	Lk	Lk: Lake influence index[2] > 0.033

1. Effective precipitation = annual rainfall – annual potential evapotranspiration

2. See Snelder and Biggs (2002) for description.

2.4 Lake predictor data

The FENZ database provides characteristics of 3821 lakes greater than one hectare in area occurring across the North and South Islands, and some of the smaller outlying islands. Details of these variables and their derivation are provided by Snelder *et al.* (2006). Characteristics include descriptors of climatic, geological, topographic, bathymetric, land cover, and hydrological conditions in New Zealand lakes and their catchments.

The FENZ dataset includes estimates of average lake depth that were made using a geospatial statistical model (Snelder *et al.*, 2006). We also had measured maximum depth for all monitored lakes. We tested including maximum lake depth in our spatial models. However, because we used our models to make predictions for all lakes, we used the estimated average lake depth in our spatial models.

Table 4. Predictor variables used in the spatial models of lake water quality.

Predictor	**Abbreviation**	**Description**	**Unit**
Lake	lkArea	Lake surface area	m^2
	lkDistCoast	Straight line distance to coast	km
	lkDepth	Estimated average lake depth	m
	lkElev	Lake elevation	m ASL
Catchment topography	catSlope	Catchment average slope	Degrees
	catArea	Catchment area	m^2
	catElev	Catchment elevation	m ASL
Climate and flow	lkDecSolRad	Lake summer (December) solar radiation	W m^{-2}
	lkJuneSolRad	Lake winter (June) solar radiation	W m^{-2}
	lkDecTemp	Lake average summer (December) air temperature	Degrees
	lkJunTemp	Lake average winter (June) air temperature	Degrees
	lkFetch	Lake wind fetch	m
	lkSumWind	Lake summer (December) wind speed	m s^{-1}
	lkWinWind	Lake winter (June) wind speed	m s^{-1}
	catSumTemp	Catchment average summer (December) air temperature	Degrees
	catWinTemp	Catchment average winter (June) air temperature	Degrees
	catFlow	Catchment average annual discharge	m^3 yr^{-1}
Geology	catPhos	Catchment average phosphorous	Ordinal*
	catCalc	Catchment average calcium	Ordinal*
	catHard	Catchment average induration or hardness value	Ordinal*
	catPsize	Catchment average particle size	Ordinal*
	catPeat	Proportion of catchment occupied by peat	Proportion
	catAlluv	Proportion of catchment occupied by alluvium	Proportion
Land cover	catGlacial	Proportion of catchment occupied by permanent	Proportion
	catIndigForest	Proportion of catchment occupied by indigenous forest	Proportion
	catBare	Proportion of catchment occupied by bare ground	Proportion
	catExoticForest	Proportion of catchment occupied by exotic forest	Proportion
	catPastoral	Proportion of catchment occupied by pasture	Proportion

*Geological variables are based on regolith, using averages of ordinal values assigned to LRI top-rock categories by (Leathwick *et al.*, 2003). The variables catHard and catPsize characterise physical regolith conditions; and catPhos and catCalc characterises regolith fertility.

3 Methods

3.1 Analysis of water quality data

3.1.1 Rivers

For each site we first calculated four annual descriptor variables for each site; the median and 95th percentile *E.coli* concentrations and the percentage of samples that exceeded 260 and 540 *E.coli* 100 mL^{-1} (referred to as PercGT260 and PercGT540). These four descriptors of *E.coli* at the SOE sites were used as the response variables for the spatial models.

We assessed the relationship between flow and *E.coli* concentration and the timing of exceedance of *E.coli* thresholds at each site. Strong and consistent patterns of *E.coli* with flow or with the timing of exceedances would be important information to include in a strategic assessment of freshwaters from a human health perspective. Conversely weak and/or inconsistent patterns would indicate that annual statistics (i.e. the median and 95th percentile *E.coli* concentrations and PercGT260 and PercGT540) do not obscure strategically important temporal patterns in the indicator organism.

For each site we investigated relationships between *E.coli* and flow by fitting linear models to the $\log_{10}$-transformed *E.coli* values and their corresponding flow observations. We extracted the coefficients of determination (i.e. R^2 values) and the slope parameter from these models for each site and used these as measures of the strength of the relationship and its direction respectively. We plotted these data to visualise the relationships and investigated spatial patterns in these by parsing the results by REC Source-of-flow class.

For each site we assessed the timing of exceedance of *E.coli* thresholds by first identifying the month corresponding to each sample. We then counted the proportion of samples in each month that exceed concentration thresholds of 260 and 540 *E.coli* 100 mL^{-1}. We expressed the number of samples exceeding the threshold as a percentage of the total samples for each month. We plotted these data to visualise the seasonal pattern of exceedances and investigated spatial patterns in these by parsing the results by REC Source-of-flow class. We tested if month, within REC classes, explained the proportion of samples exceeding the threshold using Kruskal–Wallis tests. The Kruskal–Wallis test is a non-parametric test of whether the medians of data representing different groups are equal. In our analysis, significant results indicate that the median of site values of exceedances varies between months.

3.1.2 Lakes

Our analysis aimed to estimate cyanobacterial biovolume in lakes nationally. Our strategy to do this was to 1) use the SOE dataset and spatial modelling to predict chlorophyll *a*, TN, TP, Secchi depth, and TLI3 for all lakes nationally, and 2) to establish relationships between cyanobacterial biovolumes and their concurrent water quality parameters that could be used to transform the spatial model predictions into national estimates of cyanobacterial biovolume. To do this we required robust spatial model predictions (i.e. predictions from models that had satisfactory performance) and a robust model relating cyanobacterial biovolume to one or more of the predicted values (chlorophyll *a*, TN, TP, Secchi depth and TLI3). We did not know *a priori* which combination of SOE variables would satisfy these criteria and therefore proceeded with spatial models of all five SOE variables. For each lake and variable in the SOE dataset we calculated the median of the sampled values. These median values were used as the response variables for the spatial models.

For all cyanobacterial biovolume datasets, except Bay of Plenty (see below), we only maintained data points where concurrent cyanobacterial biovolume and water quality variable samples were available. In some instances, biovolume and water quality sample collection dates varied by up to one week. In these cases, biovolume values were matched with the closest available water quality data. When water quality parameters were below analytical detection limits, half the value was used.

Cyanobacterial biovolume and water quality data obtained from Bay of Plenty Regional Council were not collected at the same location (water quality – mid lake, and cyanobacterial biovolumes – at multiple sites around the lake edge) or date. For this dataset a mean water quality and cyanobacterial biovolume per lake was calculated and aligned for each month.

The median chlorophyll *a*, TN, TP, and Secchi depth and 80th percentile of cyanobacterial biovolumes were calculated for each lake. Relationships between log-transformed cyanobacterial biovolume + 0.001, and log-transformed median TP, TN, chlorophyll *a* and Secchi depth water quality variables were investigated using linear and multiple linear regressions. We checked that predictor variables were not collinear based on the variance inflation factor < 5 (Zuur *et al.*, 2010). Models were selected using a stepwise procedure based on the Akaike Information Criteria and were validated by inspecting residuals.

Predicted values of chlorophyll *a*, TN, TP, and Secchi depth made using spatial models were included as 'new data' in the cyanobacterial biovolume multiple linear regression equation and used to predict cyanobacterial biovolumes for all New Zealand lakes. Back-transformation (by exponentiation) was required to convert the predicted biovolume values back to the original units (mm^3 L^{-1}). When these values are back-transformed, the model error term no longer has a mean of zero. Ignoring this results in retransformation bias, i.e. predictions that systematically underestimate the response. We corrected the retransformation bias using the smearing estimate (*S*; Duan, 1983):

$$S = \frac{1}{n}\sum_{i=1}^{n} e^{\widehat{\varepsilon_i}} \qquad \text{(Equation 1),}$$

where $\hat{\varepsilon}$ are the residuals of the cyanobacterial biovolume model. The predictions were back-transformed by exponentiation, then corrected for retransformation bias by multiplying by *S.*

3.2 Spatial modelling

3.2.1 Random forest models

We defined spatial models for the four descriptors of the river SOE sites (median, and 95th percentile *E.coli* concentrations and percentage of samples exceeding 540 and 260 *E.coli* 100 mL^{-1}) and the median values of the five water quality variables for the lake SOE sites. Models were fitted to each variable as a function of predictor variables using Random Forest (RF) models (Breiman, 2001; Cutler *et al.*, 2007).

An RF model is an ensemble of individual classification and regression trees (CART). In a regression context, CART partitions observations (in this case the individual water quality variables) into groups that minimise the sum of squares of the response (i.e. assembles groups that minimise differences between observations) based on a series of binary rules or splits that are constructed from the predictor variables. CART models have several desirable features including requiring no distributional assumptions and the ability to automatically fit non-linear relationships and high order interactions. However, single regression trees have the limitations of not searching for optimal tree structures, and of being sensitive to small changes in input data (Hastie *et al.*, 2001). RF models reduce these limitations by using an ensemble of trees (a forest) and making predictions based on the average of all trees (Breiman, 2001). An important feature of RF models is that each tree is grown with a bootstrap sample of the fitting data (i.e. the observation dataset). In addition, a random subset of the predictor variables is made available at each node to define the split. Introducing these random components and then averaging over the forest increases prediction accuracy while retaining the desirable features of CART.

An RF model produces a limiting value of the generalization error (i.e. the model maximises its prediction accuracy for previously unseen data; Breiman, 2001). The generalization error converges asymptotically as the number of trees increases, so the model cannot be over-fitted. The number of trees needs to be set high enough to ensure an appropriate level of convergence, and this value depends on the number of variables that can be used at each split. We used default options that included making one third of the total number of predictor variables available for each split, and 500 trees per forest. Some studies report that model performance is improved by including more than ~ 50 trees per forest, but that there is little improvement associated with increasing the number of trees beyond 500 (Cutler *et al.*, 2007). Our models took less than a minute to fit when using the default of 500 trees per forest.

Unlike linear models, RF models cannot be expressed as equations. However, the relationships between predictor and response variables represented by RF models can be represented by importance measures and partial dependence plots (Breiman, 2001; Cutler *et al.*, 2007). During the fitting process, RF model predictions are made for each tree for observations that were excluded from the bootstrap sample; these excluded observations are known as out-of-bag (OOB) observations. To assess the importance of a specific predictor variable, the values of the response variable are randomly permuted for the OOB observations, and predictions are obtained from the tree for these modified data. The importance of the predictor variable is indicated by the degree to which prediction accuracy decreases when the response variable is randomly permuted. Importance is defined in this study as the loss in model performance (i.e. the increase in the mean square error; MSE) when predictions are made based on the permuted OOB observations compared to those based on the original observations. The differences in MSE between trees fitted with the original and permuted observations are averaged over all trees, and normalized by the standard deviation of the differences (Cutler *et al.*, 2007).

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

RF models include any of the original set of predictor variables that are chosen during the model fitting process. However, marginally important predictor variables may be redundant (i.e. their removal does not affect model performance) and their inclusion complicates model interpretation. We used a backward elimination procedure to remove redundant predictors from the initial ‘saturated’ models (i.e. models that included any of the original predictor variables). The procedure first assesses the model mean square error (MSE) using a 10-fold cross validation process. The predictions made to the hold out observations during cross validation are used to estimate the MSE and its standard error. The model’s least important predictor variables are then removed in order, with the MSE and its standard error being assessed for each successive model. The final, ‘reduced’ model is defined by the “one standard error rule” as the model with the fewest predictor variables whose error is within one standard error of the best model (i.e. the model with the lowest cross validated MSE) (Breiman *et al.*, 1984). Importance levels for predictor variables were not recalculated at each reduction step to avoid over-fitting (Svetnik *et al.*, 2004).

Although RF models do not depend on distributional assumptions, transformation of the response variable to an approximately symmetric distribution can improve model performance. We investigated transformations of the modelled water quality (i.e. response) variables on the model performance. Where performance was improved we made predictions using these models.

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

3.2.2 Model performance

Model performance was assessed by comparing observations with independent predictions (i.e. sites that were not used in fitting the model), which were obtained from the OOB observations. We summarised the model performance using five statistics; regression R^2, Nash-Sutcliffe efficiency (NSE), percent bias (PBIAS) and the relative root mean square error (RSR) and the root mean square deviation (RMSD). The regression R^2 value is the coefficient of determination derived from a regression of the observations against the predictions. The R^2 value indicates the proportion of the total variance explained by the model but is not a complete description of model performance (Piñeiro *et al.*, 2008). NSE indicates how closely the observations coincide with predictions (Nash and Sutcliffe, 1970). NSE values range from $-\infty$ to 1. An NSE of 1 corresponds to a perfect match between predictions and the observations. An NSE of 0 indicates the model is only as accurate as the mean of the observed data and values less than 0 indicate the model predictions are less accurate than using the mean of the observed data. Bias measures the average tendency of the predicted values to be larger or smaller than the observed values. Optimal bias is zero, positive values indicate underestimation bias and negative values indicate overestimation bias (Piñeiro *et al.*, 2008). PBIAS is computed as the sum of the differences between the observations and predictions divided by the sum of the observations (Moriasi *et al.*, 2007). RSR is calculated as the ratio of the root mean square error to the standard deviation of the observations a measure of the characteristic model uncertainty and is estimated as the mean deviation of predicted values with respect to the observed values divided by the standard deviation of the observations (Moriasi *et al.*, 2007). The normalization associated with PBIAS and RSR allowed the performance of models to be compared across all the modelled water quality variables. A rule of thumb is that model predictions are satisfactory if NSE > 0.50, RSR < 0.70, and if PBIAS < ±25% and are good if 0.65 < NSE > 0.75, 0.5 < RSR < 0.60, and if 25% < PBIAS < ±40% (Moriasi *et al.*, 2007).

The RMSD is a measure of the characteristic model statistical error or uncertainty. RMSD is mean deviation of predicted values with respect to the observed values (distinct from the standard error of the regression model).

3.2.3 Representativeness of monitoring sites used in RF models

A graphical comparison was used to gauge how well the monitoring sites used to fit the RF models represented environmental variation at the national scale. Here, representativeness refers to the degree to which the distribution of monitoring sites over the range of an environmental predictor variable matches the distribution of all network segments over the range of the same environmental variable. Poor representativeness can reduce the reliability of the model predictions because certain sets of environmental conditions are not represented in the fitting data.

Histograms of the proportions of monitoring site numbers over the ranges of the most important predictor variables in the RF models (i.e. the predictors with the greatest explanatory power) were visually compared with histograms of the proportions of all network segments over the same predictor variables. Note that representativeness of monitoring sites is different from model bias, which is defined in Section 3.2.2.

3.2.4 Model predictions

Predictions are made with RF models by “running” new cases down every tree in the fitted forest and averaging the predictions made by each tree (Cutler *et al.*, 2007). Some of the models in this study were fitted to $\log_{10}$-transformed data and when the model predictions were back-transformed, we corrected for retransformation bias using the smearing estimate (Duan, 1983; Equation 1, but using base 10, not base *e*). The back-transformed predictions were used to produce national maps depicting the variation in each water quality variable and the predictions were exported so that they could be displayed in a geographic information system (GIS).

4 Results - Rivers

4.1 *Escherichia coli* data

4.1.1 Site *E.coli* variables

The values of *E.coli* variables differed markedly between sites (Figure 4). The values were generally high in low-elevation REC Source-of-flow classes (CD/L, CW/L, CX/L, WD/L, WW/L,WX/L), were low in the mountain Source-of-flow classes (CD/M, CW/M, CX/GM, CX/M) and were intermediate in the hill Source-of-flow classes (CD/H, CW/H, CX/H, WW/H, WX/H).

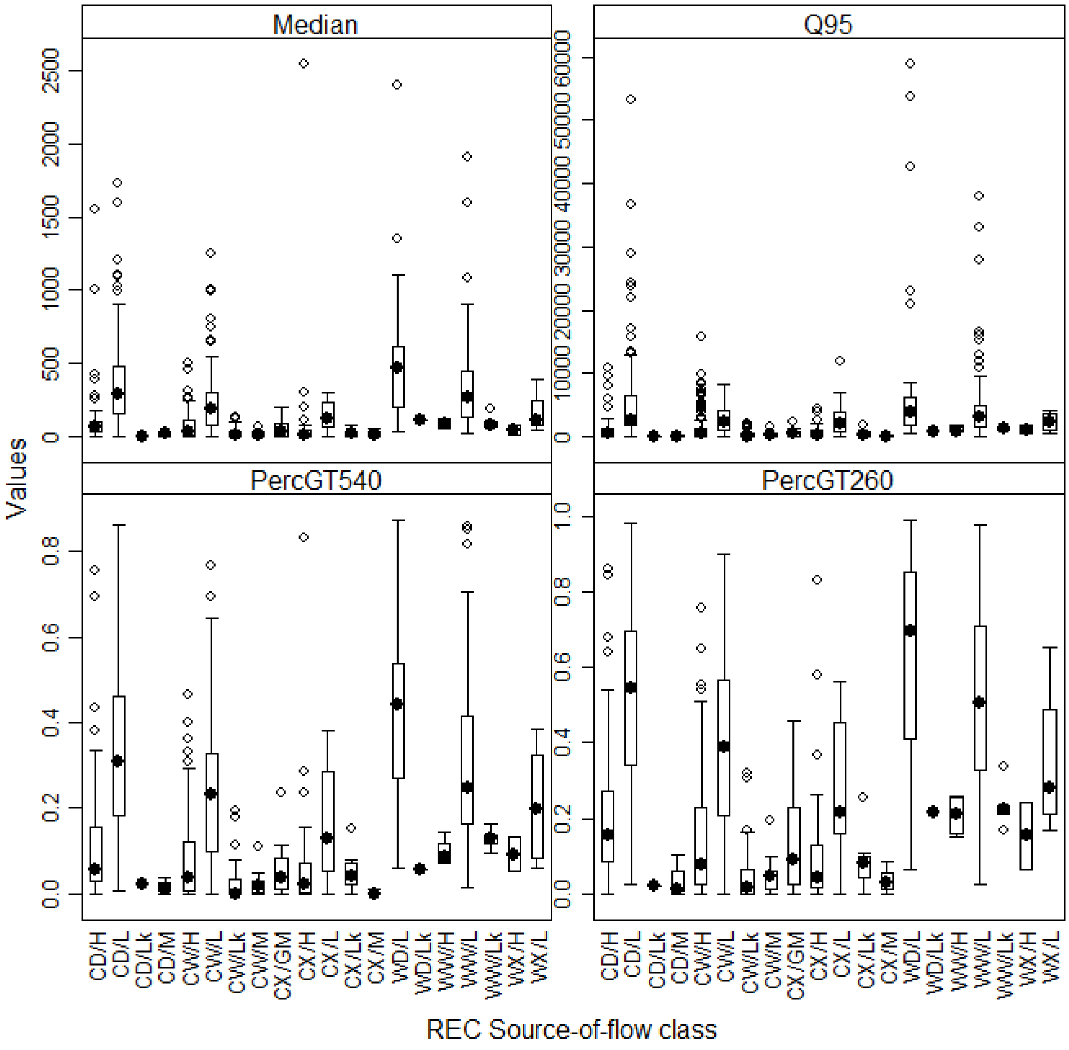

Figure 4. Distributions of the site values of the* Escherichia coli *variables. *The plots show for each REC Source-of-flow class, the Median and 95th percentile* E.coli *concentrations (Median and Q95) and the percentage of samples that exceeded 260 and 540* E.coli *100 mL^{-1} (PercGT260 and PercGT540). The top and bottom lines of the rectangle represent the 3rd and 1st quartiles and the solid dot represents the median. The whiskers extend to the 3rd and 1st quartiles plus and minus 1.5 times the interquartile range. The open circles represent data beyond 1.5 times the interquartile range. REC classes are defined in Table 3.*

4.1.2 Relationships with flow

Relationships between *E.coli* concentrations and flow were generally weak (i.e. R^2 values were low) (Figure 5). Only 6% of sites had R^2 values greater than 0.3 and 69% of sites had R^2 values less than 0.1. The direction of the relationships between *E.coli* concentrations and flow were variable (positive or negative; Figure 5). The slope of the *E.coli* concentration - flow relationship was positive at 71% of sites and negative for the remaining 29%. No obvious spatial patterns in the *E.coli* concentrations and flow relationships were revealed by the REC Source-of-flow classification (Figure 5).

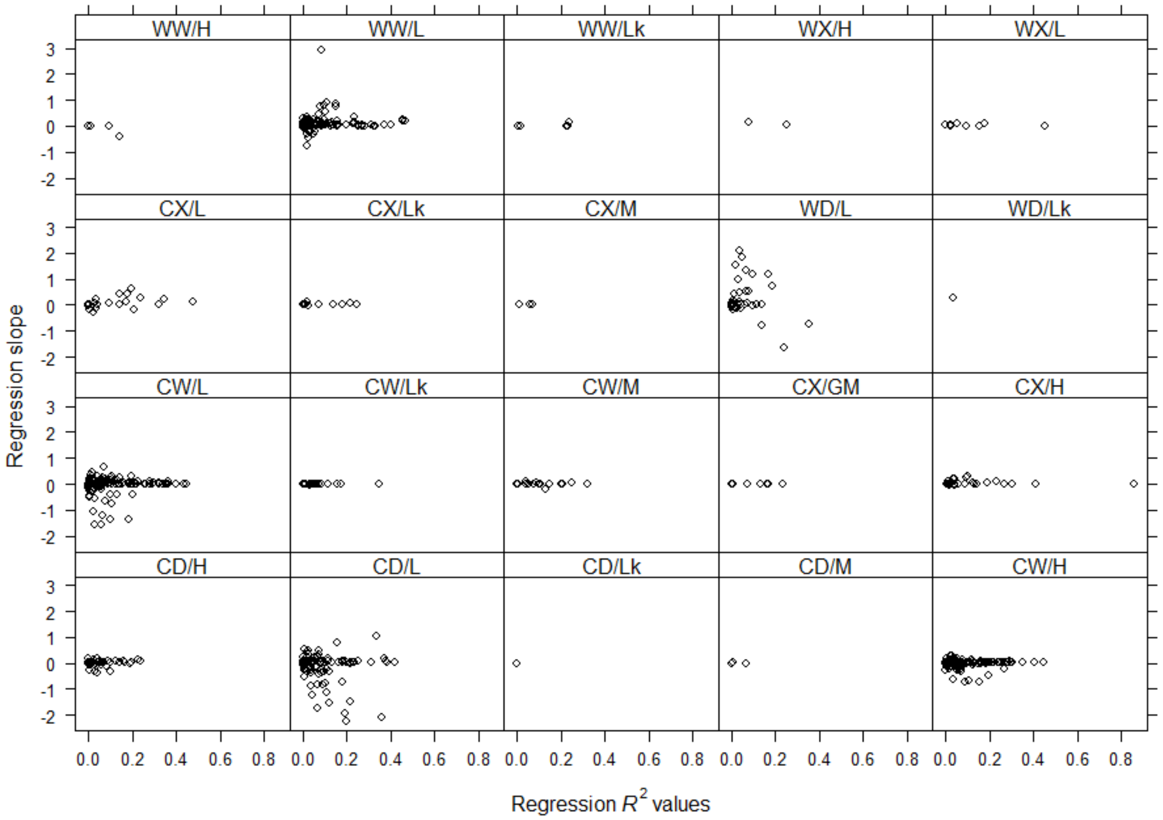

Figure 5. Relationships between* Escherichia coli *and flow. *The plots show the R^2 and slope values for regression between E. coli and flow for each REC Source-of-flow class. REC classes are defined in Table 3.*

4.1.3 Timing of exceedances of the *E.coli* threshold value

The analysis of timing of exceedances of the *E.coli* thresholds indicated that exceedances were generally higher in low-elevation REC Source-of-flow classes (as also shown in Figure 4). The Kruskal–Wallis tests indicated that the proportion of exceedances differed significantly by month in some REC classes (Figure 6, Table 5). In those classes with significant differences in the proportion of exceedances by month, there was a seasonal trend. Exceedances tended to be a highest in the autumn (March-May) followed by spring (Oct -Nov) (Figure 6). Although there were seasonal trends associated with the median values of site exceedance in class, there was also considerable variation between sites within classes. For example, in many classes there were sites in each month that had either no samples exceeding the threshold or 100% exceeding. The patterns associated with a threshold of 540 *E.coli* 100 mL^{-1} were similar but less pronounced (data not shown).

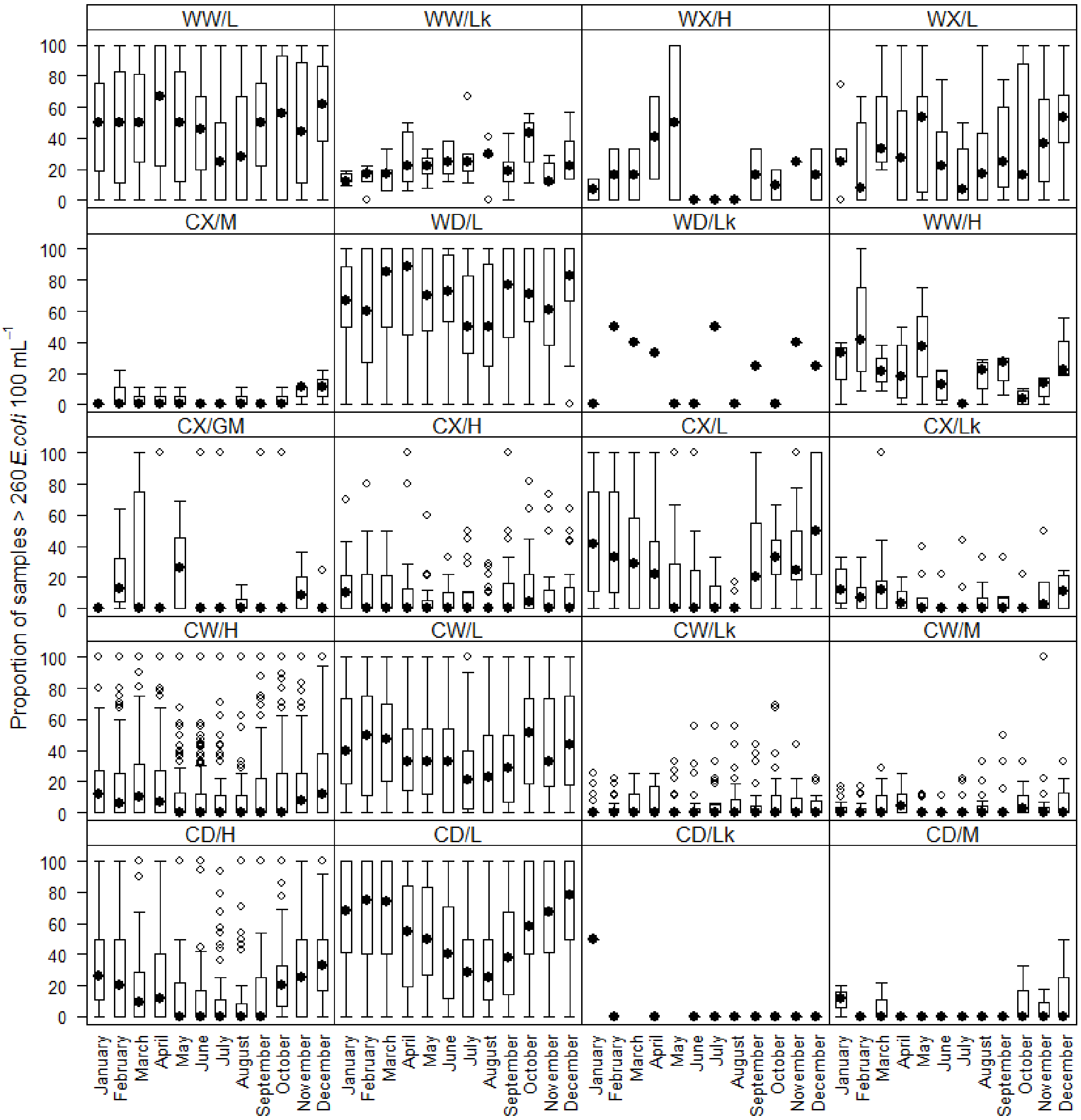

***Figure 6. Results of analysis of timing of exceedances of the* Escherichia coli *threshold value.** The plots show the distributions of the site proportions of samples that exceed 260* E. coli *100 mL^{-1} in each month. The data have been plotted separately for each REC Source-of-flow class. The top and bottom lines of the rectangle represent the 3rd and 1st quartiles and the solid dot represents the median. The whiskers extend to the 3rd and 1st quartiles plus and minus 1.5 times the interquartile range. The open circles represent data beyond 1.5 times the interquartile range. REC classes are defined in Table 3.*

Table 5. Results of Kruskal–Wallis tests of timing of exceedances of the* Escherichia coli *threshold of 260 E. coli 100 mL^{-1}. *Bold P-values indicate the tests that were significant at the 5% level. REC classes are defined in Table 3.*

REC Source–of-flow class	**Number of sites**	**Kruskal–Wallis statistic**	***P* value**
CD/H	53	84	**<0.0001**
CD/L	107	180	**<0.0001**
CD/Lk	1	9	0.437
CD/M	3	11	0.450
CW/H	155	74	**<0.0001**
CW/L	122	74	**<0.0001**
CW/Lk	23	6	0.884
CW/M	16	11	0.458
CX/GM	8	16	0.133
CX/H	28	14	0.237
CX/L	25	69	**<0.0001**
CX/Lk	10	20	**0.050**
CX/M	3	9	0.639
WD/L	39	15	0.171
WD/Lk	1	11	0.443
WW/H	4	19	0.065
WW/L	140	82	**<0.0001**
WW/Lk	5	11	0.431
WX/H	2	8	0.684
WX/L	8	9	0.623

4.2 *Escherichia coli* spatial models

4.2.1 Model performance

We used a log$_{10}$-transformation of median and Q95 to make the distributions of these variables more symmetric and improved model performance. We tried logit transforming the PercGT260 and PercGT540, which had values between zero and one. While the logit transformation produced more symmetric distributions for these variables, it did not improve model performance and we left these variables untransformed.

The *E.coli* RF models of median, Q95 and PercGT260 had generally good performance as indicated by the following statistics: R^2 > 0.65, NSE > 0.65, RSR < 0.60 (Table 6 and Figure 7; Moriasi *et al.*, 2007). The *E.coli* RF models of PercGT540 had satisfactory performance with a slightly higher RSR and lower NSE than the other three models (Table 6; Moriasi *et al.*, 2007). All four models had very low bias (PBIAS; Table 6).

Table 6. Performance of the Escherichia coli spatial models. *Performance was determined using independent predictions (i.e. sites that were not used in fitting the models) generated from the out-of-bag observations. Regression R^2 = coefficient of determination of observation versus predictions, NSE = Nash-Sutcliffe efficiency, PBIAS = percent bias, RSR = relative root mean square error, RMSD = root mean square deviation. Units for RMSD for Median and Q95 are the log$_{10}$-transformed units of the respective water quality variables.*

Model (*E.coli* variable)	**Regression *R*^{2}**	**NSE**	**PBIAS**	**RSR**	**RMSD**
Median	0.73	0.72	-0.40	0.53	0.35
Q95	0.70	0.69	-0.09	0.55	0.38
PercGT260	0.67	0.67	-1.02	0.58	0.16
PercGT540	0.58	0.58	-2.07	0.65	0.13

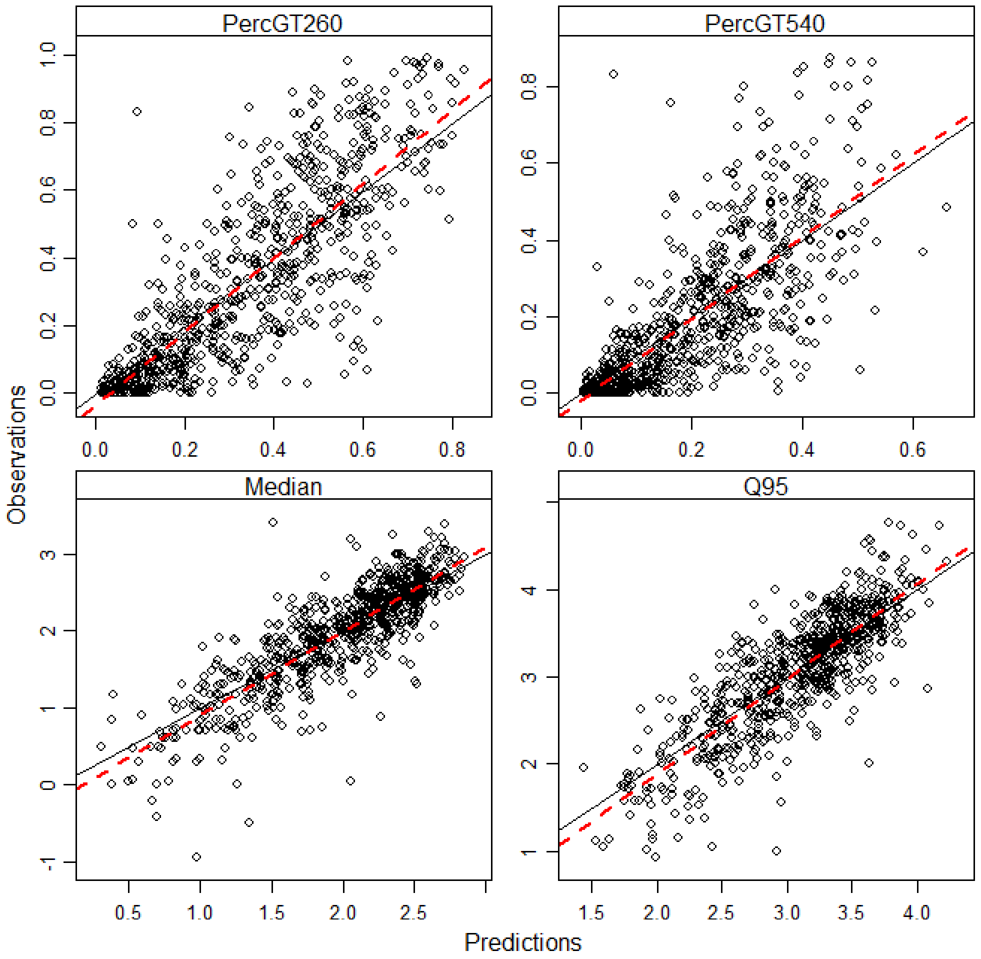

Figure 7. Comparison of observed* Escherichia coli *descriptor variable versus values predicted by the Random Forest models*. Note that the observed values are plotted on the Y-axis and predicted values on the X-axis, following Piñeiro* et al. *(2008). Red dashed line: best fit linear regression of the observed and predicted values. The solid black line is one-to-one. Units for the variables Median and Q95 are the log$_{10}$-transformed units and for PercGT260 and PercGT540 are non-transformed values.*

4.2.2 Modelled relationships

A large number of predictor variables were retained in the reduced *E.coli* models (Table 7). The models of PercGT260 and median retained all 24 predictors and the PercGT540 and Q95 models retained 17 predictors (Table 7).

The retained predictor variables with high importance in all RF models reflected associations between *E.coli* and catchment elevation and slope (usCatElev, usAveSlope), the proportion of the catchment occupied by high producing exotic grassland and bare ground (usPastoral and usBare) and climatic variables (usAveTCold, usAveTWarn and usAnnRainVar) (Figure 8). These relationships were consistent with expectations. For example, the values of all four *E.coli* response variables increased with increasing pastoral land cover and decreased with increasing catchment elevation and slope (Figure 8). The association of the *E.coli* variables with pastoral land cover and elevation are consistent with recent evaluations of environmental patterns in river water quality (e.g., Larned *et al.*, 2016, 2004; Unwin *et al.*, 2010).

Some relationships may represent correlations rather than causative processes. For example, there were strong relationships between *E.coli* and annual variability in rainfall (usAnRainVar). The usAnRainVar predictor strongly differentiates between the western (low values) and eastern (high values) aspects of New Zealand. The modelled relationships between *E. coli* and usAnRainVar may reflect differences in the temporal distribution of rainfall, or some other climatic factor, that are correlated with this pattern, rather than annual variability in rainfall *per se*.

***Table 7. Predictors retained by the reduced Random Forest models* of Escherichia coli.** *The values indicate the rank importance of the predictor for the individual models. NA indicates that the predictor was not included in the reduced model. Predictor variables are defined in Table 2.*

Predictor	PercGT540	PercGT260	Median	Q95
usCatElev	1	1	1	2
usPastoral	2	2	2	1
usAveSlope	3	3	3	3
usBare	NA	7	7	NA
usAnRainVar	4	4	5	4
usAvTWarm	6	5	6	8
usAvTCold	5	6	9	6
segAveElev	9	9	4	5
usParticleSize	NA	10	13	16
usHard	8	15	10	7
usIndigForest	10	8	12	11
usLake	NA	23	8	10
usScrub	NA	21	22	NA
usFlow	11	18	15	NA
usWetland	NA	22	23	NA
usArea	12	17	18	NA
usGlacial	NA	24	24	NA
usPhos	13	13	11	15
segAveTCold	NA	20	19	13
usUrban	7	12	17	17
usRainDays20	15	11	16	12
usExoticForest	16	16	14	9
usRainDays100	17	19	21	NA
usRainDays10	14	14	20	14

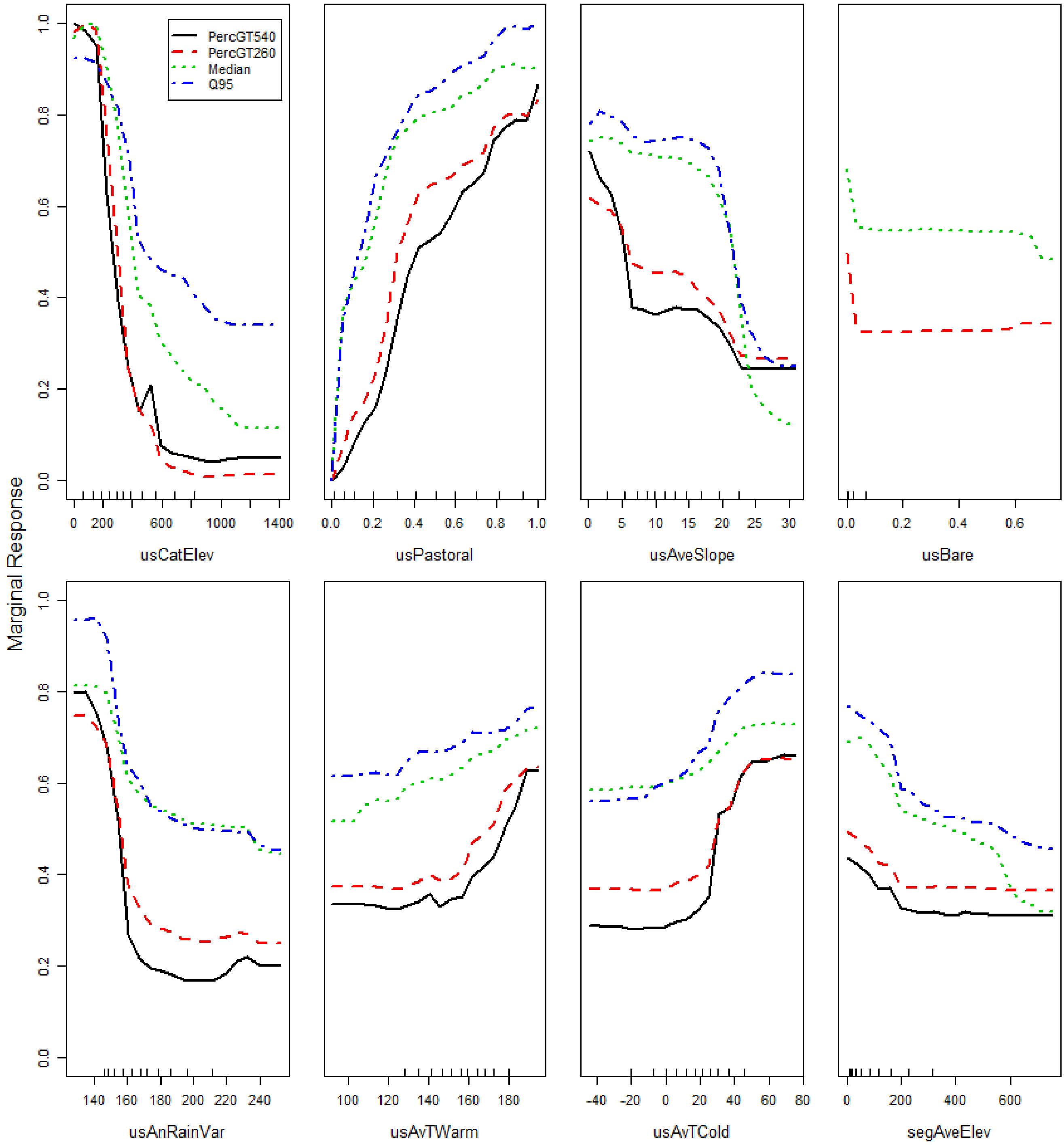

Figure 8. Partial plots for the eight most important predictor variables in Random Forest models of the* Escherichia coli *variables. *Each panel corresponds to one predictor. The Y-axis is the standardised value of the marginal response for each of the eight modelled variables. In each case, the original marginal responses over all eight predictors were standardised to have a range between zero and one. Plot amplitude (the range of the marginal response on the Y-axis) is directly related to a predictor variable's importance; amplitude is large for predictor variables with high importance. Legend in top left panel applies to all panels. Predictor variable are defined in Table 2.*

4.2.3 Representativeness of *E.coli* monitoring sites

The distributions of *E.coli* monitoring sites across the environmental gradients defined by the 12 most important predictor variables were generally consistent with the distribution of all segments in the river network across the same gradients (Figure 9). The predictor variables shown in the histograms in Figure 9 were those that were important in the RF models.

For some environmental gradients, there was moderate over- and under-representation of monitoring sites compared to the river network. *E.coli* monitoring sites were over-represented in environments characterised by low segment elevations (segAveElev), low catchment elevations (usCatElev) and low catchment slopes (usAveSlope) (Figure 9). *E.coli* monitoring sites were under-represented in catchments with very high proportions of indigenous forest land cover (usIndigForest), and catchments with low proportions of intensive agricultural land cover (usPastoral). There was also under-representation of sites with low catchment average summer air temperature (usAvTWarm),low values of catchment average variation of annual rainfall (usAnRainVar) and sites in high altitude catchments (usCatElev). The under-represented parts of these three predictor gradients reflect locations on the southern parts of the West Coast of the South Island in particular (Figure 1).

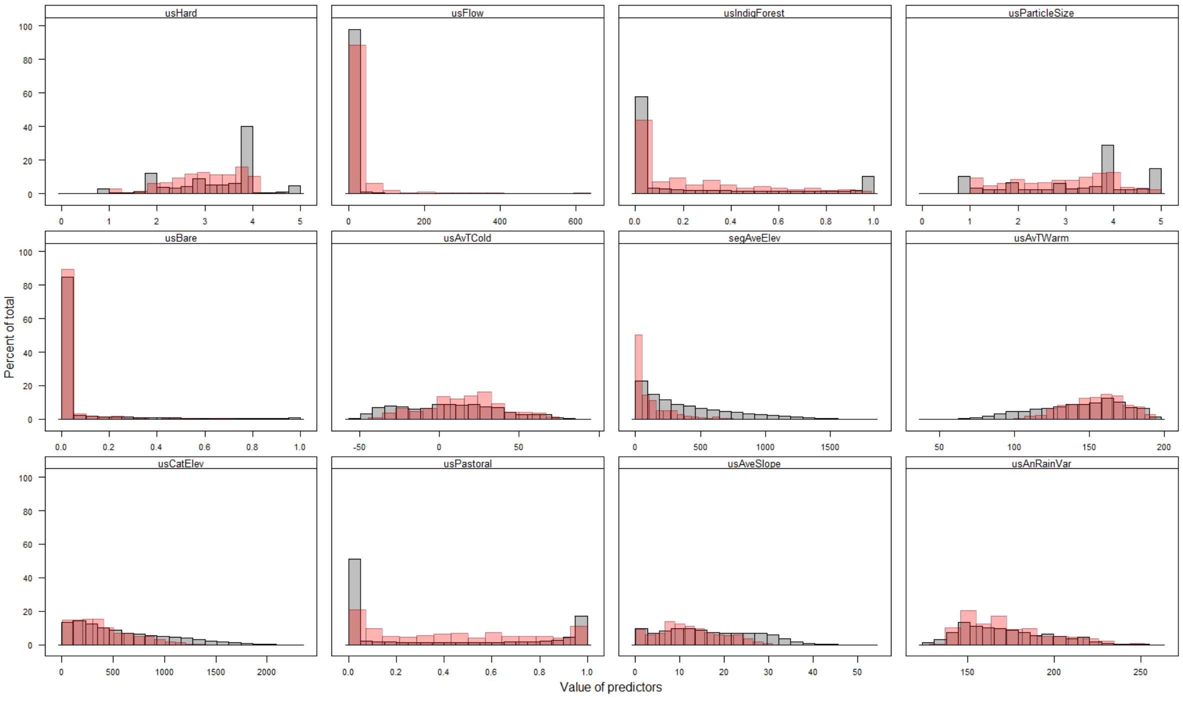

Figure 9. Histograms comparing the distributions of predictor variable values for all river segments nationally and the* Escherichia coli *monitoring sites. *The national pool of river segments is represented by the grey histograms and the monitoring sites used for Random Forest (RF) models are represented by the red histograms. Similarities in the distributions shown in the two histograms in each panel provide an indication of the degree to which environmental variation across the monitoring sites represents environmental variation across all river segments in New Zealand; complete representativeness would be indicated by exact matches between the histograms. The 12 predictor variables shown in the figure were the most important overall predictors in the RF models. Predictor variable are defined in Table 2.*

4.2.4 Model predictions

The mapped predictions for all four *E.coli* variables (Median, Q95, PercGT260, PercGT540) had similar spatial patterns, with high values in low-elevation areas on the east coasts of the North and South Island, and in the inland Waikato, Ruamahunga Valley, Rangitikei-Manawatu coastal plain, Taranaki Ring Plain, and Auckland Region (Figure 10). Predicted values for all four *E.coli* variables were generally low in major mountain ranges (e.g. Southern Alps, Kahurangi, Kaimanawa, and Tararua Ranges), in large areas of the Department of Conservation estate (e.g. Fiordland, Westland, Te Urewera, Egmont, Whanganui and Tongariro National Parks), and in smaller, native forest-dominated areas of Northland and the Coromandel Peninsula.

The low elevation areas characterised by high values of the *E.coli* variables coincide with land used for intensive agriculture and with most of New Zealand’s urban centres. High-intensity agricultural and urban land currently account for 60% of the land area below 350 m elevation (Larned *et al.*, 2016). Within these areas, there are some finer scaled differences in predicted values of all four *E.coli* variables. For example, the Canterbury Plains were characterised by slightly lower Q95 concentrations, than similarly intensively farmed regions such as Southland and the Waikato-Hauraki Plains area (Figure 10). These differences reflect regional variation in the important predictor variables that were included in the models other than usPastoral.

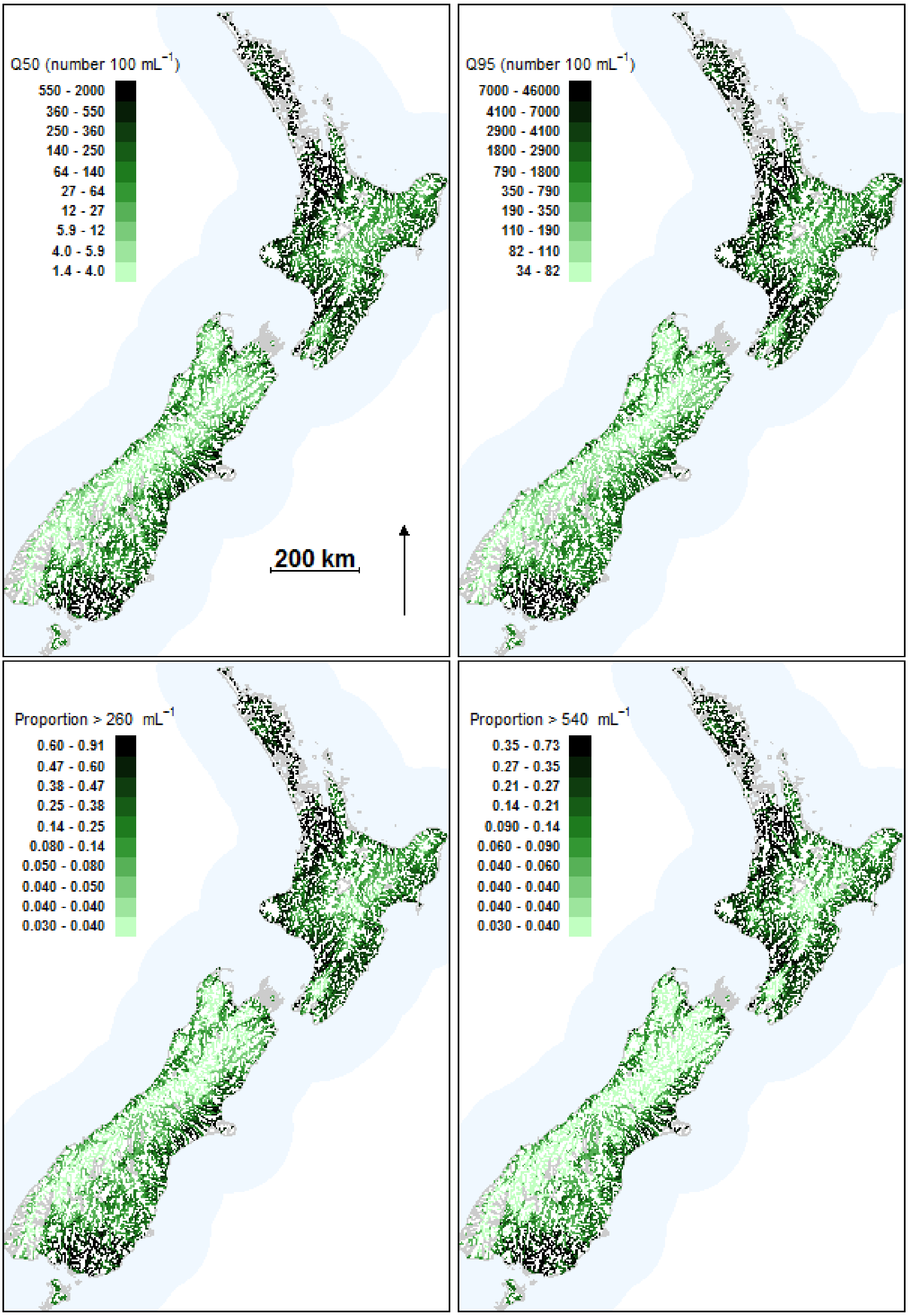

Figure 10. Predicted **Escherichia coli** ***response variables in New Zealand rivers***. *Order 1 to 3 rivers have been omitted to make river networks distinguishable but predictions for all network segments have been made. Note that concentration scales vary between each map. Q50 = median.*

5 Results - Lakes

5.1 Lake data

The median values of the lake SOE data are described by Larned *et al.* (2015).

A significant positive relationship was observed among cyanobacterial biovolume and chlorophyll *a* (R^2 = 0.60, P < 0.001), TP (R^2 = 0.46, P < 0.001), and TN (R^2 = 0.41, P < 0.001) (Figure 11a, b, d) and a significant negative relationship with Secchi depth (R^2 = 0. 23, P < 0.01) (Figure 11c).

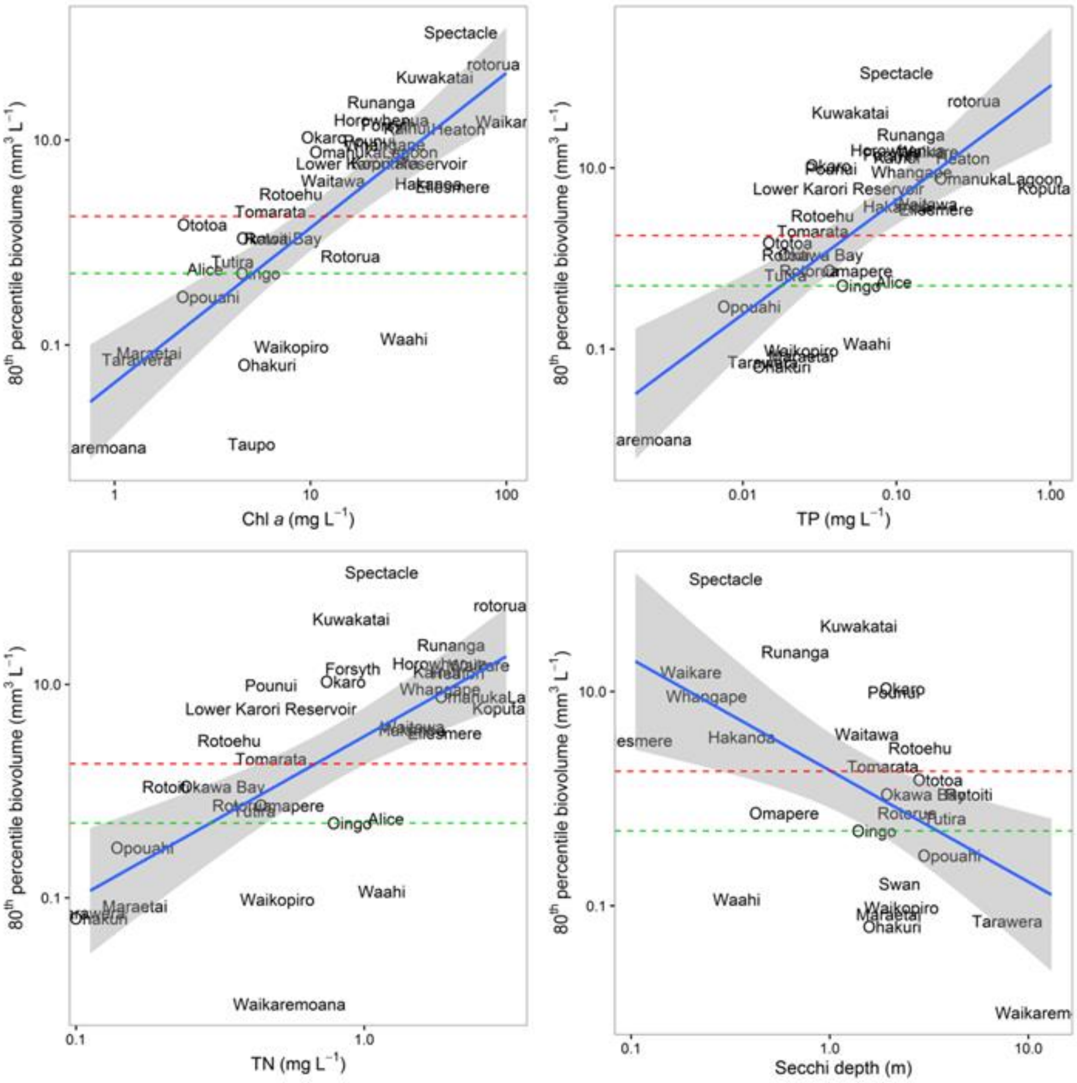

Figure 11. Individual relationships between total cyanobacterial biovolumes (80th percentile) explanatory variables in 37 New Zealand lakes. The explanatory variables are median (a) chlorophyll a (Chl a), (b) total phosphorus (TP), (c) total nitrogen (TN), and (d) Secchi disk. The blue line is a linear regression, and grey shading represents pointwise 95% confidence interval of the fitted values. The green and red horizontal lines indicate the NPS-FM thresholds of 0.5 (band C above line) and 1.8 (band D above line) mm^3 L^{-1}.

The most parsimonious multiple regression model included chlorophyll *a*, TP, and Secchi depth, and had a R^2 of 0.64 (P < 0.001, Equation 2).

$$\begin{aligned}\log\ (\text{cyanobacterial volume} + 0.001) &\qquad \text{(Equation 2)}\\ = 0.964 + 1.467 * \log\ (\text{chlorophyll } a) + 1.389 * \log\ (\text{TP})\\ + 1.290 * \log\ (\text{Secchi depth})\end{aligned}$$

5.2 Lake spatial models

5.2.1 Model performance

The performance of all models were improved by $\log_{10}$-transformation of the site median values of the water quality variables (the model responses). The raw variable distributions were strongly right skewed and the transformations made these more symmetric.

The RF models of CHLA, SECCHI, TN, TP and TLI3 had satisfactory to good performance as indicated by the following statistics: NSE > 0.5, RSR < 0.7 (Table 8, Figure 12; Moriasi *et al.*, 2007). All five of these models had very low bias (PBIAS; Table 8, Figure 12).

Table 8. Performance of the lake water quality models. *Performance was determined using independent predictions (i.e. sites that were not used in fitting the models) generated from the out-of-bag observations. R^2 = coefficient of determination of observation versus predictions, NSE = Nash-Sutcliffe efficiency, PBIAS = percent bias, RSR = relative root mean square error, RMSD = root mean square deviation. RMSD units are the $\log_{10}$-transformed original units.*

Model	N	R^2	NSE	PBIAS	RSR	RMSD
CHLA	92	0.57	0.56	-1.11	0.66	0.37
SECCHI	64	0.61	0.60	-2.19	0.63	0.30
TLI3	76	0.63	0.61	0.77	0.62	0.09
TN	82	0.68	0.66	0.22	0.58	0.25
TP	99	0.55	0.55	1.22	0.67	0.33

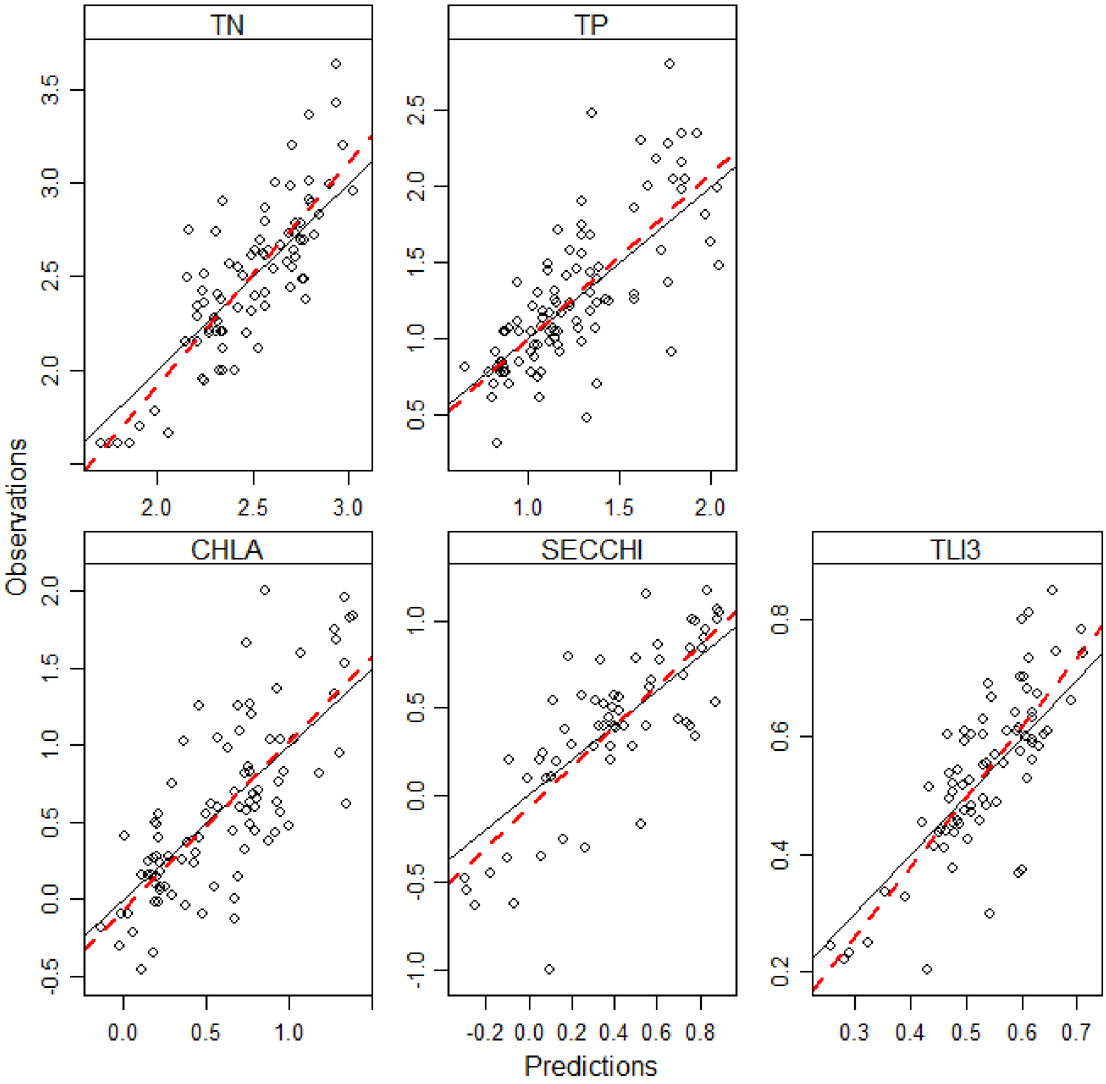

Figure 12. Comparison of observed water quality versus values predicted by the Random Forest models. *Note that the observed values are plotted on the Y-axis and predicted values on the X-axis, following Piñeiro et al. (2008). Red dashed line: best fit linear regression of the observed and predicted values. The solid black line is one-to-one. Units for the variables are the log$_{10}$ of the original units.*

5.2.2 Modelled relationships

The reduced RF models retained only a small subset of the original set of predictors (Table 9). The eleven retained predictors reflected associations between water quality and lake and catchment elevation, geological and climatic factors (Table 9).

The lake water quality variables had logical relationships with many of the individual predictor variables included in the reduced RF models (Figure 13). Nutrient concentrations and chlorophyll *a* decreased and Secchi depth increased with increasing lake and catchment elevation (lkElev, catElev) and decreasing wintertime catchment air temperature (catWinTemp). This is consistent with an observed gradient in trophic conditions for lakes that is associated with altitude and climate (Sorrell *et al.*, 2006). Predictors describing catchment land cover were not retained in any of the RF models that had satisfactory performance (Figure 13). Some relationships may represent correlations rather than causative processes. For example, the strong relationship between elevation and catchment climate probably indicates that the inclusion of these predictors is partly because they are correlated with increased nutrient supply from agriculture. TLI3 and TN decreased with lake fetch (lkFetch), which may be a reflection of the generally lower trophic status of larger lakes rather the effect of wind mixing on lakes.

Table 9. Predictors retained by the reduced Random Forest models of lake water quality variables. *The values indicate the rank importance of the predictor for the individual models. NA indicates that the predictor was not included in the reduced model. Predictor variables are defined in Table 4.*

Predictor	CHLA	SECCHI	TLI3	TN	TP
lkElev	1	1	2	1	1
catWinTemp	2	2	3	2	4
catElev	3	3	1	3	3
catCalc	NA	NA	NA	NA	2
lkFetch	NA	NA	4	4	NA
lkDistCoast	4	NA	6	NA	8
catSlope	NA	NA	5	NA	6
lkSumWind	5	NA	NA	NA	7
catAlluv	NA	NA	8	5	NA
catPhos	NA	NA	9	NA	5
lkDecSolRad	NA	NA	7	NA	9

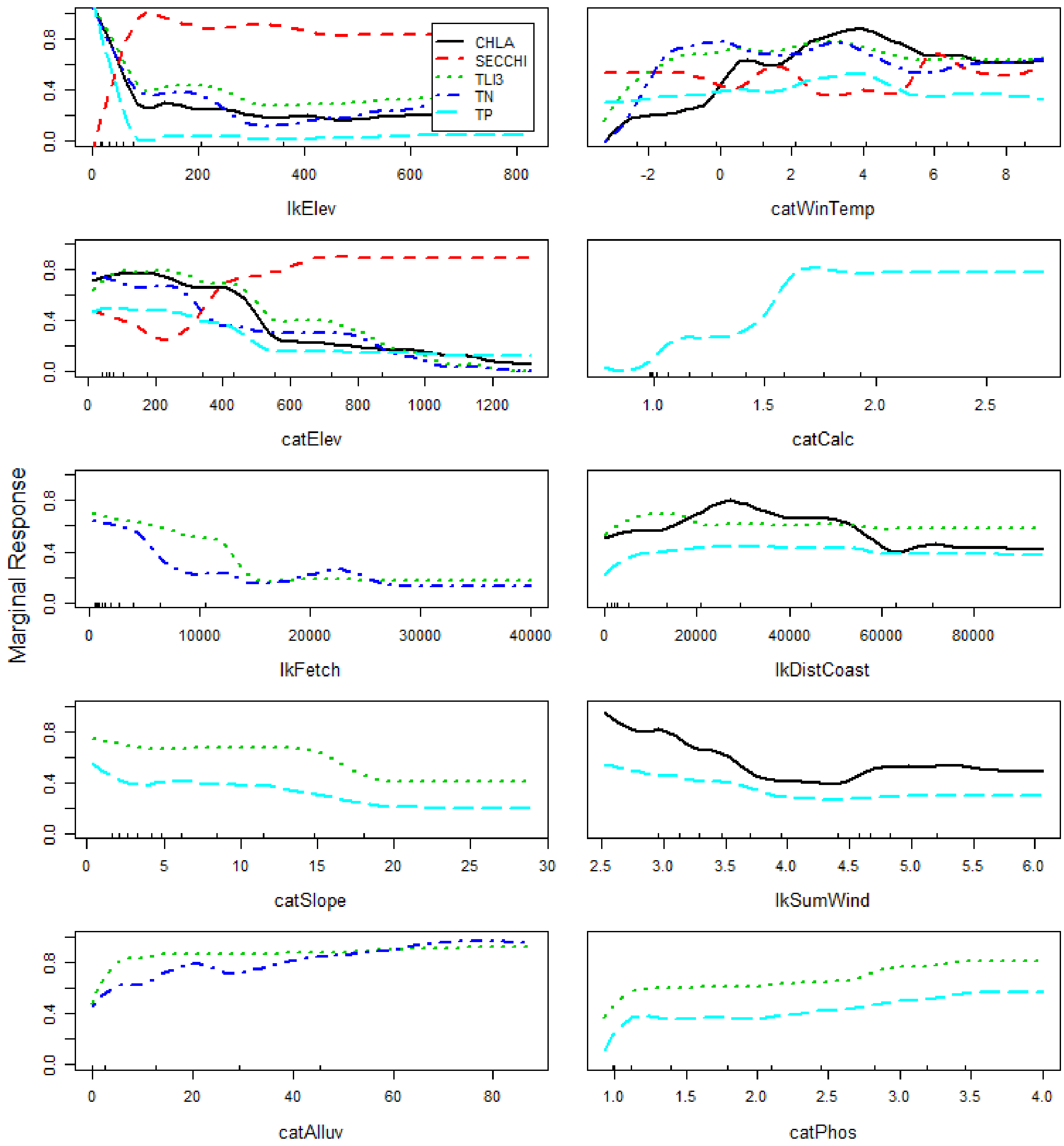

Figure 13. Partial plots for the ten most important predictor variables in the Random Forest models of lake water quality. *Each panel corresponds to one predictor. The Y-axis is the standardised value of the marginal response for each of the ten modelled variables. In each case, the original marginal responses over all ten predictors were standardised to have a range between zero and one. Plot amplitude (the range of the marginal response on the Y-axis) is directly related to a predictor variable's importance; amplitude is large for predictor variables with high importance. Legend in top left panel applies to all panels.*

5.3 Representativeness of monitored lakes

The distributions of monitored lakes across the environmental gradients retained in the reduced RF models (CHLA, SECCHI, TN, TP and TLI3) were generally consistent with the distribution of all lakes nationally across the same gradients (Figure 14). For some environmental gradients, there was moderate over- and under-representation. Monitored lakes were slightly over-represented in environments characterised by low elevations (lkElev, catElev), and low catchment slopes (catSlope) and catchments with high alluvium (catAlluv) (Figure 14). Monitored lakes were under-represented in lakes with low fetch (lkFetch) and high altitude (lkElev). They were also under-represented in lake catchments with very low wintertime temperature (catWinTemp). For example, there were no lakes in our dataset with values of catWinTemp < -3.2 ^{o}C, however, 11% of lakes nationally have values of catWinTemp in this category. In addition, the monitored lakes were over represented in lake catchments with very high wintertime temperature.

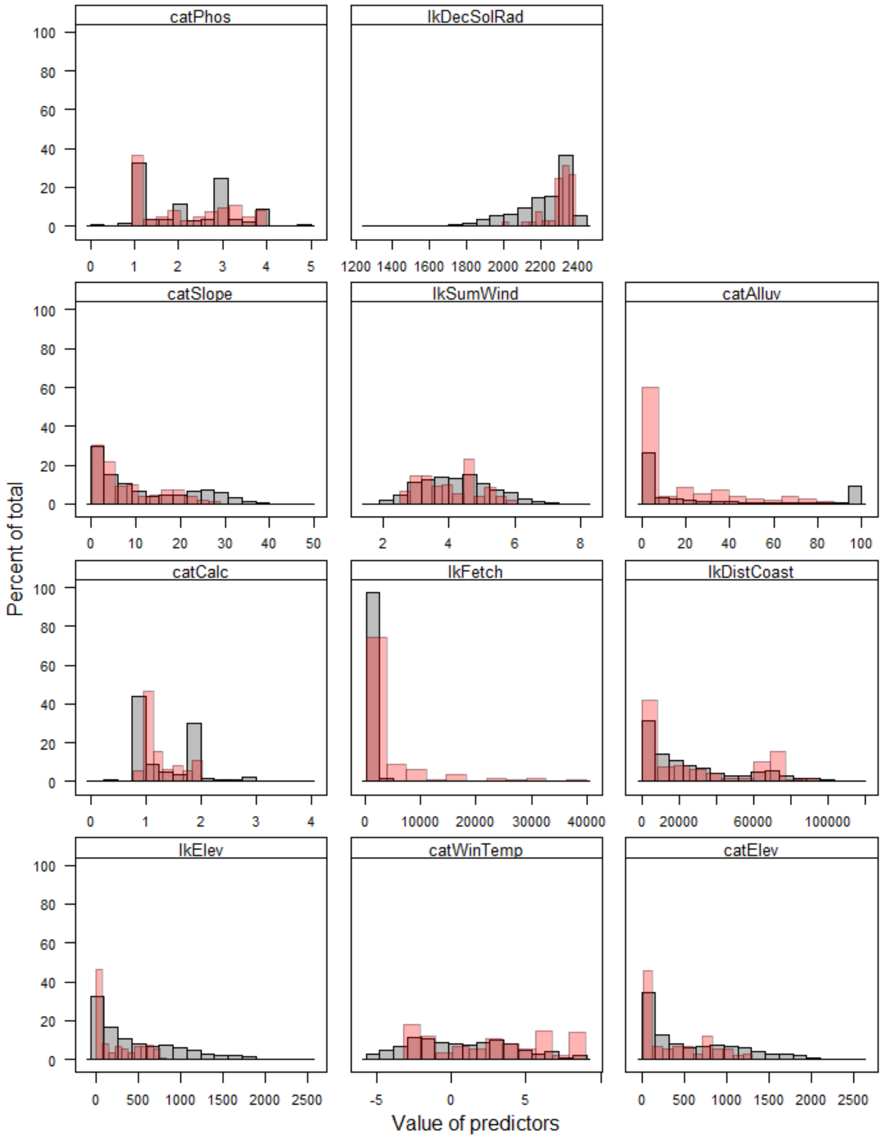

Figure 14. Histograms comparing the distributions of predictor variables for all lakes and the monitored lakes used to build the Random Forest models. *The national pool of lakes is represented by the grey histograms and the monitored lakes used for RF models that had satisfactory performance (CHLA, SECCHI, TN, TP and TLI3) are represented by the red histograms. Similarities in the distributions shown in the two histograms in each panel provide an indication of the degree to which environmental variation across the monitoring sites represents environmental variation across all lakes in New Zealand;*

complete representativeness would be indicated by exact matches between the histograms. The figure shows the 11 predictors (defined in Table 4) retained in the reduced RF models.

5.4 Model predictions

Predictions for CHLA, SECCHI, TN, TP and TLI3 are shown in Figure 15 and Figure 16. The figures show only the 895 lakes with shoreline length greater than 1500 m for clarity however predictions were made for the 3802 lakes that had complete data in the FENZ dataset. The mapped predictions for all five variables had similar spatial patterns, with high values of CHLA, TN, TP and TLI3 and low values of SECCHI, in low-elevation areas on the coasts of the North and South Island, apart from areas with little or no pastoral land cover (e.g., Fiordland). Values of CHLA, TN, TP and TLI3 were also high and values of SECCHI were low further inland in areas of both islands that are dominated by agricultural land use such as Southland, parts of Otago, Hawkes Bay, Bay of Plenty, Waikato and Northland (Figure 15 and Figure 16). Values of all of CHLA, TN, TP and TLI3 were generally low and SECCHI high in inland areas of the South Island.

There was an apparent geographical trend with high predicted cyanobacterial biovolume in lakes closer to the coastline of both South and North Island (Figure 17). High cyanobacterial biovolume values were also predicted in some inland areas, such as Otago, parts of Southland, Hawkes Bay and Manawatu (Figure 17). Otherwise, predicted cyanobacterial biovolume was generally low or medium in the inland areas of both the South and North Island.

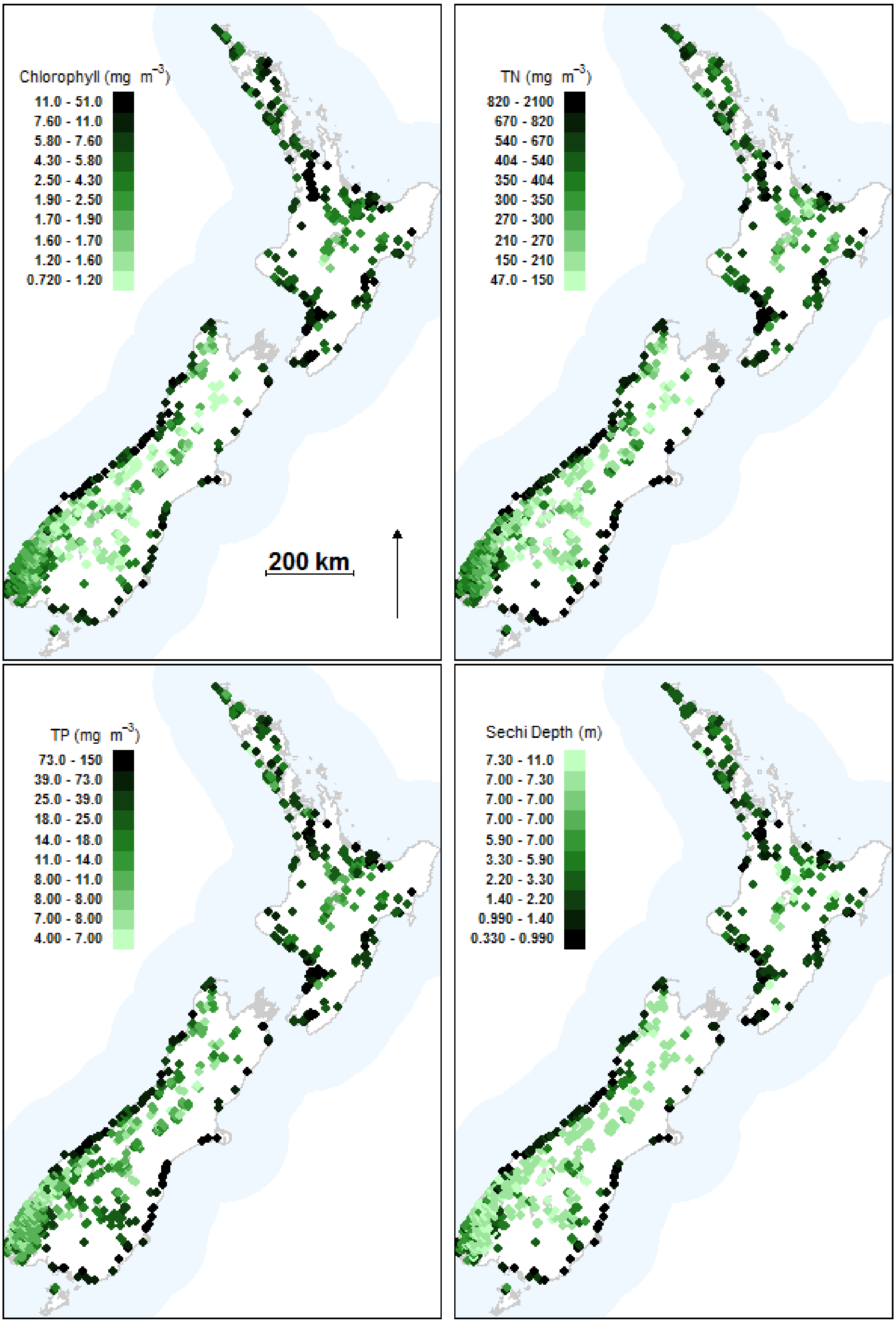

Figure 15. Predicted water quality for New Zealand lakes. *The 895 lakes with shoreline length greater than 1500 m are indicated by points located at the lake centre.*

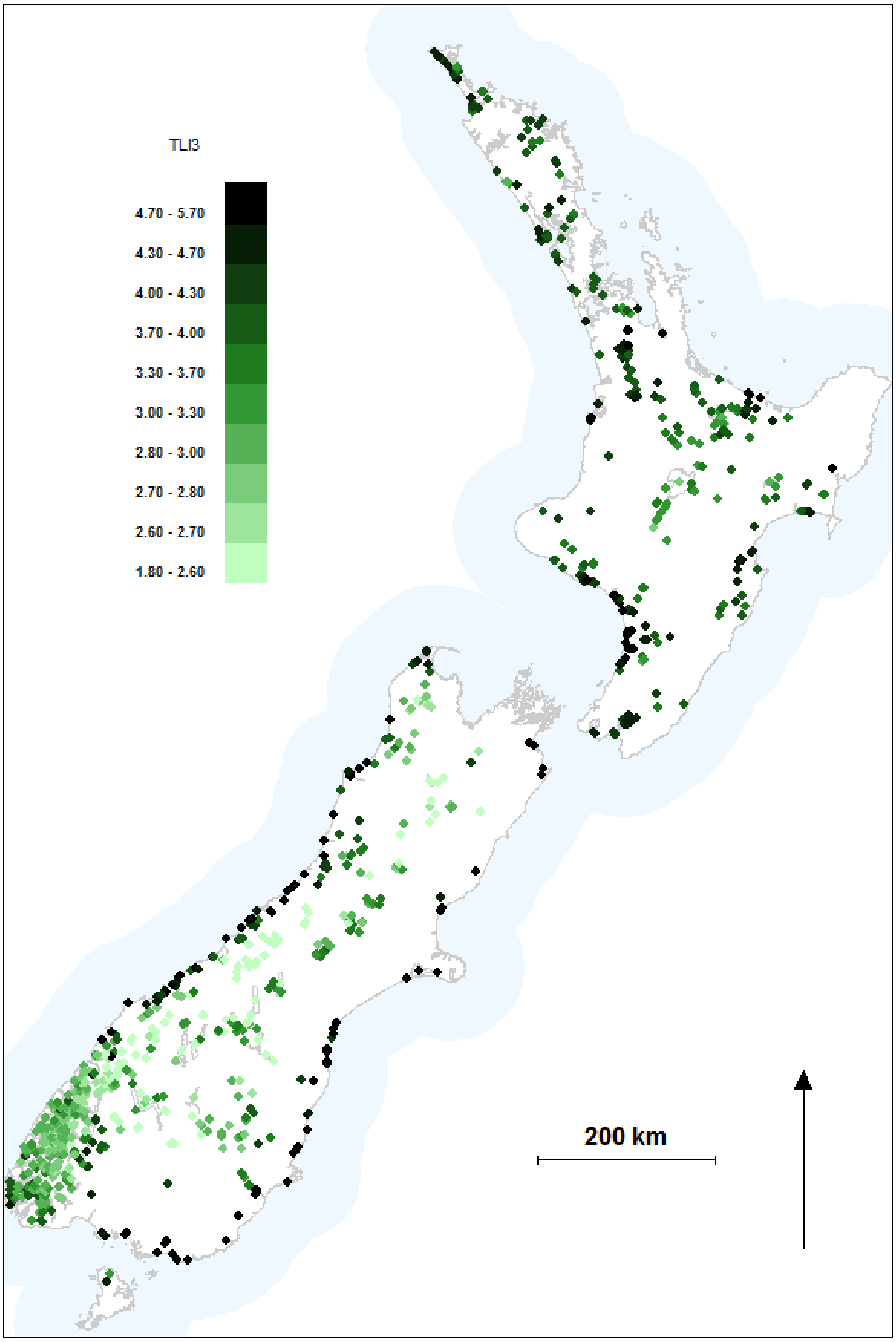

Figure 16. Predicted Trophic Level Index 3 for New Zealand lakes. *The 895 lakes with shoreline length greater than 1500 m are indicated by points located at the lake centre.*

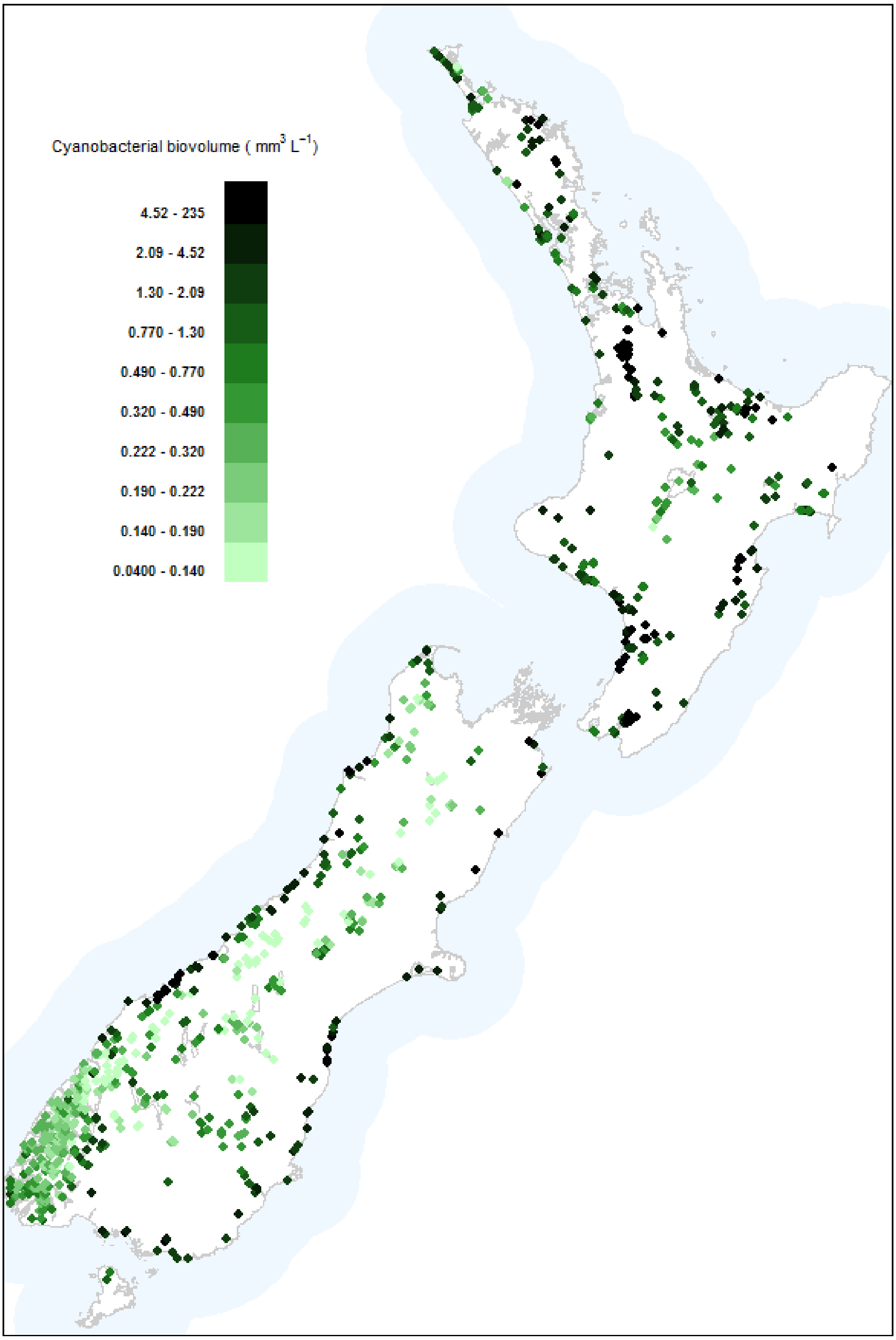

Figure 17. Predicted cyanobacterial biovolumes for New Zealand lakes. *The 895 lakes with shoreline length greater than 1500 m are indicated by points located at the lake centre.*

6 Discussion

6.1 *Escherichia coli* temporal behaviour and relationships with flow

The generally weak and inconsistent relationships between *E.coli* concentrations and flow (Figure 5) is probably due to at least three factors. First, there are likely to be differences in the dominant processes driving *E.coli* concentrations at different sites. For example, dilution of concentrations with increasing flows may occur at some sites and spikes in concentrations associated with wash off of *E.coli* sources with increasing flows at others. Second, at individual sites the relationships between concentration and flow may be inconsistent due to hysteresis[2]. This occurs when concentrations increase with flow during rising flows but are considerably lower at the same flow during reducing flows (sometimes referred to as first flush concentrations). The third reason for the lack of consistent relationships is likely to be the poor representation of high flows at most sites in the SOE dataset. Most samples were taken at flow close to the median with few samples representing high flows.

Strong relationships between flow and *E.coli* concentrations have been shown for some sites in New Zealand (Davies-Colley, 2013). However, these studies have been based on sampling at much higher temporal resolution and have indicated that relationships are complex. For example McBride (2011) showed that *E.coli* relationships with flow in some rivers is associated with hysteresis such that concentrations reach a maximum value ahead of the peak flow. In another study, Wilkinson *et al.* (2011) showed that it was more difficult to model *E.coli* response to storm flows in smaller catchments. That study suggests that spatial heterogeneity in rainfall-run-off and faecal sources probably contribute to considerable variability in relationships between flow and *E.coli* concentrations in small catchments in particular. The analysis of the timing of *E.coli* exceedances of the NPS-FM thresholds (260 *E.coli* 100 mL^{-1} and 540 *E.coli* 100 mL^{-1}) indicated there was significant spread in the data such that, in any REC class and month, there were sites with both very large proportion of exceedances and those with no exceedances (Figure 6). It is likely that site scale analysis might reveal that exceedance of *E.coli* thresholds at specific sites is more reliably seasonal. However, the analysis indicates that, at broad scales (i.e. national scale to the regional scales associated with REC Source-of-flow classes), exceedance of *E.coli* thresholds can occur in any month.

The conclusion from these two analyses is that, given SOE data and at the national scale to regional scales associated with REC Source-of-flow classes, the frequency of *E.coli* exceedances, and therefore risk to human health, cannot be determined by flow state or timing (i.e. does not depend on month or season). This indicates that the four *E.coli* statistics (site median, 95th percentile, PercGT260 and PercGT540), which are calculated from all observations, are relevant descriptions of human health risk. Using the SOE data available for the present study, accounting for flow or time of the year would not strongly influence conclusions about broad scale patterns drawn from these statistics.

6.2 Predictions of river *E.coli*

The monitoring sites used for fitting our *E.coli* spatial models were reasonably representative of the full range of environmental characteristics of New Zealand's rivers (Figure 9). The *E.coli* dataset was also relatively large and covered a significant proportion of the geographic domain (Figure 1). The good representation of the rivers of New Zealand, coupled with the

[2] Hysteresis is the time-based dependence of a system's state based on the present and past. For example, the concentration of *E.coli* at a given flow may depend on whether the flow at the time is increasing or decreasing. Therefore, to predict the *E.coli* concentration at a point in time it is not only necessary to know the flow, but also to know what the flow was previously.

good performance of the spatial models, means the predictions provide a useful description of regional to national scale patterns in *E.coli* for strategic purposes such as quantifying the proportion of New Zealand's rivers that have high and low human health risks associated with different types of activities.

There was a lack of SOE sites on the southern parts of the west coast of the South Island and in high altitude locations generally (Figure 1). This lead to under-representation of some environmental gradients that were important in the *E. coli* models, in particular sites with low values of catchment average summer air temperature (usAvTWarm) and catchment average variation of annual rainfall (usAnRainVar) (Figure 9). There are also likely to be other parts of the national environmental space that are poorly represented by the monitoring data. Prediction error in these poorly represented environments may be larger than the quantified uncertainties for the models and therefore, it is important to be aware of this limitation when using model predictions.

Lack of fit of the model may, in part, also reflect locations that are affected by point sources. These locations may have predictions that are inconsistent with observed concentrations because they are based on the characteristics of the catchment and do not explicitly account for point sources of faecal contamination. Site data, rather than model predictions are therefore the best basis for identifying specific locations with *E.coli* issues, and for assessing the effectiveness of response actions such as improving treatment of point source discharges.

6.3 Predictions of lake water quality

The lake dataset was smaller and had a restricted geographic coverage compared to the *E.coli* data set (Figure 2). In particular there was no or very limited data available for the Taranaki and Gisborne regions in the North Island, the top and west coast of the South Island and Southland. Monitored lakes were slightly over-representative of low elevations and lakes in regions with warmer climates and were under-representative of lakes in regions with colder climates (Figure 14). A somewhat surprising result was that the lake models included no predictors that directly described catchment land cover. It is well established that the proportion of the catchment occupied by pastoral land cover is strongly associated with magnitude of nutrient loads from agricultural source at the national scale (Larned *et al.*, 2016). However, the elevation predictors (catElev and lkElev) and the mean wintertime temperature predictor (catWinTemp) are likely included in the models partly because they are correlated with the catchment land use. Low elevation catchments and those in warmer regions are commonly associated with greater pastoral land use intensity than catchments at higher elevations and in colder regions.

The correlative rather than causative nature of the relationships between these predictors and nutrient loads to lakes is not relevant when considering the statistical measures of predictive performance of the models. However, it does mean that the lake model predictions are unrealistic in situations where the relationship between catElev, lkElev and catWinTemp and the actual causative variables (catchment nutrient loads) is significantly different to the fitting dataset. The most obvious situations where this is likely are lakes at low elevations whose catchments are largely unmodified, and lakes with cold climates (i.e. low catWinTemp) but low elevation. The model predictions are therefore likely to be less reliable in geographic regions that have low elevation lakes and lake catchments with relatively unmodified catchment land cover such as the West Coast of the South Island, Fiordland and Stewart Island.

6.4 Uncertainties associated with the spatial predictions

RF model performance differed between modelled variables and this variation may be attributable to differences in the biophysical processes that control the variables. Some biophysical processes may be poorly represented by our catchment-averaged spatial predictor variables. For example, concentrations of TN and TP in lakes are influenced to differing degrees by adsorption-desorption processes, deposition and suspension, and biological assimilation, transformation and removal; these mechanisms are not explicitly represented in the RF models. The absence of predictors that account for these and other processes means that some level of unexplained variation is inevitable.

All of the *E.coli* models and lake water quality models performed acceptably well to be used to make national scale predictions. Because the processes determining water quality at any location are complex, some unexplained variation in our models is to be expected. Predictions made for individual locations are associated with uncertainties that are characterised by model RMSD values (Table 6, Table 8). These uncertainties mean that predictions for individual river segments and lakes can be quite large. However, the level of model bias (i.e. systematic error) was low. This indicates that the predicted patterns are reliable descriptions of broad scale relative differences in the modelled variables between locations.

6.5 Cyanobacteria

Cyanobacteria are one component of lake plankton. While in eutrophic systems they are usually volumetrically dominant, eukaryotic algae such as chlorophytes (green algae) and bacillariophytes (diatoms), are often more abundant in mesotrophic and oligotrophic systems (Paul *et al.*, 2012). These other organisms contribute to total nutrient and chlorophyll *a* concentrations in lakes. The water quality parameters related to cyanobacterial biovolume in the present study (chlorophyll *a*, TN, TP, Secchi depth) can't be used to distinguish cyanobacteria from other organisms. Despite this caveat we observed significant relationships among cyanobacterial biovolumes and chlorophyll *a*, TN, TP and Secchi depth using samples collected from lakes spanning a range of trophic categories (Figure 11). These patterns are consistent with previous national and international studies. For example, Smith *et al. (*2016) showed a strong (R^2=0.66) relationship between TP and cyanobacterial biovolume in 71 Northern Hemisphere temperate lakes and four New Zealand lakes. Because of these strong relationships, when the modelled lake water quality data was transposed into cyanobacterial biovolumes using the multiple linear regression equation, eutrophic lakes were predicted to have the highest cyanobacterial biovolumes. This aligns with current knowledge, with many studies both nationally and internationally showing that high cyanobacterial biomass in lakes is strongly associated with eutrophication (e.g., Paerl and Otten, 2013; Wood *et al. 2016*). A recent New Zealand study that assessed samples from 143 lakes found that cyanotoxins were only detected in water samples from eutrophic lakes (Wood *et al.*, In review). This was largely due to the presence of specific cyanobacterial species in the eutrophic lakes, which differ from those found in mesotrophic or oligotrophic lakes. This knowledge adds further evidence to support the predictions of the present study that eutrophic lakes are likely to pose the highest risk to human health.

Both the cyanobacteria attribute specified by the NPS-FM and New Zealand Guidelines for Recreational waters suggest a two-tiered approach to trigger a D band (80th percentile over 12 samples for three years) or alert mode. When potentially toxic species are present a threshold of >1.8 mm^3 L^{-1} is suggested. A second threshold of >10 mm^3 L^{-1} of total cyanobacteria can be used when no toxins are detected, or no known toxin producing species are present. For the purpose of this study we have taken the conservative approach and assumed that all cyanobacterial species are toxic. We recommend this approach when assigning human health risk to lakes where there is high biovolumes. Many bloom forming taxa (which will be abundant when biovolumes are high) in New Zealand lakes are known toxin producers e.g. *Cylindrospermopsis*, *Anabaena, Aphanizomenon* and *Microcystis* (Wood *et al.*, 2006).

For the reasons discussed in Section 6.3, it is likely cyanobacterial biovolumes in lakes are over estimated in some areas of New Zealand. This is particularly likely to apply to the coastal regions of the West Coast of the South Island and Southland. However, we note that there is very little available cyanobacterial data for these regions. This highlights a need for more routine monitoring of lakes in these regions. Additionally, many water quality monitoring programmes do not include cyanobacteria/algal identification and enumeration. Although this is unlikely to have affected the cyanobacterial analysis undertaken in this study (see Section 2.2), it should be an essential part of lake monitoring programmes and would improve knowledge on national algal/cyanobacterial distribution patterns and therefore assist in improved predictions of potential health risk to humans.

The cyanobacterial dataset used to develop the multiple recreation equation was in some cases biased towards summer values and hence may overestimate cyanobacterial biovolumes. To establish the relationship between cyanobacterial biovolumes and water quality parameters we only used data points with corresponding year-month-site data. In many New Zealand lakes cyanobacterial analysis is only undertaken during summer when biomass is usually highest and there is greatest human contact. This may have resulted in an overestimation of cyanobacterial biovolumes, because the relationship was established during periods when cyanobacterial biovolume was likely to be highest, and not using long-term median values. As more frequent and consistent cyanobacterial datasets are obtained this approach could be revisited.

6.6 *Escherichia coli* in lakes

As noted in Section 2.2 data on *E. coli* in lakes in the SOE dataset were limited and prevented statistical analysis in this study. We are aware that there are other datasets, in particular samples collected as part of recreational monitoring programmes. Unlike cyanobacteria we do not anticipate *E. coli* concentration will be solely related to water quality variables due to the influence of other variables such as localised point sources, e.g. leaking septic tanks, rainfall, wind and inflows (Dada and Hamilton 2016). We recommend collation and analysis of available *E. coli* datasets to further understand the risks posed by this organism to human health in New Zealand lakes.

Acknowledgements

Regional councils are acknowledged for provision of all data. We thank Laura Biessy (Cawthron) for assistance with collating cyanobacterial datasets. Wendy Paul, Richard O'Rouke, Marcus Vandergoes (GNS) and Andrew Rees (Victoria University) are thanked for provision of datasets for cyanobacterial biovolume analysis. The NIWA and Cawthron micro-algal laboratories are acknowledged for provision of cyanobacterial cell concentration data. We thank Roger Young and Joanne Clapcott of Cawthron Institute for their helpful review of our draft report.

References

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

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

Burns, N., G. Bryers, and E. Bowman 2000. Protocol for Monitoring Trophic Levels of New Zealand Lakes and Reservoirs. Lakes Consulting Client Report: 99/2. 122 p. http://202.36.137.86/sites/default/files/media/Fresh%20water/Protocol%20for%20monitoring%20trophic%20levels%20of%20New%20Zealand%20lakes%20and%20reservoirs.pdf. Accessed 25 Aug 2016.

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

Dada, A.C., and D.P. Hamilton, 2016. Predictive Models for Determination of *E. coli* Concentrations at Inland Recreational Beaches Water Air Soil Pollution. 227: 347. doi:10.1007/s11270-016-3033-6

Davies-Colley, J.R., 2013. River Water Quality in New Zealand: An Introduction and Overview. Ecosystem Services in New Zealand: Conditions and Trends. Manaaki Whenua Press, Lincoln:432–447.

Duan, N., 1983. Smearing Estimate: A Nonparametric Retransformation Method. Journal of the American Statistical Association 78:605–610.

Hastie, T., R. Tibshirani, and J.H. Friedman, 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York.

Helsel, D.R., 2012. Reporting Limits. Statistics for Censored Environmental Data Using Minitab® and R, Second Edition:22–36.

Larned, S.T., M.R. Scarsbrook, T. Snelder, N.J. Norton, and B.J.F. Biggs, 2004. Water Quality in Low-Elevation Streams and Rivers of New Zealand. New Zealand Journal of Marine & Freshwater Research 38:347–366.

Larned, S.T., T.H. Snelder, M. Unwin, and G.B. McBride, 2016. Water Quality in New Zealand Rivers: Current State and Trends. New Zealand Journal of Marine and Freshwater Research. 50:389-417.

Larned, S.T., T.H. Snelder, M. Unwin, G.B. McBride, P. Verburg, and H.K. McMillan, 2015. Analysis of Water Quality in New Zealand Lakes and Rivers. NIWA CLIENT REPORT, NIWA, Christchurch, New Zealand.

Larned, S.T. and M. Unwin, 2012. Representativeness and Statistical Power of the New Zealand River Monitoring Network. NIWA Report, NIWA, Christchurch, New Zealand.

Leathwick, J.R., J.M. Overton, and M. McLeod, 2003. An Environmental Domain Analysis of New Zealand, and Its Application to Biodiversity Conservation. Conservation Biology 17:1612–1623.

McBride, G.B., 2011. Explaining Differential Sources of Zoonotic Pathogens in Intensively-Farmed Catchments Using Kinematic Waves. Water Science and Technology 63:695–703.

Ministry for the Environment, 2014. National Policy Statement for Freshwater Management.

Moriasi, D.N., J.G. Arnold, M.W. Van Liew, R.L. Bingner, R.D. Harmel, and T.L. Veith, 2007. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. Transactions of the ASABE 50:885–900.

Nash, Je. and J.V. Sutcliffe, 1970. River Flow Forecasting through Conceptual Models Part I—A Discussion of Principles. Journal of Hydrology 10:282–290.

Paerl, H.W. and T.G. Otten, 2013. Harmful Cyanobacterial Blooms: Causes, Consequences, and Controls. Microbial Ecology 65:995–1010.

Paul, W.J., D.P. Hamilton, I. Ostrovsky, S.D. Miller, A. Zhang, and K. Muraoka, 2012. Catchment Land Use and Trophic State Impacts on Phytoplankton Composition: A Case Study from the Rotorua Lakes' District, New Zealand. Hydrobiologia 698:133–146.

Piñeiro, G., S. Perelman, J. Guerschman, and J. Paruelo, 2008. How to Evaluate Models: Observed vs. Predicted or Predicted vs. Observed? Ecological Modelling 216:316–322.

Smith, V.H., S.A. Wood, C. McBride, J. Atalah, D. Hamilton, and J. Abell, 2016. Phosphorus and Nitrogen Loading Restraints Are Essential for Successful Eutrophication Control of Lake Rotorua, New Zealand. Inland Waters 6:273–283.

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., U. Shankar, K. Dey, and H. Hurren, 2006. Development of Variables for Freshwater Environments of New Zealand (FWENZ): Lakes. Christchurch.

Sorrell, B., M. Unwin, K. Dey, and H. Hurren, 2006. A Snapshot of Lake Water Quality. NIWA, Christchurch.

Svetnik, V., A. Liaw, C. Tong, and T. Wang, 2004. Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. Multiple Classifier Systems:334–343.

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

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

Wilkinson, R.J., L.A. McKergow, R.J. Davies-Colley, D.J. Ballantine, and R.G. Young, 2011. Modelling Storm-Event E. Coli Pulses from the Motueka and Sherry Rivers in the

South Island, New Zealand. New Zealand Journal of Marine and Freshwater Research 45:369–393.

Wood, S.A., H. Borges, J. Puddick, L. Biessy, J. Atalah, I. Hawes, D.R. Dietrich, and D.P. Hamilton, 2016. Contrasting Cyanobacterial Communities and Microcystin Concentrations in Summers with Extreme Weather Events: Insights into Potential Effects of Climate Change. Hydrobiologia:1–19.

Wood, S.A., D.P. Hamilton, W.J. Paul, K.A. Safi, and W.M. Williamson, 2009. New Zealand Guidelines for Managing Cyanobacteria in Recreational Freshwaters—interim Guidelines. Wellington, Ministry for the Environment and Ministry of Health Wellington.

Wood, S.A., P.T. Holland, D.J. Stirling, L.R. Briggs, J. Sprosen, J.G. Ruck, and R.G. Wear, 2006. Survey of Cyanotoxins in New Zealand Water Bodies between 2001 and 2004. New Zealand Journal of Marine and Freshwater Research 40:585–597.

Wood, S.A., M. Maier, J. Puddick, X. Pochon, A. Zaiko, D.R. Dietrich, and D.P. Hamilton, In review. Trophic State and Geographic Gradients Influence Planktonic Cyanobacterial Diversity and Cyanotoxin Distribution. FEMS Microbiology Ecology.

Zuur, A.F., E.N. Ieno, and C.S. Elphick, 2010. A Protocol for Data Exploration to Avoid Common Statistical Problems. Methods in Ecology and Evolution 1:3–14.

Appendices

Appendix 1. List of lakes where biovolume data was obtained and number of observations for each variable. TP = total phosphorus, TN = total nitrogen, Chl-*a* = chlorophyll *a*.

Lake	Region	Biovolume	TP	Chl-*a*	TN	Secchi	Total
Alice	Manawatu	7	7	7	7	0	7
Ellesmere	Canterbury	475	473	473	473	431	475
Forsyth	Canterbury	122	95	94	95	0	122
Hakanoa	Waikato	58	39	58	38	37	58
Heaton	Manawatu	8	7	7	7	0	8
Horowhenua	Manawatu	18	11	11	11	0	18
Kainui	Waikato	10	10	10	10	0	10
Koputara	Manawatu	8	7	7	7	0	8
Kuwakatai	Auckland	12	12	12	1	11	12
Lower Karori Reservoir	Wellington	58	58	58	58	0	58
Maraetai	Waikato	84	84	84	84	82	84
Ohakuri	Waikato	85	85	85	85	85	85
Oingo	Hawkes Bay	12	8	8	7	8	12
Okaro	Bay of Plenty	103	89	90	88	102	103
Okawa Bay	Bay of Plenty	113	111	112	108	105	113
Omanuka Lagoon	Manawatu	7	6	6	6	0	7
Omapere	Northland	113	99	0	101	76	113
Opouahi	Hawkes Bay	12	7	7	7	7	12
Ototoa	Auckland	8	8	8	0	7	8
Pounui	Wellington	22	22	21	22	22	22
Rotoehu	Bay of Plenty	113	113	112	110	110	113
Rotoiti	Bay of Plenty	115	115	111	109	115	115
Rotorua	Canterbury	90	90	37	90	0	90
Rotorua	Bay of Plenty	115	111	109	109	108	115
Runanga	Hawkes Bay	13	9	9	9	9	13
Spectacle	Auckland	14	14	14	2	14	14
Swan	Northland	7	0	0	0	7	7
Tarawera	Bay of Plenty	33	32	32	31	32	33
Taupo	Waikato	27	0	27	0	0	27
Tomarata	Auckland	10	10	10	2	10	10
Tutira	Hawkes Bay	87	58	49	58	50	87
Waahi	Waikato	63	63	63	63	52	63
Waikare	Waikato	49	49	49	49	49	49
Waikaremoana	Hawkes Bay	12	9	9	9	8	12
Waikopiro	Hawkes Bay	13	9	9	9	9	13
Waitawa	Wellington	23	23	23	23	23	23
Whangape	Waikato	58	58	58	58	58	58

Appendix 2. List of lakes where biovolume data was obtained but the data was not used in this study as the number of observation was considered too low.

Lake	Region	Number of samples
Dudding	Manawatu	5
Hatuma	Hawkes Bay	5
Herbert	Manwatu	5
Horseshoe	Hawkes Bay	2
Kohata	Manawatu	3
Koitata	Manawatu	3
Manapouri	Southland	1
Ngaroto	Waikato	6
Opunake	Taranaki	3
Pauri	Manwatu	6
Pukepuke Lagoon	Manawatu	4
Pupuke	Auckland	1
Ratapiko	Taranaki	3
Rotokare	Taranaki	4
Rotomanu	Taranaki	3
Te Anau	Southland	1
The Reservoir	Southland	3
Vincent	Southland	3
Waihoropita	Northland	1
Waikareiti	Hawkes Bay	6
Wainamu	Auckland	5
Waiparera	Northland	1
Waipu	Manawatu	3
Westmere	Manawatu	2
Whakaki	Hawkes Bay	4
Whakaki	Hawkes Bay	2
Whakaki First Bluff	Hawkes Bay	6
William	Manawatu	3
Wiritoa	Manawatu	5