Earth Sciences New Zealand

Spatial modelling of lake water quality state

Incorporating monitoring data for the period 2020 to 2024

Prepared for the Ministry for the Environment

November 2025

Prepared by:
David Wood
Doug Booker
Rachel Smith

For any information regarding this report please contact:

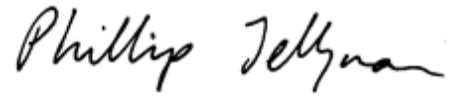
David Wood
Water Quality Scientist
Freshwater Modelling Group
+64 3 343 8050
david.wood@niwa.co.nz

New Zealand Institute for Earth Science Limited
PO Box 8602
Riccarton
Christchurch 8440

Phone +64 3 348 8987

Client Report No: 2025340CH
Report date: November 2025
Project No: MFE26501

Revision	Description	Date
Version 1.0	Final version sent to client	6 November 2025

Quality Assurance Statement		
	Reviewed by:	Aidin Jabbari
	Formatting checked by:	Terry Smith
	Approved for release by:	Phillip Jellyman

Contents

Tables

Figures

Executive summary

Purpose

The New Zealand Ministry for the Environment (MfE) and Stats NZ Tatauranga Aotearoa (Stats NZ) use lake water quality data originating from state-of-the-environment (SoE) monitoring to inform policy development and meet their requirements for environmental reporting on the freshwater domain under the Environmental Reporting Act 2015. Analysis of monitored data can describe water quality at SoE monitoring sites but does not provide insight into conditions at unmonitored sites. Raw monitored data for each water quality variable are often summarised over a specified period using a specified summary statistic. The combination of variable and statistic is called an attribute. This report supplements analysis of monitored data by providing model-based predictions of current water quality attribute states for all lakes across New Zealand.

Methods

Predictions were generated for each of approximately 4500 unique lakes represented within the Freshwater Ecosystems of New Zealand (FENZ) database. These predictions are based on the observed state of water quality in lakes for the period 2020–2024. Comparable reports were produced in 2016, 2019, and 2022 using data for the periods 2009–2013, 2013–2017 and 2016–2020, respectively. This report is the second in a series of reports prepared for the Ministry for the Environment on the topic of national-scale state and trends in lake water quality. The first report described data collation and analysis that provided site-specific water quality state and trends for approximately 150 lake monitoring sites operated by Regional Councils. The lake water quality data acquired and processed for the first report were used as input to the predictions presented in the current report.

The predicted water quality attribute states presented in this report were generated using random forest models, an advanced form of regression-tree modelling. Unlike single regression trees, which can be unstable and sensitive to small changes in data, random forests build many trees and base their predictions on the average outcome, improving accuracy and reliability. This approach is particularly well-suited to water quality analysis because it can handle complex and variable datasets, works well when predictors are intercorrelated, and does not require strict statistical assumptions about the data. Random forests are also less prone to overfitting than many other modelling methods, making them a robust tool for predicting water quality states.

Random forest models were developed for eleven water quality variables: Secchi depth (Secchi), chlorophyll-*a* (Chl-*a*), ammoniacal nitrogen (NH_4-N), pH-adjusted ammoniacal nitrogen (NH_4-N-adj), nitrate + nitrite-nitrogen (NNN), total nitrogen (TN), dissolved reactive phosphorus (DRP), total phosphorus (TP), *Escherichia coli* (*E. coli*), and Trophic Level Index 3 (TLI3) and Trophic Level Index 4 (TLI4). These variables were combined with a selection of statistics (e.g., median and 95th percentile) describing aspects of the distribution of the observed values. Several of the attributes, including Chl-*a* (median and annual maximum), TN (median), TP (median), NH_4-N-adj (median and 95th percentile) and *E. coli* (median and 95th percentile), are specified in the National Policy Statement for Freshwater Management (NPS-FM). Model predictors comprised 38 variables associated with FENZ lakes or their upstream catchments. These predictors were selected to represent climatic, geological, topographic,

land cover, land use intensity, and hydrological conditions across New Zealand lakes and their catchments.

The observational data used in the random forest models consisted of site attribute states from monthly and quarterly measurements for the period 2020–2024. These data came from 26–122 monitored lakes (depending on the modelled attribute state). They were distributed across the North and South Islands, although there were some significant gaps. To assess the degree to which the monitoring sites used for observational data represent the range of environmental conditions present in New Zealand, we compared density plots of the distributions of predictor values for the monitoring sites with the distributions of the same predictors for all lakes in FENZ in New Zealand. The monitoring sites were reasonably representative, with moderate over-representation of low-elevation, low-gradient catchments.

Results

The random forest models generally performed well in predicting attribute states, as indicated by high variance explained, close observed versus predicted agreement, low bias, and low uncertainty. However, there were exceptions: the models for the proportion of *E. coli* samples exceeding 260 or 540 per 100 mL^{-1}, as well as those for median DRP and median NNN concentrations, showed much poorer performance when compared with other attributes.

Model performance, importance/selection of predictors and predicted values representing the period 2020–2024 produced for this report were commensurate with those produced by previous work representing the period 2016–2020. National maps of the predicted attribute states are presented for each lake water quality attribute whose model exhibited satisfactory predictive performance. Maps indicated that lake water quality is expected to vary across the country. For example, results for Secchi indicated that higher water clarity would be expected at higher elevation lakes located inland compared to lower elevation lakes located nearer to the coast.

1 Introduction

State-of-the-environment (SoE) monitoring for lake water quality involves collecting, analysing, and reporting on physical, chemical, and biological data to assess health and trends over time. SoE monitoring involves regularly measuring a suite of standard variables at representative sites to provide a comprehensive and unbiased understanding of water quality condition (state) and its changes (trends).

Only a minority of New Zealand lakes are monitored as part of the SoE process. In the companion report, Booker et al. (2025) noted that SoE monitoring took place in approximately 115 lakes whereas about 4,500 large (>1 ha area) lakes are represented in the FENZ lake database after having excluded lakes associated with artificial construction of mines (FENZ 2024). The New Zealand Ministry for the Environment (MfE) and Stats NZ Tatauranga Aotearoa (Stats NZ) asked Earth Sciences New Zealand to estimate the state of water quality in all large lakes in New Zealand, excluding those whose origin is associated with artificial construction of mines (Booker and Snelder 2025). These estimates can then be used to inform policy and for other purposes.

Specifically, Earth Sciences NZ has been asked to update the lake water quality spatial modelling report prepared by (Snelder et al. 2022). This update would take into account the summary statistic representing conditions up to the end of 2024 from Booker et al. (2025). Other model inputs, referred to as predictors, would come from the Freshwater Ecosystems of New Zealand (FENZ), the Land Cover Database (LCDB), Fundamental Soil Layers (FSL), Digital Network (DN), Digital Elevation Model (DEM) and Agriculture Production Statistics (APC) from Stats NZ as described by Booker and Wilkins (2025). For each of the core water quality variables, the state would be predicted for all of the approximately 4,500 large lakes (>1 ha area) in the FENZ 2024 database that are not classified as either Artificial Constructed or Mine. MfE requested that predictive performance be assessed, including estimates of modelling uncertainty.

This report details the methods for using random forest models to predict current lake attribute states across New Zealand's diverse environments. We describe the data preparation, predictor selection, site representativeness assessment, modelling process, and model performance evaluation. The results section presents national predictions, identifies key predictors, and quantifies model performance. The discussion compares these models with previous ones, addresses uncertainties, and explores alternative methods. We also provided the Ministry for the Environment (MfE) with the model outputs for each of the large lakes (>1 ha area) in the FENZ (2024) database that are not classified as either Artificial Constructed or Mine, and the prediction data is available in the supplementary file “LakeRF_WQModel_Predictions_2025-11-04.csv”.

2 Data

Predictive modelling aims to establish a relationship between response variables (such as lake attribute state data) and predictor variables (which include input data). Once this relationship is defined, it can be used to estimate attribute states for lakes where direct measurements are not available, provided that predictor data are accessible.

To achieve this, response and predictor data must be combined into a training dataset. This integration is facilitated by using unique lake identifiers from the FENZ database, ensuring that data from each lake are correctly matched.

The remainder of this section describes the lake attribute state data and the predictor variables used in the modelling process.

2.1 Lake attribute state data

Lake attribute state data, for a five-year period (January 2020 to December 2024), were compiled from monitoring results collated by Booker et al. (2025). The attribute state data consisted of physical, chemical, or biological variables summarised over the five-year period using a specified statistic (median, etc) from lake monitoring sites in council SoE networks (Table 2-1). Detailed methods for processing the water quality observations are given in Booker et al. (2025).

Table 2-1: Lake attributes, measurement units and site numbers used to develop random forest models. A random forest model was developed for each combination of variable and statistic (referred to as an attribute state). Sites = the number of monitoring sites with observations available for the attribute state.

Variable type	Variable	Abbreviation	Statistic	Units	Sites#	Lakes#	Modelled Lakes
Physical	Secchi depth	Secchi	Median	M	112	89	88
Biological	Chlorophyll *a*	Chl-*a*	Median, Annual Maximum	mg L^{-1}	148	114	113
	Trophic Level Index 3	TLI3	Median	Unitless	154	123	122
	Trophic Level Index 4	TLI4	Median	Unitless	124	99	98
	Escherichia coli	*E. coli*	Median, 95th percentile	cfu or MPN 100 mL^{-1}	119	94	93
			>260	% exceedances	119	94	38
			>540	% exceedances	119	94	26
Chemical	Dissolved reactive phosphorus	DRP	Median	mg L^{-1}	145	114	113
	Total phosphorus (unfiltered)	TP	Median	mg L^{-1}	147	115	114
	Ammoniacal nitrogen	NH_4-N	Median	mg L^{-1}	145	114	113
	Ammoniacal nitrogen adjusted for pH	NH_4-N-adj	Median, 95th percentile	mg L^{-1}	144	113	112
	Nitrate + nitrite nitrogen	NNN	Median	mg L^{-1}	102	79	78
	Total nitrogen (unfiltered)	TN	Median	mg L^{-1}	147	115	114

Data from Booker et al. (2025).

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

Five different nutrient species (NNN, NH_4-N, DRP, TN and TP) were included because they influence the growth of planktonic, epiphytic and benthic algae and vascular plants (macrophytes) in lakes, and because ammonia can be toxic to lake organisms at high

concentrations. Nutrient enrichment can promote proliferations of planktonic algae (phytoplankton) and epiphytic algae on the surfaces of lake macrophytes. These algae can inhibit macrophyte growth by reducing light penetration. At elevated concentrations, free ammonia (NH_3) can be toxic to animals, including lake fish and invertebrates (Randall and Tsui 2002). The concentration of free ammonia and consequent risk to fish and invertebrates is determined by water temperature, pH and salinity, as well as the concentration of total ammonium plus ammonia ($NH_4^+ + NH_3$), in this report termed ammoniacal nitrogen and abbreviated NH_4-N.

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

The Trophic Level Index (TLI) is an indicator variable that summarises data related to lake trophic state and potential primary production. The TLI is used to classify New Zealand lakes into trophic classes (e.g., oligotrophic, eutrophic); TLI scores increase with increasing eutrophication. There are two versions of TLI in use in New Zealand, one with three variables (TLI3) and one with four variables (TLI4) (Burns et al. 2000; Verburg et al. 2010). TLI3 scores are derived from log-transformed concentrations of chlorophyll *a*, TN and TP. TLI4 uses Secchi data in addition to chlorophyll *a*, TN and TP concentrations. However, Secchi data were not available for all lakes in the current study. Moreover, Secchi data are strongly influenced by factors that are independent of trophic state, such as fine glacial sediment and tannins. To ensure consistent calculations, we calculated both TLI3 and TLI4 scores for all lakes in our national dataset (using the formulae given by Burns et al. (1999)) and used these scores in lieu of TLI scores provided in council datasets.

The concentration of the bacterium *Escherichia coli* (*E. coli*) is used as an indicator of human or animal faecal contamination, which is associated with the risk to humans arising from infection or illness from waterborne pathogens during contact-recreation.

We used attributes for lakes that have been defined by the NPS-FM to provide context to the water quality state analyses. Five of the nine variables used in the current report are also attributes in the NPS-FM: phytoplankton (as chlorophyll *a*), TN, TP, NH_4-N and *E. coli*.

The lake attribute dataset comprised 16 combinations of variables and statistics, gathered from 158 sites across 123 lakes. Not all attributes were measured at each monitoring site, and not all lakes were monitored for every attribute. Some lakes had multiple monitoring sites. In such cases, an overall estimate of the state of each attribute was derived from the median value of the attribute in that lake.

Predictive performance of random forests can be improved by transforming the response variable before modelling. Transformation can be particularly beneficial in cases where the response variable is strongly skewed, which was the case for many of the water quality variables. Transformation of skewed variables is beneficial because predictions from random forests are derived from a weighted mean of observations. Attribute data were therefore transformed before inclusion in the modelled dataset. Most attribute states were log-transformed (log_{10}), except for TLI3 and TLI4 which were untransformed, and *E. coli*

exceedances (proportion of samples exceeding 260 or 540 cfu or MPN 100 mL^{-1})[1], which were logit-transformed.

Data transformation led to the exclusion of some lake attribute combinations from the training dataset where data transformation created numeric values that cannot be modelled. Specifically, values of minus infinity were generated when the logit-transformation was applied to *E. coli* concentration data from sites that did not exceed 260 or 540 cfu or MPN 100 mL^{-1} within the analysis period. Sites that did not exceed the *E. coli* concentration thresholds within the analysis period were therefore excluded from the modelling process because numeric values of minus infinity cannot be modelled with random forests. Additionally, three sites without a lake identifier were excluded from the training dataset because predictor variables could not be identified. The resulting count of lakes is provided in the modelled lakes column in Table 2-1.

The frequency of water quality monitoring can vary depending on the variable and location. Water quality monitoring often occurs monthly or quarterly. The number of water quality observations per site collected over a five-year period ranged from 19 to 60, except for Trophic Level Index (TLI3 and TLI4), where the counts ranged from 4 to 5 over a five-year period.

The geographic distribution of lake monitoring sites used for modelling is shown in Figure 2-1. There is a high degree of overlap among the sites used for physical, chemical, and biological water quality monitoring, as many or all of the corresponding variables are measured at each site in council SoE programmes. The maps provide the location of sites where attribute data were available after the data were transformed (see Section 3.1).

There are noticeably more monitoring sites in the North Island than in the South Island; a difference particularly pronounced for NNN. It also appears that some regions are not represented at all in the monitoring data, while other regions (notably the far north of the North Island) have a dense network of sites and therefore may be over-represented.

[1] There are two methods for counting *E. coli* bacteria in a sample. The first method involves directly counting visible colonies on a plate, reported as colony-forming units (cfu). The second method uses a statistical approach, with results expressed as the Most Probable Number (MPN)

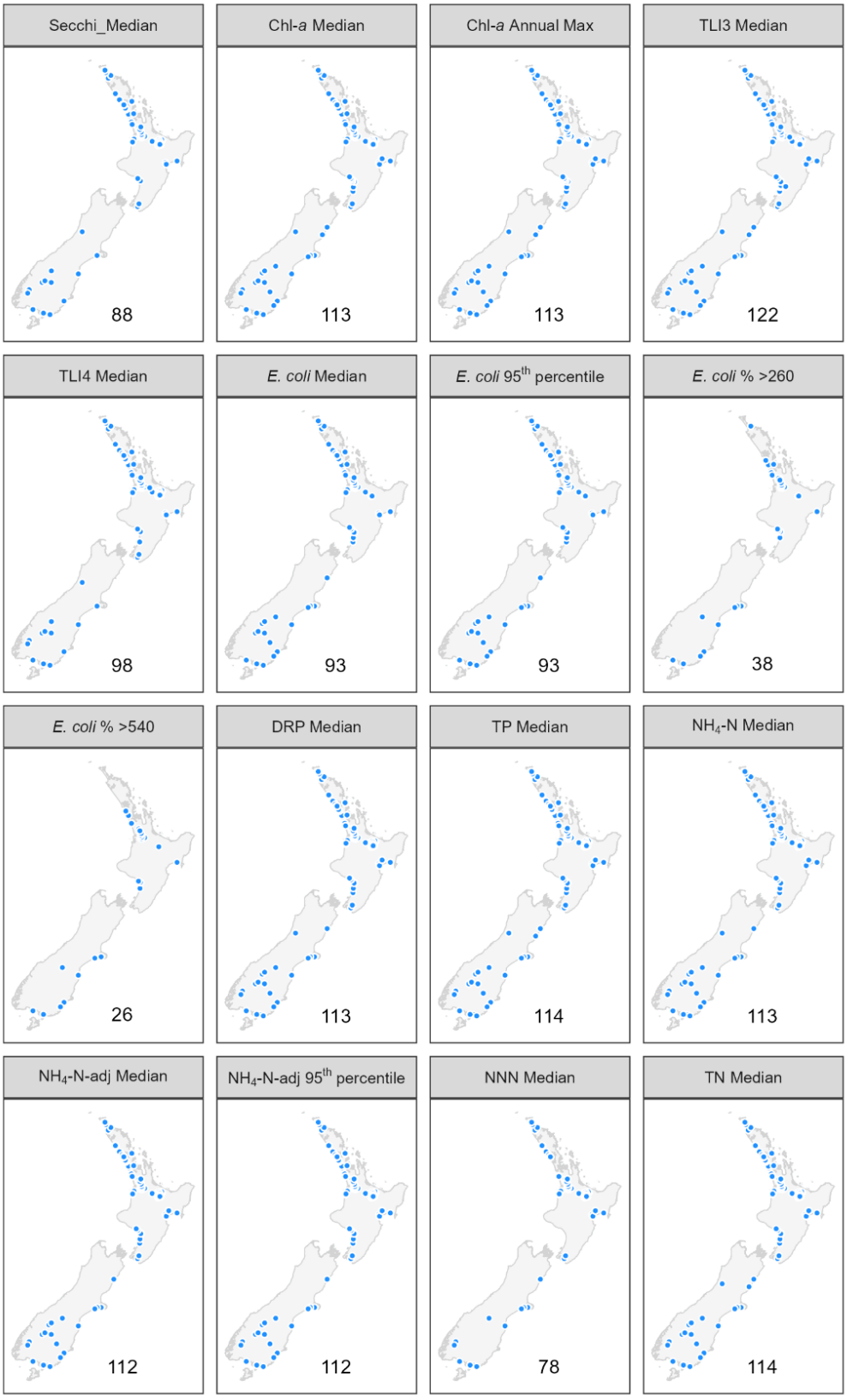


Figure 2-1: Locations of lake monitoring sites used for modelling the current state of the 16 attributes. The number in the lower right of each panel corresponds to the number of lakes included in each attribute state model after data transformation.

2.2 Predictor data

We used the FENZ database[2] to provide a spatial framework for the random forest models of the lake attribute state. The FENZ 2024 database includes information on approximately 4500 natural or dammed lakes with a surface area of one hectare or more. Lakes associated with mines were excluded from the analysis following the suggestion of Booker and Snelder (2025).

Spatial data layers describing the climate, topography, geology, vegetation, land use, infrastructure and hydrology of New Zealand were used to calculate predictor variables and were mapped onto FENZ. This mapping process generates variables that describe the environmental characteristics of the lake and its catchment. The predictor variables and data source used to calculate each predictor variable are described in Booker and Wilkins (2025) and summarised in Table 2-2.

[2] https://data.mfe.govt.nz/layer/120714-freshwater-ecosystems-of-new-zealand-fenz-lakes-november-2024/

We selected 38 spatial variables (Table 2-2) for predictors in spatial models of the 16 attribute states listed in Table 2-1. The predictors were selected based on their expected statistical relationships with water quality and experience from previous data-driven modelling studies, as well as data availability. Kuczynski et al. (2024) noted that lakes in agricultural and urban areas exhibited elevated levels of nutrients, *E. coli*, Chl-*a,* and TLI scores, along with low Secchi depths. Elevation serves as a proxy for land cover: low elevations are mostly agricultural, while high elevations are forested or alpine. Elevation and climate are also related, for example higher altitudes are associated with lower temperature. Abell et al. (2020) highlighted the influence of geology on lake water quaity, noting the volcanic zone in the North Island having elevated phosphorus, whereas sedimentary geology is associated with lower phosphorus. Consequently, land cover, land use intensity, climate, topography and geology are all expected to be associated with lake water quality. Previous experience with national-scale modelling of water quality using data-driven approaches (Unwin et al. 2010; Larned et al. 2016; Whitehead 2018; Snelder et al. 2022; Whitehead et al. 2022), identified similar factors and informed our selection of predictors. Predictors were generated from the most recent available data, including the recently published Land Cover Database version 6.0 (LCDB6[3]).

Table 2-2: Predictors used in random forest models of lake attribute states. See Booker and Wilkins (2025) for further details.

Predictor class	Predictor Description	Abbreviation	Unit
Lake	Lake surface area	lkArea	m^2
	Straight line distance to coast	lkDistCoast	km
	Lake elevation	lkElev	m
	lkDepth[#]	lkDepth	m
Catchment topography	Mean catchment slope	catSlope	Degrees
	Catchment area	catArea	km^2
	Catchment elevation	catElev	m ASL
	Lake wind fetch	lkFetch	m
Climate and flow	Mean summer (December) solar radiation	lkSolarRadSum	W/m^2
	Mean winter (June) solar radiation	lkSolarRadWin	W/m^2
	Mean summer (December) air temperature	lkAirTempSum	degC
	Mean winter (June) air temperature	lkAirTempWin	degC
	Mean summer (December) wind speed	lkWindSpeedSum	m/s
	Mean winter (June) wind speed	lkWindSpeedWin	m/s
	Mean catchment summer (December) air temperature	catAirTempSum	degC
	Mean catchment winter (June) air temperature	catAirTempWin	degC
	Mean catchment rain days >10 mm	catRainDays	Days/year

[3] LCDB v6.0 - Land Cover Database version 6.0, Mainland, New Zealand | LRIS Portal

Predictor class	Predictor Description	Abbreviation	Unit
	Mean catchment coefficient of variation of annual rainfall	catRainVar	Ratio
	Catchment average discharge	catFlow	m^3/s
Geology	Mean catchment phosphorus content of regolith	catPhosphorus	Ordinal
	Mean catchment calcium content of regolith	catCalcium	Ordinal
	Mean catchment induration (hardness) of regolith	catHardness	Ordinal
	Mean catchment particle size of regolith	catPsize	Ordinal
	Proportion of catchment occupied by peat	catPeat	Proportion
	Proportion of catchment occupied by alluvial FSL types K, S, L, Z, Ts and Tl	catAlluvial	Proportion
Land cover	Proportion of catchment that is bare (LCDB6 classes 12, 14, 15, 16)	usBare	Proportion
	Proportion of catchment in exotic forest (LCDB6 class 64 and 71)	usExoticForest	Proportion
	Proportion of catchment in indigenous forest (LCDB6 class 54 and 69)	usIndigenousForest	Proportion
	Proportion of catchment occupied LCDB6 classes 10, 68, 70	usMisc	Proportion
	Proportion of catchment occupied by combination of high producing exotic grassland short-rotation cropland, orchard, vineyard and other perennial crops (LCDB6 classes 40, 30, 33)	UsPastoral	Proportion
	Proportion of catchment occupied in scrub (LCDB6 classes 50, 51, 52, 55, 56, 58)	usScrub	Proportion
	Proportion of catchment in low producing grassland (LCDB6 class 41, 43, 44)	usTussockGrassland	Proportion
	Proportion of catchment in built-up areas urban parkland, surface mines, dumps and transport infrastructure (LCDB6 classes 1, 2, 6, 5)	usUrban	Proportion
	Proportion of catchment occupied by wetlands (LCDB6 classes 45, 46, 47)	usWetlands	Proportion
Land use intensity	Catchment density of total stock units (SU)	TotalSUDensity	SU/m^2
	Proportion of total stock units attributable to beef cattle in catchment	usBeef	Proportion
	Proportion of total stock units attributable to dairy cows in catchment	usDairy	Proportion
	Proportion of total stock units attributable to deer in catchment	usDeer	Proportion
	Proportion of total stock units attributable to sheep in the catchment	usSheep	Proportion

[#] Excluded from modelling due to a large number of missing items.

3 Modelling methods

3.1 Random forest models

We modelled each attribute state as a function of predictors using random forest models (Breiman 2001). Most attribute states were log-transformed ($\log_{10}$), except for *E. coli* exceedances (proportion of samples exceeding 260 or 540 counts/100 mL), which were logit-transformed, and TLI3 and 4, which used untransformed data.

A random forest model is an ensemble of individual classification and regression trees (CART). In regression, CART partitions observations into groups that minimise response variance, based on binary splits derived from predictor variables. CART models require no distributional assumptions and can automatically fit non-linear relationships and high-order interactions. However, single regression trees are prone to instability and may not find globally optimal splits (Hastie et al. 2009). Random forests overcome these issues by averaging predictions from many trees grown on bootstrap samples of the data, with a random subset of predictors considered at each split. This randomisation and averaging improve prediction accuracy while retaining CART's flexibility.

Each random forest produces an asymptotic generalisation error, the prediction error for unseen data, which stabilises as the number of trees increases, reducing the likelihood of overfitting (Breiman 2001). We used default settings: 500 trees per forest, with one-third of the total predictors available at each split as used in Snelder et al. (2022) previous study. Performance typically improves with tree number, most gain occurring in the first 100 trees, with computational limits being the main constraint (Probst and Boulesteix 2018).

Unlike linear models, random forest models cannot be expressed as a simple equation, so it can be challenging to see the relationship, if any, between the predictor and response variable. To assess predictor importance, random forest models permute response values for out-of-bag (OOB) observations and measure the resulting loss in prediction accuracy. Importance is quantified as the increase in mean squared error (MSE) between predictions based on permuted versus original OOB observations, averaged across all trees and normalised by the standard deviation of these differences (Cutler et al. 2007). OOB observations are observations not used to fit the mode.

Partial dependence plots (PDPs) depict the marginal effect of a predictor on the response when other predictors are held constant (typically at their mean values). Although correlations or interactions among predictors can distort the interpretation, PDPs provide a useful approximation of modelled relationships (Cutler et al. 2007).

Random forest models can include any of the available predictors selected during fitting. Including marginally important or correlated predictors does not reduce predictive performance but may complicate interpretation. Random forest models produce an estimate of the importance of each predictor based on the OOB approach. Then 10-fold cross validation was used for model selection. In brief, we applied a backward elimination process to remove the least important predictors from the initial saturated models. The mean squared error (MSE) and its standard error were estimated using 10-fold cross-validation. The least important predictors were removed iteratively, with MSE recalculated at each step. The final "reduced" model was defined as the simplest model whose error was within one standard error of the minimum error

("one standard error rule"; Breiman et al. 2017). This approach yields a parsimonious model with equivalent predictive performance. Importance values were not recalculated at each step to avoid overfitting (Svetnik et al. 2004), though Speiser et al. (2019) noted that Svetnik's approach is computationally intensive; this approach was retained to be consistent with previous reports.

All calculations were conducted in R (R Core Team 2022) using the randomForest package (Liaw and Wiener 2002), with supporting packages including tidyverse for data manipulation and graphing (Wickham et al. 2019), sf for spatial data handling (Pebesma and Bivand 2023), and the pdp package for partial dependence plots (Greenwell 2017).

3.2 Model performance

A key question about any model is how well it performs in predicting unmeasured observed values. A key advantage of random forests is that model error can be estimated directly from the data using out-of-bag (OOB) observations, eliminating the need for a separate test dataset as required by many other modelling approaches. For each tree, predictions are made for samples excluded from its bootstrap selection (the OOB data). Prediction accuracy for these OOB samples provides an unbiased estimate of model error and variable importance.

Model performance was assessed using predictions for out-of-bag (OOB) samples (i.e. data not used to train each tree), which provide an internal, approximately independent validation. We summarised the models using four statistics: regression R^2, Nash-Sutcliffe Efficiencies (NSE), per cent bias (PBIAS) and 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 shows the proportion of the total variance explained by the regression model (Piñeiro et al. 2008). However, the regression R^2 is not a complete description of model performance.

The NSE (Nash and Sutcliffe 1970) provides a measure of overall model performance by indicating how closely a plot of observed versus predicted values lies to the 1:1 line (i.e., the degree to which two sets of values coincide). NSE values range from $-\infty$ to 1. An NSE of 1 corresponds to a perfect match between predictions and the observed data, an NSE of 0 indicates that the model predictions are as accurate as the mean of the observed data; and an NSE less than 0 indicates that the observed mean is a better predictor than the model.

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 2007). Model predictions were evaluated to be very good, good, satisfactory or unsatisfactory, following the criteria proposed by Moriasi et al. (2015), outlined in Table 3-1.

Table 3-1: Performance ratings for statistics used in this study. From Moriasi et al. (2015).

Performance Rating	R^2	NSE	PBIAS
Very good	$R^2 \geq 0.70$	NSE > 0.65	|PBIAS| <15
Good	$0.60 < R^2 \leq 0.70$	$0.50 < NSE \leq 0.65$	$15 \leq$ |PBIAS| < 20
Satisfactory	$0.30 < R^2 \leq 0.60$	$0.35 < NSE \leq 0.50$	$20 \leq$ |PBIAS| < 30
Unsatisfactory	$R^2 < 0.30$	$NSE \leq 0.35$	|PBIAS| ≥ 30

The root mean square deviation (RMSD) is a measure of the characteristic model statistical error or uncertainty. RMSD is the mean deviation of predicted values with respect to the observed values (distinct from the standard error of the regression model). RMSD can be used to evaluate the prediction intervals of the expected value of the observation.

The 95% prediction intervals for values predicted by our models for individual segments can be obtained using the following approximations. Equation 1 can be used for calculating the intervals for the TLI predictions. Equation 2 can be used for calculating the intervals for the water quality variables for which the variables were $\log_{10}$ transformed prior to model fitting, and the prediction uncertainty (RMSD) values have been reported in the $\log_{10}$ transformed space. Equation 3 can be used for calculating the intervals for attribute states for which the variables were logit-transformed prior to model fitting (*E. coli* % >260, *E. coli* % >540) and the prediction uncertainties are reported in the logit-transformed space.

$$95\%\ PI = x \pm 1.96 \times RMSD \quad (1)$$

$$95\%\ PI = 10^{[\log_{10}(x) \pm 1.96 \times RMSD]} \quad (2)$$

$$95\%\ PI = \frac{e^{[logit(x) \pm 1.96 \times RMSD]}}{(1+e^{[logit(x) \pm 1.96 \times RMSD]})}, \text{ where } logit(x) = log\left(\frac{x}{1-x}\right) \quad (3)$$

In all equations, x is the estimated value in the original units, RMSD is the model error in the transformed space, and 1.96 is the standard normal deviate or Z-score for probability ($0.025 \leq Z \geq 0.975$). The prediction intervals for the $\log_{10}$-transformed variables, when expressed in the original units, are asymmetric, and their values vary in proportion to the predicted water quality value. For example, if we let x If a predicted value for Secchi depth of 1 m, the lower and upper 95% confidence intervals are 0.37 and 2.7 m, respectively, assuming RMSD = 0.22.

3.3 Representativeness of monitoring sites used in random forest models

A graphic comparison was used to assess the fit of the monitoring sites to the random forest models, which represented environmental variation at the national scale. Here, representativeness refers to the degree of similarity of the distribution of predictors at monitoring matches the distribution of non-monitored lakes. Poor representativeness can reduce accuracy in model predictions because certain combinations of environmental conditions are not represented in the fitting data.

Density plots illustrating the distribution of the 12 most important predictors in the random forest models (i.e., the predictors with the greatest explanatory power) for sites where

monitoring took place and sites (segments) that were not monitored. The plot represented data from all sites that monitored at least one water quality attribute. Note that the representativeness of monitoring sites is different from model bias, which is defined in Section 3.2. Model bias is a measure of systematic error in model predictions (i.e., over- or under-estimation).

3.4 Model predictions

Predictions are made with random forest models by "running" new cases down every tree in the fitted forest and averaging the predictions made by each tree (Cutler et al. 2007). The models in this study were fitted to $\log_{10}$-transformed variables, except for TLI, which used non-transformed data and *E. coli* exceedances, which were logit-transformed. When the predictions made by models fitted to $\log_{10}$-transformed variables are back-transformed, the model error term no longer has a mean of zero. Ignoring this introduces retransformation bias, meaning the predictions systematically underestimate the response. To correct for this retransformation bias, we use the smearing estimate (*S*) developed by Duan (1983):

$$S = \frac{1}{n}\sum_{i=1}^{n} 10^{\hat{\varepsilon}_i} \quad (1)$$

where $\hat{\varepsilon}$ are the residuals of a random forest model. The predictions were back-transformed by raising them to the power of 10, then corrected for retransformation bias by multiplying by *S*. Predictions of *E. coli* exceedance values (%>260, %>540) were back-transformed using the inverse-logit function. The back-transformed and corrected predictions for all lakes in the FENZ database were projected on a national map for each attribute state.

4 Results

4.1 Model performance

Assessment of OOB predictions from random forest models showed that four of the 16 attribute states were considered as very good (Secchi Median, TLI4 Median and TN Median) or good (TLI3 Median) based on R^2, NSE and PBIAS and the criteria of Moriasi et al. (2015) (Table 4-1). Eight of the remaining twelve models (Chl-*a* Median, Chl-*a* Annual Max, *E. coli* Median, *E. coli* 95th percentile, TP Median, NH_4-N Median, NH_4-N-adj 95th percentile) were satisfactory, and four (*E. coli* % >260, *E. coli* % >540, DRP Median, NNN Median) were unsatisfactory. All 16 models had very low bias (PBIAS; Table 4-1) as illustrated by the fact that the line representing the regression of the observed versus OOB predicted values (blue line in Figure 4-1 is neither systematically above or below the one-to-one line (red dashed line in Figure 4-1) for all the predicted values.

Table 4-1: Performance of the 16 attribute state models. Performance was determined using independent predictions (i.e., sites that were not used in fitting the models) generated from the out-of-bag (OOB) observations. R^2 = coefficient of determination, NSE = Nash-Sutcliffe efficiency, PBIAS = percent bias, RMSD = root mean square deviation. Units for RMSD are the $\log_{10}$ or logit-transformed units of the attribute state, except for TLI, which were not transformed. The colours indicate the performance ratings indicated in Table 3-1.

Attribute	N	R^2	NSE	PBIAS	RMSD	Rating
Secchi Median	88	0.72	0.70	-0.75	0.27	Very good
Chl-*a* Median	113	0.42	0.41	-0.37	0.44	Satisfactory
Chl-*a* Annual Max	113	0.55	0.53	-0.27	0.46	Satisfactory
TLI3 Median	122	0.67	0.67	0.57	0.09	Good
TLI4 Median	98	0.72	0.72	0.49	0.08	Very good
E. coli Median	93	0.60	0.59	-0.81	0.4	Satisfactory
E. coli 95th percentile	93	0.59	0.58	0.25	0.45	Satisfactory
E. coli % >260	38	0.13	0.12	0.88	0.92	Unsatisfactory
E. coli % >540	26	<0.01	-0.12	1.23	0.78	Unsatisfactory
DRP Median	113	0.21	0.20	0.63	0.56	Unsatisfactory
TP Median	114	0.52	0.52	0.38	0.42	Satisfactory
NH_4-N Median	113	0.46	0.46	-0.35	0.38	Satisfactory
NH_4-N-adj Median	112	0.35	0.35	-0.37	0.43	Satisfactory
NH4N_adj 95th percentile	112	0.45	0.45	-0.58	0.53	Satisfactory
NNN Median	78	0.27	0.27	-0.50	0.63	Unsatisfactory
TN Median	114	0.71	0.70	0.08	0.24	Very good

Figure 4-1 displays the prediction intervals, the range where a future observation is expected to lie for any predicted value, as illustrated in the upper and lower grey lines (prediction interval). In some cases, the prediction intervals cover a small range relative to the observed values, such as the case of TN Median and the very good rating. In some cases, the prediction interval is very wide, such as for *E. coli* % >260, covering the range of observed values and justifying the rating

of unsatisfactory. The prediction intervals show that the models cannot explain some of the observed variation.

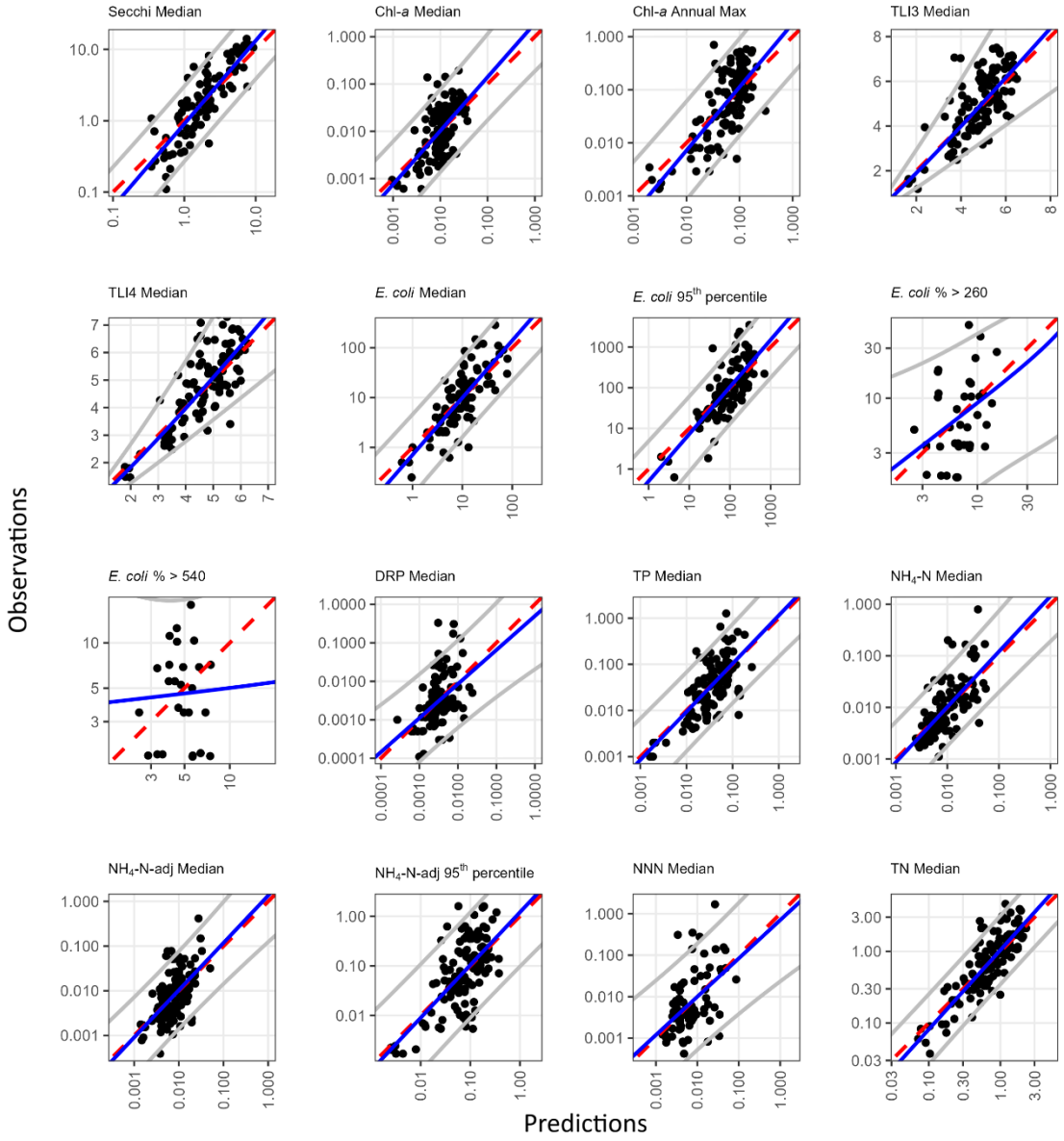


Figure 4-1: Comparison of observed attribute state versus values predicted by each of the random forest models. Note that the observed values are plotted on the Y-axis and predicted values on the X-axis, following Piñeiro et al. (2008). Blue solid line: best fit linear regression of the observed and predicted values. Red dashed line: one-to-one line. The grey lines represent the upper and lower prediction intervals, the regions where the actual value is expected to lie for a given prediction value. Note that all the axes are plotted on a log scale except for TLI3 and TL4. All the numbers are in the appropriate units for each variable.

4.2 Modelled relationships

In all the random forest models, the predictors with high importance reflected strong associations between water quality and land cover, catchment topography, land use intensity, and climate (Table 4-2). The details of the relationships varied by attribute. These associations

are broadly similar to the findings in previous studies of water quality state (Snelder et al. 2022b; Whitehead et al. 2022).

The highest-ranked predictor (Table 4-2) was the proportion of land upstream occupied by grassland, crops, vineyards and orchards (usPastoral), a variable related to land cover. It was among the top ten ranked predictors, except for three attributes with poor model performance ratings (*E. coli* % >260, *E. coli* % >540, DRP Median). The partial plots (Figure 4-2) indicated that lakes with a higher proportion of pastoral land (usPastoral) tend to have worse water quality. Worse water quality corresponds to higher numerical values for all attributes except Secchi depth, where lower values indicate poorer quality.

The second and third-highest-ranked predictors were lake (lkElev) and catchment (catElev) elevation. Predictors related to topography. Higher elevations are associated with better water quality, except for one attribute, NNN, which had a poor model performance rating.

Climate-related predictors, such as wind speed (lkWindSpeedWin) and number of rainy days (catRainDays), were ranked fourth and twelfth, respectively. Low lake wind speeds were associated with higher TLI3 and TN values and lower median NH_4-N and NH_4-N-adj levels than higher wind speeds. A lower number of rainy days was associated with higher DRP, median NH_4-N and NH_4-N-adj levels.

Geology-related predictors relating to the induration (hardness) of regolith (catHardness), the portion of the catchment occupied by alluvial regolith (catAlluvial) and the calcium content of regolith (catCalcium) were ranked sixth, eighth and eleventh. Higher hardness and lower calcium levels were associated with better water quality.

Together with the proportion of land upstream occupied by grassland, crops, vineyards, and orchards (usPastoral), which was the highest rank predictor. In addition, several other predictors associated with land cover and land use intensity were ranked in the top twelve. These included the proportion of the catchment that is classified as bare (usBare), wetlands (usWetlands) as well as the proportion of stock units attributable to beef cattle in the catchment, which were ranked fifth, ninth and tenth respectively.

Table 4-2: Rank order of importance of predictors retained in the random forest models for at least one attribute state. Blank cells indicate that a predictor was not included in the reduced model. The predictors are listed in descending order of the median rank importance over all 16 modelled attribute states.

Predictor	Secchi Median	Chl-*a* Median	Chl-*a* Annual Max	TLI3 Median	TLI4 Median	*E. coli* Median	*E. coli* 95th percentile	*E. coli* % >260	*E. coli* % >540	DRP Median	TP Median	NH_4-N Median	NH_4-N-adj Median	NH_4N-adj 95th percentile	NNN Median	TN Median
usPastoral	9	1	1	1	1	7	8	-	-	19	1	2	3	4	3	2
lkElev	1	2	3	2	2	1	2	-	-	7	3	20	8	7	-	3
catElev	2	3	4	6	3	3	9	1	12	4	6	6	1	6	17	1
lkWindSpeedWin	5	9	-	5	-	2	5	-	-	14	-	3	-	10	-	4
usBare	-	4	8	4	5	-	-	-	-	-	2	27	5	-	-	-
catHardness	-	-	-	-	-	6	1	-	8	22	-	4	13	1	5	-
lkArea	3	10	6	7	4	8	6	-	-	27	-	13	4	9	6	6
catAlluvial	-	21	-	-	-	-	-	-	1	-	-	7	-	-	-	-
usWetlands	-	7	2	-	-	-	-	-	-	-	-	37	-	-	-	-
usBeef	-	-	-	-	-	9	4	-	6	-	-	30	-	23	2	-
catCalcium	7	8	12	3	6	15	-	2	-	11	4	18	18	26	-	5
catRainDays	-	-	-	-	-	-	-	-	-	1	-	8	-	15	-	-
usScrub	-	-	-	-	-	11	3	-	2	-	-	9	-	5	8	9
usIndigenousForest	-	11	10	-	-	-	-	-	-	26	-	1	9	3	-	7
catAirTempWin	-	12	9	8	-	5	-	-	-	10	-	10	2	20	9	-
usExoticForest	-	-	-	-	-	-	-	-	9	6	-	12	-	-	10	-
lkSolarRadWin	-	23	-	-	-	-	-	-	10	5	-	11	10	17	-	-
catSlope	-	22	-	11	-	-	7	-	5	20	-	16	12	2	1	-
TotalSUDensity	8	26	-	9	-	19	11	-	-	9	5	34	-	-	19	-
lkFetch	4	13	7	12	-	4	10	-	-	-	-	19	7	13	15	-
catFlow	6	17	5	10	-	-	-	-	-	-	-	28	11	12	12	-
lkAirTempWin	-	6	-	-	-	14	-	-	7	2	-	21	17	19	11	-
catAirTempSum	-	19	11	13	-	-	-	-	13	3	-	14	6	27	4	-
catRainVar	-	18	-	-	-	-	-	-	-	12	-	26	14	-	7	8
lkWindSpeedSum	-	5	-	-	-	17	-	-	-	23	-	5	-	14	-	-
usDairy	-	-	-	-	-	12	-	-	-	24	-	23	-	11	14	-
catPhosphorus	-	15	-	-	-	-	13	-	3	8	-	38	-	-	16	-
usMisc	-	16	-	-	-	-	-	-	-	16	-	15	15	22	-	-
catPsize	-	-	-	-	-	18	12	-	4	18	-	17	-	16	-	-

Predictor	Secchi Median	Chl-*a* Median	Chl-*a* Annual Max	TLI3 Median	TLI4 Median	*E. coli* Median	*E. coli* 95th percentile	*E. coli* % >260	*E. coli* % >540	DRP Median	TP Median	NH_4-N Median	NH_4-N-adj Median	NH_4N-adj 95th percentile	NNN Median	TN Median
catArea	-	20	13	-	-	-	-	-	11	-	-	22	16	-	18	-
usUrban	-	-	-	-	-	-	-	3	-	-	-	35	-	18	-	-
lkDistCoast	-	-	-	-	-	13	-	-	-	25	-	24	19	8	-	-
lkAirTempSum	-	14	-	-	-	-	-	-	-	13	-	32	-	25	-	-
usTussockGrassland	-	24	-	-	-	10	-	-	-	17	-	29	-	-	-	-
lkSolarRadSum	-	27	-	-	-	16	-	-	-	-	-	25	-	21	-	-
catPeat	-	-	-	-	-	-	-	-	-	-	-	36	-	-	13	-
usSheep	-	25	-	-	-	-	-	-	-	15	-	31	-	24	-	-
usDeer	-	-	-	-	-	-	-	-	-	21	-	33	-	-	-	-

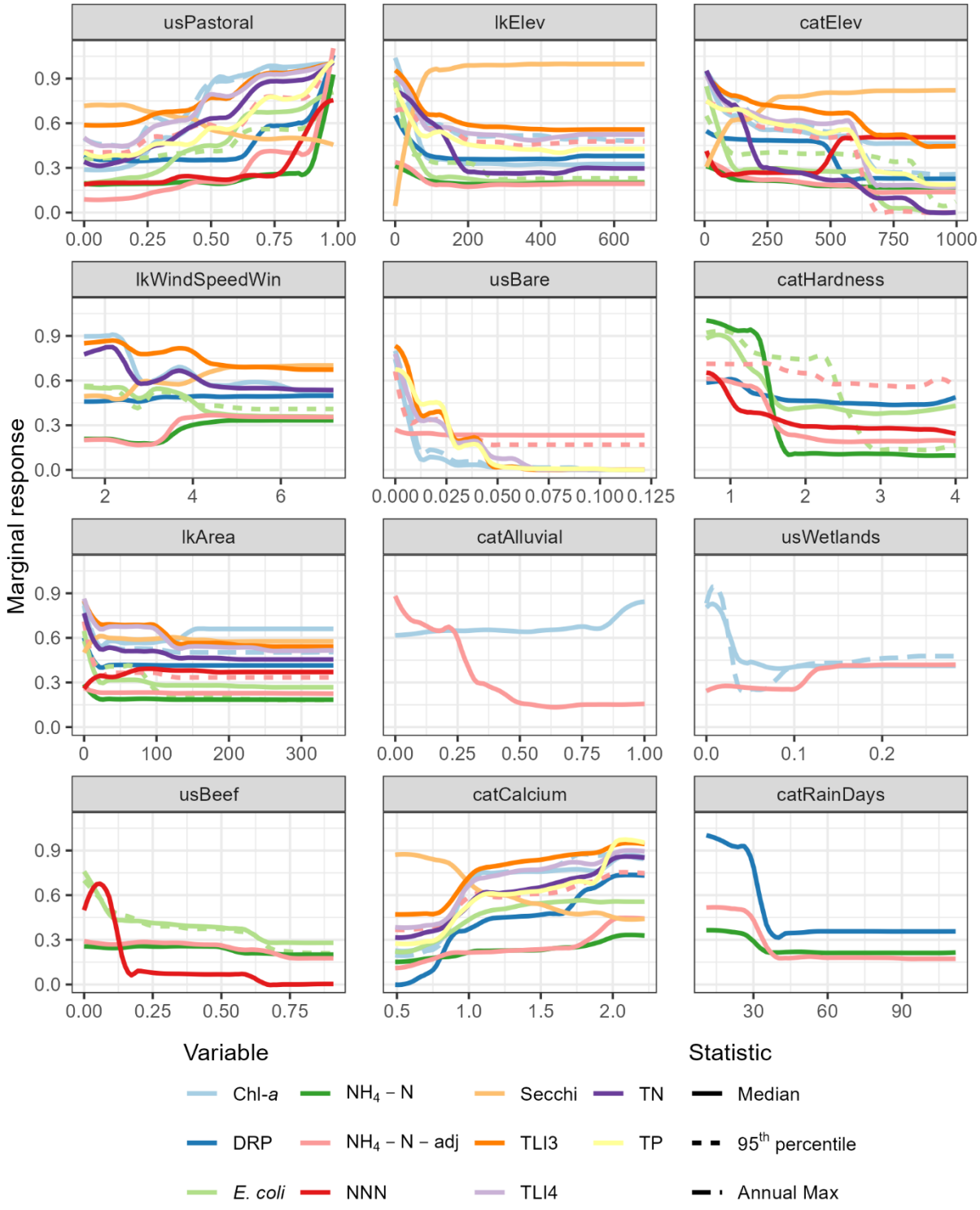


Figure 4-2: Partial plots for the 12 most important predictors in random forest models of current attribute state. Colours represent water quality variables, with the statistic indicated by line type (i.e., the combination of colour and line type represents an attribute). Each panel corresponds to one predictor, with predictors ordered by overall importance from most (top left) to least (bottom right) important. Y-axis scales represent marginal response standardised across all modelled attribute states. Plot amplitude (the range of the marginal response on the Y-axis) is directly related to a predictor's importance, with amplitude larger for predictors with higher importance. Units on X-axes are in Table 2-2. Only combinations of variable and statistics that form the studies attributes are shown.

4.3 Monitoring site representativeness

The distributions of the top 12 ranked predictors were similar for lakes with water quality monitoring sites compared to non-monitored lakes (Figure 4-3). Though there were some differences and several cases of moderate over- and under-representation of monitoring sites compared to the lakes in the FENZ database

The monitored lakes were overrepresented at lower elevations (lkElev, catElev) and underrepresented at higher elevations.

In terms of climate, monitored lakes were underrepresented in terms FENZ lakes with high numbers of rainy days (catRainDays) and slightly overrepresented in terms of wind speed around 4 m/s.

In terms of land use intensity, monitored lakes were overrepresented in the proportion of total stock units attributable to beef cattle in the catchment (usBeef).

For the landcover predictor, pastoral (usPastoral) and the geology predictor, alluvial regolith (catAlluvial), the monitored lakes are overrepresented in the intermediate proportions when unmonitored lakes tend to have either very high or very low proportions of these predictors.

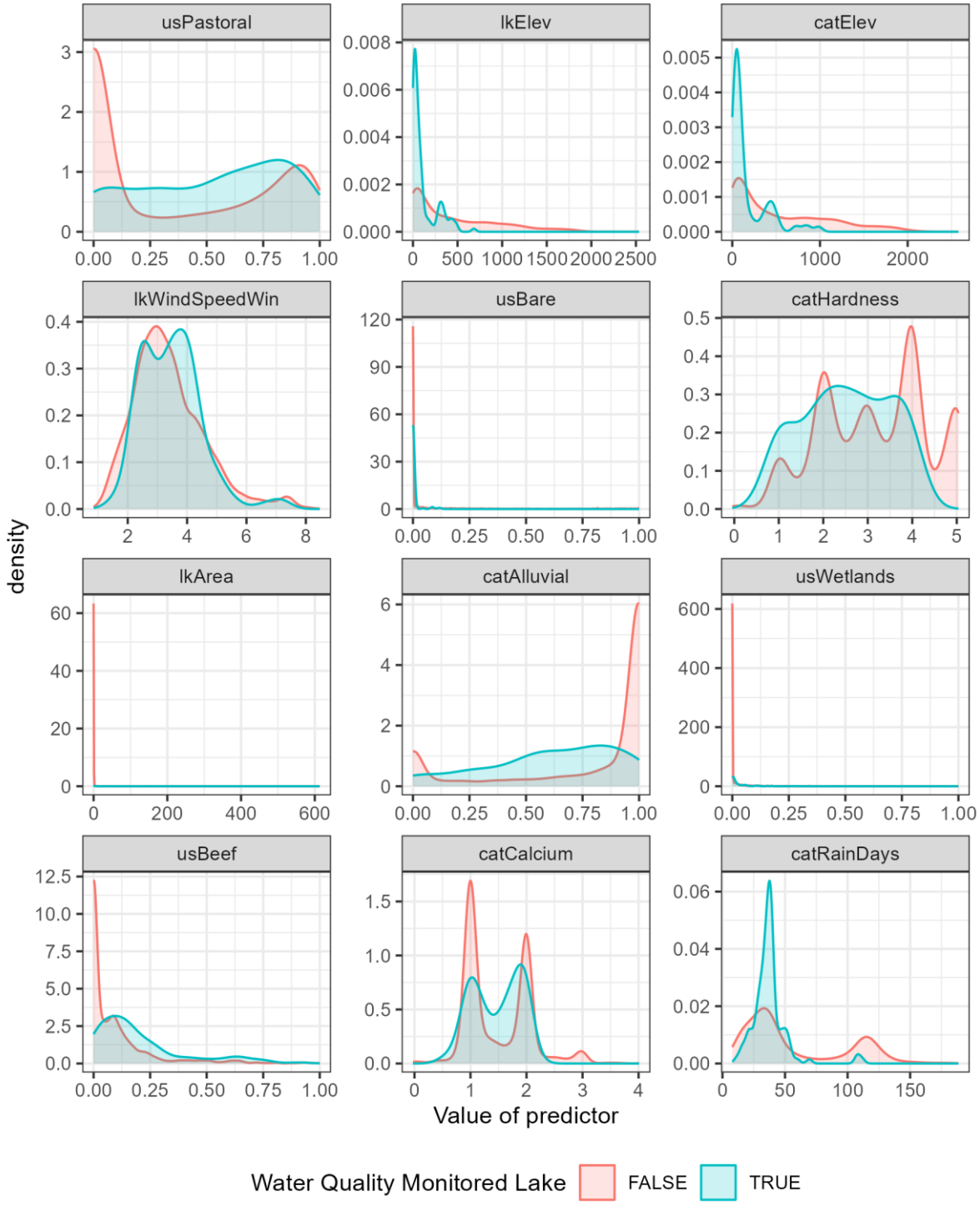


Figure 4-3: The distributions of predictors across monitored and non-monitored lakes in the FENZ database. Similarities in the distributions shown in the two density plots in each panel indicate the degree to which environmental variation across the monitoring sites represents environmental variation across the New Zealand lakes in the FENZ database; exact matches between the density plots would indicate complete representativeness. These 12 predictors were the most important overall predictors in the water quality random forest models and are ordered from most (top left) to least (bottom right) important (Blue is for the monitored lakes and pink is for non-monitored).

4.4 Model predictions

The minimum values predicted by the random forest models were always somewhat larger than the minimum of the observed values and the maximum predicted values were always somewhat smaller than the maximum observed values (Table 4-3). This is an expected outcome of random forest models, which are based on partitioning the observed data, and predictions are derived from a weighted mean of observations that are assigned to a particular partition. As a consequence, the predictions for each attribute state were always within the range of the observations.

Table 4-3: Comparisons of the minimum and maximum observed and predicted values of water quality attribute states.

		Observed Values		Predicted Values	
Attribute State	Units	Minimum	Maximum	Minimum	Maximum
Secchi Median	m	0.11	14.05	0.25	10.66
Chl-*a* Median	mg L^{-1}	0	0.19	0	0.09
Chl-*a* Annual Max	mg L^{-1}	0	0.70	0	0.39
TLI3 Median	Unitless	1.19	7.51	1.54	6.8
TLI4 Median	Unitless	1.46	7.38	1.68	6.52
E. coli Median	cfu or MPN 100 mL^{-1}	0.25	285	0.55	129.25
E. coli 95th percentile	cfu or MPN 100 mL^{-1}	0.62	3480	1.80	1098.82
E. coli % >260	% exceedances	1.75	50.00	2.57	23.07
E. coli % >540	% exceedances	1.75	17.86	2.25	10.09
DRP Median	mg L^{-1}	0	0.33	0	0.07
TP Median	mg L^{-1}	0	1.26	0	0.36
NH_4-N Median	mg L^{-1}	0	0.41	0	0.14
NH_4-N-adj Median	mg L^{-1}	0	1.60	0	0.83
NH4N_adj 95th percentile	mg L^{-1}	0	0.79	0	0.20
NNN Median	mg L^{-1}	0	1.69	0	0.31
TN Median	mg L^{-1}	0.04	4.60	0.06	2.88

The mapped predictions (Figure 4-4 to Figure 4-15) for attribute states describing biological (Chl-*a*, TLI, *E. coli*) and nutrients (NH_4-N, NH_4-N-adj, TN, TP) have similar coarse-scale spatial patterns, with relatively higher values in lower elevation areas. The physical attribute (Secchi) shows the reverse trend, with higher water clarity at higher elevations.

Mapped predictions for *E. coli* exceedances (*E. coli* % >260 and *E. coli* % >540), DRP median, and NNN median are now shown in this report because our models of these attributes exhibited unsatisfactory predictive performance. Unsatisfactory predictive performance indicates that the accuracy of predictions for *E. coli* % >260, *E. coli* % >540, DRP median, and NNN median was expected to be poor. Poor predictive accuracy at low values for *E. coli* % >260 and *E. coli* % >540 was particularly likely due to the exclusion of observed values of zero within models of those two attributes.

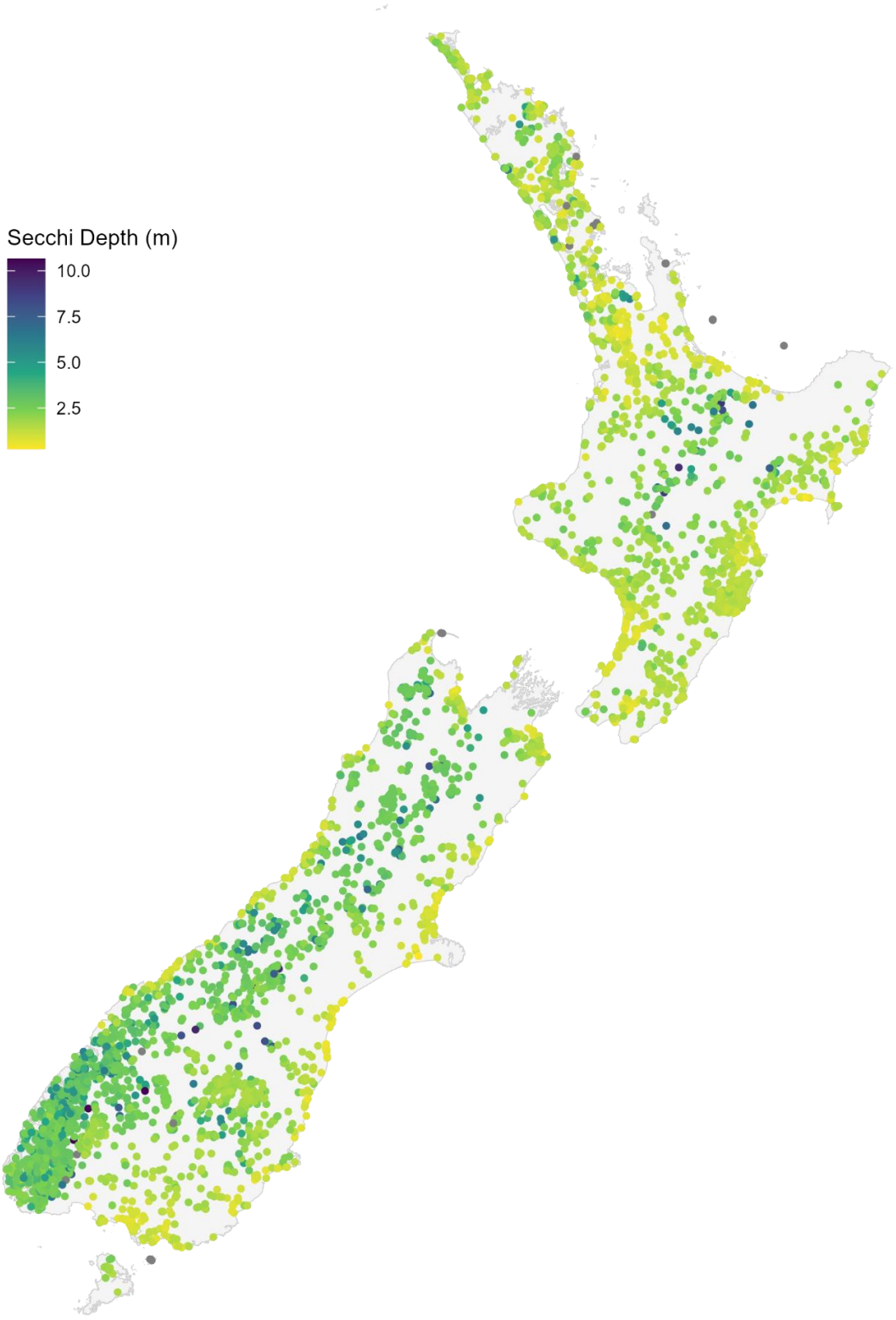


Figure 4-4: Predicted median Secchi depth in New Zealand lakes. Grey dots indicate lakes with missing predictors.

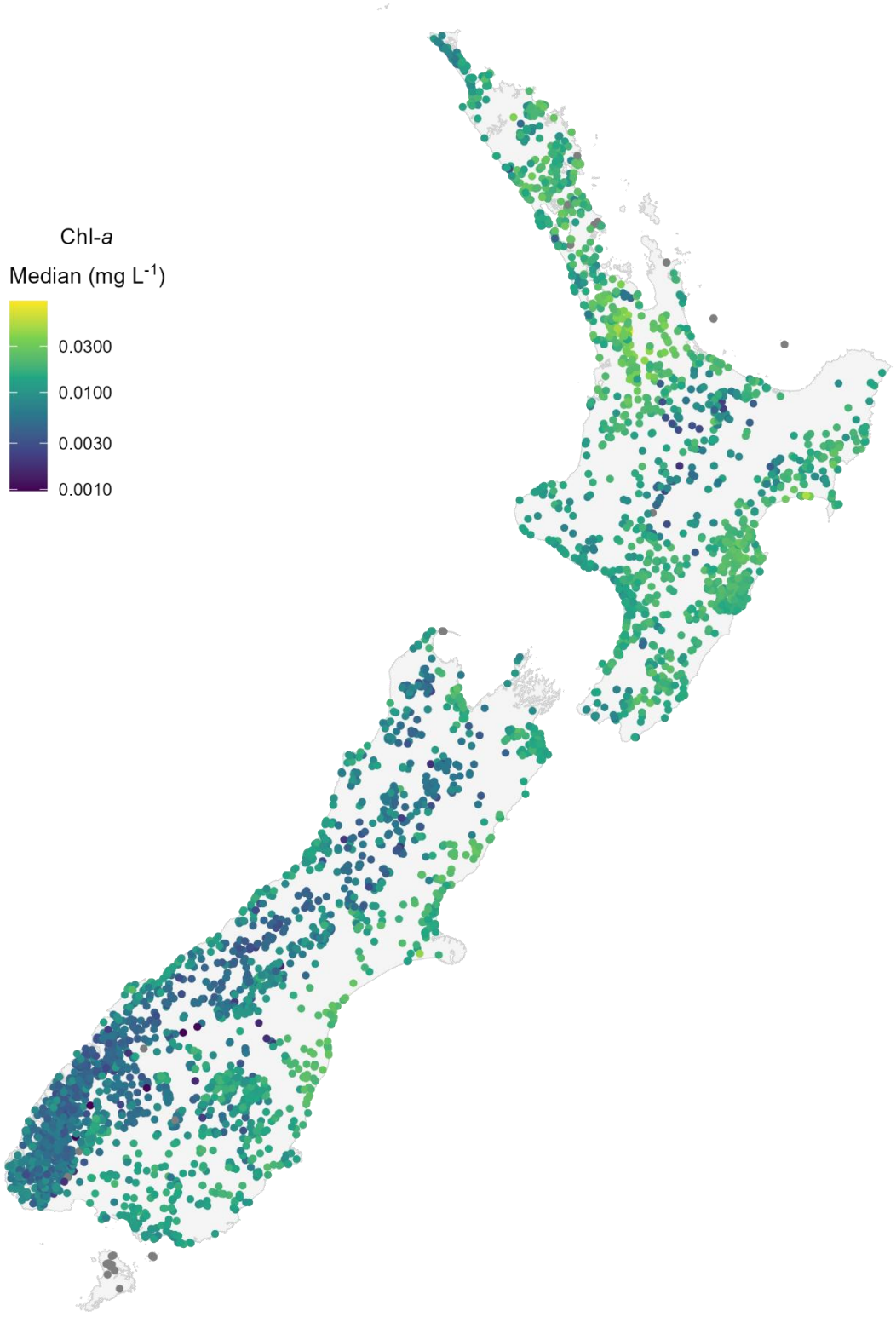


Figure 4-5: Predicted median Chlorophyll *a* concentration in New Zealand lakes. Grey dots indicate lakes with missing predictors.

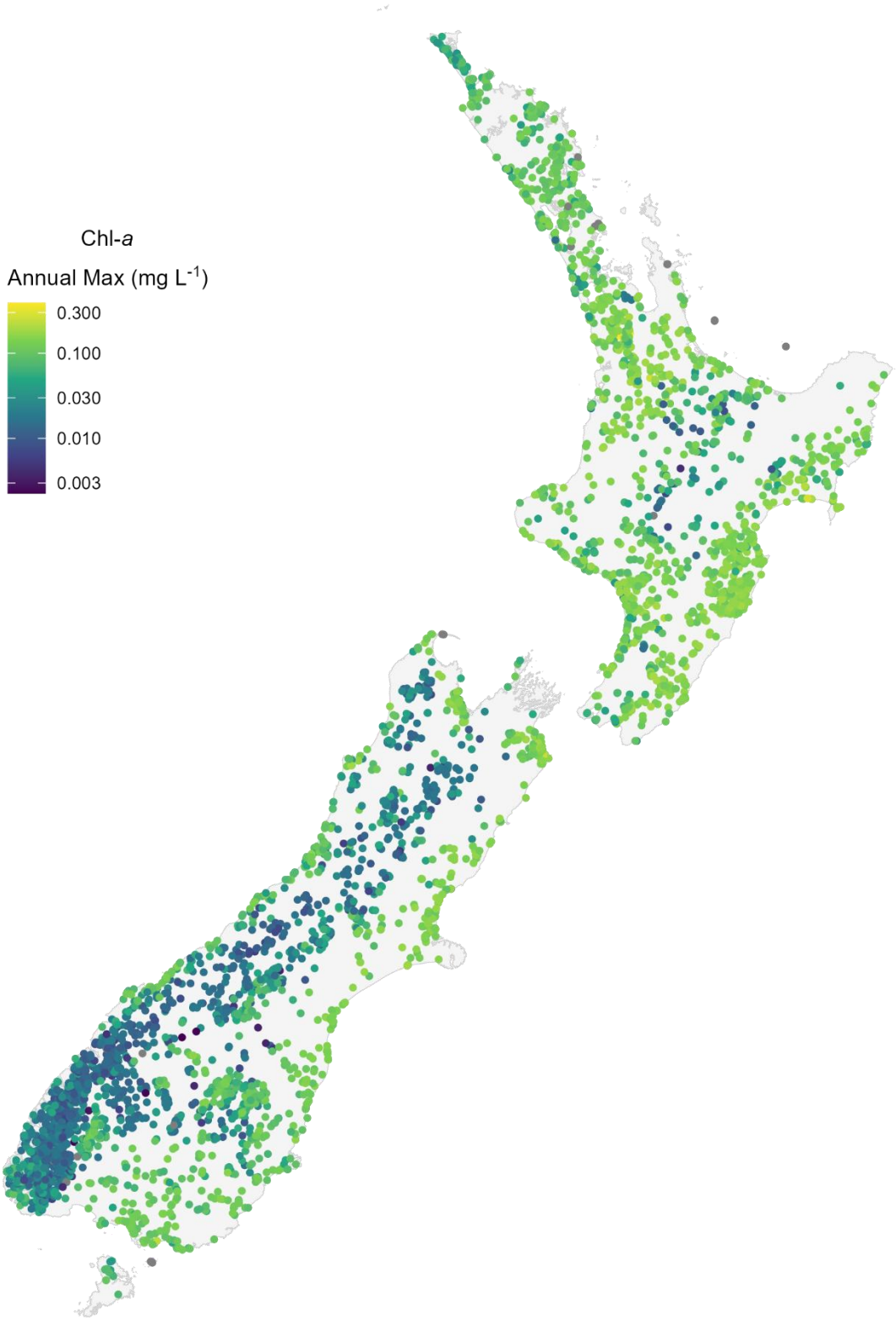


Figure 4-6: Predicted annual maximum Chlorophyll *a* concentration in New Zealand lakes. Grey dots indicate lakes with missing predictors.

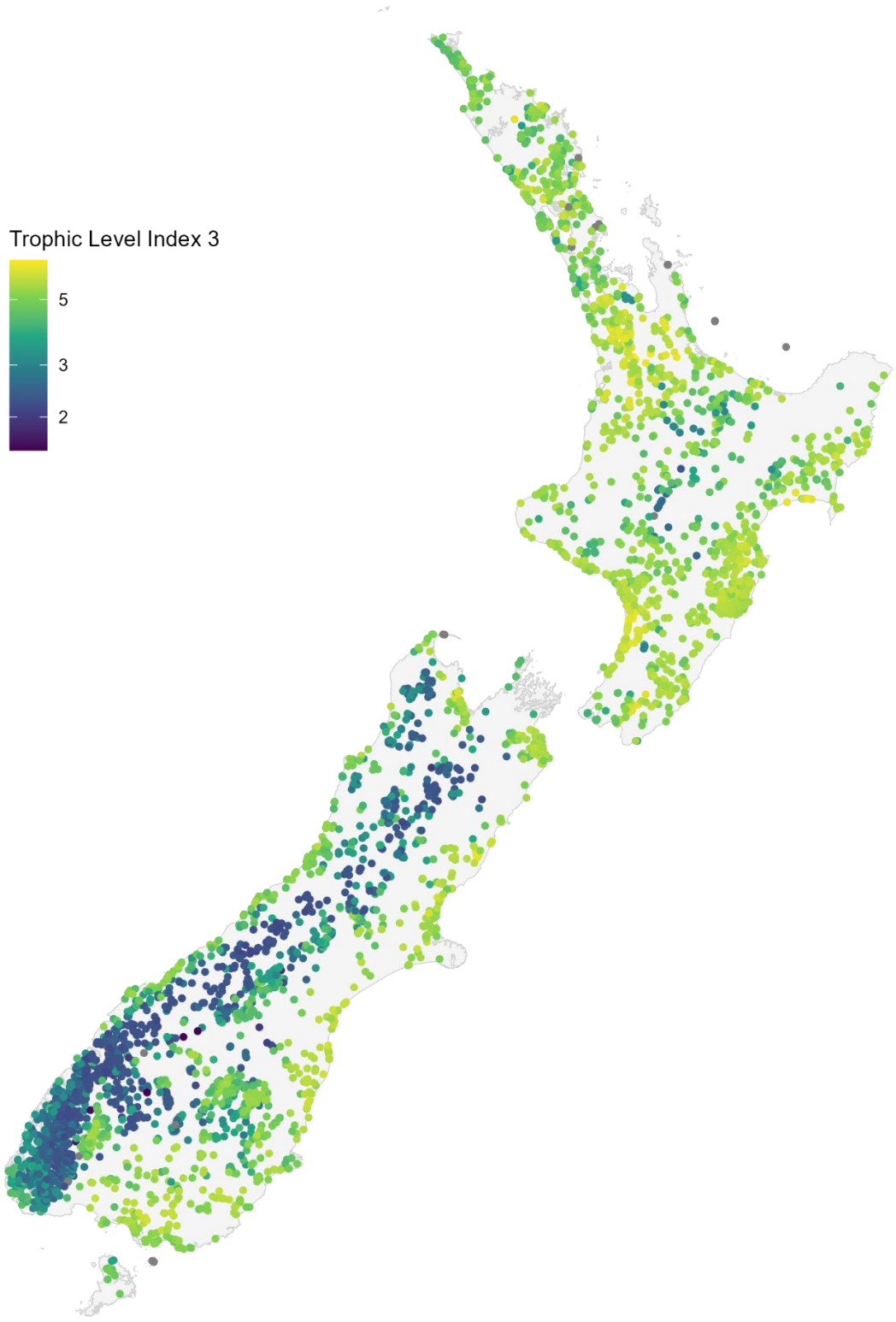


Figure 4-7: Predicted median Trophic Level Index 3 (TLI3) in New Zealand lakes. Grey dots indicate lakes with missing predictors.

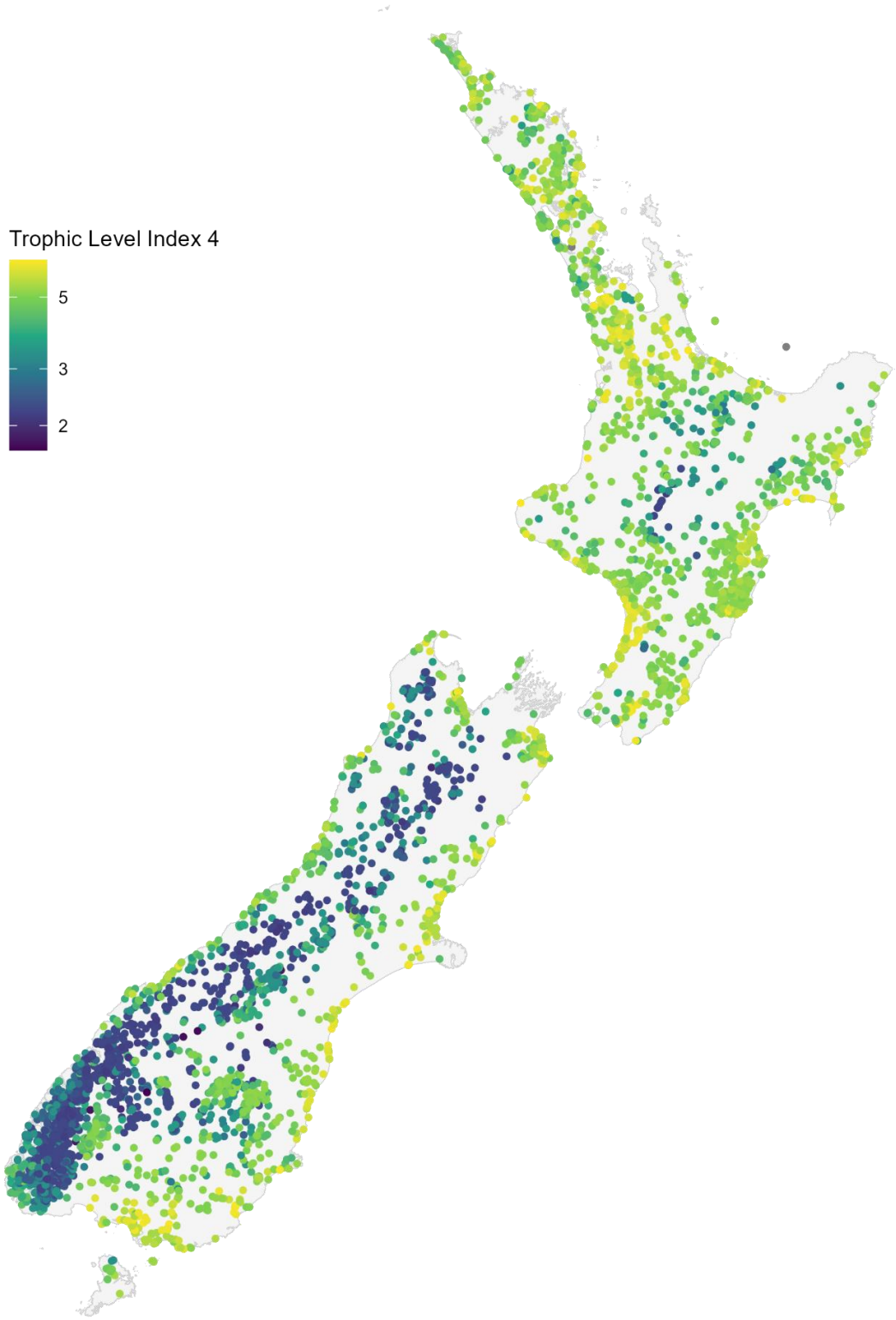


Figure 4-8: Predicted median Trophic Level Index 4 (TLI4) in New Zealand lakes. Grey dots indicate lakes with missing predictors.

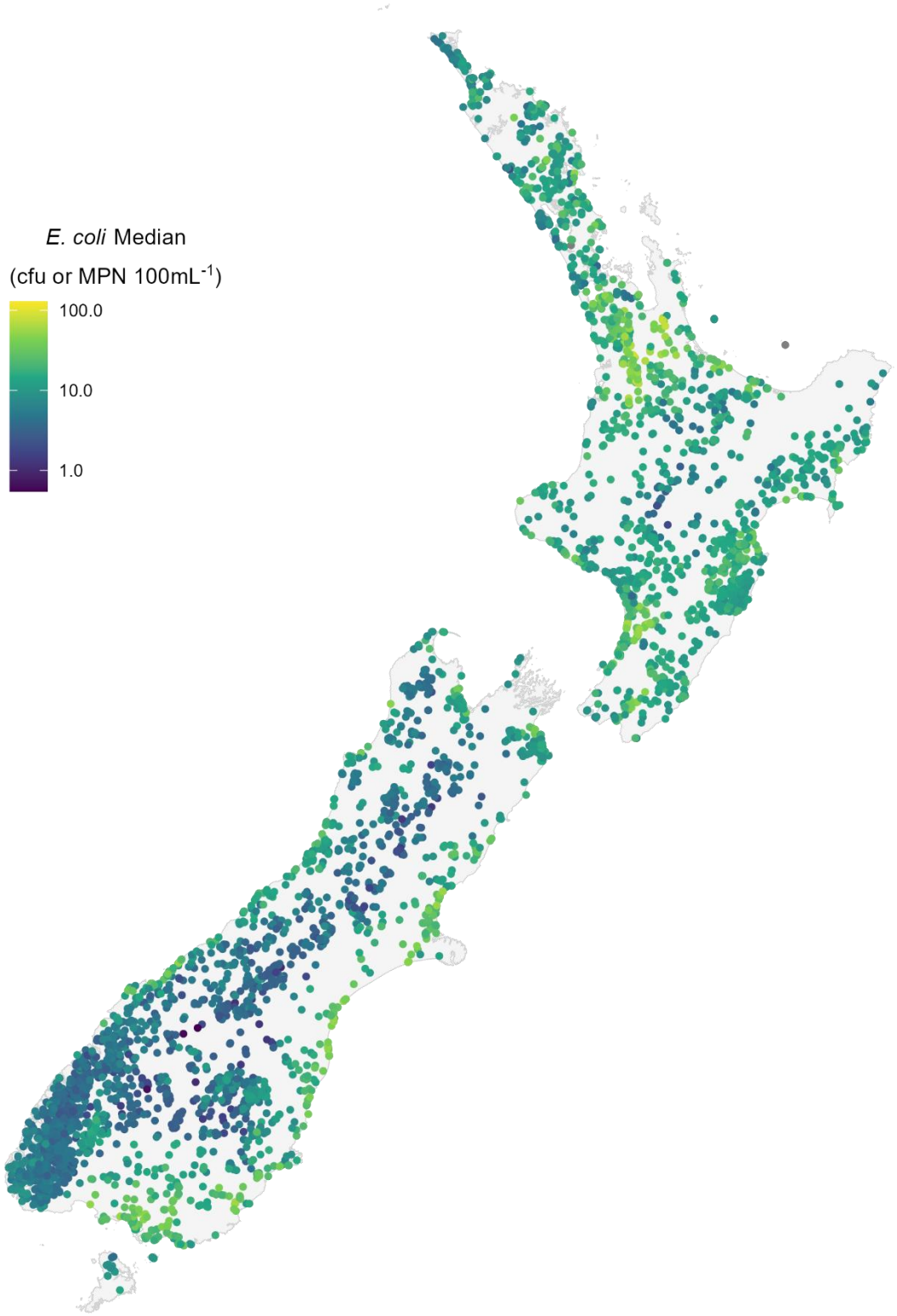


Figure 4-9: Predicted median *E. coli* concentration in New Zealand lakes. Grey dots indicate lakes with missing predictors.

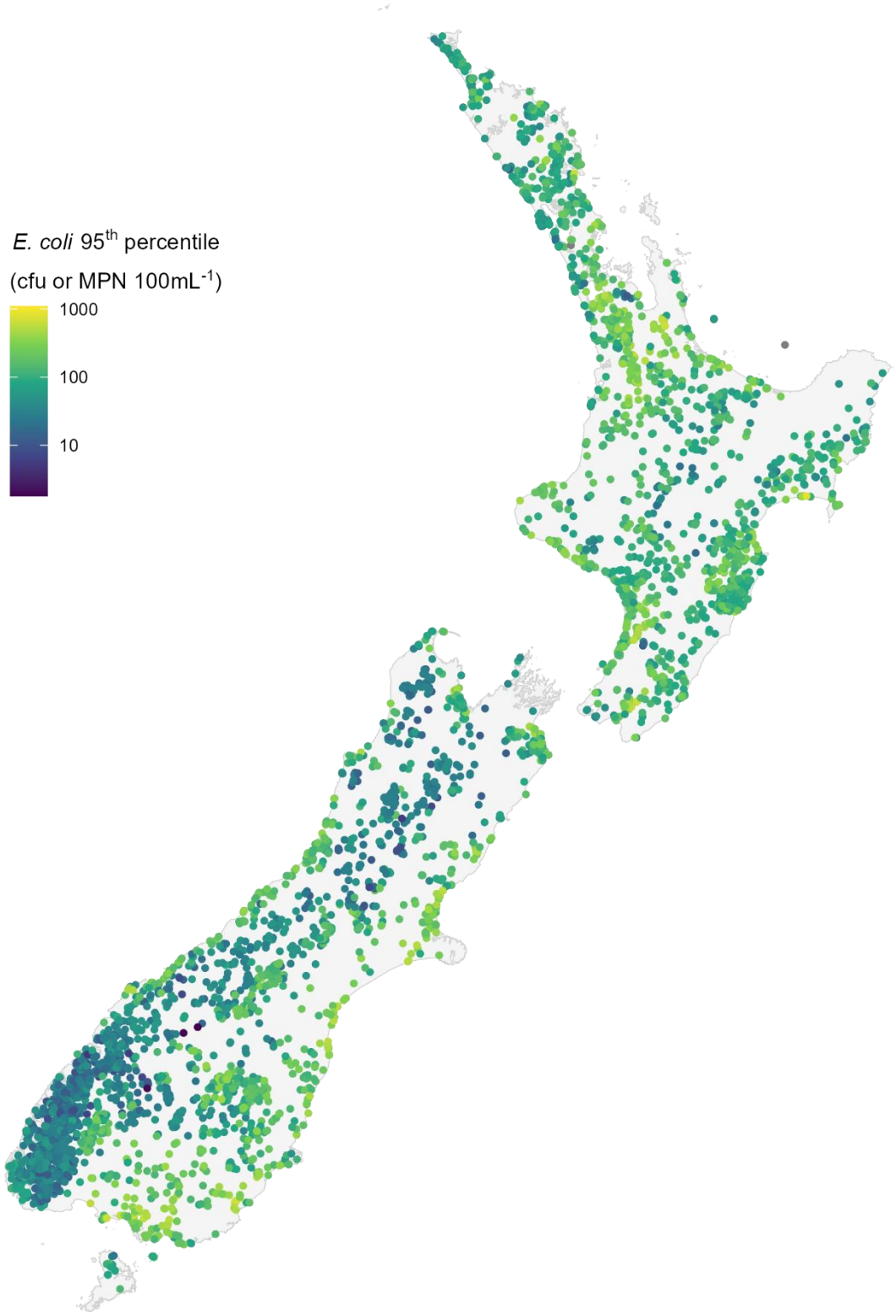


Figure 4-10: Predicted 95th percentile *E. coli* in New Zealand lakes. Grey dots indicate lakes with missing predictors.

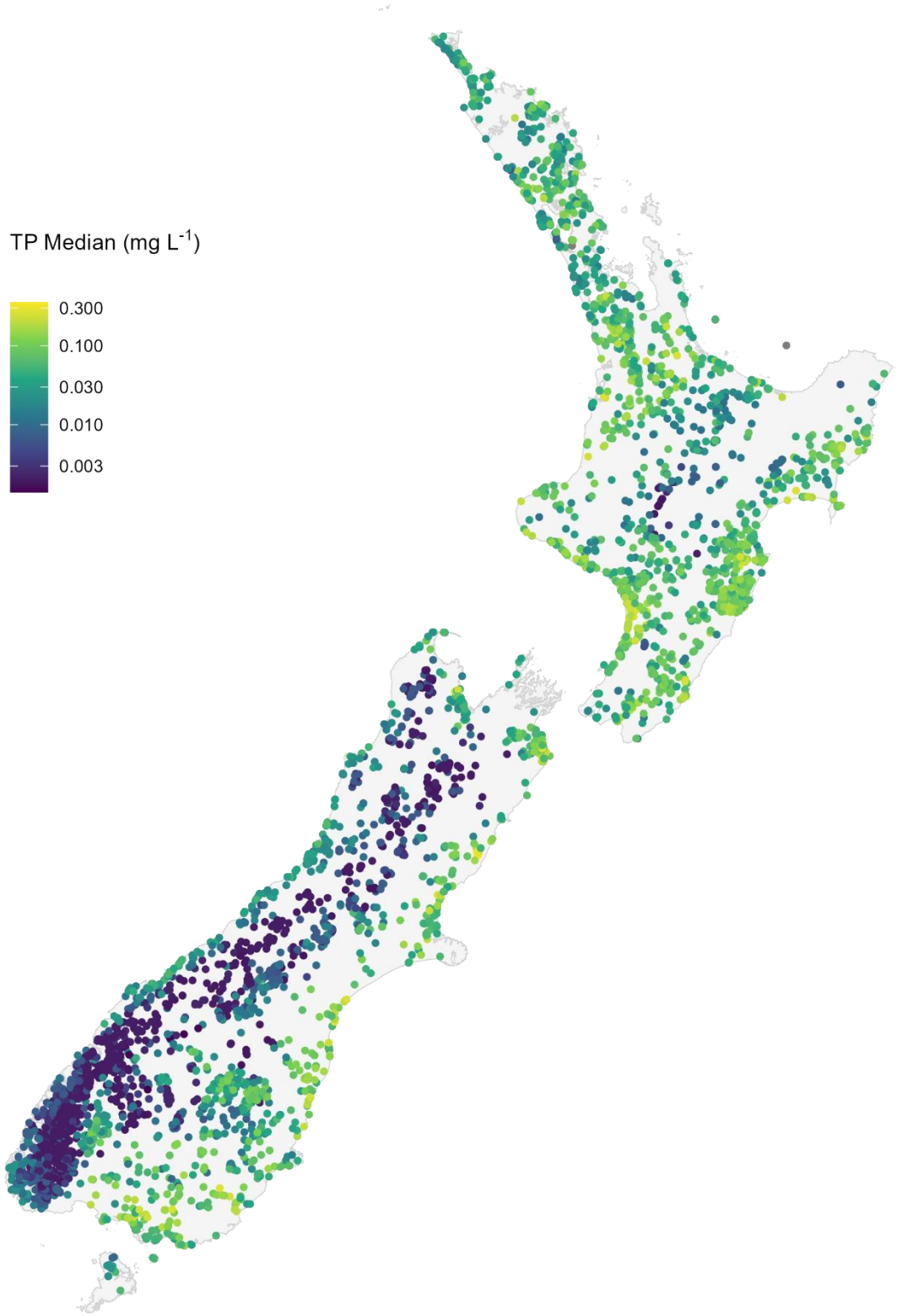


Figure 4-11: Predicted median Total Phosphorus (TP) concentration in New Zealand lakes. Grey dots indicate lakes with missing predictors.

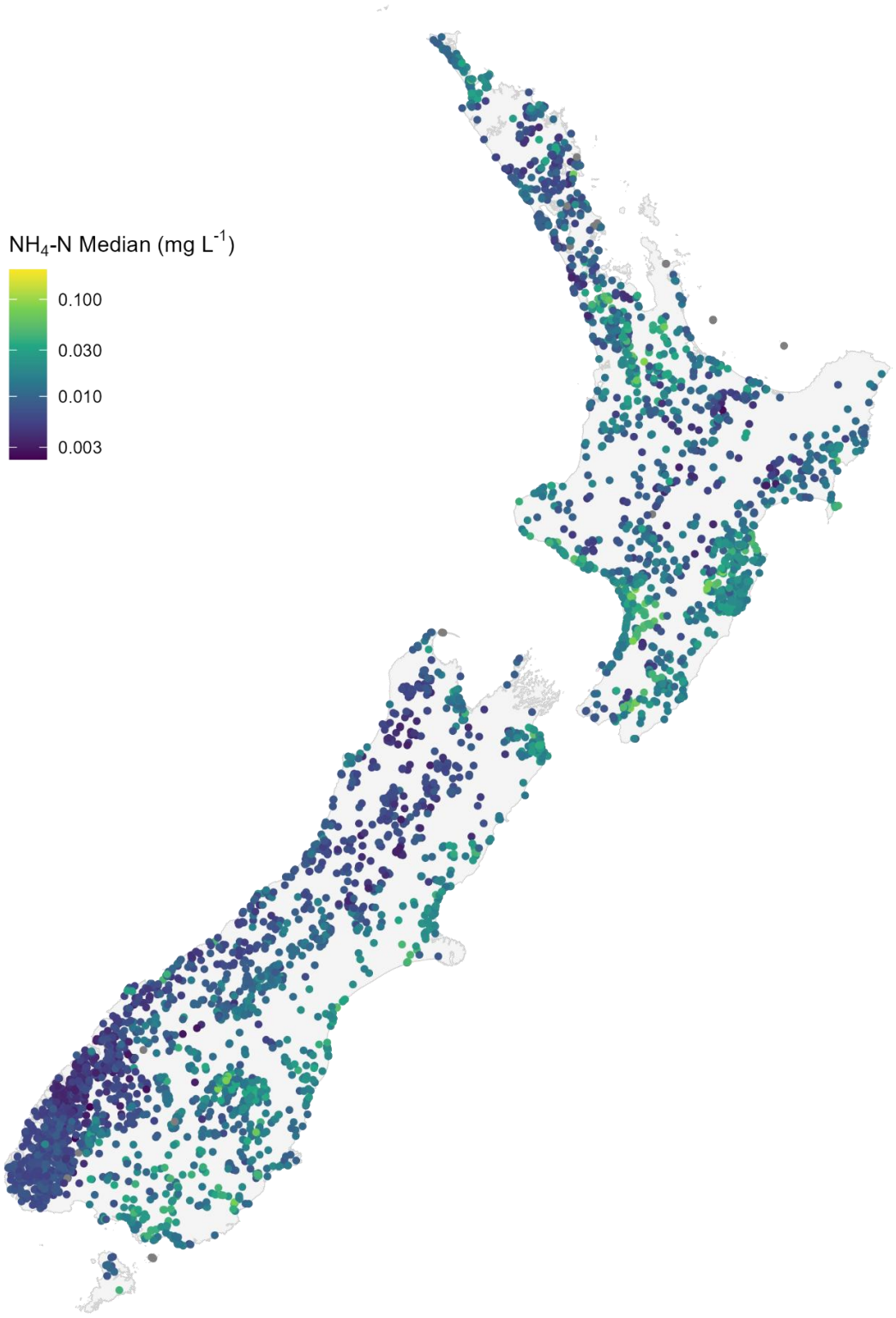


Figure 4-12: Predicted median Ammoniacal nitrogen (NH_4-N) concentration in New Zealand lakes. Grey dots indicate lakes with missing predictors.

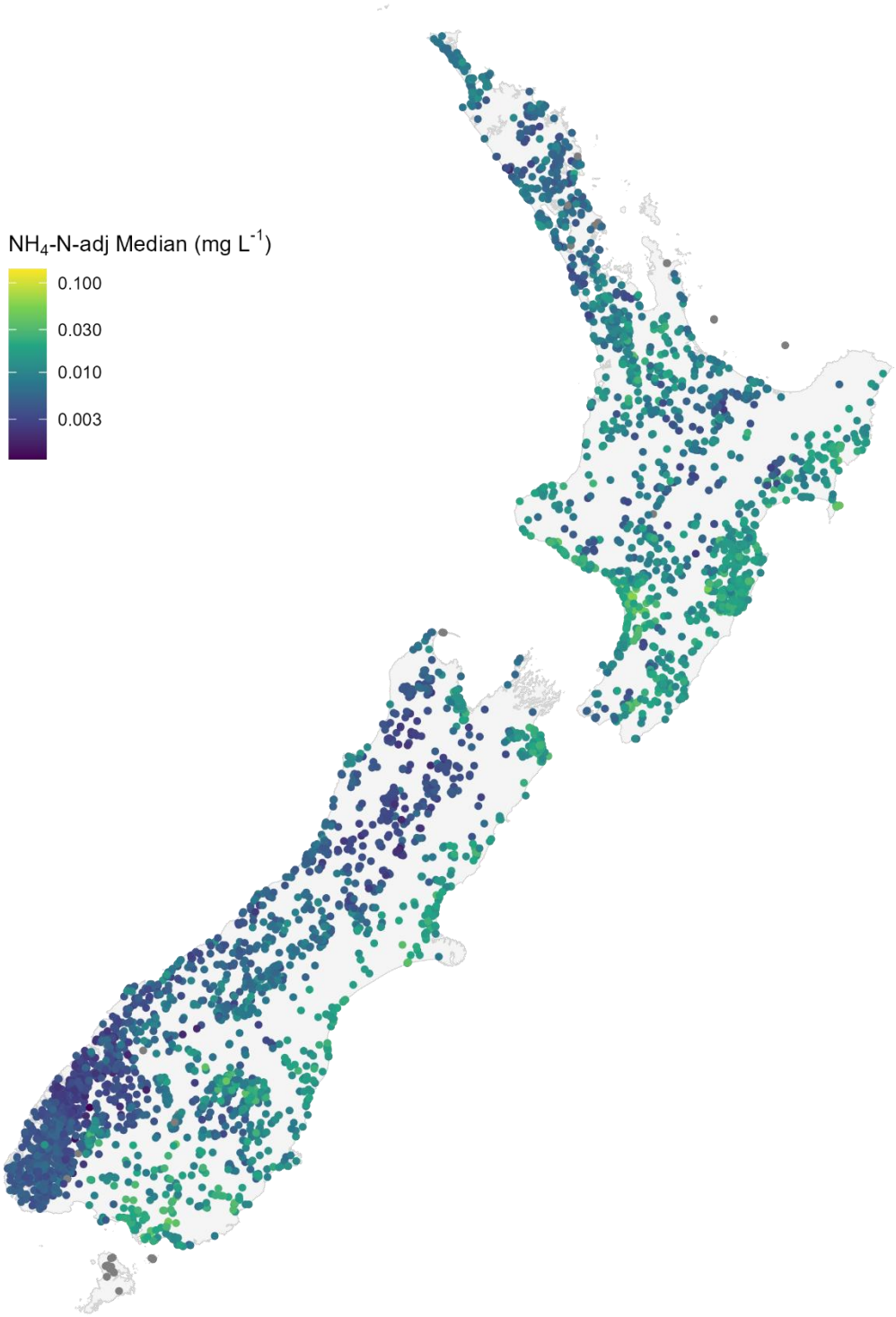


Figure 4-13: Predicted median Ammoniacal nitrogen adjusted for pH (NH_4-N-adj) concentration in New Zealand lakes. Grey dots indicate lakes with missing predictors.

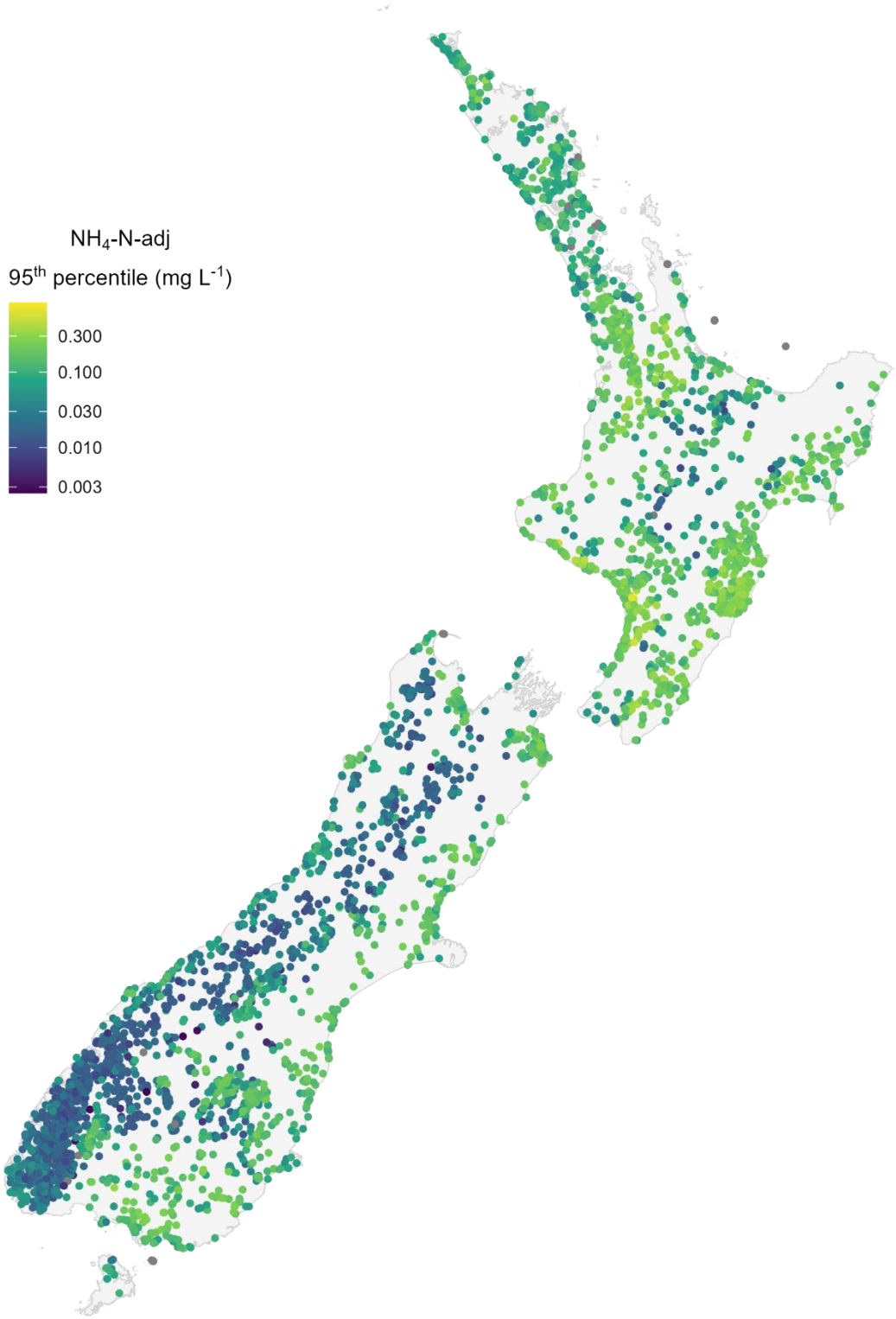


Figure 4-14: Predicted 95th percentile Ammoniacal nitrogen adjusted for pH (NH_4-N-adj) concentration in New Zealand lakes. Grey dots indicate lakes with missing predictors.

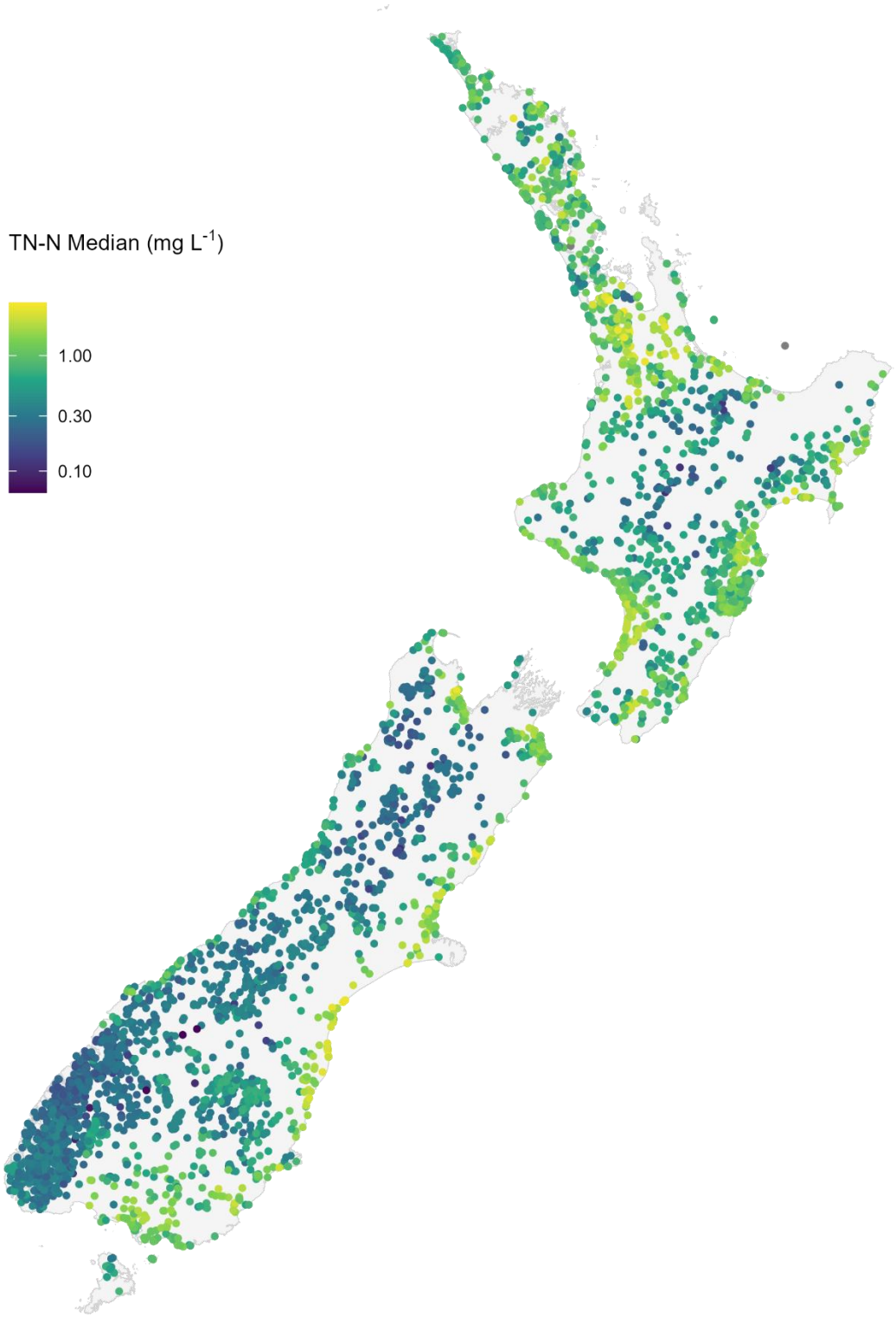


Figure 4-15: Predicted median Total nitrogen (TN) in New Zealand lakes. Grey dots indicate lakes with missing predictors.

5 Discussion

5.1 Comparison with previous studies

The models of the current lake attribute state presented in this study update previous modelling carried out by (Fraser and Snelder 2019; Snelder et al. 2022). These previous reports were based on 2013–2017 and 2016–2020, whilst the current report is based on data from 2020–2024. Broadly speaking the results are generally consistent with those of the earlier studies despite being generated from updated observed and predictor data.

There were similarities in the types of predictors that were important in this and earlier work. In addition, the partial plots showed similar water quality responses with respect to the important predictors, which were similar, though with details of the relationships differed.

The performance of the models in the present case study (as indicated by the ratings) was similar to the study of Snelder et al. (2022) for seven attributes, improved for four attributes (Secchi Median, TLI4 Median, NH_4-N-adj Median, NH_4N-adj 95th percentile) and worse for three (*E. coli* Median, *E. coli* 95th percentile, TP Median). In addition, we have modelled two extra attributes (NNN and DRP), though the random forest performance of these new attributes was unsatisfactory and poor.

The same modelling methodology was applied as previous studies, using similar predictor variables. Though there have been improvements to predictor data. We generated new landcover predictors using 2023 landcover data (LCDB6), instead of earlier versions used in previous reports. In addition, the number of sites has increased as councils improve their environmental monitoring, resulting in an increase in the volume of training data. Improvements in the FENZ data enable us to estimate water quality in more lakes.

5.2 Model uncertainty

In this study, we modelled broad-scale patterns in attribute state using catchment characteristics and lake-scale descriptors as predictors. Because the processes determining lake attribute state are complex, some unexplained variation in our models is to be expected. The R^2 value ranges from <0.01 to 0.72, indicating the random forest models could explain none of the observed variation for *E. coli* % >540 to 72% of the observed variation for the Secchi median attribute. However, the bias was low, indicating the predicted values reflect the overall broad-scale patterns of New Zealand lakes.

Random forest model performance varied across modelled attributes, and this variation may be attributable to differences in the biophysical processes that control distinct aspects of lake water quality. Our catchment-averaged spatial predictor variables may poorly represent some biophysical processes. 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 random forest models. The absence of predictors that account for these and other processes means that some level of unexplained variation is inevitable.

We used the criteria of Moriasi et al. (2015) to categorise the performance of the 16 models as ‘very good’, ‘good’, ‘satisfactory’ and ‘unsatisfactory’ (Table 3-1). These criteria are subjective and are used as an indication of the quality of the predictions. The actual acceptability of the

predictions depends on their use and needs to be considered in the context of each application. The models for *E. coli* % >260, *E. coli* % >540 DRP Median and NNN Median had "unsatisfactory" performance, indicating that the predictions from these models should not be used without applying a high level of caution.

5.3 Alternative modelling approaches

The random forest method that we used to model attribute state is well-suited to the generation of spatial predictions using data from monitoring sites that represent a wide range of environmental conditions. However, it is not the only method available. Alternative statistical models include generalised additive models (GAMs; Hastie et al. 2009), artificial neural networks (Joy and Death 2004), and boosted regression trees (e.g., Elith et al. 2008). We did not employ these alternatives, but it is possible that some water quality applications would be better served by models developed by one of the alternative methods. In particular, if it is important to identify areas with potentially extreme water quality values, models such as GAMS that can extrapolate beyond the range of the fitting data would be useful, although such predictions may lead to spurious and misleading results. Alternative methods for modelling *E. coli* exceedances, which would allow inclusion of all available observed data, include: a) develop a model whose response is classes indicating whether sites had any exceedances within the assessment period; b) application of Generalised Linear Models (GLM) with an appropriately link function to represent proportion data; or c) application of hurdle models that first models whether there is an exceedance and then model the proportion of exceedances.

Models that incorporate biophysical processes (e.g., Elliott et al. 2016) are available. In some circumstances, process models are better suited than machine learning statistical models (e.g., random forests) to inform a particular aspect of environmental policy. We considered random forest models to be a reliable and robust tool for predicting attribute state for national-scale reporting for three general reasons:

1. Spatial data that correspond to land cover and other environmental characteristics are widely available in New Zealand. These data are suitable for investigating associations between water quality and environmental characteristics, and empirical models are appropriate tools for identifying those associations. In contrast, process models require measurements or estimates of catchment processes (e.g., erosion, contaminant transport, and transformation), and these data are far scarcer. In addition, process models are generally more time-consuming and complex to calibrate or parameterise than purely empirical models.

2. Random forest model predictions can be mapped across spatial scales, from individual lakes to the entire country. These maps provide a useful description of spatial patterns in water quality attribute state for environmental reporting purposes.

3. Among empirical modelling methods that generate associations between water quality and environmental characteristics, random forest models have several advantages: they are minimally affected by multi-collinearity among predictor variables, they are robust against over-fitting, they are unaffected by variation in data distributions, and it is possible to use techniques such as a variable

importance plot to assist with model interpretability. Random forest models cannot predict beyond the range of the observations, which may limit their utility in some applications. In the present study, limiting model predictions to the range of observations was a positive attribute as it ensured that those predictions were conservative. See Booker and Whitehead (2018) for further discussion of this topic.

6 Glossary of abbreviations and terms

Ammoniacal nitrogen adjusted for pH (NH_4-N-adj)	The NPS-FM values for ammoniacal nitrogen are set at a pH of 8 and a temperature of 20 °C. The raw values have to be adjusted to allow for comparison.
Ammoniacal nitrogen (NH_4-N)	Ammoniacal nitrogen is a form of nitrogen in water that exists as either ammonia (NH_3) or ammonium (NH_4) and is usually derived from human or animal waste, and at high concentrations can be toxic. NH_4-N is the concentration of ammoniacal nitrogen measured as nitrogen (N).
Attribute states	The combination of water quality variables and statistics (e.g., 95th percentile of dissolved reactive phosphorus).
Chlorophyll *a* (chl-*a*)	Chlorophyll-*a* is the predominant type of chlorophyll used by algae and cyanobacteria. It can therefore be used to measure the quantity of these organisms in a lake.
DEM	Digital Elevation Model.
Digital network	Digital representation of New Zealand rivers.
Dissolved reactive phosphorus (DRP)	Phosphate, which is dissolved and available for aquatic plants and algae growth. High levels of DRP can cause rapid weed growth or algal blooms, which can choke aquatic life.
DN2.4	Version 2.4 of the digital network.
Escherichia coli (*E. coli*)	A bacterium found in the faeces of warm-blooded animals. Indicates the level of faecal contamination and risk to human health due to drinking or accidentally ingesting water contaminated with faecal matter.
FENZ	Freshwater Ecosystems of New Zealand – a database of lakes.
FSL	Fundamental Soil Layers- a national soil database.
monitoring site	Location on the lake where samples are taken for the purpose of monitoring water quality.
Nitrate + nitrite-nitrogen (NNN)	Sum of nitrate-nitrogen (NO_3-N), nitrite-nitrogen (NO_2-N). At high concentrations, it can become toxic to fish and macroinvertebrates. It can also cause excessive algal growth.
NPS-FM	National Policy Statement for Freshwater Management. Contains attribute tables to define and measure the health of freshwater bodies.
random forest	An advanced form of regression-tree modelling that builds many decision trees and combines their predictions to improve accuracy and reduce overfitting.
Secchi Depth	A measure of water clarity, determined by lowering a black and white disk into the water until it disappears from view.
SoE	State-of-the-environment.

Total nitrogen (TN)	The sum of all nitrogen species in water, including nitrate-nitrogen (NO3-N), nitrite-nitrogen (NO2-N), ammoniacal-nitrogen (NH4-N) and organic-nitrogen, measured as nitrogen (N). At high concentrations, it can become toxic to fish and macroinvertebrates. It can also cause excessive algal growth.
Total phosphorus (TP)	The combined amount of all phosphorus forms in water, including both dissolved reactive phosphorus (DRP), which is easily taken up by plants, and phosphorus bound to sediments or particles.
Trophic Level Index	The Trophic Level Index (TLI) is an indicator variable that summarises data related to lake trophic state and potential primary production. The TLI is used to classify New Zealand lakes into trophic classes (e.g., oligotrophic, eutrophic); TLI scores increase with increasing eutrophication. There are two versions of TLI in use in New Zealand, one with three variables (TLI3) and one with four variables (TLI4)
Virtual climate station network (VCSN)	Daily estimates of rainfall and other climate variables across New Zealand

7 Acknowledgements

Many thanks are due to the staff across the various organisations who provided water quality data. We also thank Matt Wilkins for support in providing predictor data, and to Caroline Frazer, who prepared the attribute data.

8 References

Abell, J.M., van Dam-Bates, P., Özkundakci, D., Hamilton, D.P. (2020) Reference and current Trophic Level Index of New Zealand lakes: benchmarks to inform lake management and assessment. *New Zealand Journal of Marine and Freshwater Research*, 54(4): 636–657. 10.1080/00288330.2020.1726974

Booker, D., Snelder, T. (2025) Prefeasibility study for improved spatial lake water quality modelling. *NIWA Client Report*, 2025031CH: 36.

Booker, D., Wilkins, M. (2025) Calculation of lake and lake catchment data layers for objects in the national lake database. Memo prepared for the Ministry for the Environment.: 10.

Booker, D.J., Smith, R.G., Wood, D., Fraser, C.E., Snelder, T.H. (2025) Water quality state and trends in New Zealand lakes; analysis of national data ending in 2024. *ESNZ Client Report*, 2025334CH: 90.

Booker, D.J., Whitehead, A.L. (2018) Inside or Outside: Quantifying Extrapolation Across River Networks. *Water Resources Research*, 54(9): 6983–7003. https://doi.org/10.1029/2018WR023378

Breiman, L. (2001) Random forests. *Machine Learning*, 45(1): 5–32.

Burns, N., Bryers, G., Bowman, E. (2000) Protocol for monitoring lake trophic levels and assessing trends in trophic state. *Client Report*, 99(2).

Burns, N.M., Rutherford, J.C., Clayton, J.S. (1999) A Monitoring and Classification System for New Zealand Lakes and Reservoirs. *Lake and Reservoir Management*, 15(4): 255–271. 10.1080/07438149909354122

Cutler, D.R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J. (2007) Random forests for classification in ecology. *Ecology*, 88(11): 2783–2792. https://doi.org/10.1890/07-0539.1

Elith, J., Leathwick, J.R., Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4): 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Elliott, A.H., Semadeni-Davies, A.F., Shankar, U., Zeldis, J.R., Wheeler, D.M., Plew, D.R., Rys, G.J., Harris, S.R. (2016) A national-scale GIS-based system for modelling impacts of land use on water quality. *Environmental Modelling & Software*, 86: 131–144. 10.1016/j.envsoft.2016.09.011

FENZ (2024) Freshwater Ecosystems of New Zealand (FENZ) 'Lakes' November 2024. In: MfE (Ed).

Fraser, C., Snelder, T. (2019) Spatial modelling of lake water quality state: Incorporating monitoring data for the period 2013 to 2017, LWP Client Report Number: 2018-16.

Greenwell, B.M. (2017) pdp: An R package for constructing partial dependence plots. *The R Journal*, 9: 421–436. https://doi.org/10.32614/RJ-2017-016

Hastie, T., Tibshirani, R., Friedman, J.H. (2009) *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York. https://go.exlibris.link/C2wM0hC2

Joy, M.K., Death, R.G. (2004) Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. *Freshwater Biology*, 49(8): 1036–1052. https://doi.org/10.1111/j.1365-2427.2004.01248.x

Kuczynski, A., Smith, R.G.R., Fraser, C.E., Larned, S.T. (2024) Environmental indicators of lake ecosystem health in Aotearoa New Zealand: current state and trends. *Ecological Indicators*, 165: 112185. https://doi.org/10.1016/j.ecolind.2024.112185

Larned, S., Snelder, T., Unwin, M. (2016) Water quality in New Zealand rivers: Modelled water quality state. *NIWA Client Report*, CHC2016-070: 40. Q:\LIBRARY\ClientRept\E-copies Client reports\CHRISTCHURCH

Liaw, A., Wiener, M. (2002) Classification and Regression by randomForest. *R news*, 2(3): 18–22. https://CRAN.R-project.org/doc/Rnews/

Moriasi, D.N. (2007) Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Transactions of the ASABE*, 50(3): 885–900. 10.13031/2013.23153

Moriasi, D.N., Gitau, M.W., Pai, N., Daggupati, P. (2015) Hydrologic And Water Quality Models: Performance Measures And Evaluation Criteria. *Transactions of the ASABE*, 58(6): 1763–1785. <Go to ISI>://WOS:000368341900024

Nash, J.E., Sutcliffe, J.V. (1970) River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3): 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Pebesma, E., Bivand, R. (2023) *Spatial data science: With applications in R*. Chapman and Hall/CRC.

Piñeiro, G., Perelman, S., Guerschman, J.P., Paruelo, J.M. (2008) How to evaluate models: Observed vs. predicted or predicted vs. observed? *Ecological Modelling*, 216(3): 316–322. https://doi.org/10.1016/j.ecolmodel.2008.05.006

Probst, P., Boulesteix, A.-L. (2018) To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, 18(181): 1–18.

Snelder, T., Fraser, C., Whitehead, A. (2022) Spatial modelling of lake water quality state: Incorporating monitoring data for the period 2016 to 2020, LWP Client Report Number: 2021-15.

Svetnik, V., Liaw, A., Tong, C., Wang, T. (2004) Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. *Multiple Classifier Systems*, Berlin, Heidelberg, 2004//.

Unwin, M., Snelder, T., Booker, D., Ballantine, D., Lessard, J. (2010) Predicting water quality in New Zealand rivers from catchment-scale physical, hydrological and land use descriptors using random forest models. *NIWA Client Report*, CHC2010-037: 50.

Verburg, P., Hamill, K., Unwin, M., Abell, J. (2010) Lake water quality in New Zealand 2010: status and trends. *NIWA Client Report*, HAM2010-107: 49.

Whitehead, A. (2018) Spatial modelling of river water-quality state Incorporating monitoring data from 2013 to 2017, 2018360CH: 41.

Whitehead, A., Fraser, C., Snelder, T. (2022) Spatial modelling of river water quality state: Incorporating monitoring data from 2016 to 2020. *NIWA Client Report*, 2021303CH: 47.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D.A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H. (2019) Welcome to the tidyverse. *Journal of Open Source Software*, 4(43): 1686.