Statistical models, indicators and trend analyses for reporting national-scale river water quality)

(NEMAR Phase 3)

Prepared for Ministry for the Environment

May 2013

Authors/Contributors:
M J Unwin
S T Larned

For any information regarding this report please contact:
Scott Larned

+64-3-343 3834
scott.larned@niwa.co.nz

National Institute of Water & Atmospheric Research Ltd
10 Kyle Street
Riccarton
Christchurch 8011
PO Box 8602, Riccarton
Christchurch 8440
New Zealand

Phone +64-3-348 8987
Fax +64-3-348 5548

NIWA Client Report No: CHC2013-033
Report date: April 2013
NIWA Project: MFE13502

Contents

Tables

Figures

Reviewed by

Graham McBride

Approved for release by

Jochen Schmidt

Summary

The New Zealand Ministry for the Environment (MfE) is required to report on river water quality and ecological state, and on temporal trends in these metrics. In 2011 MfE initiated the National Environmental Monitoring and Reporting (NEMaR) project, which aims to establish more consistent and dependable monitoring procedures for national reporting. This report concerns three components of NEMaR:

1. compiling national datasets on water quality and stream macroinvertebrate communities;
2. developing predictive models relating water quality state and trends to metrics of catchment-scale geography, climate, geology, and land cover; and
3. trialling the use of composite indices, based on selected water quality and invertebrate community metrics, as a tool for characterising national-scale environmental variation.

We assembled data from regional council State of Environment and NIWA's National River Water Quality Network programmes into two parallel databases, representing water quality and macroinvertebrate communities, respectively. Water quality data were screened to ensure consistency of measurement procedures and units, and to remove gross outliers. Variables used for this report were water clarity (CLAR), suspended sediment (SS), and turbidity (TURB); temperature (TEMP); dissolved oxygen concentration (DO) and percent saturation (DOSAT); *Escherichia coli* concentration (ECOLI); and five measures of dissolved nutrients including total nitrogen (TN) and total phosphorus (TP).

Invertebrate data were pre-processed to ensure taxonomic resolution and counting procedures were consistent across all data sets, and used to generate four invertebrate community metrics. These were total number of taxa (TAXA); number of taxa from the insect orders Ephemeroptera, Plecoptera and Trichoptera (EPTtaxa); the percentage of EPT individuals (%EPTabund); and the Semi-Quantitative Macroinvertebrate Community Index for hard-bottom streams (SQMCI-hb). We also calculated a suite of river condition indices which combine multiple water quality metrics into a single index, and have recently been trialled in New Zealand at regional scales. These were the Canadian Council of Ministers for the Environment water quality index (CCME WQI), the Victorian Index of Stream Conditions (VISC-WQI), and various composite indicators based on combinations of the CCME-WQI, VISC-WQI, and SQMCI-hb.

The resulting data set included 789 water quality sites and 519 invertebrate sites. We obtained catchment-level descriptors for all sites by selecting relevant variables from two classification systems which are commonly used to classify New Zealand rivers: the River Environment Classification (REC), and the Freshwater Environments of New Zealand (FWENZ). We used data collected since 1 January 2006 to characterise current state, but used all available data from 1 January 2000 for trend analysis so as to provide more robust estimates of trend strength and direction.

We used random forests, a form of multivariate regression, to model 50 water quality variables (log-transformed where appropriate), indices, and trends using 28 site-specific catchment descriptors as predictor variables. We then used these models to predict each

variable for all mainland New Zealand rivers represented by the REC at a mean spatial scale of 740 m.

The most successful models, for TN, ECOLI, TP, and SQMCI-hb, explained over 70% of the observed site-to-site variation, with another five models (TEMP, CLAR, TURB, nitrate nitrogen (NO3N), EPTtaxa) explaining 65%-69% of observed variation. Leading predictors varied among models, but generally included measures of catchment topography (e.g., elevation, mean slope); climate (e.g., rainfall variability, mean temperature); and catchment land-cover (particularly the percentage of the catchment covered by indigenous forest or heavy pastoral agriculture). Predictions based on these models yielded credible representations of regional-scale variation in water quality, which was predicted to be highest in elevated catchments along the main axis of both islands, intermediate in more impacted hill-country or lowland catchments, and lowest in intensively developed lowland areas such as Waikato, Manawatu, Canterbury, and Southland.

Trends in water quality variables since 2000 suggest that water clarity has declined (and turbidity has increased) in Waikato; that NO3N has increased in Waikato and Southland; and that both NO3N and TP have decreased in the lower North Island. Trends in other variables (e.g., ECOLI, DRP) were less consistent, with increasing and decreasing trends often apparent at neighbouring sites in the same region. RF models with trend strength as the dependent variable generally performed poorly, with none explaining more than 42% of observed site to site variation, and most explaining less than 20%. However, our analyses were confounded by limited availability of data, and by the large number of sites for which no significant trend was apparent. We therefore caution readers that the trend modelling results should be interpreted as illustrating the difficulty of obtaining credible fits rather than providing information for State of the Environment reporting.

Similar comments apply to our analysis of trends in invertebrate community metrics, for which the available data at each site were limited to a 10-year time series of annual sample data. Consistent regional trends were sometimes apparent for metrics related to the number of invertebrate taxa present, but RF models for these trends did not yield credible results and their relevance to variation in water quality is unclear. Large-scale spatial patterns in predicted trend direction were apparent in some regions, but bore little if any relationship to climatic or environmental gradients.

Results for the CCME WQI, VISC-WQI, and related composite indicators, were also problematical. Applying these indices nationally, as opposed to the region-based contexts in which they were previously trialled, involved making numerous subjective choices driven by arbitrary factors such as data availability; choice of variables on which each index was to be based; the time period covered by each index; the reference conditions used to specify threshold values; and regional variation in reporting procedures. CCME WQI calculations were particularly hampered by regional variation in the water quality variables measured, requiring a trade-off between more robust indices incorporating a large subset of variables but with very limited spatial coverage, and less robust indices with broader spatial coverage.

Index scores tended to be highly clumped, with scores at most sites falling within a relatively narrow range. This behaviour is partly driven by the choice of reference conditions, which – for the CCME WQI – were so stringent that most sites scored below 45 (ranked as “poor”) on a scale of 0-100. However, all index calculations involve multiple levels of averaging, so that

sites tend to clump together in all but the most extreme cases. RF models gave plausible fits (>50% explained site-to-site variation) for three of the six indices we investigated, but performed worse when used to predict index scores for all rivers represented by the REC. All predictions suggested a tendency for water quality to be highest in upland catchments and lowest in lowland catchments, particularly those dominated by heavy pastoral agriculture. These results are consistent with models for the individual water quality variables used to calculate each index, but add no further insight into regional-scale trends.

The results of this study have significant implications for the NEMaR project. First, our findings confirm that viable multivariate models can be developed for predicting the current state of commonly used water quality variables and invertebrate community metrics. A priority for the future is to reduce prediction uncertainty, and hence improve stakeholder uptake of results, by establishing new monitoring sites to fill gaps in the environmental gradients used for modelling, and also by optimising the model fitting process. Options for achieving this include identifying optimal predictor sets for each water quality variable; trialling new or updated predictors incorporating recent revisions to the REC and LCDB; and exploring alternatives to RF for model-fitting.

Second, the water quality data sets available for this study appear to be of limited value for predicting long term water quality trends at a national scale. Significant trends were apparent for some variables when analysed at local and regional scale, but the data did not yield credible models when used to predict trends at national scale. Possible reasons for this result include the limited number of sites available for long-term trend analysis; the paucity of trends at these sites that were both significant and meaningful; and the potential for regional-scale trends to be confounded by regional variation in management practices. It may also reflect our reliance on explanatory variables which potentially change with time (e.g., catchment land cover), but for which the available descriptors were measured at a single point in time. In particular, we predicted monotonic changes based on explanatory variables that are either constant (e.g., altitude), or change over temporal scales much longer than the 10-year analysis period (e.g., annual rainfall). A possible remedy would be to use data from LCDB1 (1996-1998) and the recently created LCDB3 (2008-2009) to develop predictors more accurately representing contemporary changes in land-cover.

Third, considerable further work is necessary before multi-metric and composite water-quality indices can be consistently generated and interpreted at a national scale. Our experience with the CCME WQI and VISC-WQI shows them to be highly context-dependent, with values that depend strongly on multiple decisions made during the analysis process. These decisions include the choice of variables to be used for index calculations; the period for which the index is required; and the reference conditions used to define baseline values. Indices also tend to dampen rather than enhance variation among sites, because their calculation typically involves multiple levels of averaging (e.g., averages of medians). Consequently, they tend to perform poorly, relative to their individual components, when used to characterise national variability.

Many of the problems associated with calculating indices and modelling temporal trends were caused by data limitations rather than the nature of the indices or trends themselves. Our calculations were often based on minimal subsets of variables and sites, due to the lack of consistent data for all required variables across sites, compromising our ability to characterise and analyse aquatic conditions at the national scale. If, in the future, water-

quality indices can be calculated using multiple variables, consistently measured at sites that adequately span the entire range of river environments in New Zealand, we would almost certainly obtain more tractable results. A major goal of the NEMaR project is to ensure that better and more representative datasets will become available in the future.

1 Introduction

As part of its National Environmental Reporting Programme, the New Zealand Ministry for the Environment (MfE) reports on the water quality and ecological state of rivers, and on temporal trends in river water quality and ecological conditions. In previous reports, the core variables used for assessments were dissolved and total nutrient concentrations (e.g., nitrate-nitrogen), indicator bacteria concentrations (e.g., *Escherichia coli*), physical attributes (e.g., water clarity), and metrics that describe aquatic invertebrate communities (e.g., Macroinvertebrate Community Index).

In most of the previous MfE reports on state and trends in river water quality and ecological condition, the analyses were based on data from 77 monitoring sites in the National River Water Quality Network (NRWQN), or from the aggregated network of several hundred sites used for regional and district council monitoring programmes, plus the NRWQN sites (e.g., Ballantine *et al.* 2010, Ballantine & Davies-Colley 2010). Results of analyses of NRWQN sites alone are generally reported on a site-by-site basis, with some comparisons between paired sites intended to represent impaired and baseline conditions. Results of analyses of the aggregate monitoring network are generally extrapolated to the river segments in several environmental classes, with each class represented by multiple monitoring sites. Two classification systems have been used to classify river reaches for extrapolation: the River Environment Classification (REC) and the Freshwater Environments of New Zealand (FWENZ). Both systems group river reaches on the basis of environmental attributes such as climate, geology, and land-cover/land-use. The implicit assumption is made that water quality and ecological state/ trends will be similar among the river reaches in a given class; this assumption is the basis for extrapolation from monitoring sites.

MfE's National Environmental Monitoring and Reporting (NEMaR) project commenced in 2011, with the goal of establishing consistent and dependable monitoring of New Zealand's freshwater resources as a foundation for national reporting. NEMaR has several broad objectives, including trialling alternative approaches for river monitoring, data analysis, and reporting on state and trends. Within these objectives, several tasks were identified that are the topics of this report:

4. Trial the use of water quality, macroinvertebrate and multi-metric indices as an alternative to the core variables listed above.
5. Trial the use of statistical models to extrapolate water quality and ecological conditions and trends from monitoring sites to unmonitored reaches as an alternative to classification-based extrapolation;
6. Trial a combination of Tasks 1 and 2, in which statistical models are used to extrapolate water quality, macroinvertebrate and multi-metric indices and trends in those indices to unmonitored reaches.

The current report relates to Deliverables 1-6 of NEMaR Phase 3, as follows:

1. Compile a national dataset of physical and chemical water quality variables (TN, NO3-N, NH4-N, TP, DRP, *E. coli*, clarity, temperature, DO, TSS, turbidity).

2. Produce random forest models of each physical and chemical water quality variable for all river reaches in New Zealand using the national dataset.
3. Estimate values of two water quality indices for all river reaches in New Zealand using output from the random-forest models in Deliverable 2 and an existing dataset on physical-chemical modelled reference conditions.
4. Estimate values of multi-metric (composite) indices for all river reaches in New Zealand using the output from Deliverable 3 and modelled current state and reference state macroinvertebrate data.
5. Analyse trends in macroinvertebrate indices and extrapolate to all river reaches in New Zealand using random forest models.
6. Analyse trends in physical and chemical water quality variables and extrapolate to all river reaches in New Zealand using random forest models.

2 Methods

2.1 Compiling monitoring data

The dataset used in this report consists of physical-chemical water quality and macroinvertebrate data from the regional council State of Environment (SoE) and National River Water Quality Network (NRWQN) programmes. The physical-chemical water quality variables are water clarity (CLAR), *Escherichia coli* concentration (ECOLI), nitrate-nitrogen concentration (NO3N[1]), ammoniacal nitrogen concentration (NH4N), total nitrogen concentration (TN), dissolved reactive phosphorus concentration (DRP), total phosphorus concentration (TP), temperature (TEMP), dissolved oxygen concentration (DO[2]) and percent saturation (DOSAT), suspended sediment (SS), and turbidity (TURB). The invertebrate variables are taxa lists and counts or coded-abundance classes for each taxon. The raw invertebrate data were post-processed to generate four variables: total number of taxa in a sample (TAXA), the number of taxa from the insect orders Ephemeroptera, Plecoptera and Trichoptera (EPTtaxa), the percentage of individuals in a sample from EPT taxa (%EPTabund), and the Semi-Quantitative Macroinvertebrate Community Index for hard-bottom streams (SQMCI-hb). Details of invertebrate data post-processing are given below.

In this report, we use "water quality" as a general term to refer to some or all of the preceding variables. Unless otherwise stated, we make no distinction between data collected at regional council sites and NRWQN sites, and we refer to the sites collectively as the "river monitoring network".

Most of the data used in this report were compiled in 2011-2012 for a previous NEMaR project on statistical power and representativeness in the monitoring network (Larned & Unwin 2012). The geographic locations and corresponding NZReach numbers of the monitoring sites used in the current study were determined and verified in the 2012 study. Our final dataset included sites on 1,021 reaches, comprising 789 water quality sites and 519 invertebrate sites. Water quality sites tended to be on larger rivers than invertebrate sites, consistent with the tendency for invertebrate samples to be collected from streams small enough to be waded safely.

In the dataset used for the 2012 study, the starting dates for all monitoring site records were 1 January 2006 or earlier, and the ending dates ranged from June 2009 to February 2012. The range of ending dates poses some potential problems due to temporal variation in water quality. Further, we carried out temporal trend analyses in the current study and recent data were needed to ensure that the analyses corresponded to recent conditions. For these reasons we requested updated physical-chemical water quality data from five regional councils, to fill the most severe gaps in recent data. Each of the five regional councils provided updates, and the ending dates in the current dataset range from January 2011 to December 2012. Note that starting and ending dates can vary among sites within councils, and among variables within sites.

[1] Including NO_2-N, usually a minimal component.
[2] TEMP, DO, and DOSAT were data taken as received, and were rarely consistently recorded at the same time of day. See Section 3.2.1 for further discussion of this problem.

2.2 Data processing

Documentation of quality control standards vary widely among datasets provided by different councils. The initial data processing step used in the previous study (Larned & Unwin 2012) was repeated with the updated data to produce an internally consistent dataset for the current study. Briefly, we used quantile plots to visually (and subjectively) identify and remove gross outliers for each variable (e.g., CLAR = 735 m; DO = 34.5 mg/l; TEMP = 110 °C); otherwise all updated data were loaded into the database as received. A total of 47 records were identified as outliers, out of 2,240,524 individual measurements. Where necessary, individual variables were multiplied by an appropriate scale factor (e.g., 10, 1000) to ensure consistent units of measurement across all datasets. Records identified in the source dataset as being less than a specified detection limit (e.g., CLAR < 0.2) were replaced with a numerical value equal to half the detection limit. Records identified as exceeding some upper limit (e.g., ECOLI > 2400) were interpreted as equal to that limit. Almost all such records (761 of 772) were for either CLAR (394 records) or ECOLI (367 records). These adjustments are unlikely to have affected our analyses of current state, which were based on medians for each variable, but could potentially have influenced our analysis of temporal trends.

The updated data were also checked for multiple procedures used to measure single variables; alternative field and laboratory procedures for a given variable can be an extraneous source of data variability. We used the rules set out in the previous study (Larned & Unwin 2012) to pool data from comparable methods. Data measured using non-comparable methods were omitted.

The second data-processing step was to calculate four invertebrate indices: Ntaxa, EPTtaxa, SQMCI-hb and %EPTabund. These indices are widely used in New Zealand to summarise invertebrate data (Boothroyd & Stark 2000). The intent of the indices is to convey information about the integrated environmental conditions at monitoring sites, based on the assumption that invertebrate abundance, diversity and composition vary in response to multiple environmental variables. Some regional councils provide values for one or more of these indices as part of their SoE datasets. However, we calculated each index from raw data to ensure consistency in taxonomic resolution and MCI tolerance values. The range of taxonomic levels (from phylum to species) varies among regional councils, and it was necessary to standardise taxonomic levels to make sites comparable. For a standard taxa list, we used Table 1 of the User's Guide to the Macroinvertebrate Community Index (Stark & Maxted 2007).

We used the MCI tolerance scores for hard-bottom streams in all SQMCI calculations, regardless of the channel condition at the monitoring site. Soft-bottom SQMCI was not used for two reasons. First, splitting the sites into two groups based on substrate would have required two parallel analyses with fewer sites in each, which would have affected the representativeness and power of the statistical models. Second, there was insufficient information about site substrate in council-supplied datasets to consistently assign sites to hard- and soft-bottomed groups.

The SQMCI was developed to meet the need for an invertebrate index that is more informative than the presence-absence-based MCI, and less costly than the quantitative QMCI, which requires full or fixed counts of samples (Stark 1998). The SQMCI uses five

abundance categories, with corresponding abundance ranges and coded abundances: rare (1–4 individuals; coded abundance = 1), common (5-19 individuals; coded abundance = 5), abundant (20-99 individuals; coded abundance = 20), very abundant (100-499 individuals; coded abundance = 100), very very abundant (≥500 individuals; coded abundance = 500). We used SQMCI-hb in lieu of QMCI-hb because there were too few samples in the invertebrate dataset consisting of full or fixed counts. Approximately 37% of the invertebrate records in our dataset consisted of coded abundances, with the remaining 63% based on full or fixed counts. For consistency, all full or fixed counts were converted to coded abundance, and SQMCI-hb was calculated for every sample using the equation

$$\mathrm{SQMCI} = \sum_{i=1}^{i=S} \frac{n_i \times a_i}{N},$$

where S is the total number of taxa in the sample, n_i is the coded abundance for the i^{th} taxon (using the abundance values listed above), a_i is the tolerance value for the i^{th} taxon, and N is the total of the coded abundances for the sample (Stark 1998).

For both EPTtaxa and %EPTabund, algivorous caddisflies in the Family Hydroptilidae were excluded because these taxa often proliferate in algal blooms (Collier 2008). Calculations of %EPTabund were affected by the mixture of coded abundance and fixed or full count data, as discussed in the previous paragraph for SQMCI-hb. To be consistent, we converted all fixed or full counts to code abundances, then calculated %EPTabund as the sum of coded abundances for EPT taxa divided by the sum of coded abundances for all taxa. We used the same coded abundance values used for SQMCI-hb calculations: 1, 5, 20, 100, 500.

We originally proposed to use a recent, multi-metric invertebrate index, Average Score Per Metric (ASPM), in addition to the four metrics listed above. ASPM was developed in New Zealand in response to concerns that single metrics such as EPTtaxa fail to capture community-level responses to multiple gradients in environmental conditions (Collier 2008, Stoddard *et al.* 2008). Like other multi-metric indices, the ASPM combines the standardised scores from several traditional indices to produce a single value for a sample. The first and only published case study of the ASPM used full-count invertebrate samples from the Environment Waikato river monitoring programme (Collier 2008). The use of ASPM with a new, national-scale dataset would require several steps: 1) calculating values for a suite of candidate metrics; 2) reducing the number of candidate metrics by eliminating highly correlated metrics; 3) standardising the values of the remaining metrics (e.g., by dividing by the maximum value in the dataset; and 4) identifying the optimum subset of the remaining metrics based on maximum discrimination between degraded environmental conditions (e.g., urban, pastoral) and reference conditions (e.g., native forest). This last step requires standardised values of candidate metrics from reference sites and impacted sites. After discussions with the ASPM developer, it was clear that using ASPM with a national-scale, multi-year dataset would require a substantial amount of redevelopment, for several reasons. The reduced number of candidate metrics and the optimum set of metrics may vary geographically and with changing spatial scales, and this will influence whether and how the national dataset is subdivided. The maximum values and reference sites or conditions used for standardisation and optimisation would need to be determined. Finally, it is not yet clear how to incorporate multiple years of data. For these reasons, ASPM was omitted as an invertebrate index.

2.3 Rules for including monitoring sites in models and trend analyses

To identify sites suitable for our analyses, we applied date-filtering rules to the pooled data provided by councils and the NRWQN. Our aim was to maximise the number of sites for which sampling duration and frequency were sufficient to calculate robust medians and trends. Start and end dates varied among datasets, so our selection rules involved choosing a time interval long enough to yield sufficient data, short enough to capture as many sites as possible, and with an ending date that was relatively recent, but not so recent that many sites were excluded. For modelling current state for each variable, we restricted the analyses to data collected since 2006; the results were based on data collected over the most recent 5-6 years of record. Water quality data were generally available up to at least December 2011 and often extended into 2012, so we used all available data from 1 January 2006 for sites with records for at least 16 calendar quarters in at least 5 calendar years. Invertebrate data were rarely available beyond December 2010, and generally came from samples collected once per year.

For modelling trends in water-quality variables and water-quality indices, we restricted our analyses to sites with data for at least 8 years since 2000. Trend analyses for specific data sets (e.g., NRWQN) have previously been conducted over periods as short as five years, but for the purposes of this report we opted for a longer period on the assumption that the resulting estimates of trend strength and direction would be more robust, and hence better suited to national-scale modelling. For sites monitored four times per year, we excluded those with data for less than 32 calendar quarters. For modelling trends in invertebrate community metrics calculated using data from annual invertebrate samples, we used all available data, but trialled eight different date-filtering rules to identify the most viable compromise between data currency, length of record, and goodness of fit. These were:

i) At least 5 years from 1 January 2006;
ii) At least 6 years from 1 January 2006;
iii) At least 5 years from 1 January 2000;
iv) At least 7 years from 1 January 2000;
v) At least 10 years from 1 January 2000;
vi) At least 10 years from 1 January 1995;
vii) At least 15 years from 1 January 1995;
viii) At least 20 years from 1 January 1990.

2.4 Random forest models for predicting physical-chemical water quality and invertebrate community metrics

We used random forest (RF) regression to model each water quality and invertebrate community variable using a set of 28 catchment and land-cover descriptors as predictor variables. These predictor variables were selected in a recent study of environmental factors influencing local water quality (Unwin *et al.* 2010). An RF model is an ensemble of individual classification and regression trees fitted via an algorithm that is free from distributional assumptions, and can automatically fit non-linear relationships and high order interactions. Each tree is grown with a bootstrap sample of the input data, using random subsets of the available predictor variables to grow the tree. Introducing these random components and

then averaging over the forest increases prediction accuracy while retaining many statistically desirable features. We refer readers to (Unwin *et al.* 2010) and references therein for further details about RF models.

We modelled median values for each variable and NZReach, using raw (i.e., untransformed) data for TEMP, DO, DOSAT, and the four invertebrate community variables, and log-transformed medians (i.e., the log of the median of the untransformed raw data) for all other variables (Figure 1). To examine the nature of the resulting models we calculated importance scores for all predictor variables, and examined and partial dependence plots (Breiman, 2001; see also Unwin *et al.* 2010). Importance scores typically range from 10 or less for predictor variables of little or no importance to 30 or more for those of highest importance; scores of 20 or above can be taken to indicate strong relationships between predictor and response variables. Partial dependence plots show the marginal effect of a predictor variable on the response after accounting for the average effects of the other predictor variables in the model. These plots do not perfectly represent the effects of each variable, particularly when predictors are highly correlated or strongly interacting, but provide useful information for interpretation. We note that, because estimating a RF model involves randomly selecting observations and predictors throughout the fitting process, successive models fitted to the same dataset will exhibit subtle differences in structure and diagnostics such as total explained deviance, mean square error, partial dependence plots, and the rank order of predictors of similar importance.

We used a jack-knife procedure to estimate confidence intervals for the each model. In this step we withheld one water quality site from the sites used to fit the RF models. The fitted model was then used to predict the variable of interest at the withheld site. Confidence intervals were then derived from these data using quantile regression. These confidence intervals will tend to overestimate confidence for rivers with combinations of catchment characteristics that are not well represented (cf. Snelder *et al.* 2009). We used normal Q-Q (quantile) plots to assess the distribution of residuals for each model, and to characterise the extent to which these deviate from normality. Further details are given in Appendix 2 of Unwin & Snelder (2010).

All calculations were performed using Version 2.12.1 of the software environment R (R Development Core Team 2010) via the *randomForest* function library, using the *predict.randomForest* function to estimate variable levels for all REC segments throughout New Zealand.

Model results are summarised in Appendices 1 (water quality variables), 2 (invertebrate community metrics) and 3 (water quality indicators), as a multi-panel figure for each variable showing model diagnostics (a scatterplot of the observed value vs. the corresponding jack-knife prediction; a normal Q-Q plot for each model; and partial residual plots for the six leading predictors), together with mapped representations of model predictions for all river segments in New Zealand.

2.5 Random forest models for predicting composite water quality indices

We generated RF models for three types of river condition indices: the Canadian Council of Ministers for the Environment water quality index (CCME WQI), the Victorian Index of Stream

Conditions (VISC-WQI), and a composite indicator based on the CCME-WQI, VISC-WQI, and SQMCI-hb (Collier 2008). River condition indices are still under development, but are a potential method for concisely summarising river conditions spanning a range of metrics as single indices. In this section we briefly summarise the key features of the indices investigated to date; for further details see Ballantine (2012). We review the steps involved in calculating these indices for the present dataset, and identify some of the assumptions and caveats underlying the calculations.

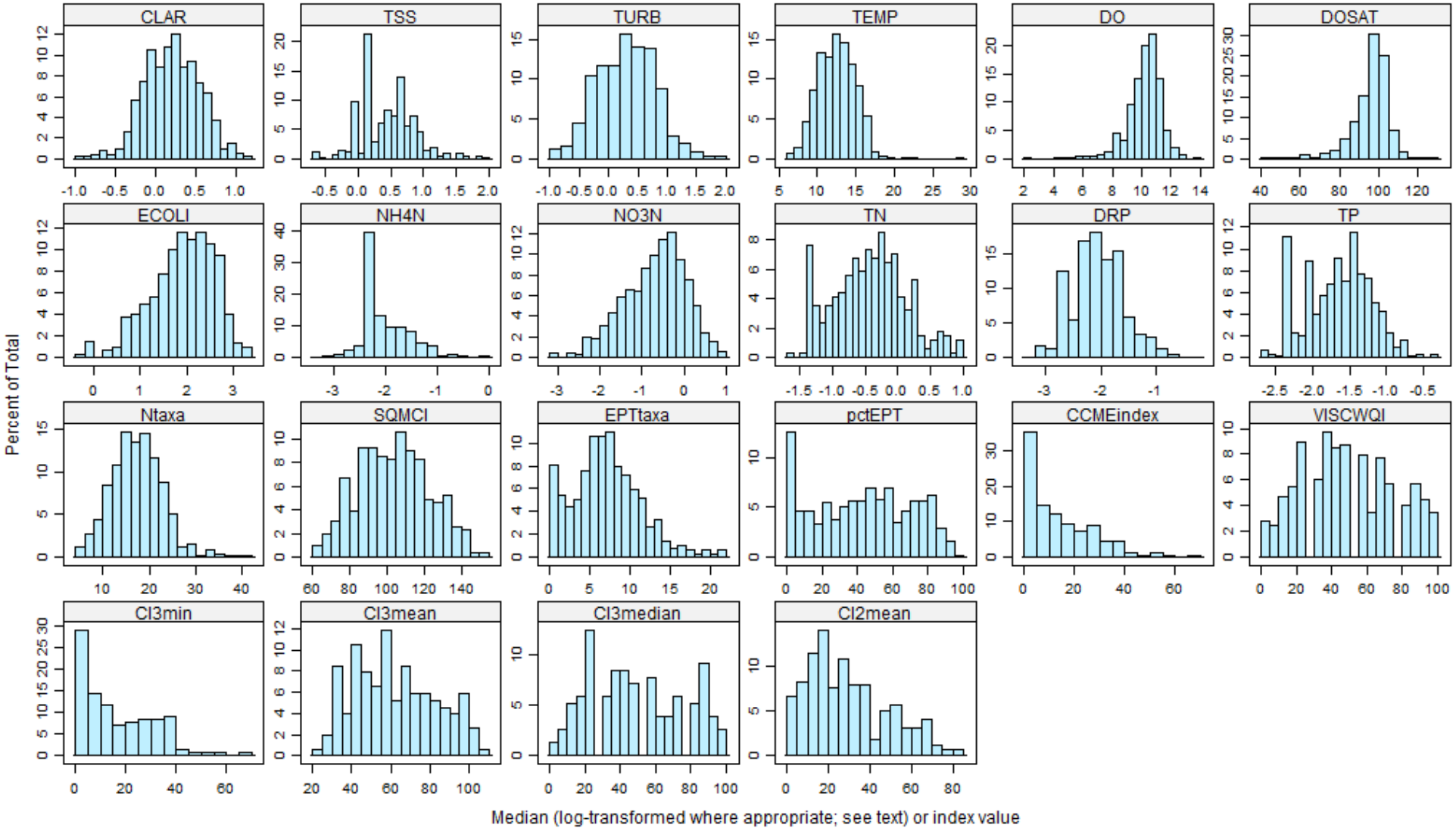


Figure 1: Distribution histograms for the 12 water quality variables, 4 invertebrate community metrics, and 6 indices considered in this study.

2.5.1 CCME-WQI

The CCME-WQI is used both as a stand-alone water quality index, and as a sub-index for use in a composite river condition indicator (Collier 2008). Given a monitoring site where a fixed set of variables is measured consistently over a period of time, the CCME index is calculated as the average of three components relating to a pre-defined set of "objectives" or reference conditions for each variable: the proportion of variables which exceed these objectives (scope: F1); the proportion of samples in which these objectives are exceeded (frequency: F2); and the total amount by which these objectives are exceeded (amplitude: F3). The resulting index is scaled to yield a number between 0 (worst water quality) and 100 (best water quality).

The specific water quality variables, objectives, and time periods used for CCME-WQI calculations are not specified a priori. The NEMaR expert panel on indicators recommended a core set of seven variables: CLAR, ECOLI, NO3N, NH4N, TN, DRP and TP, at least four of which are required to calculate a valid index. Objectives were based on reference conditions for all core variables (McDowell *et al.* 2013), supplied by MfE. These reference conditions were specified separately for all significant REC climate/source-of-flow classes, so that objectives for each NZReach were specific to the REC class associated with that reach.

Selecting core variables for the current study required us to make a compromise between robustness and coverage in environmental space. Using all seven core variables would maximise robustness, but sufficient data for the seven variables were only available for 116 of 784 reaches, over half of which were in one region. After inspecting the database we identified a subset of five variables (ECOLI, NH4N, NO3N, DRP, TP) for which there were sufficient data for 398 reaches, and a subset of four variables (ECOLI, NO3N, DRP, TP) for which there were sufficient data for 525 reaches. We chose the latter set, on the assumption that, for RF modelling, maximising geographical coverage and minimising gaps in environmental gradients was more important than maximising index robustness.

The freedom to define the time periods variables used to calculate the CCME-WQI means that the resulting indices for any given dataset are context-dependent. This feature has some potentially undesirable effects. For example, the F1 component represents the percentage of variables that exceed their objectives at least once during the time period under consideration; for variables that exceed their threshold values only rarely, F1 will tend to increase with time regardless of any changes in water quality. The F3 component depends on the amount by which each threshold is exceeded, expressed as a ratio, and is sensitive both to outliers and to variation in reporting standards between regions. In the present dataset this variation was particularly noticeable for ECOLI, for which upper reporting limits varied from 2,400 MPN/100 ml in some regions to > 100,000 MPN/100 ml in others.

CCME-WQI scores for the 525 sites with sufficient data were calculated as per Ballantine (2012), and then used as dependent variables in RF models based on the same predictor set as for the water quality and invertebrate community variables (Section 2.4). Index scores were not normally distributed (Figure 1), but we did not log-transform them as 5% (26 of 525) values were zero. The distribution of index scores is also strongly influenced by the choice of reference conditions, and could easily (and more naturally) be normalised by adjusting the reference scores (see Section 3.3.1).

2.6 VISC WQI

The VISC WQI as originally developed was based on four variables (TP, TURB, pH, conductivity), which were considered relevant for reporting water quality issues in rivers in Victoria, Australia. For this study, we adopted the version developed for the Hawkes Bay and Greater Wellington Regional Councils, which was based on DRP, NO3N, CLAR, and ECOLI by Ballantine (2012). Site availability for the present dataset was limited by the lack of CLAR for several regions (e.g., Auckland, Canterbury, Otago), but viable index calculations were possible for a total of 401 sites. Other permutations of variables may have increased the number of available sites, but investigating the properties of the resulting indices was beyond the scope of this study.

The VISC-WQI is based on percentiles, via a three-step calculation in which medians for each variable are converted to percentiles across all sites within each REC climate/source-of-flow class, the resulting percentiles are converted into levels spanning successive percentile bands (0-20%, 20-40% etc.), and the resulting levels are averaged to yield an index on a scale from 0-100. As with the CCME-WQI the resulting index scores are context-dependent, in that the score for a given site depends on the characteristics of other sites within the same REC class, but – because they are based on site medians – were much less sensitive to outliers than the CCME-WQI. The resulting scores were approximately normally distributed (Figure 1), and were used as dependent variables in RF models as for the CCME-WQI.

2.7 Composite index

A composite index for conveying information about environmental conditions combines the standardised values of multiple indices (referred to here as sub-indices). The ASPM described in Section 2.2 is an example of a composite index, as it uses several invertebrate indices as sub-indices. Composite indices are intended to integrate information about multiple aspects of environmental conditions into single values.

We trialled several composite indices in this study, based on combinations of two water quality sub-indices, VISC-WQI and CCME-WQI, and one invertebrate sub-index. We originally planned to use ASPM as the invertebrate sub-index, but after evaluating ASPM (see Section 2.2) we used SQMCI-hb instead. SQMIC-hb scores were standardised by converting scores to percentiles and expressing the results on a scale from 0 to 100.

The number of sites suitable for calculating composite index scores were limited by the relatively small number of sites that were regularly sampled for both water quality and invertebrates (see Section 2.1). We identified 153 suitable sites, scattered thinly throughout most of New Zealand with clusters of sites in the lower North Island and Southland.

We trialled four composite indices; three based on VISC-WQI, CCME-WQI and SQMIC-hb, and one based on VISC-WQI and CCME-WQI alone (see Table 5, p. 24 for details). The three composite indices based on VISC-WQI, CCME-WQI and SQMIC-hb were denoted CI3mean, CI3median and CI3min; these names refer to the use of the mean, median and minimum value of the three sub-indices as the composite indicator score. The composite index based on VISC-WQI and CCME-WQI alone was denoted CI2mean, as the composite indicator scores were the means of the two sub-indices. Dropping the SQMCI-hb sub-index in CI2mean increased the number of suitable sites from 153 to 316, at the expense of

information about invertebrate communities. In addition to CI2 based on VISC-WQI and CCME-WQI, there were two other possible combinations of two sub-indices, SQMCI-hb with VISC-WQI, for which there were 215 suitable sites, and SQMCI-hb with CCME-WQI, for which there were 177 suitable sites. Since RF model performance depended on maximising the number of sites, only CI2 based on VISC-WQI and CCME-WQI was calculated and modelled. The untransformed CI3mean, CI3median and CI2mean scores were normally distributed, but CI3min scores were left-skewed (Figure 1). All four composite indices were used as dependent variables in random forest models.

2.8 Temporal trends in macroinvertebrate community indices

We used correlation analysis and linear regression to estimate the magnitude, direction, and significance of trends in the four macroinvertebrate community indices (Ntaxa, EPTtaxa, SQMCI-hb and %EPTabund). This contrasts with our analyses for trends in water quality variables (Section 2.9), which were based on seasonally adjusted non-parametric Mann-Kendall tests (cf., Ballantine *et al.* 2010) to allow for seasonal (i.e., monthly or quarterly) signals in the underlying time series, and departures from normality in the underlying distributions. Neither of these considerations were relevant for the invertebrate data, which were approximately normally distributed (Figure 1), and did not require seasonal adjustment because the data were from annual samples.

For each site and each invertebrate index, we regressed the observed values against date (measured in decimal years), and divided the slope of the fitted line by the median observed value for all sites to express the trend as percentage annual change. Associated diagnostics were the attained significance (P value) and 95% confidence intervals for the corresponding Pearson correlation coefficient, with trends being assessed as significant only if the correlation coefficient differed significantly (i.e., $P < 0.05$) from zero. Even in the absence of any significant correlations, approximately 5% of fitted trends will fall below $P = 0.05$ purely by chance, so our results may overestimate the number of correlations we report as "significant".

We then fitted RF models to the estimated trends in invertebrate indices using the procedure summarised in Section 2.4. The utility of each model was assessed on the basis of the percentage of explained variance and the associated diagnostic plots. To characterise model performance in relation to the lengths of time-series, we used various combinations of starting year and time-series lengths in the RF models, then focused our assessments on the datasets we judged to be most robust.

2.9 Temporal trends in water quality variables

We used the non-parametric Seasonal Kendall Sen Slope Estimator (SKSE; Sen 1968) to estimate the magnitude and direction of temporal trends for each site and water quality variable. The trend values were normalised by dividing by the raw data median to give the relative SKSE (RSKSE), allowing for direct comparison between sites measured as per cent change per year. RSKSE calculations were accompanied by a Seasonal Kendall test of the null hypothesis that there is no monotonic trend. If the associated P-value is "small" (i.e. $P < 0.05$), the null hypothesis can be rejected (i.e., the observed trend or any larger trend, either upwards or downwards, is unlikely to have arisen by chance). As with our calculations of trends in macroinvertebrate indices, our assessments of significance levels do not account for analyses of multiple sites, and may thus overestimate the number of significant trends.

The raw water quality data were not flow-adjusted prior to trend analyses, as they were in some previous analyses (e.g., Ballantine et al. 2010). The decision to omit flow-adjustments was made following discussions between NIWA and MfE, which considered (a) the resources needed to obtain updated flow data to match the latest water quality data; (b) the need to predict flows at water quality sites that are not associated with flow-gauging sites; and (c) the potential for flow-adjustment to introduce additional noise into the raw data (c.f., Ballantine *et al.* 2010). In the absence of flow-adjustment, there is a risk that some trend estimates are due to flow variability alone. However, the effects of flow-dependent trends on RF models are likely to be minimal, possibly causing some affected sites to appear as outliers but having little effect on the overall fit.

We estimated trends for two parallel datasets, based on monthly and quarterly data, respectively. Monthly analyses were possible only for a subset of regions with monthly sampling programmes. Monthly data provided more robust and powerful trend analyses than quarterly data, but there were more sites and greater geographical coverage when quarterly data were used. Quarterly data were available from all regions; monthly (and occasionally bimonthly) datasets from some regions were converted to quarterly datasets by calculating quarterly medians.

To categorise trends of differing strength, trends for each water-quality variable into grouped into one of three categories. These categories are intended to differentiate between trends which are statistically significant ($P < 0.05$) but not necessarily meaningful in a environmental management context, and those which are both significant and meaningful (c.f., Ballantine *et al.* 2010, Davies-Colley & Nagels 2008). The categories are:

i. not statistically significant ($P \geq 0.05$), i.e., the null hypothesis for the Seasonal Kendall test was not rejected;

ii. statistically significant but not meaningful ($P < 0.05$, RSKSE < 1%), i.e., the null hypothesis for the Seasonal Kendall test was rejected, but the trend is unlikely to be meaningful in an environmental-management context;

iii. statistically significant and meaningful ($P < 0.05$, RSKSE ≥ 1%), i.e., the null hypothesis for the Seasonal Kendall test was rejected, and the trend is potentially meaningful in a management context.

We fitted RF models to the estimated trends using the same procedures and diagnostic methods as for all other variables. As with the models for trends in invertebrate community indices, detailed results are presented only for the most robust.

3 Results

3.1 Deliverable 1: Data compilation

The water quality and macroinvertebrate community data on which this report is based are contained in two independent but closely related MS-Access databases, each of which is, in turn, based on parent databases compiled for one or more previous projects. Progenitor databases for water quality were compiled in 2009 (Unwin *et al.* 2010) and 2011 (Larned & Unwin 2012). The macroinvertebrate database is an updated version of a database developed for a Department of Conservation TFBIS project, and subsequently used as the basis for a MfE-funded compilation of macroinvertebrate data at a nationally consistent level of taxonomic resolution. Both databases have structures in common, including tables of all sampling sites in each region and the results (water quality measurements or invertebrate counts) for each sampling date. Differences mostly relate to the ancillary information needed to establish national consistency, such as lookup tables allowing individual variables to be uniformly interpreted (for water quality), or to associate individual taxa with measures such as MCI score and EPT status (for macroinvertebrates). In addition, we accessed data from the REC, FWENZ, and other related databases, copies of which are held by NIWA, to obtain NZReach-level data as necessary for the RF models.

Including data for sites and time periods not used in this study, the pooled water quality database currently holds 2,240,524 individual records, representing 692 measured variables at 1154 sites in 16 regions. After discarding variables provided by some regions but irrelevant to the current study (e.g., heavy metals, pH), and applying our site and date filtering criteria (Section 2.3), our final working datasets consisted of 681,948 measurements (for the 2000-2012 trend analyses), and 399,585 measurements (for the 2006-2012 RF analyses). Including data for sites and time periods not used in this study, the macroinvertebrate database holds 226,787 taxonomic occurrence records, representing 444 distinct taxa at 1,501 sites in 16 regions.

The databases are essentially data repositories, and no attempt has been made to provide a user interface beyond the standard MS-Access table and query design tools. Standard queries for exporting raw water quality and invertebrate community data as Excel or flat text files are available, but users with more advanced data extraction needs should be conversant with the basics of query design.

With very few exceptions, data have been retained exactly as received from the original source (i.e., regional council or NRWQN). Gross outliers (e.g., CLAR = 82.2 m, TEMP = 204 °C) have been flagged so that they can be discarded at query time, but in borderline cases (e.g., CLAR = 20-30 m, TEMP ~ 32 °C) we have opted to retain the data.

We have eliminated errors in cross-referencing site locations to the REC, and all sites should be matched to the correct NZReach. The current databases include recent updates to the master site list for macroinvertebrate samples, and are believed to be error-free. As with the raw sample data, however, developing a fully definitive site list will require consultation with regional council staff who are familiar with their sites, particularly with regard to recent shifts in location.

3.2 Deliverable 2. Random forest models for predicting physical-chemical water quality

Diagnostic plots for all models include quantile regression and quantile plots (Appendix 1). These diagnostics show the extent to which the residuals deviate from normality, and help to assess model performance. The most extreme residuals for all models were more dispersed (at one or both ends of their range) than if they were distributed normally, suggesting a general tendency for the models to over-predict sites with low values for each variable, and under-predict sites with the highest values. Partial dependence plots for the leading six predictors for each model are also shown on a common vertical scale, so that responses for each predictor can be compared directly.

RF model predictions for all NZReaches are included in Appendix 1 as a series of maps, colour coded so that predicted water quality is highest at the indigo/blue end of the scale. High water quality is presumed to correspond to low values of TEMP, TURB, ECOLI, TSS and nutrient concentrations, and high values of DO and DOSAT. The maps also show the distribution of monitoring sites on which each RF model was based, thereby indicating gaps in geographical coverage. Predictions for regions or REC climate/land-cover classes which are poorly represented in the raw data are potentially less reliable than those for regions which are well-represented.

3.2.1 Model performance

Model performance (indicated by percent explained variance) varied widely among variables, but the results were comparable to the preceding RF modelling study (Unwin *et al.* 2010) for variables common to both studies (Table 1). The strongest improvement on the 2010 study was for TSS: 39.7% of variance was explained in the previous study, and 50.9% was explained in the current study. This improvement is largely due to an increase in the number of sites with suitable data (from 225 sites in the previous study to 466 in the current study), with a corresponding increase in geographical coverage. The most notable decline in model performance was for TN (from 77.8% of variance explained in the previous study to 73.8% in the current study). This small decline was probably due to the removal of over 200 sites in the current study, after we became aware of inconsistencies in measurement procedures (see Larned & Unwin 2012 for a full discussion of this issue).

Of the four new variables in this study, TURB and TEMP were reasonably well modelled (66.7% and 68.8% explained variance, respectively), but models for DO and DOSAT (i.e., DO as % saturation) were weaker, with DOSAT (43.2% explained variance) the most poorly modelled of the 12 water quality variables. Residuals for both the DO and DOSAT models were strongly over-dispersed (Appendix 1), indicating that the respective models were poor at predicting extreme values. A possible confounding factor for TEMP, and hence for DO and DOSAT (which are temperature-dependent), is the time of day at which field measurements were made. Time of day varies arbitrarily among samples and can also vary systematically between sites. For example, the NRWQN includes two sites on the Waimakariri River near Christchurch, with the upstream site (Waimakariri Gorge, 60 km above the mouth) generally sampled between 07:00 and 09:00, and the lower site (6 km above the mouth) generally sampled between 13:00 and 16:00. Measured spot temperatures at the downstream site are, on average, 2-6 ^{o}C warmer than at the upstream site, with the mean difference over the five

months from October to February consistently exceeding 5 °C. The extent to which this difference reflects time of measurement rather than longitudinal warming is unknown.

Table 1: Diagnostic statistics and random forest model performance for each water quality variable. Figures in parentheses refer to the corresponding values for the 2010 random forest models.

Variable	Nsites	Median	Mean	±1 SD	% variance explained
CLAR	507 (382)	1.60	1.61	0.74 - 3.51	67.9 (62.2)
TSS	466 (225)	3.00	2.85	1.12 - 7.30	50.9 (39.7)
TURB	714 (-)	2.00	1.97	0.62 - 6.29	66.7 (–)
TEMP	748 (-)	12.60	12.57	10.10 - 15.05	68.8 (–)
DO	713 (-)	10.40	10.20	9.00 - 11.41	55.9 (–)
DOSAT	666 (-)	98.1	95.9	85.9 - 105.8	43.2 (–)
ECOLI	738 (396)	98.4	81.7	17.1 - 390.0	72.3 (69.8)
NH4N	459 (553)	7.50	10.15	3.45 - 29.8	56.9 (57.0)
NO3N	682 (552)	251.0	201.8	37.6 – 1084	65.9 (68.6)
TN	344 (526)	382.5	367.9	102.1 - 1325	73.8 (77.8)
DRP	722 (565)	9.00	8.91	3.05 - 26.1	56.6 (58.9)
TP	593 (528)	25.00	22.45	8.06 - 62.5	71.8 (72.4)

3.2.2 Predictor variables

Averaging across all water quality variables, the predictor variables with the highest importance scores were % heavy pastoral (average importance score 21.2), catchment elevation (average importance score 17.8), mean slope (average importance score 17.2), minimum annual air temperature (average importance score 15.3), and maximum annual air temperature (average importance score 15.2) (Table 2). The high-scoring predictors for each water quality variable were the same as those identified in the previous RF model study (Unwin *et al.* 2010), with minor changes in order. Given that predictor order for the same dataset can vary subtly each time the model is estimated, due to the random component built into the RF model fitting algorithm, such variation is to be expected. Examples of variables for which the leading predictors were identical in the two studies include ECOLI (catchment elevation, %heavy pastoral land-cover, rain variability); NH4N (catchment elevation, %heavy pastoral land-cover); and TP (mean catchment slope, % indigenous forest).

Model fits for CLAR, TSS, and TURB were similar, as expected, given that all three variables are related to water clarity and sediment load. CLAR and TSS were also modelled in 2010, but datasets for both variables were sparse and geographically limited in the 2010 study. The new models for both variables benefitted from the inclusion of additional sites, although the distributions of sites remain patchy. This patchiness reflects the tendency of some councils to measure only one of these two variables (e.g., absence of CLAR in Canterbury and Otago, absence of TSS in Waikato, Tasman, and West Coast). TURB was not modelled in 2010, but is widely available throughout New Zealand and was included in the current study. All three variables show strong dependences on the percent of catchment land-cover classified as heavy pastoral, rain variability, and (particularly for TURB) other flow- and climate-related variables such as temperature, mean flow, and annual rain days (Table 2).

Table 2: Importance scores[1] from RF models for predictors of water quality variables. The percent of variability in each variable explained by the RF model is in the top row, in parentheses.

Predictor	CLAR (67.9%) [2]	TSS (50.9%)	TURB (66.7%)	TEMP (68.8%)	DO (55.9%)	DOSAT (43.2%)	ECOLI (72.3%)	NH4N (56.9%)	NO3N (65.9%)	TN (73.8%)	DRP (56.6%)	TP (71.8%)
Reach elevation	18.6	4.4	11.9	**20.5**	9.5	7.9	15.2	8.7	10.9	10.3	10.2	8.7
Catchment elevation	17.7	10.7	17.9	10.3	14.9	**20.0**	**33.4**	**26.7**	16.6	16.9	14.8	13.6
Mean slope	16.7	18.1	13.1	12.1	16.2	15.4	19.5	16.8	18.5	12.3	19.6	**28.5**
Catchment area	13.9	16.3	16.5	16.7	9.7	7.7	12.4	9.2	10.5	7.1	12.2	12.5
Lake index	7.0	4.2	10.1	5.6	6.7	3.2	15.1	2.6	16.5	2.3	10.8	6.1
Mean flow	16.1	**20.3**	17.2	15.1	9.6	7.7	10.9	6.3	9.9	6.0	12.4	12.1
Rain variability	**20.6**	**20.3**	**27.1**	14.4	17.4	10.6	**27.0**	9.8	14.0	8.9	14.4	19.8
Min temperature	12.7	12.5	**24.9**	**20.1**	**27.3**	10.8	14.7	7.9	14.0	10.8	12.1	15.7
Max temperature	15.7	17.2	**26.0**	13.9	18.3	8.4	14.4	8.1	14.4	9.5	16.4	19.5
Rain days > 10	13.7	15.4	17.7	12.3	13.8	14.2	14.1	8.9	14.9	6.9	17.4	11.7
Rain days > 50	17.9	12.5	**22.9**	15.6	14.3	17.3	14.0	10.1	12.0	8.5	17.9	16.7
Rain days > 200	19.1	10.8	**21.1**	16.5	10.2	12.7	12.9	8.8	11.6	8.5	12.8	16.9
Evapotranspiration	10.2	13.5	17.2	13.9	9.8	8.1	14.8	6.9	11.3	8.9	14.2	13.6
%alluvium	16.2	12.5	14.7	8.2	10.8	8.9	15.9	8.9	15.7	14.7	11.0	14.0
%glacial	2.2	2.6	2.8	-1.8	1.9	-1.1	0.7	0.9	3.1	1.0	4.2	0.1
%peat	15.7	2.6	10.5	3.6	8.9	13.3	7.1	7.3	3.0	5.5	5.6	10.9
Calcium	7.6	8.4	8.2	9.0	10.2	5.1	9.5	6.3	9.8	6.2	12.5	14.3
Hardness	12.2	14.2	16.0	6.8	9.1	8.1	12.7	9.7	11.8	8.4	17.2	**20.0**
Particle size	12.9	11.2	11.6	8.9	**21.7**	12.0	12.3	6.8	14.5	7.6	**26.9**	19.3
Phosphorous	12.5	10.7	12.9	8.8	11.7	11.6	15.9	7.6	14.7	11.9	14.6	13.7
%bare	12.9	13.7	14.0	12.3	8.6	6.7	12.9	6.8	8.6	6.8	18.4	14.7
%exotic forest	5.8	10.1	10.9	7.7	8.2	13.1	13.8	5.3	18.4	8.9	9.7	7.9
%indigenous forest	13.7	10.6	13.6	9.3	7.0	8.8	14.4	9.8	19.1	11.8	14.1	**22.2**
%pastoral heavy	**24.5**	**20.3**	**21.6**	15.7	10.4	9.3	**34.5**	19.9	**34.5**	**30.2**	15.2	17.8
%pastoral light	6.2	9.9	14.7	6.8	7.4	5.5	11.5	6.4	10.2	7.4	13.7	14.9
%scrub	5.6	7.6	9.4	8.8	7.6	8.5	12.4	5.5	18.0	9.7	10.8	6.4
%urban	8.2	4.7	9.7	11.0	0.2	2.0	15.3	14.8	18.3	14.0	12.7	9.3
%wetland	10.2	6.4	15.7	5.5	7.6	6.8	6.1	4.2	9.6	6.9	10.6	7.8

[1] Importance score (IS) is highlighted so as to identify IS ≥ 30.0 (pink shading); 20 ≤ IS < 30 (bold red); 15 ≤ IS < 20 (orange); and 10 ≤ IS < 15 (blue). See Appendix 1 for a more detailed description and interpretation of each predictor. Scores are indicative only, particularly for lower ranked predictors in weak models for which predictor order can vary each time the model is fitted due to the random component built into the RF model process.
[2] Columns are ordered so as to facilitate comparisons between variables representing physical parameters, bacteria counts, and nutrients.

The leading predictor variables for TEMP were reach elevation and minimum annual temperature, although partial dependence for plots suggest that minimum annual air temperature was by far the more important of the two (Appendix 1). Minimum annual air temperature was also the most important predictor for DO, although for DOSAT the leading predictor was mean catchment elevation. Predictor variables related to land-cover were only weakly significant for TEMP, and even less so for DO and DOSAT. As noted in Section 3.2.1, Intrinsic diurnal and season variation in TEMP, DO and DOSAT may obscure the effects of land-cover and other geographic factors.

3.2.3 Model predictions

Mapped predictions of physical and chemical water quality variables for all REC river segments were always plausible, and often compelling (Appendix 1). All models were broadly consistent with mesoscale New Zealand geography, reflecting features such as the axial mountain ranges in both islands, and latitudinal/longitudinal variation in climate. In practice, however, we expect the most robust predictions to emerge from models which are both well-fitted (i.e., have a high percentage of explained variance), and have no major geographical gaps in the underlying distribution of sites. Variables for which the model meets both these criteria are limited to TURB, TEMP, ECOLI, and NO3N. For the other variables, prediction accuracy is potentially constrained by limited number and geographical distribution of sites (e.g., CLAR, TN), low explained variance (e.g., DRP), or both (e.g., TSS, NH4N).

3.2.4 Random forest models for invertebrate community metrics

We generated RF models for the four macroinvertebrate community variables (NTaxa, SQMCI-hb, EPTtaxa, %EPTabund) using the same procedures as for the water quality models. Results and diagnostic plots are shown for all four models (Appendix 2), together with maps of predicted values for each NZReach. Briefly, all models appeared to yield credible fits to the observed data, with little if any tendency towards over- or under-prediction (Table 3). The most successful model (explained variance 71.5%) was for SQMCI-hb, followed by EPTtaxa (explained variance 64.6%). Leading predictors shared by at least three models were the percentage of indigenous forest cover, catchment and/or site elevation, and the percentage of heavy pastoral land-cover (Table 4).

Table 3: Diagnostic statistics and random forest model performance for four macroinvertebrate community variables.

Variable	Nsites	Median	Mean	SD	% variance explained
NTaxa	519	17	17.1	5.2	54.6
SQMCI-hb	519	105	105.2	18.9	71.5
EPTtaxa	519	7	7.4	4.1	64.6
%EPTabund	519	45.0%	42.7%	27.1%	55.5

Table 4: Importance scores from RF models for predictors of invertebrate community metrics and water quality indices. The percent of variability in each variable explained by the RF model is in the top row, in parentheses. See Table 2 for formatting conventions used to indicate relative importance.

Predictor	Ntaxa (54.6%)	SQMCI (71.5%)	EPTtaxa (64.6%)	pctEPT (55.5%)	CCMEindex (54.6%)	VISCWQI (52.3%)	CI3min (46.2%)	CI3mean (54.6%)	CI3median (33.0%)	CI2mean (55.4%)
Reach elevation	16.7	**21.5**	**21.5**	**22.6**	6.6	6.1	-1.4	15.5	8.7	8.6
Catchment elevation	13.2	**22.4**	18.0	**22.0**	12.4	7.7	8.7	14.3	4.2	12.7
Mean slope	16.0	18.1	11.9	14.0	**21.9**	14.1	14.9	12.9	7.1	8.7
Catchment area	10.3	14.6	12.1	12.0	7.0	7.9	4.9	6.5	2.5	9.7
Lake index	2.5	5.9	1.1	-0.8	8.5	3.5	3.2	2.0	3.0	4.0
Mean flow	11.5	11.2	10.5	11.2	7.3	8.4	5.0	6.2	2.3	10.7
Rain variability	10.3	11.4	9.8	10.8	13.4	13.2	6.4	7.1	5.3	6.4
Min temperature	11.4	13.8	12.2	15.4	9.4	9.7	2.2	6.6	5.3	9.9
Max temperature	14.2	12.1	11.2	13.2	12.4	9.0	3.7	5.1	6.1	11.5
Rain days > 10	17.7	12.2	15.0	10.7	8.6	8.4	1.4	4.9	1.3	9.2
Rain days > 50	14.7	12.2	16.2	10.6	13.8	8.1	5.0	3.1	4.3	11.3
Rain days > 200	16.1	10.3	13.1	6.6	14.4	10.8	8.2	5.3	7.8	9.6
Evapotranspiration	9.9	15.2	10.7	11.8	9.8	10.4	0.2	1.4	7.1	6.3
%alluvium	**20.7**	13.0	12.7	8.6	9.5	8.0	3.2	4.6	2.3	5.3
%glacial	4.4	2.5	5.8	-0.8	-2.3	2.6	1.7	2.8	1.7	0.7
%peat	5.1	4.7	5.2	0.5	3.0	3.0	0.6	6.6	0.9	5.2
Calcium	16.2	16.0	9.9	13.9	7.8	9.4	5.3	2.3	4.5	7.7
Hardness	11.7	13.6	10.7	12.4	12.2	11.2	9.5	10.1	8.4	4.8
Particle size	11.8	12.1	13.0	11.9	**21.4**	**21.2**	9.7	5.5	18.1	12.5
Phosphorous	18.6	10.9	14.2	11.7	10.3	6.9	4.1	5.4	4.4	7.9
%bare	8.8	7.3	9.6	3.2	8.2	10.0	9.1	5.4	3.5	6.8
%exotic forest	7.9	11.1	8.7	11.1	7.4	6.2	1.8	3.6	4.1	6.2
%indigenous forest	16.2	**30.4**	**26.3**	**24.7**	10.9	14.5	9.5	16.3	6.1	6.4
%pastoral heavy	12.5	**23.7**	19.9	**20.7**	**22.6**	**20.9**	9.7	12.9	6.3	9.5
%pastoral light	11.1	5.5	9.7	6.8	13.0	4.0	2.5	4.9	3.2	9.4
%scrub	10.5	10.3	9.3	8.8	11.5	13.1	3.0	2.7	5.1	11.5
%urban	13.0	**23.1**	**20.4**	18.1	11.7	18.9	5.7	4.2	11.6	8.3
%wetland	7.0	4.0	3.6	5.6	5.6	4.1	3.5	4.7	2.5	5.4

3.3 Deliverable 3. Random forest model for predicting CCME-WQI and VISC-WQI scores

3.3.1 Raw CCME index scores

The raw CCME-WQI scores for the 525 sites with suitable data were generally low, and were not evenly distributed across the water quality categories associated with this index: excellent, good, fair, marginal, poor (Ballantine 2012). Most sites had CCME-WQI scores below 45 (categorised as "poor"). Only ten sites attained scores of 45-64 (categorised as "marginal"), and only two (Waikato River at Reids Farm, 2 km above Huka Falls; Monowai River below Lake Monowai) scored over 65 (categorised as "fair"). High-profile and highly regarded rivers rated as "poor" included the Clutha River at Luggate; Hurunui River at Mandamus; Waitaki River at Kurow; Ruamahanga River at McLays (2 km upstream of Mt Bruce); Ngaruroro River at Kuripapango; Akatarawa River; Oreti River at Three Kings; and Buller River at Longford.

The CCME-WQI categories are arbitrary, and could be altered so as to be more relevant to New Zealand conditions (Ballantine 2012). In our view, however, a more fundamental problem is that – at least for the current dataset – the index appears to be naturally and strongly skewed towards very conservative values. Mean and median index values for the 525 available sites were 13.8, and 10.1, respectively, and 78 sites (15% of the total) scored below 1. Given that the index is defined to range from 0 to 100, with scores of 95-100 categorised as "close to pristine", the skewed distribution cannot be remedied simply by rescaling or transforming the data.

Several factors may have contributed to this result. Some tendency towards low scores is to be expected, given that regional councils prioritise sites where water quality is likely to be compromised (Ballantine 2012). As noted earlier, we would also expect index scores to vary depending on the choice of core variables, as well as the time period over which exceedances are tallied. However, the most important factor influencing CCME-WQI scores is the choice of reference conditions used to define exceedances for each variable, which directly affect all three components (F1, F2, F3) on which the index is based. The reference values used for this study were based on median reference conditions estimated for each nutrient and ECOLI in the absence of agricultural landcover (McDowell *et al.* 2013). These median reference values are stringent because, by definition, exceedances for each variable are expected to occur 50% of the time. More lenient reference values would increase the CCME-WQI scores at most sites.

3.3.2 Model performance and predictor variables

RF model performance for the CCME index was fair but not exceptional, with the percentage of explained variance (57.5%; Table 4) similar to that for some individual water quality variables (e.g., NH4N, DRP, and %EPTabund) but less than the percentage explained in the best models (e.g., ECOLI, TP). The corresponding Q-Q plot (Appendix 3) suggests that the RF model performed well for predicting the lowest CCME-WQI scores, but systematically (and significantly) under-predicted scores at the upper end of the observed range. Sites with sufficient data for CCME-WQI calculations were distributed across much of New Zealand, but were sparse along the West Coast, in Hawkes Bay and East Cape, and in Taranaki.

Leading predictors and response curves for the CCME index directly reflected those of the component variables, with importance scores for the top three predictor variables (percentage of heavy pastoral land-cover, mean catchment slope, and mean catchment particle size (a geological surrogate for substrate size) ranging from 21.4 - 22.6 (Appendix 3). These results are consistent with those for the individual components, with % heavy pastoral cover the dominant predictor for ECOLI and NO3N, mean slope important for all components and the dominant predictor for TP, and particle size the dominant predictor for DRP. More generally, we would expect the best predictors to be almost completely determined by the water quality variables used in CCME-WQI, so that – assuming all variables were available at all sites – parallel models of indices based on different subsets of variables would not necessarily share the same predictors.

3.3.3 Model predictions

Predicted CCME-WQI scores for all NZReaches were uniformly low, with 99.7% of reaches predicted to have scores below 45 (i.e., category “poor”), and only 26 reaches (0.004%) exceeding 50. The maximum predicted score was 53.3, with mean and median predicted scores of 22.8 and 21.9, respectively. When presented in map form (Appendix 3) the predictions appear to divide New Zealand into two regions corresponding to lowlands and uplands, with most lowland reaches predicted to have indices below 15-20, and most upland reaches predicted to score above 25.

3.3.4 VISC-WQI index

We fitted a RF model to the VISC-WQI indices for all available sites, and predicted VISC-WQI values for all NZReaches; diagnostic plots and mapped predictions are in Appendix 3. The RF model fit was comparable to that for the CCME-WQI, with 52.3% percentage explained variance. However, VISC-WQI values (mean = 61.8, median = 63.2) were higher than for the CCME-WQI (mean = 13.8, median = 10.1), and spanned a broader range (standard deviation = 26.4; CCME standard deviation = 13.2). The corresponding Q-Q plot also suggest a markedly better fit than for the CCME index, with little if any difference between the sampled and theoretical quantiles (Appendix 3).

Leading predictors for were similar to those for the CCME index, with the percentage of heavy pastoral land-cover, rain variability, particle size, and mean catchment slope among the top six predictors for both models, with similar response curves (Appendix 3). Percentage heavy pastoral land-cover (IS = 26.04) was the most important predictor, well ahead of the next two (evapotranspiration, IS = 21.22; rain variability, IS = 20.85) (Table 4). As with the CCME-WQI, the response curve for percentage heavy pastoral land-cover suggest a rapid decline as pastoral land-cover in the upstream catchment increased from zero to 20%, with relatively little change thereafter.

3.4 Deliverable 4. Random forest models for predicting composite index scores

3.4.1 Model fits

Model fits for the four version of the composite index ranged from fair to poor, with percentage explained variance ranging from 33.0 – 55.4% (Table 5). The corresponding diagnostic plots and maps of predicted values for the composite indices are shown in

Appendix 3. As was the case for CCME-WQI and VISC-WQI, the leading predictor variables for the composite indices were largely predetermined by the choice of components used to calculate each sub-index. In the remainder of this section we focus on the characteristics of the composite indices: the number of sites for which it could be calculated, its range and distribution, and the strength of the underlying RF model.

Table 5: Summary statistics for three base indices (CCME, VISC-WQI, SQMCI-hb) and four composite indicators used to estimate random forest models. SQMCI-hb scores have been converted to percentiles to conform to the 0-100 scale used for the other sub-indices. The last column shows the percent of variance explained by each model.

Index	Description	Site number	Median	Mean	SD	% variance explained
CCME-WQI	CCME-WQI	525	10.1	13.8	13.2	54.6
VISC-WQI	VISC-WQI	401	50	49.9	26.4	52.3
SQMCI	semi-quantitative MC index	519	105	105.2	18.9	71.5
CI3min	min(CCME-WQI, VISC-WQI, SQMCI-hb)	153	12.3	16.5	14.3	46.2
CI3mean	mean(CCME-WQI, VISC-WQI, SQMCI-hb)	153	59.3	62.6	21.2	54.6
CI3median	median(CCME-WQI, VISC-WQI, SQMCI-hb)	153	43.8	49.8	26.5	33.0
CI2mean	mean(CCME-WQI, VISC-WQI)	316	26.5	29.7	19.5	55.4

Random forest models of the CI3 indices (composite indices based on CCME-WQI, VISC-WQI and SQMIC-hb) were constrained by the limited number of sites used in the models. These site numbers were in turn limited by the scarcity of sites for which both water quality and macroinvertebrate data were available and met the rules in Section 2.3. There were 153 sites with suitable data for composite index calculations. Of these, 59 (39%) are in the lower North Island (Wellington, Wairarapa, Manawatu), and 29 (19%) are in Southland (Appendix 3). The remainder of New Zealand is represented by only 65 sites (42% of the total), of which 45 are derived from the NRWQN and 20 from a further five regional councils. In terms of REC climate/land-cover classes representing more than 5% of New Zealand (i.e., 30,000 – 95,000 reaches), the number of suitable sites per class ranged from 1 to 42, with WD/L and CX/M severely under-represented, and CW/L and CW/H strongly over-represented. Spatial coverage was better for CI2mean, and the number of suitable sites more than doubled to 316 when SQMCI-hb was omitted, but the site distribution was still affected by clusters in Waikato, the lower North Island, and Southland (Appendix 3).

Comparisons of the percentage of variance explained by each composite index model suggests that using means of the sub-index scores produced better fits than the minima or medians (Table 5). We attribute this to the relatively discrete ranges of values for the three available components, particularly the tendency for CCME-WQI (range ~0-40) to be much lower than either of the other two (range ~ 30-80). Consequently, CI3min (the minimum of the three available indices) is usually identical to the CCME-WQI score, and so inherits its underlying distribution along with any of its undesirable features. In particular, CI3min has the lowest standard deviation of any of the composite indicators, suggesting that it provides less discrimination between individual reaches than the remaining three indices. Conversely, CI3median has the broadest distribution (standard deviation = 26.5), but the corresponding RF model gave the poorest fit (explained variance 33.0%).

3.4.2 Model predictions

Mapped predictions of the four composite indicators appear, at first glance, to be strikingly different (Appendix 3), but are in fact broadly similar apart from arbitrary changes in colour resulting from their different numerical ranges. Mapped values generally range from 5-35 for CI3min; from 45-90 for CI3mean; from 35-75 for CI3median; and from 15-60 for CI2mean. In all cases, index values are highest in elevated inland regions, and lowest in lowland areas.

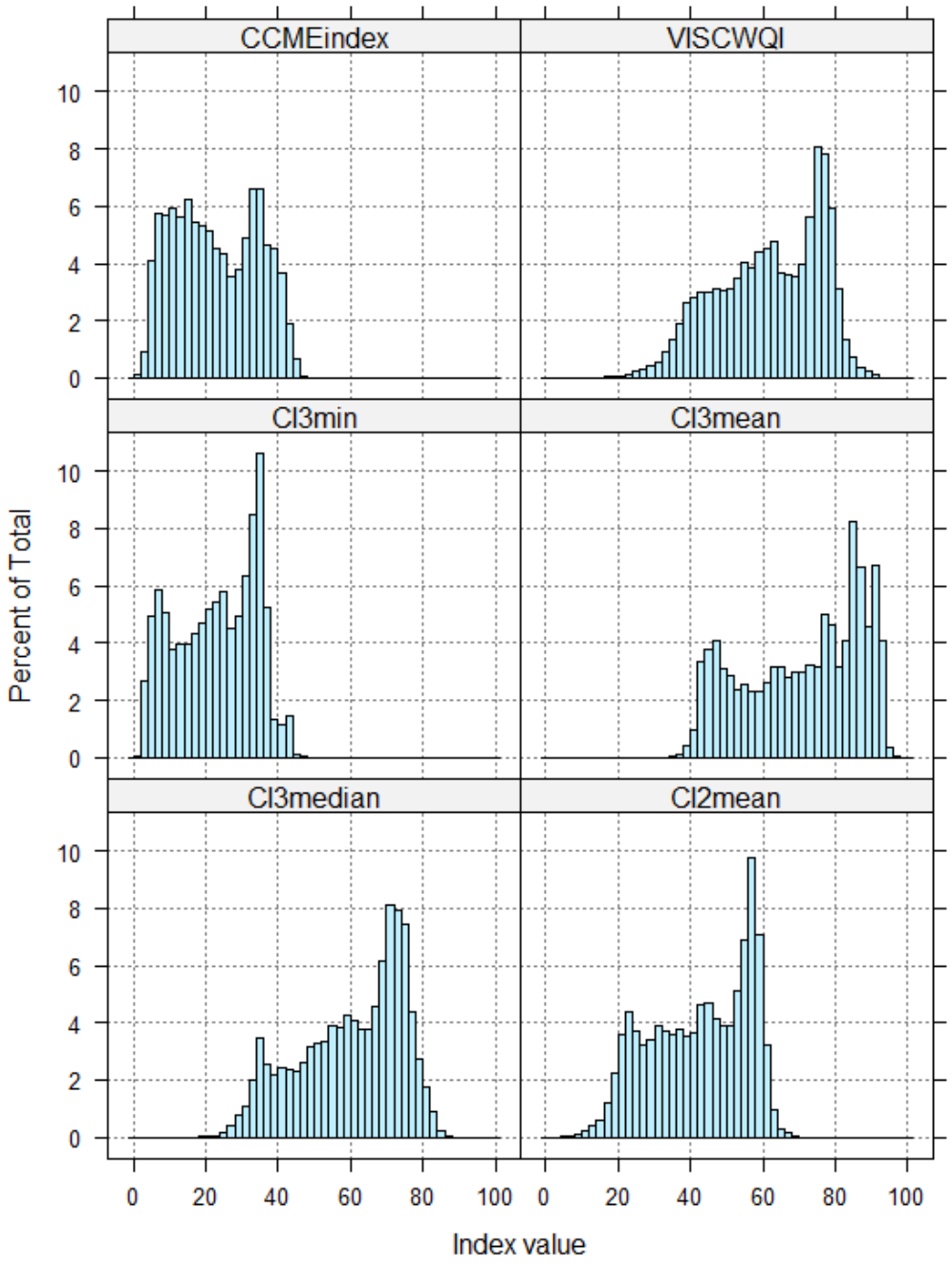


Figure 2: Distribution of predicted values for six water quality indices over all NZReaches (N = 574,502). CI3min, CI3mean, and CI3median refer to the minimum, mean, and median of CCME, VISC-WQI, and SQMCI, respectively. CI2mean is the mean of CCME and VISC-WQI.

3.5 Deliverable 5. Random forest models for predicting temporal trends in macroinvertebrate community indices

3.5.1 Observed trends

After reviewing the data yield for each date-filtering rule (Section 2.3), we restricted our analysis of trends in invertebrate indices to 331 sites for which at least 10 annual records were available since 1 January 2000[3]. Shorter periods of record (e.g., at least 5 years from 1 January 2005) yielded a large number of sites with apparently significant trends, but most of these trends were spurious. In particular, a 5-year analysis of the same 331 sites used for our 10-year analysis produced 247 statistically significant trends (pooled across NTaxa, SQMCI-hb, EPTtaxa, and %EPTabund). Only 38 of these trends remained significant for the corresponding 10-year analysis, suggesting that trends based on only 5 years of data are not informative. Longer periods of record would have yielded more robust trends for individual sites, but such sites were too few (15 years: 101 sites; 20 years: 58 sites) and too unevenly distributed to justify developing a national-scale predictive model. Sites with data that were suitable for the 10-year analyses were distributed across New Zealand, except for a gap in Auckland. The sites used in the trend analyses included clusters in Taranaki, Nelson City, Kaikoura, Banks Peninsula, Timaru, and (to a lesser extent) Southland (Appendix 2).

Trends that were significantly different from zero at the 95% level were present at 59 (18%) of sites for NTaxa; 66 (20%) of sites for SQMCI-hb; 62 (19%) of sites for EPTtaxa; and 60 (18%) of sites for %EPTabund (Appendix 2). EPTtaxa and %EPTabund showed little evidence of consistent regional patterns, but trends in NTaxa and SQMCI-hb were generally consistent (although in opposite directions) across multiple sites in Southland, Taranaki and Northland. These results suggest that NTaxa has increased slightly in Northland and Southland, and decreased in Taranaki. Conversely, SQMCI-hb appears to have decreased slightly in Northland and Southland, and increased in Taranaki.

3.5.2 Modelled trends

Model fits for 10-year trends were fair for NTaxa (explained variance: 45.1%); poor for EPTtaxa (explained variance: 25.0%); and minimal for SQMCI-hb and %EPTabund (explained variance: 10.7% and 9.7%, respectively). Only the model for NTaxa is discussed further in this report.

Most of the variance explained in trends in NTaxa trends was associated with rainfall and climate, with annual rainfall variability (IS = 28.6) and maximum annual temperature (IS = 20.1) the only two predictor variables for which the IS exceeded 20 (Appendix 2). Predictor variables relating to catchment land-cover made little contribution to the fit, with only one such predictor (percentage of heavy pastoral cover) in the top ten (importance score = 8.9; ninth in overall importance). The mapped predictions suggest some large-scale spatial patterns in the trend direction, but in many areas these bear little relationship to environmental or climatic gradients. For example, boundaries between regions where the 10-year trend changes from increasing (blue shading) to decreasing (red shading), in areas such as Central Otago, northwest Nelson, the Southern Alps, and Manawatu/Rangitikei,

[3] We refer to these as "10-year trends, although the data set includes 38 sites (12% of the total), and 25 sites (8% of the total), for which data were available up to 2011 and 2012, respectively.

suggest environmental gradients which are inconsistent with the generally homogenous landscapes and land-cover in these regions.

3.6 Deliverable 6. Random forest models for predicting trends in physical and chemical water quality variables

3.6.1 Observed trends

The parallel analyses of 10-year trends using monthly and quarterly time-series highlighted the trade-off between data availability and the ability to detect trends. There were fewer sites in the monthly analyses than in the quarterly analyses (monthly: N = 124 to 448 sites; quarterly: N = 175 to 531 sites), but there was a slightly higher percentage of meaningful trends in the monthly analyses (monthly: 9% to 51% of sites with meaningful trends in a water quality variable; quarterly: 11% to 41% of sites with meaningful trends; Table 6). However, we repeat the caution that these figures have not been adjusted to allow for multiple trend analyses, and may overestimate the number of significant trends. Results for the monthly analyses are summarised in Appendix 1; most of the results for the quarterly analyses are almost identical to the monthly analyses (apart from the presence of additional sites) and are not shown.

The results in Table 6 show some evidence of consistent trends in water quality, notably for CLAR, NH4N, DRP, and TP, all of which appear to be decreasing significantly more often than they are increasing. Taken at face value, these results suggest that water clarity may be declining, which represents a deterioration in water quality, and concentrations of some nutrients are also declining, which represents improving water quality. However, in addition to potentially overestimating the number of significant trends, simple tabulation of positive and negative trends gives no insight into their underlying spatial distribution. Coherent trends at regional or sub-regional scale are likely to be far more meaningful, from an SoE perspective, than trends with substantial variation at smaller spatial scales (e.g., between adjacent monitoring sites).

To help evaluate these spatial patterns, each water quality plot in Appendix 1 includes a panel at lower left highlighting site by site variation in trends for each variable. Coherent regional-scale trends are apparent for CLAR, TURB, NO3N, and TP (Appendix 1). Interpretation of these results is limited by the lack of data for some regions, but they suggest that CLAR has declined (and TURB has increased) in Waikato; that NO3N has increased in Waikato and Southland; and that both NO3N and TP have decreased in the lower North Island. Trends in other variables (e.g., ECOLI, DRP) were less consistent, with increasing and decreasing trends often apparent at neighbouring sites in the same region. For three variables (TEMP, DO, and DOSAT) there was no evidence of spatially coherent increasing or decreasing trends. These results are consistent with those reported by Ballantine et al. (2010) for variables common to both datasets.

Table 6: Number of sites showing significant and meaningful trends in 12 water quality variables, 2000-2010, based on monthly and quarterly time series. Successive columns for each time-series show the total number of sites, and the percentage of these sites for which a significant (P < 0.05) and meaningful increasing or decreasing trend was detected. For variables marked §, increasing trends imply improving water quality. For other variables, negative trends imply improving water quality. The percentages in the table are likely to overestimate the true number of significant trends; see text for further details. The "significance" column for each variable is the binomial probability that the numbers of increasing and decreasing trends are equal, assuming both to be equally likely. Thus, for the monthly CLAR analyses, "significant and meaningful" trends were apparent at 35% (24% + 11%) of 340 sites, i.e., 81 and 37 (of 118) sites, respectively. The binomial probability of this occurring by chance is 0.00003.

Water quality variable	Monthly time series				Quarterly time series			
	Number of sites	% significant & meaningful			Number of sites	% significant & meaningful		
		decrease	increase	significance		decrease	increase	significance
CLAR §	340	24%	11%	<0.0005	405	20%	7%	<0.0005
TSS	124	15%	7%	0.044	175	15%	7%	0.012
TURB	418	15%	15%	0.500	509	13%	10%	0.132
TEMP	448	3%	6%	0.010	531	4%	8%	0.007
DO §	436	3%	17%	<0.0005	508	3%	16%	<0.0005
DOSAT §	398	2%	12%	<0.0005	481	1%	11%	<0.0005
ECOLI	301	13%	8%	0.065	385	11%	5%	0.002
NH4N	346	21%	5%	<0.0005	355	17%	4%	<0.0005
NO3N	420	21%	26%	0.088	494	14%	26%	0.000
TN	164	24%	14%	0.021	194	15%	12%	0.244
DRP	442	40%	11%	<0.0005	517	32%	10%	<0.0005
TP	400	30%	10%	<0.0005	428	25%	6%	<0.0005

3.6.2 Modelled trends

Estimated 10-year trends in water quality variables from RF models were essentially identical for the monthly and quarterly datasets. Model fits were fair for NO3N (explained variance: 41.8%); poor for CLAR, DO, and TN (explained variance: 26.3 – 36.1%); and very poor (explained variance < 20%) for all other variables (Table 7). Importance scores were low for most of the predictor variables in each model, with only five scores exceeding 15, and one exceeding 20. The results are also notable for the rarity of significant predictors representing land-cover, for which the maximum importance score was 12.2. In particular, the percentage of heavy pastoral land-cover, which was a leading predictor variable in the RF models of median values for water quality variables (Section 3.2.2, Table 2), was virtually absent from the RF models of trends in the same variables; the maximum importance score for heavy pastoral land-cover in the trend models was 7.2. In the interests of fully documenting these results we provide maps and diagnostic plots for all models in Appendix 1, but emphasise that these should be interpreted as illustrating the difficulty of obtaining credible fits rather than providing information to be used for State of the Environment reporting.

One reason for the weakness of these models may be that national trends are confounded by regional variation in management practices. For example, increasing trends in one region may be partially cancelled out by decreasing trends in another region with similar environmental characteristics but contrasting management practices. If so, future models may need to take this variation into account.

Table 7: Importance scores for predictors of 10-year trends in water quality variables. The percent of variability in each variable explained by the RF model is in the top row, in parentheses. See footnotes to Table 2 for formatting conventions used to indicate relative importance.

Predictor	CLAR (36.1%)	TSS (9.6%)	TURB (19.1%)	TEMP (13%)	DO (33.3%)	DOSAT (26.4%)	ECOLI (9.4%)	NH4N (5.4%)	NO3N (41.8%)	TN (26.3%)	DRP (8.6%)	TP (18.2%)
Reach elevation	7.2	1.9	4.9	9.9	3.6	5.9	5.0	2.4	4.5	4.3	1.1	1.4
Catchment elevation	5.9	5.4	6.6	10.5	7.1	10.1	9.5	10.3	6.9	4.8	4.0	6.7
Mean slope	10.4	4.9	9.2	12.5	7.7	7.6	10.4	8.4	13.3	6.8	4.1	5.7
Catchment area	8.9	1.6	7.0	8.0	4.5	5.5	8.9	5.1	5.0	8.0	6.6	4.0
Lake index	8.1	1.7	5.7	4.5	5.4	4.8	-3.4	2.9	3.9	1.6	-0.6	6.3
Mean flow	8.9	4.0	7.3	8.8	4.3	6.0	8.7	4.4	5.2	7.2	4.0	4.4
Rain variability	9.6	0.1	5.8	8.5	11.9	9.9	5.8	4.4	12.4	9.5	4.3	6.7
Min temperature	**21.5**	5.4	11.2	12.2	9.0	8.4	10.2	10.9	10.3	7.5	4.3	7.8
Max temperature	6.9	8.9	9.1	13.0	10.5	8.8	8.2	12.9	18.9	11.6	5.6	9.3
Rain days > 10	11.0	5.9	12.4	10.7	8.2	8.7	12.4	8.9	11.1	6.3	14.5	7.7
Rain days > 50	9.1	10.0	12.1	10.6	13.9	8.6	7.7	8.0	11.5	6.6	5.5	11.2
Rain days > 200	6.5	5.5	8.5	6.6	12.4	11.2	1.5	4.4	5.2	1.9	5.7	11.6
Evapotranspiration	7.8	4.0	7.0	12.1	6.1	7.3	8.4	8.1	11.1	1.7	7.0	3.3
%alluvium	15.1	6.7	15.6	7.5	10.8	9.3	5.7	8.6	10.8	6.6	2.9	6.6
%glacial	0.6	1.5	1.3	-2.7	-0.2	2.0	1.8	0.1	2.7	3.6	-1.7	1.7
%peat	3.6	2.8	3.4	5.6	8.6	11.1	-0.3	6.3	2.7	-1.2	9.8	2.6
Calcium	6.2	3.4	4.2	8.8	7.1	5.1	5.6	6.8	7.7	3.6	3.6	4.1
Hardness	6.3	2.5	4.3	9.4	6.5	3.7	7.5	8.3	6.5	4.5	8.3	8.7
Particle size	10.9	2.7	9.4	8.3	7.2	5.7	6.3	7.1	9.1	5.0	4.5	9.1
Phosphorous	7.1	6.6	5.6	8.9	9.9	4.7	8.0	7.9	19.2	6.3	4.0	7.4
%bare	10.2	3.1	8.5	8.1	6.1	5.1	6.3	4.7	5.3	4.1	4.9	4.5
%exotic forest	9.7	3.6	4.0	7.0	4.4	4.7	3.5	3.0	7.2	8.5	2.2	6.5
%indigenous forest	6.2	8.6	7.6	8.2	9.3	6.9	7.2	4.9	8.5	4.3	8.1	7.4
%pastoral heavy	6.3	7.2	5.6	7.2	5.1	6.5	6.7	6.6	5.9	5.6	4.7	5.2
%pastoral light	6.5	1.5	6.8	12.2	6.6	5.7	9.6	7.1	8.6	4.9	6.9	5.3
%scrub	6.3	4.6	1.8	7.1	3.7	2.8	7.0	-0.1	5.7	0.9	5.1	3.7
%urban	3.3	5.9	7.9	8.5	1.3	5.7	2.3	8.8	7.3	3.4	3.7	4.2
%wetland	5.7	0.2	6.0	6.2	7.6	9.3	3.1	1.5	5.4	4.0	4.2	6.9

4 Discussion

4.1 Random forest models

The explanatory variables used in the RF models were large-scale, catchment-averaged descriptors of climate, topography, geology and land-cover. These accounted for 50% or more of the variation in all physical-chemical water-quality variables except DOSAT (43%). The same explanatory variables accounted for 55-72% of the variation in four invertebrate community indices, and 33-55% of the variation in five multi-metric indices. Some of the remaining, unexplained variation in the response variable is due to small scale processes such as nutrient uptake and regeneration, to small-scale spatial heterogeneity such as habitat patchiness, and to time-dependent processes such as flow fluctuations and diurnal variation in DO, DOSAT and TEMP. Given the large spatial scales at which the catchment-averaged variables operate, the model performance for water quality and invertebrate variables was generally good.

Several of the RF models for the physical-chemical water-quality variables were updates of models run with 2003-2007 data in the previous study (Unwin et al. 2010). For these models, the predictor variables with high importance scores were generally consistent in both studies, and model performance (as explained variance) was comparable. For three variables (CLAR, TSS, ECOLI), the newer models performed better due to increased site numbers and broader site distributions. For four variables (NH4, DRP, TP, NO3), model performance in the 2010 study and the current study was very similar. For TN, there was a moderate reduction in explained variance, from 78% in the 2010 study, to 74% in the current study. This reduction was presumably due to the elimination of > 100 monitoring sites in the current study, due to non-comparable TN measurement methods (Larned & Unwin 2012).

Model performance for the water quality variables that were not used in the 2010 study were mixed, with good, fair, and poor fits for TEMP, DO, and DOSAT, respectively. All three variables were most strongly related to geographical predictor variables such as elevation and slope, and to hydrology and climate. Catchment land-cover was a weak predictor of TEMP and had essentially no explanatory power for DO and DOSAT, possibly because the data available to us were confounded by variation in time of day. The primary controls on DO and DOSAT in rivers are temperature and flow (which control oxygen solubility), and living and decomposing organic matter and light (which control oxygen production and consumption). None of these are stable landscape-variables; rather, they are time-dependent variables that also vary at small spatial scales. Our results do not necessarily imply that DO and DOSAT are unsuitable for modelling water quality in terms of landscape variables, but it seems likely that such models will require more robust and consistent field data than were available to us for this study.

4.1.1 Commentary on random forest models for analysing and reporting

The overall goal of the RF models used in this study was to extrapolate median values and temporal trends in water quality variables and indices from monitoring sites to the entire country. Extrapolation is a fundamental step in analysing monitoring data and reporting the results for two general reasons. First, the number of monitored sites is inevitably far smaller than the total number of sites (e.g., < 0.02% of the river reaches in New Zealand are monitored). Therefore, predictions of state and trends at large spatial scales require up-

scaling from individual sites. Second, extrapolation is needed to identify unmonitored sites or areas where rivers are likely to be degraded or at risk of degradation.

While there are clearly benefits to extrapolation, there are also some risks. Statistical models with low explanatory power (due to the choices of predictor variables) can lead to high uncertainty in predictions about unmonitored sites. Predictions for river reaches in sparsely monitored and unmonitored environments can also have high uncertainty, and hence low explanatory power. High uncertainty limits our ability to reliably identify spatial patterns in water-quality and ecology or detect temporal trends. It can also reduce the likelihood that management actions will be effective. Finally, maps or tables of predicted water quality conditions that do not convey uncertainty can be misleading to stakeholders.

In the future, several approaches can be used to reduce uncertainty in predicted state and trends in river water quality and ecology. Models with different suites of predictor variables can be trialled to assess their explanatory power. New monitoring sites can be established to fill gaps in the environmental gradients used in the models. If new permanent sites are prohibitively costly, then validation data collected from temporary sites can be used to test the models.

The set of predictor variables used for the present study was identical to that used for the 2010 study (Unwin et al. 2010), and has not been optimised to identify and eliminate highly correlated variables. The RF literature (e.g., Breiman 2001) and our previous experience with RF-model studies led us to believe that they were relatively immune to over-fitting, but as we have gained experience in their use and interpretation we now acknowledge that this is not necessarily the case, and that eliminating highly correlated predictors may improve model performance. The calculations required to identify the optimal predictor set for each water quality variable are computer intensive and were beyond the scope of the present study.

4.2 CCME-WQI, VISC-WQI and composite indices

4.2.1 Comparisons among indices

The current study is the first to trial multi-metric water-quality indices and composite indices for reporting national-scale state and trend in New Zealand, and the first to use these indicators as response variables in statistical models. The NEMaR expert panel for the indicators work-stream identified TURB, CLAR, TEMP (continuous), DO (continuous), ECOLI[4], NO3N, NH4N, TN, DRP, TP and electrical conductivity as the primary or “core” variables to be included in calculations of CCME-WQI. However, we have shown that the number of monitoring sites with suitable data decreases steeply as the number of variables used in any of the multi-metric indices increases. Since site number has a strong effect on model performance and on environmental coverage, it was not realistic to use all 11 core variables.

The CCME-WQI was characterised by uniformly low scores across New Zealand. The primary reason for low scores was the choice of references conditions used as objectives for each water quality variable, as discussed in the following section. The narrow range of calculated CCME-WQI scores for monitoring sites led to a narrow range of predicted scores in the RF model. As a consequence, the CCME-WQI model was relatively uninformative; it

[4] E.coli is used in MfE’s recreational water quality indicator

divided Zealand into two general categories, one with marginal water quality (Southern Alps, Karamea-Kahurangi, Fiordland, Coromandel and East Cape) and one with poor water quality (the remainder of the country; Appendix 3).

Results of the VISC-WQI model were more informative in the sense of having a wider range of predicted scores (Appendix 3). The number of monitoring sites with suitable data for calculating VISC-WQI was 30% smaller than for CCME-WQI, but the percent of explained variance was comparable for both indices. This suggests that the VISC-WQI model could be substantially improved by including more sites. As with the CCME-WQI, the water quality variables used in VISC-WQI are not fixed, and model performance can be evaluated with VISC-WQI scores based on different sets of variables.

The composite indices that we trialled were severely constrained by the number of monitoring sites with suitable data for calculating each of the sub-indices used in each composite index. As a preliminary step in addressing this problem, we trialled composite indices with two (CI2) and three (CI3) sub-indices. For CI2, we used two water-quality sub-indices, CCME-WQI and VISC-WQI. There were 316 sites with suitable data for CI2. For CI3, we used CCME-WQI, VISC-WQI and SQMCI-hb; there were 153 sites with suitable data. The use of two water-quality sub-indices in these composite indices may be redundant, but the addition of one or more invertebrate sub-indices was constrained by the scarcity of sites at which invertebrates and multiple water quality variables are monitored. The alternative combinations of two sub-indices, SQMCI-hb with VISC-WQI and SQMCI-hb with CCME-WQI, were not modelled because the number of suitable sites dropped by 32-44% for these combinations, compared to CI2. For both CI2 and CI3, the limited number of sites resulted in low spatial densities of sites in the RF models, and very uneven site distributions (Appendix 3). We also note that fitting models with insufficient data makes the results increasingly susceptible to over-fitting, particularly when the predictor set (currently 28 variables) is large relative to the number of points to be fitted. Detailed investigation of the potential for over-fitting was beyond the scope of this study, but we suggest that the weak predictive performance for models based on less than ~300 sites, even for indices (e.g., CI3mean) for which percentage explained variance exceeded 50%, is at least partly attributable to over-fitting. We anticipate that future standardisation of monitoring site procedures will lead to an increase in the number of sites with both invertebrate and water-quality data. At that time, a new version of CI2 should be calculated using one invertebrate index and one water-quality index, and its performance reassessed.

Composite indices can be defined as the mean, median or minimum of the standardised sub-index values. We found that using the minimum was uninformative because the sub-index with the smallest value was almost always the CCME-MCI value, which made the CI3min model and the CCME-WQI model nearly identical. Using the median of the standardised sub-index values was more informative, but this approach can only be used when there are three or more sub-indices, which severely limits site numbers. Using the mean of the sub-index values was more informative than the minimum, and this approach works with two or more sub-indices.

The best-performing RF models of composite indices were for CI3mean and CI2mean; both models explained about 55% of the variance in index values. Since there were over twice as many sites with suitable data for CI2mean compared with CI3mean, it is likely that CI3mean will out-perform CI2mean with a similar number of sites.

4.2.2 Commentary on CCME-WQI, VISC-WQI and composite indices for analysing and reporting

Our trials of the CCME-WQI, VISC-WQI and composite indices demonstrate the fact that calculated and predicted values for each index are highly context-dependent. The calculated values and the categories used to group those values (e.g., excellent, marginal, poor) are strongly affected by decisions made at multiple steps in the analysis process. The steps include:

1. Number and identity of variables used in index calculations
2. Length of record
3. Reporting upper limits
4. Reference conditions or “objective values”
5. Data configuration
6. Value scaling

The effects of decisions at each of these steps are summarised below.

The number and identity of water-quality variables. The selection of water-quality variables used in the CCME-WQI and VISC-WQI, and the number and type of sub-indices used in composite indices are not predetermined; they are selected by the water quality analyst. This property of the indices influences their values and model performance in several ways. First, as the number of water quality variables increases, the proportion of variables and the proportion of samples that exceed the objectives both increase in the CCME-WQI. These exceedances drive down the index values. Second, in assessments of relationships between land-use and water quality index values, selecting variables that are insensitive to land-use variation will result in weak relationships, and vice versa. Third, the number and identity of sub-indices used in composite indices strongly effects composite index values, as discussed in Section 3.4.1. Fourth, as the numbers used to in the indices increases, the number of sites with suitable data decreases, which effects model performance as discussed below.

Length of record. Two of the three components in CCME-WQI calculations tend to increase as the length of record for a monitoring site increases, independent of water quality conditions at the site. One component is F1, the proportion of water quality variables for which an exceedance of an “objective” level occurs at least once in the record. The other component is F3, “amplitude”, or the amount by which variable measurements exceed their objectives. The probability of encountering large exceedances increases directly with the length of a data time-series. These properties will cause CCME-WQI values to decrease as the length of record increases.

Reporting upper limits. The F3 (amplitude) component of the CCME-WQI index is potentially sensitive to the presence of outliers. For most water quality variables used in the current study we were able to remove obvious outliers by inspecting quantile plots for each variable. This was not possible for ECOLI, for which upper detection limits varied markedly among regions (see Section 2.2). In the absence of a consistent code of practice for reporting water

quality variables, a possible interim solution to this problem would be to truncate all raw ECOLI counts for all sites above some fixed upper limit. However, this would require another ad hoc (and context-dependent) choice.

Reference conditions. The reference values used for this study to calculate CCME-WQI were based on median reference conditions estimated for each nutrient and ECOLI in the absence of agricultural land-cover (McDowell *et al.* 2013). These reference values are quite stringent; they are all higher than the national water quality trigger values defined by the Australian and New Zealand Environment and Conservation Council (ANZECC). Also, by using medians for reference values, we expect exceedances for each variable to occur 50% of the time at sites with no agricultural or urban land-cover. More lenient reference values, such as the corresponding 80th, 90th, or 95th percentiles above the medians or the ANZECC guideline values, would immediately generate higher CCME-WQI scores. This in turn would produce a more positive view of water-quality conditions across New Zealand. The effects of the choice of reference conditions on CCME-WQI is one example; any other metric or indicator that use reference conditions as a variable will be similarly affected. The use of observed:expected ratios for water quality assessment, where the expected values correspond to reference conditions, is also affected by the choice of reference conditions.

Data configuration. For multi-year datasets such as the ones used in the present study, the data used in calculating index values can take the form of annual, quarterly or monthly means or medians, or raw time series. If means or medians are used, extreme values will be masked and their effects on exceedences will be reduced. If raw time series are used, there will be more and larger exceedences, as discussed above for length of record.

Value scaling. When water quality or ecological index scores are uniformly low, uniformly high, or strongly skewed for sites across a large, environmentally heterogeneous area, those indices may be seen as uninformative. This problem could be addressed at two levels, the distributions of index scores among sites, and the classes used to group sites by their scores. For example, the CCME-WQI scores in the current study were generally low and highly left-skewed; scores for 98% of the sites were below 45. Low scores could be raised by re-normalising to a median of 50 and a range of 0-100. As a result of the preponderance of low scores, 98% of the sites were classed as poor and 1% were classed as marginal. To generate more information, the categories could be subdivided (e.g., poor, very poor, extremely poor), or rescaled, so that the highest scores in the marginal and poor categories become “excellent”, the next highest scores become “good” and so on. Clearly, the thresholds between categories are arbitrary. Furthermore, the distribution of sites among categories is dependent on the variables used, the length of record, and the choice of reference conditions, as discussed above.

Due to the multiple steps at which data analysts must make decisions, a single dataset analysed by different analysts is certain to produce different results. At this early stage in the development of water-quality indices for New Zealand, there are no standard calculation procedures; instead the decisions listed above are made on an ad hoc basis. This will lead to inconsistencies when updating reports between years, and between regional councils, MfE and other organisations. Resolving this problem should be straight-forward. First a systematic analysis should be carried out to identify the consequences of different choices made at each decision step (e.g., compare CCME-WQI performance using the different reference conditions listed above). Second, the results of the systematic analysis should be

used to recommend standard procedures. Since some decision steps affect conclusions about water quality status, the standard procedures should be developed in consultation with stakeholders.

Desirable features in water-quality indices that are intended to produce fine-scale predictions across New Zealand include a well distributed network of reference sites; a well-fitting model; and a distribution of modelled values which maximally spans the theoretical range (0-100, in the present study) for each index. None of the indices considered in this study satisfied all three criteria. Some of these deficiencies can be addressed by adjusting the rules used to derive each index. In particular, as noted in Section 3.3.1, the CCME-WQI could be made less punitive by adjusting the reference values associated with each REC class. Other deficiencies were caused by data limitations rather than the properties of the indices themselves. For both the CCME-WQI and VISC-WQI, we based our calculations on minimal subsets of variables and sites, due to the lack of suitable data for all variables at most sites. If in the future, water-quality indices can be calculated using multiple variables, at a large number of sites that represent the entire range of river environments in New Zealand, this would almost certainly yield more tractable results. Such a dataset does not yet exist, but one of the major goals of the NEMaR project is to ensure that better and more representative datasets will become available in the future.

The preceding discussion focused on ways to improve the use of water quality indices for summarising New Zealand monitoring data. One final caution concerns the extrapolation of those indices from monitoring sites to unmonitored river reaches. Most of the steps used to calculate each index for each site involve some form of averaging or smoothing. Raw site data are either converted to medians for a specified time period, or summarised as counts of exceedances. The resulting values are then combined in index calculations, and these index values are themselves combined in composite indices. RF models are then fitted to the index values, which introduces yet another level of smoothing, which further reduces detail.

4.3 Trend analyses

Trend analyses were carried out for water-quality variables and for the four invertebrate indices (NTaxa, SQMCI-hb, EPTtaxa, %EPTabund) for the period from 2000 to 2010/2011/2012. We reported both the observed trends at monitoring sites, and the national-scale predicted trends from RF models.

4.3.1 Observed trends

One of the practical objectives for the water quality trend analyses was to assess the effects of monthly versus quarterly data on trend detection and trend modelling. Since most regional councils monitor their SoE sites at least quarterly, the use of quarterly data can increase the number of sites used in trend analyses, and this may improve RF model performance. Conversely, the use of monthly data in lieu of quarterly data may increase our ability to detect statistically significant trends at individual monitoring sites. After filtering sites and dates based on the rules set out in Section 2.3, the differences in monthly and quarterly site numbers were modest. Across the 12 water-quality variables, the number of quarterly sites with sufficient data was 7-29% higher than the number of monthly sites. The effect of shifting from quarterly to monthly data on the detection of significant and meaningful trends was minimal. For all water quality variables, the number of sites for which trends were detected differed by less than 10% between the quarterly and monthly datasets. These observations

suggest that future trend analyses can include quarterly data without a substantial loss of statistical power.

Although there are large gaps in monitoring sites with data suitable for trend analyses, large-scale patterns in meaningful (> 1% per year) trends were apparent for CLAR, TURB, NO3N, and TP. There are multiple sites in Waikato with negative trends in CLAR and positive trends in TURB. There are also multiple sites in Waikato and Southland with positive trends in NO3N. And there are multiple sites in the lower North Island with negative trends in NO3N and TP. Most of these patterns were also observed in the previous trend analysis using data from 1998 to 2007 (Ballantine et al. 2010, Table 21).

An interesting and as-yet unexplained pattern in the observed trends is that most sites with meaningful negative trends in NTaxa also had meaningful positive trends in SQMCI-hb, and vice versa. The reason for a negative relationship between trends in NTaxa and SQMCI-hb is not clear. Assuming that this relationship is not spurious, it indicates that, when the diversity of sites increases or decreases substantially, the taxa that are appearing and disappearing over the trend period are predominately those with low MCI values (e.g., pollution-tolerant dipteran insects, crustaceans, gastropods and worms).

4.3.2 Modelled trends

RF model performance for trends in water quality variables was poor, with the exception of NO3N (42% explained variation for trends in NO3N). The importance scores for predictor variables were uniformly low across the water quality variables, which suggests that catchment-averaged variables were generally not useful for explaining and predicting temporal trends in water quality, as discussed below.

RF model performance for the invertebrate indices ranged from very poor (e.g., < 10% explained variation for trends in %EPTabund) to fair (45% explained variation for trends in NTaxa). Due to the low explanatory power, the predicted trends shown in Appendix 2 for SQMCI-hb, EPTtaxa, are %EPTabund are not reliable. The trends shown for NTaxa in Appendix 2 are more reliable, but the explanatory variables with the highest importance scores (annual rainfall variability and maximum annual temperature) are difficult to interpret. It is possible that some of these predictor variables are driving trends in invertebrate communities, but RF models only indicate broad correlative relationships, not specific causal relationships. At best, we can conclude that trends in NTaxa vary among areas with differing rainfall-runoff and temperature regimes.

The limited explanatory power of models of trends in water quality, invertebrate communities and multi-metric indices was due in part to limited site numbers and environmental coverage (Section 4.1.1), and in part to the calculation process for multimetric indices (Section 4.2.2). It was also related to the choice of explanatory variables. We used the RF models to predict monotonic changes over a 10-year period with explanatory variables that are either constant (e.g., catchment area, altitude, mean slope), or are changing gradually at long time-scales (e.g., annual rainfall, annual temperature). These time-scale mismatches may limit the power of the RF models. It is logical to expect that water quality will track changes in land-use and land-cover; this has been demonstrated in New Zealand with case studies (e.g., Hamill & McBride 2003). However, in the current and previous national-scale trend analyses, land-cover has been implicitly assumed to be constant over the period of interest. In these national-scale trend analyses, land-cover data comes from LCDB imagery captured on a

single date. A potentially more productive approach would be to use data from LCDB1 (1996-1998) and the recently created LCDB3 (2008-2009) to develop predictors representing temporal changes in land-cover, over a time scale commensurate with the water quality and ecology data.

5 Acknowledgements

We thank Lucy Baker, Ministry for the Environment, for project support, feedback, enthusiasm, and patience. We thank the many regional council staff who provided data and information about monitoring programmes. Graham Bryers provided NRWQN data. Ton Snelder and Doug Booker provided programming advice and scripts. Graham McBride reviewed the report.

6 References

Ballantine, D. (2012) Developing a composite index to describe river condition in New Zealand. *NIWA Client Report* HAM2012-131. 62 p.

Ballantine, D., Booker, D., Unwin, M., Snelder, T. (2010) Analysis of river water quality data. *NIWA Client Report* CHC2010-038. 34 p.

Ballantine, D.J., Davies-Colley, R.J. (2010) Water quality trends at NRWQN sites for the period 1989-2007. *NIWA Client Report* HAM2009-026. 39 p.

Boothroyd, I., Stark, J.D. (2000) Use of invertebrates in monitoring. In: K.J. Collier & M.J. Winterbourn, (Eds). *New Zealand Stream Invertebrates: Ecology and Implications for Management*. New Zealand Limnological Society, Christchurch, New Zealand: 344–373.

Breiman, L. (2001) Random Forests. *Machine Learning,* 45: 5-32.

Collier, K.J. (2008) Average score per metric: an alternative metric aggregation method for assessing wadeable stream health. *New Zealand Journal of Marine and Freshwater Research,* 42: 367-378.

Davies-Colley, R.J., Nagels, J.W. (2008) Predicting light penetration into river waters. *Journal of Geophysical Research ,* 113: G03028.

Hamill, K.D., McBride, G.B. (2003) River water quality trends and increased dairying in Southland, New Zealand. *New Zealand Journal of Marine and Freshwater Research,* 37: 323-332.

Larned, S.T., Unwin, M.J. (2012) Representativeness and statistical power of the New Zealand river monitoring network. *NIWA Client Report* CHC2012-079. 55 p.

McDowell, R.W., Snelder, T.H., Cox, N., Booker, D.J., Wilcock, R.J. (2013) Establishment of reference or baseline conditions of chemical indicators in New Zealand streams and rivers relative to present conditions. *Marine and Freshwater Research,* 64: 1-14.

R Development Core Team. (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Sen, P.K. (1968) Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association,* 63: 1379–1389.

Snelder, T.H., Lamouroux, N., Leathwick, J.R., Pella, H., Sauquet, E., Shankar, U. (2009) Predictive mapping of the natural flow regimes of France. *Journal of Hydrology,* 373: 57-67.

Stark, J.D. (1998) SQMCI: a biotic index for freshwater macroinvertebrate coded abundance data. *New Zealand Journal of Marine and Freshwater Research,* 32: 55–66.

Stark, J.D., Maxted, J.R. (2007) A user guide for the macroinvertebrate community index. *Cawthron Report* 1166. 58 p.

Stoddard, J.L., Herlihy, A.T., Peck, D.V., Hughes, R.M., Whittier, T.R., Tarquinio, E. (2008) A process for creating multimetric indices for large-scale aquatic surveys. *Journal of the North American Benthological Society,* 27: 878–891.

Unwin, M.J., Snelder, T., Booker, D., Ballantine, D., Lessard, J. (2010) Predicting water quality in New Zealand rivers from catchment-scale physical, hydrological and land cover descriptors using random forest models. *NIWA Client Report* CHC2010-037. 21 p.

Appendix A Graphical summaries of national state and trend analyses for 12 water quality variables.

We present graphical summaries of our analyses for the 12 water quality variables considered in this report as a series of A3 panel plots, showing one variable per page, as a separate document which accompanies this report. The four main panels for each variable show, respectively, diagnostic plots for the fitted random forest model representing current state (top left); modelled current state for all New Zealand river segments (top right); observed 2000-2010 trends for all sites based on analysis of monthly data (lower left); and modelled monthly 2000-2010 trends for all NZReaches (lower right).

Diagnostic plots for each variable are as follows, from top left:

(1) Observed vs. predicted values for all sites, using the jack knife procedure described in Section 2.4. Both axes are plotted to the same scale, with the diagonal dashed line representing agreement between observation and prediction. The number of observations and the nominal r^2 are also shown (cf. Table 3). Axes for TEMP, DO, and DOSAT are linear; axes for all other variables are logarithmic.

(2) Normal Q-Q (quantile) plot, contrasting the observed distribution of residuals for the fitted data (Sample Quantiles) to the theoretical distribution if the residuals were distributed normally (Theoretical Quantiles, diagonal line). Most models are characterised by large residuals for the most extreme values, indicating a general tendency to overestimate low analyte values and underestimate high analyte values, but perform well over the majority of the observed data range.

(3) Smoothed partial plots (using the default “3RS3R” algorithm as implemented in the smooth( ) function of R Version 2.12.1) for the six most important predictors in each model indicating the modelled response of the dependent variable to each predictor, plotted to a common vertical scale. The “rug” at the bottom of each plot represents the distribution of each predictor variable. Additional insight into the influence of each predictor can be gained by comparing the vertical response range for each plot with the vertical scale on the plot of observed vs. predicted values at top left.

Mapped predictions for each variable are constructed by plotting the centroid coordinates of each REC segment, coloured so as to represent the predicted value, with water quality decreasing from blue to red. Colours for log-transformed variables show successive percentiles in steps of ~5%, rounded as necessary to convenient integer or near-integer values. Colours for TEMP, DO, and DOSAT represent linear steps over the predicted range. Sites for which data were available to estimate each model are represented by black circles.

Appendix B Graphical summaries of national state and trend analyses for four invertebrate community metrics.

We present graphical summaries of our analyses for the four invertebrate community metrics considered in this report as a series of A3 panel plots, also as a separate document, using the same plotting conventions as for Appendix A. Axes for all metrics are linear.

Appendix C Graphical summaries of national state and trend analyses for six multi-metric and composite water quality indices.

We present graphical summaries of our analyses for the six water quality indices considered in this report as a series of A4 landscape panel plots, showing one variable per page. The two main panels for each variable show, diagnostic plots for the fitted random forest model representing current state (top left); and modelled current state for all New Zealand river segments (top right). Diagnostic plots for each variable follow the same conventions as for Appendix A.

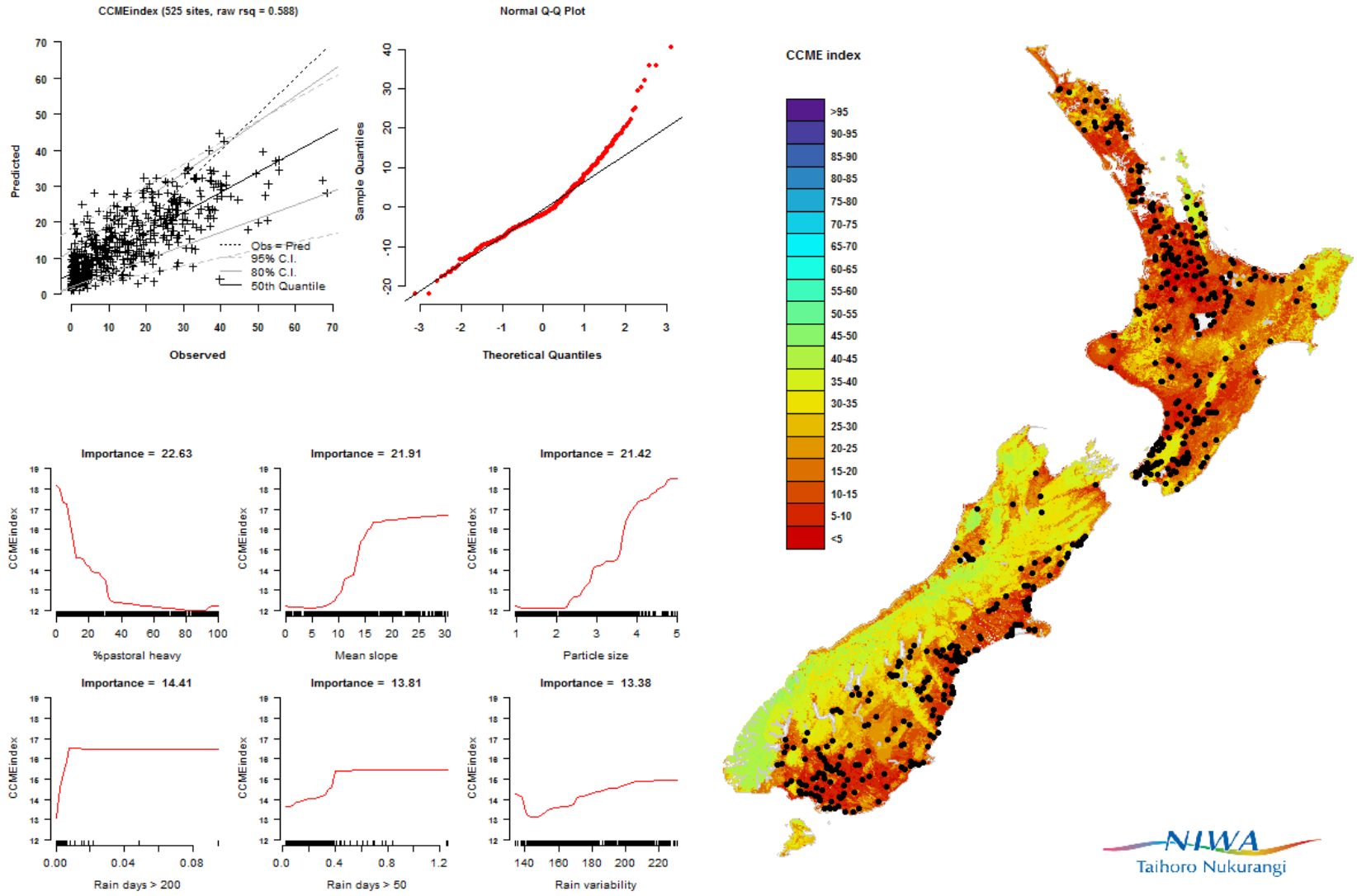


Figure C1: Modelled current state for the CCME water quality index (CCMEindex), showing diagnostic plots for the fitted random forest model (left), and modelled indices for all NZReaches (right).

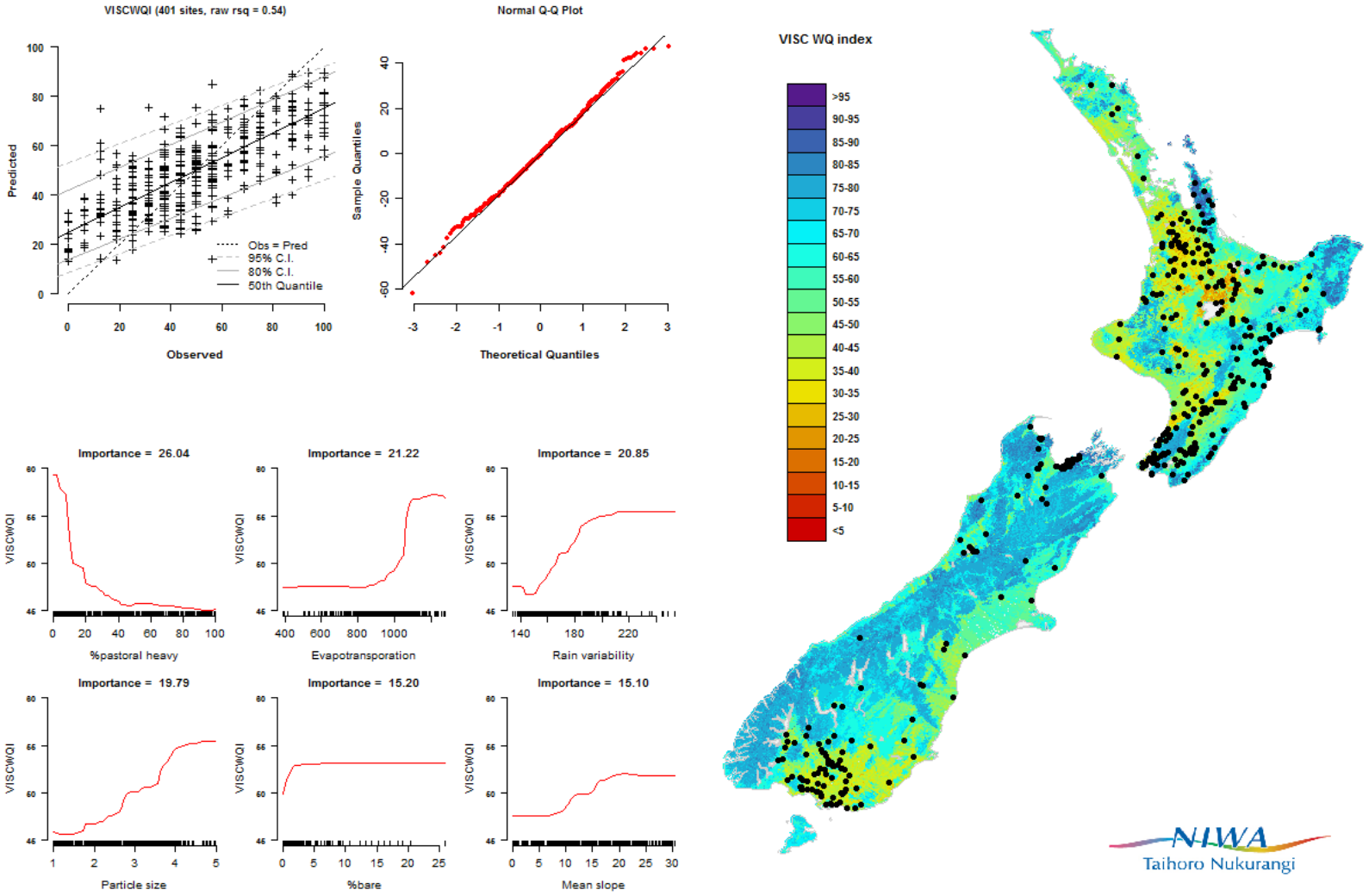


Figure C2: Modelled current state for the VISC water quality index (VISCWQI), showing diagnostic plots for the fitted random forest model (left), and modelled indices for all NZReaches (right).

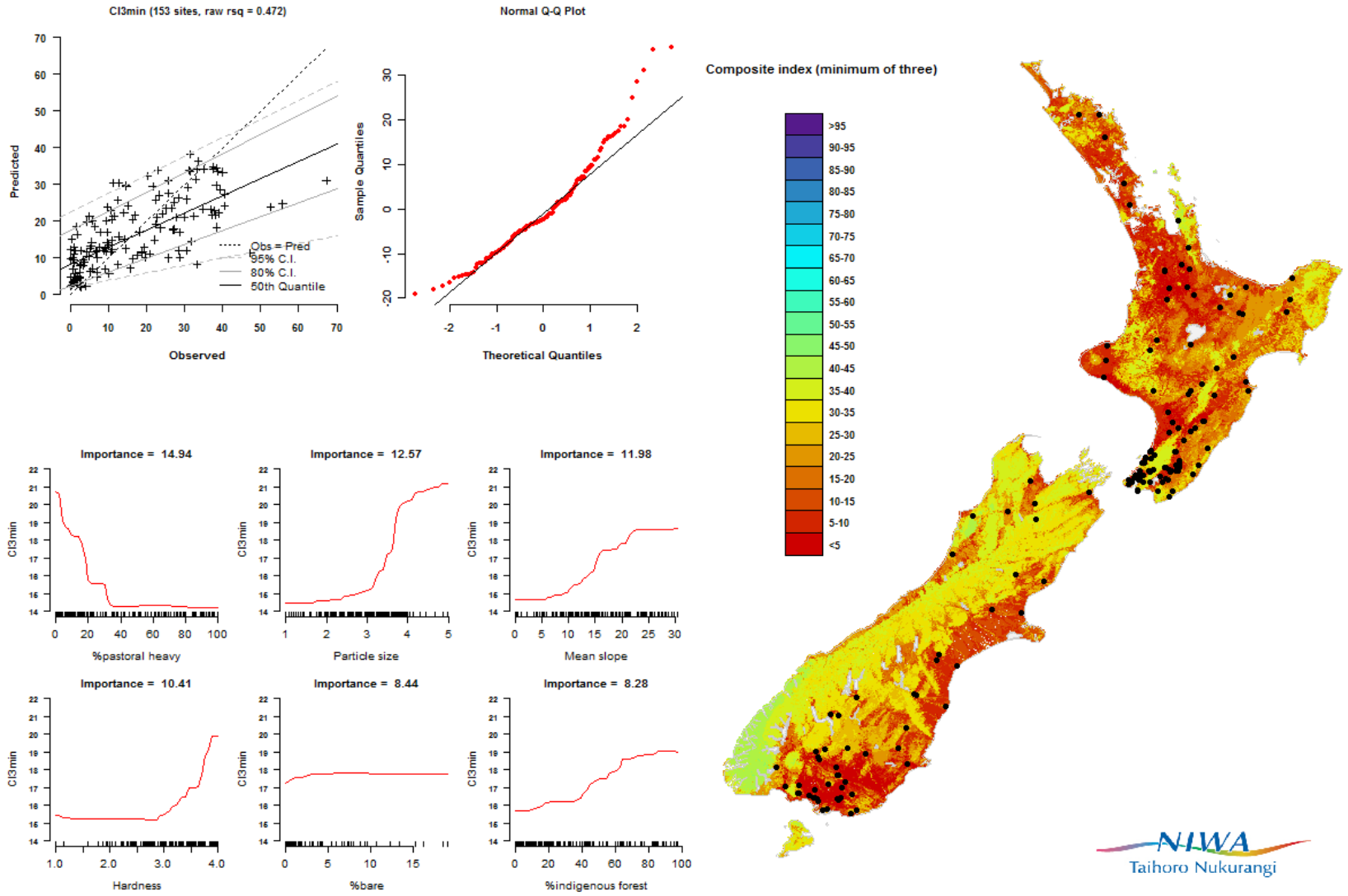


Figure C3: Modelled current state for a composite water quality index based on the minimum of three component indices (CI3min), showing diagnostic plots for the fitted random forest model (left), and modelled indices for all NZReaches (right).

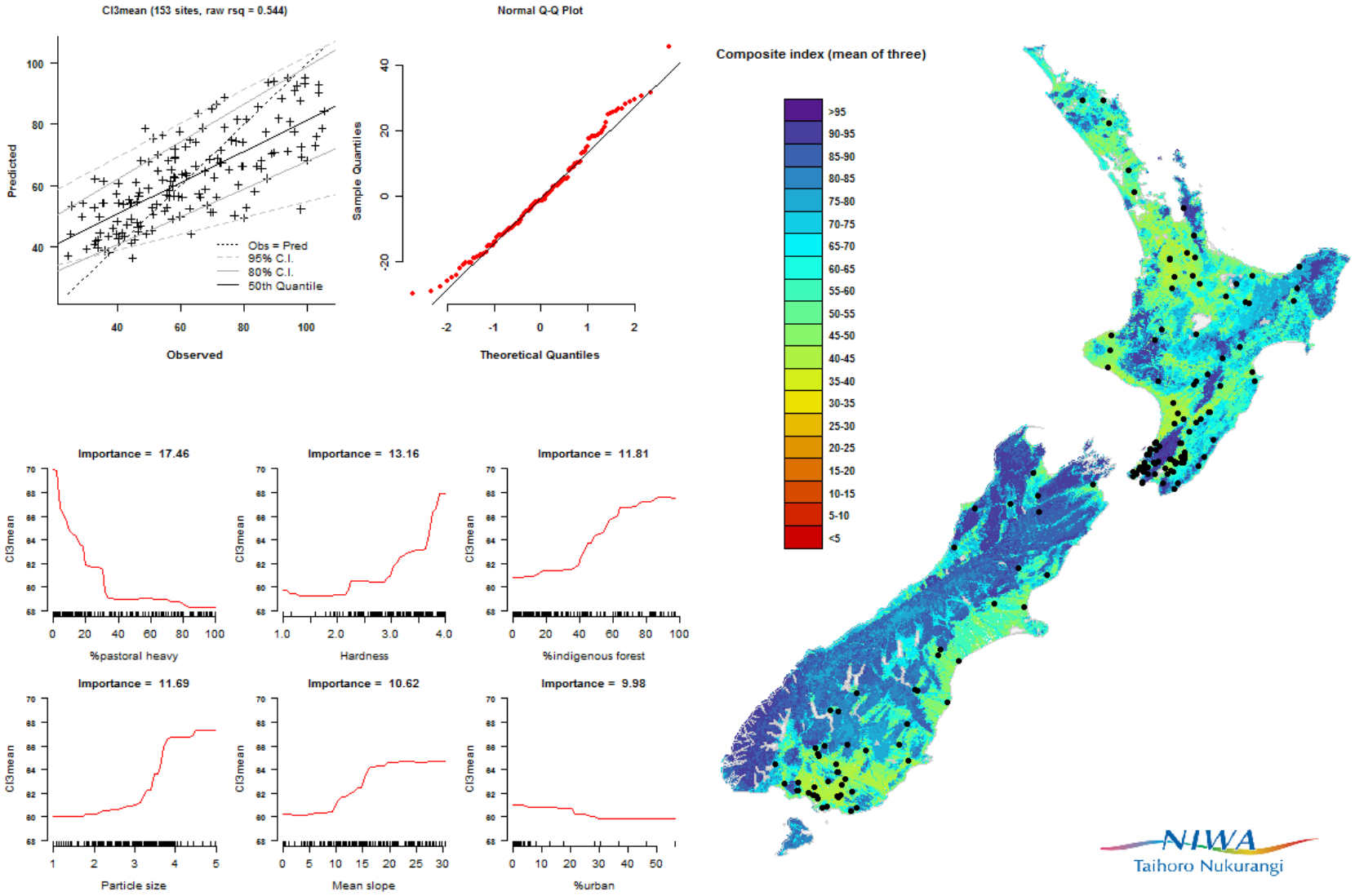


Figure C4: Modelled current state for for a composite water quality index based on the mean of three component indices (CI3mean), showing diagnostic plots for the fitted random forest model (left), and modelled indices for all NZReaches (right).

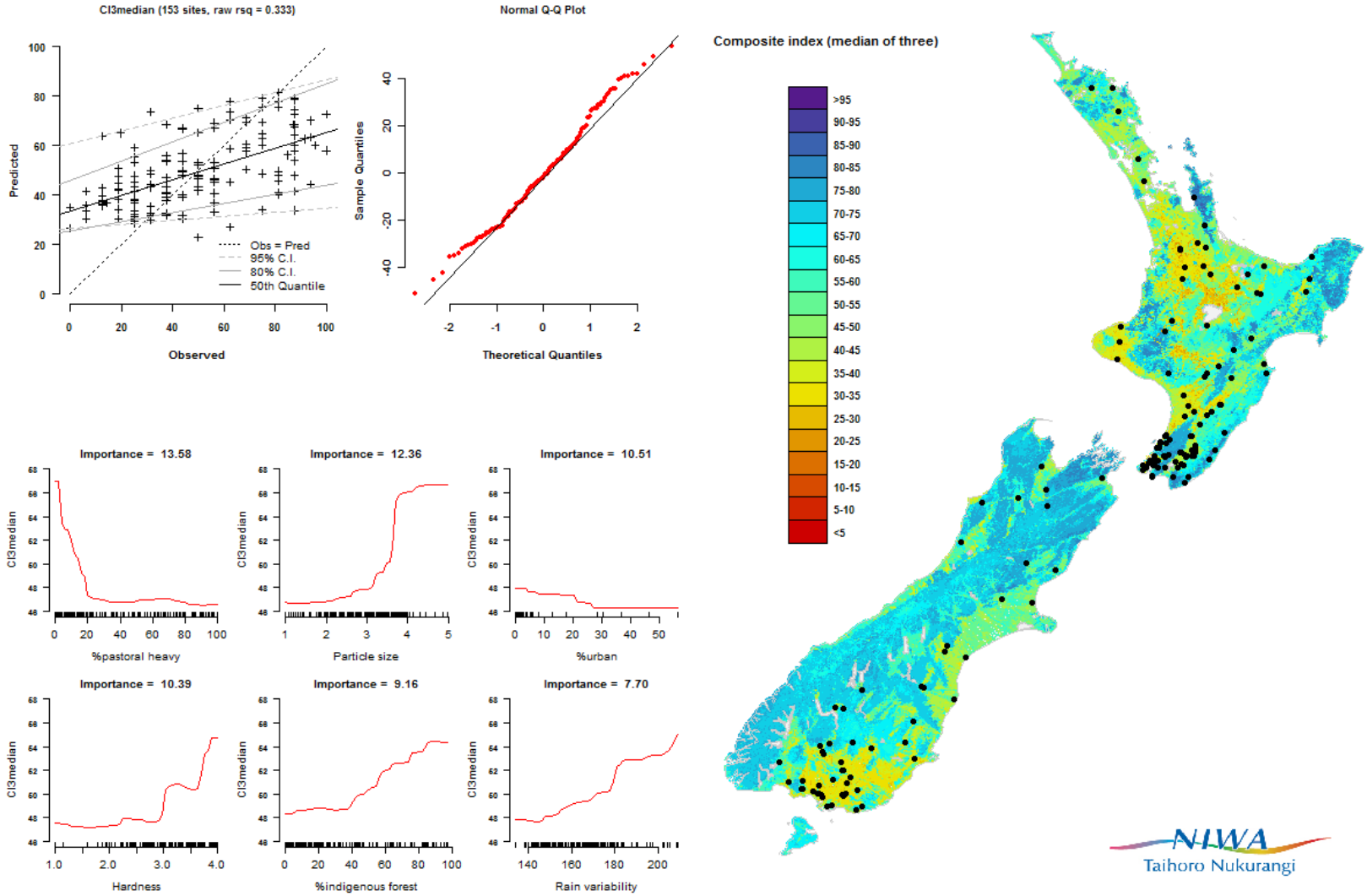


Figure C5: Modelled current state for for a composite water quality index based on the median of three component indices (CI3median), showing diagnostic plots for the fitted random forest model (left), and modelled indices for all NZReaches (right).

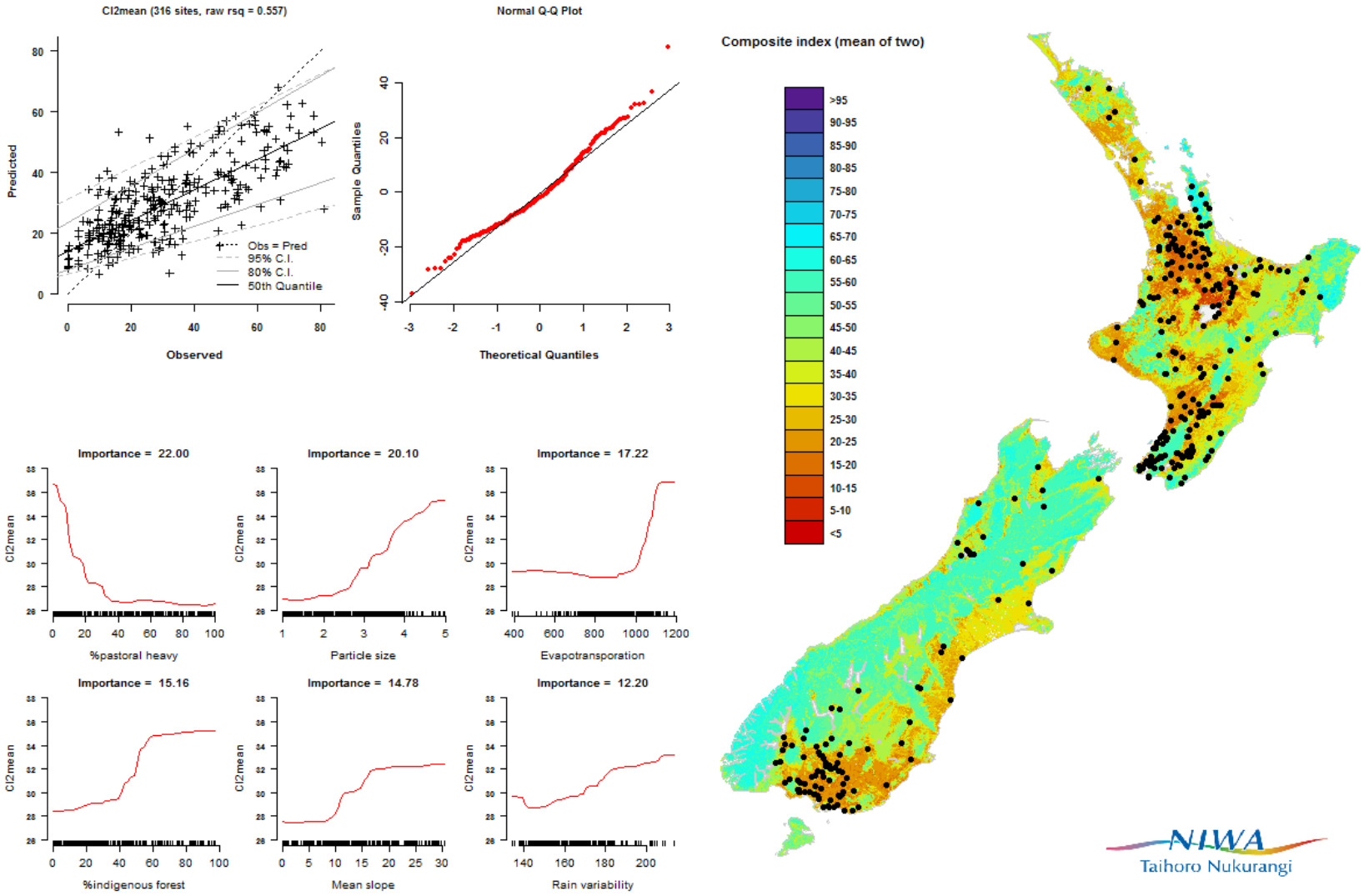
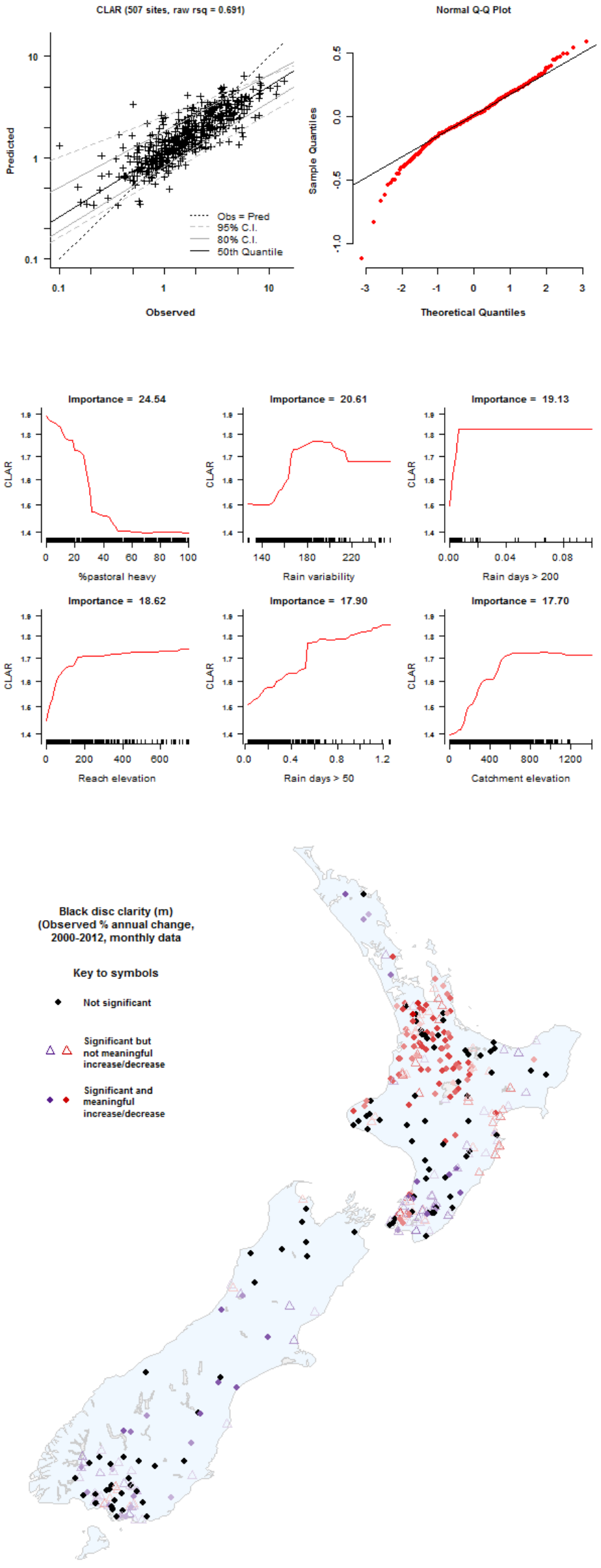
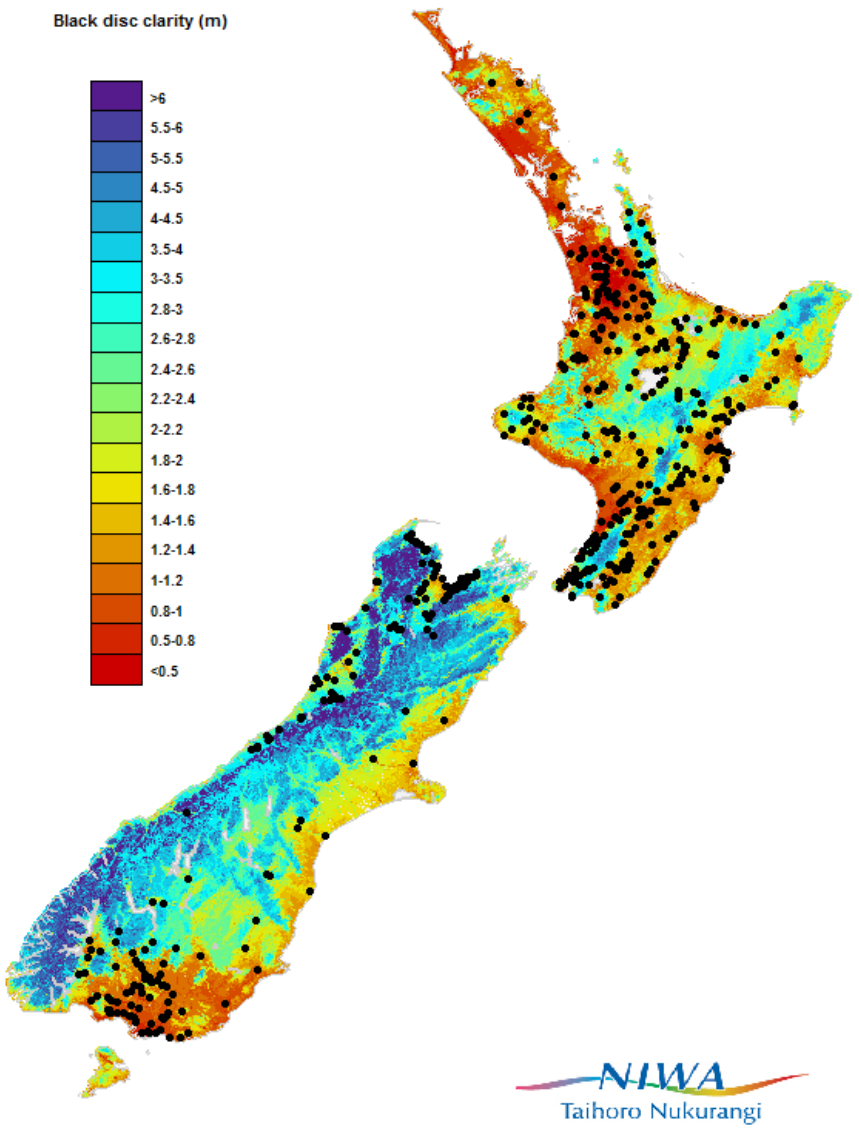


Figure C6: Modelled current state for for a composite water quality index based on the mean of two component indices (CI2mean), showing diagnostic plots for the fitted random forest model (left), and modelled indices for all NZReaches (right).





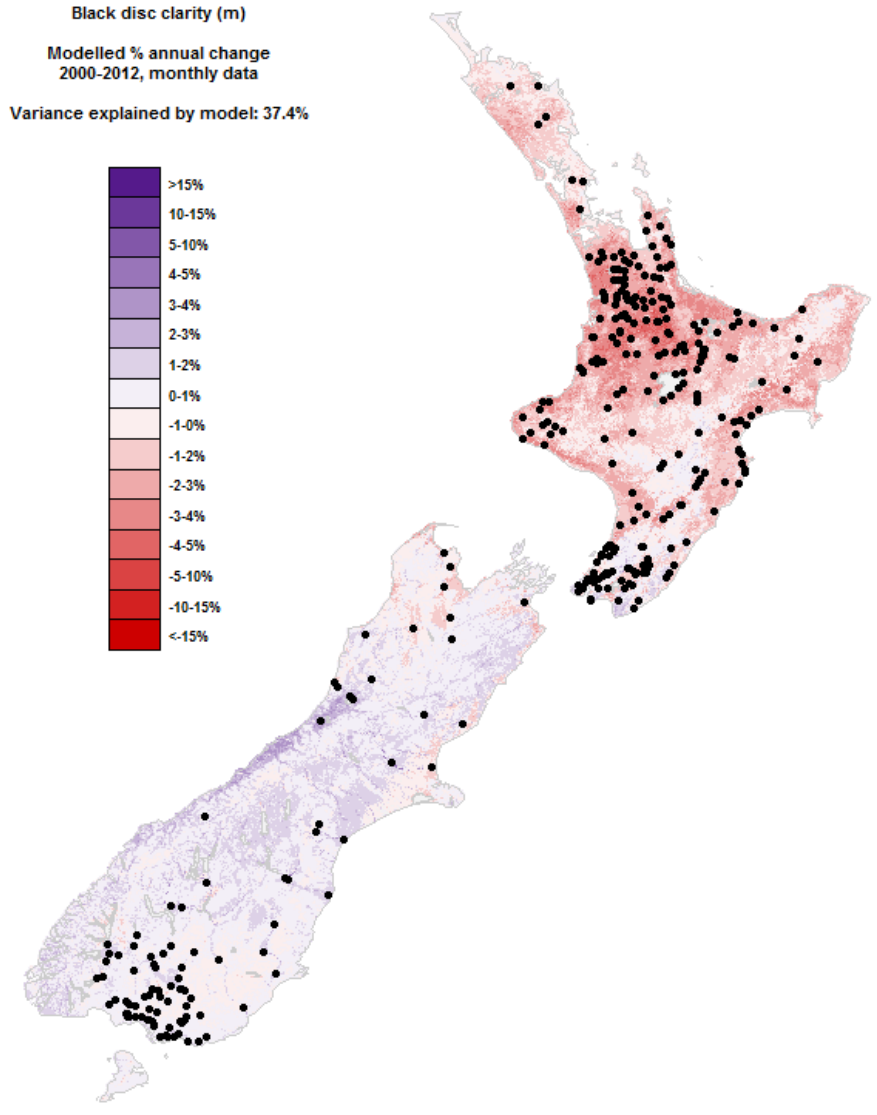
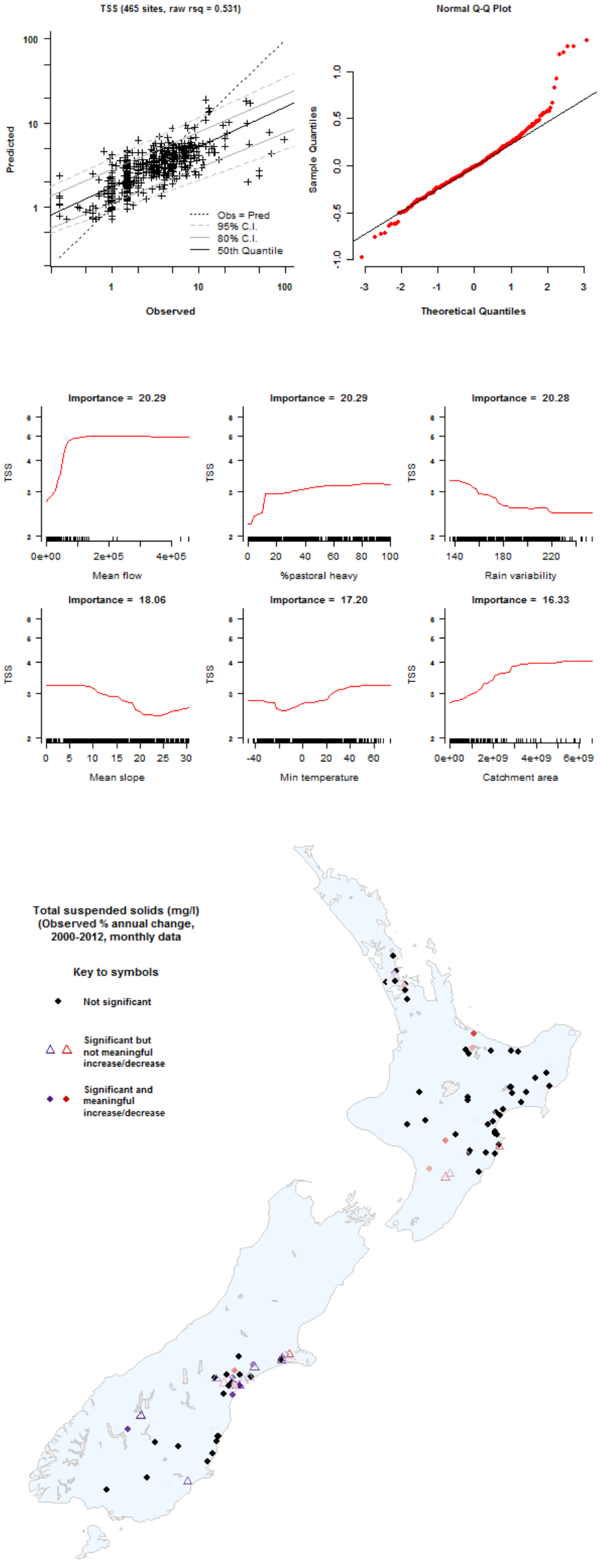
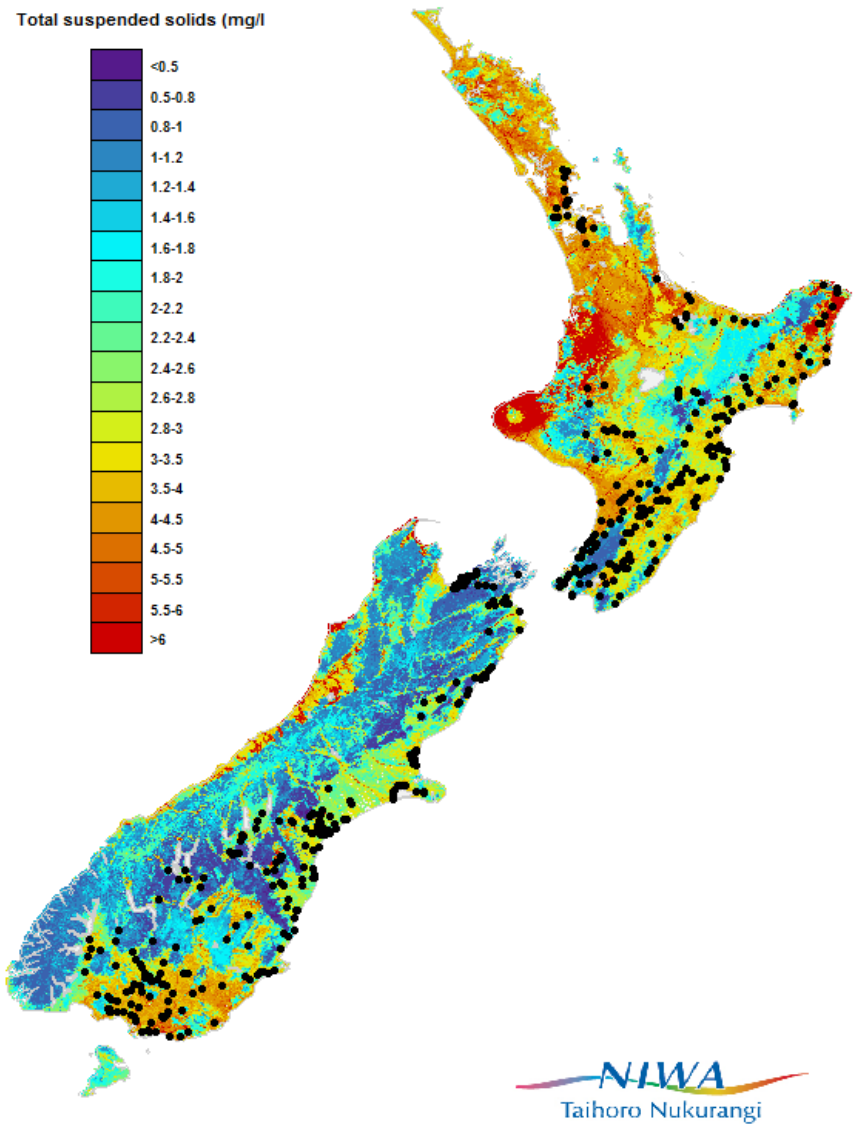


Figure A1: Modelled current state and trend data for black disc clarity (CLAR). Successive panels show diagnostic plots for the fitted random forest model representing current state (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites based on analysis of monthly data (lower left); and modelled monthly 2000-2010 trends for all NZReaches (lower right). For the trend maps, blue or red shading shows improving or decreasing water quality, respectively.





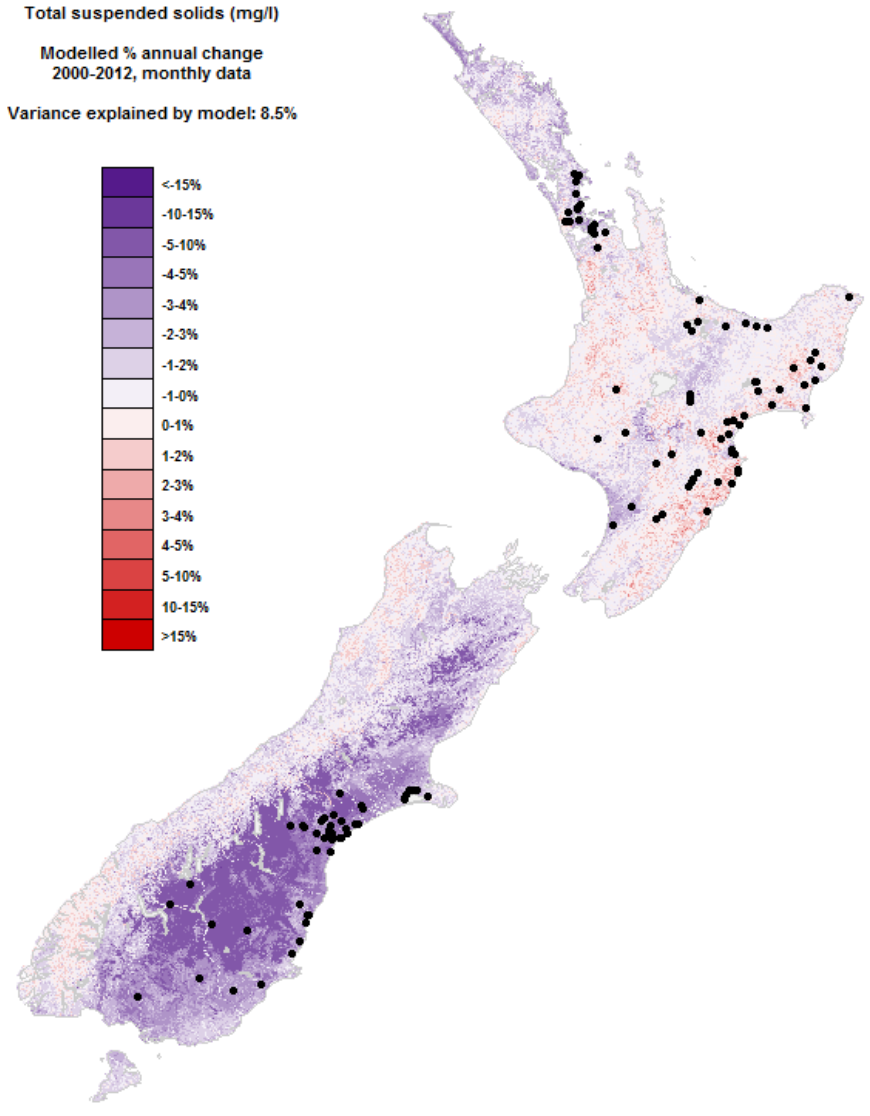


Figure A2: Modelled current state and trend data for total suspended solids (TSS). Successive panels show diagnostic plots for the fitted random forest model representing current state (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites based on analysis of monthly data (lower left); and modelled monthly 2000-2010 trends for all NZReaches (lower right). For the trend maps, blue or red shading shows improving or decreasing water quality, respectively.

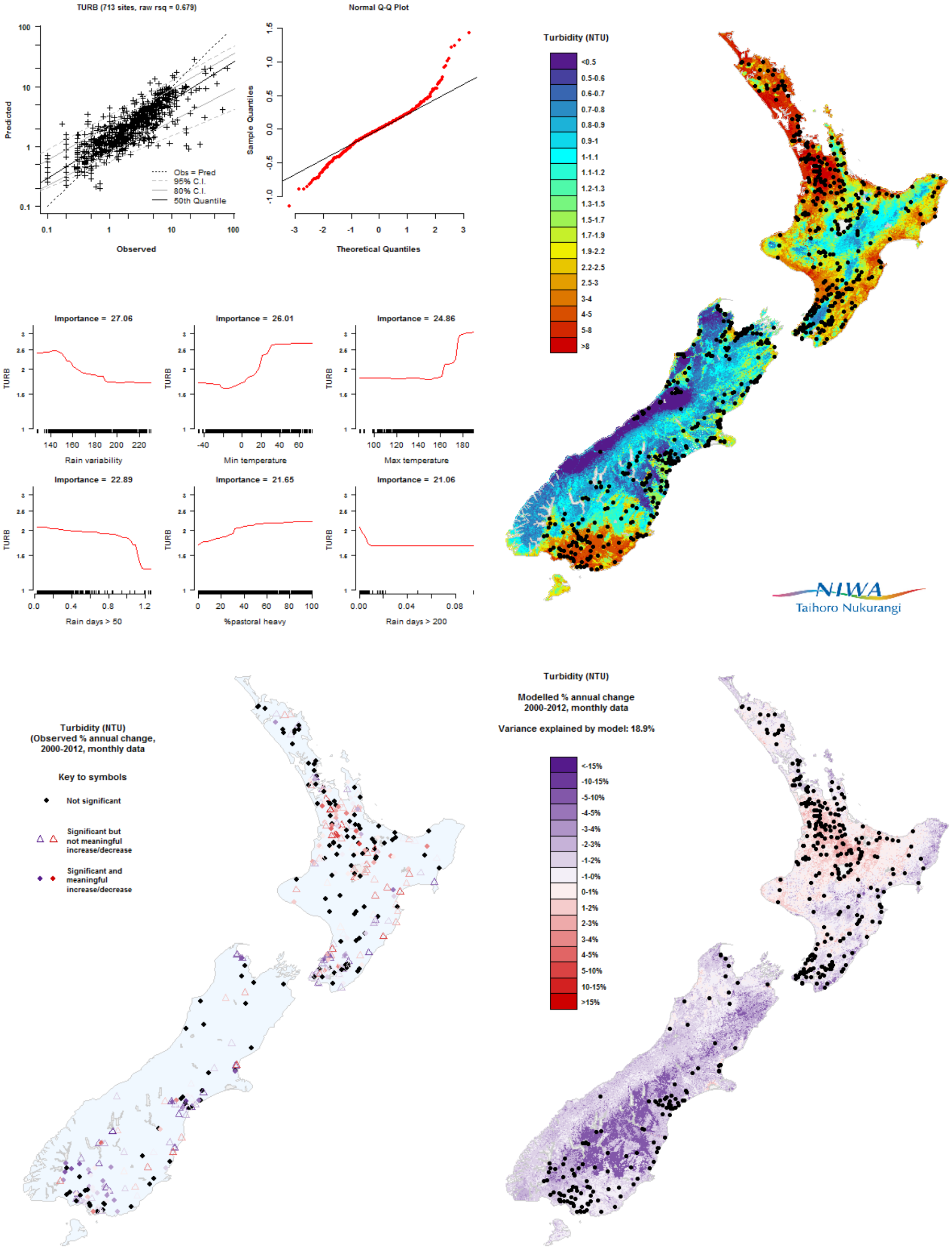


Figure A3: Modelled current state and trend data for turbidity (TURB). Successive panels show diagnostic plots for the fitted random forest model representing current state (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites based on analysis of monthly data (lower left); and modelled monthly 2000-2010 trends for all NZReaches (lower right). For the trend maps, blue or red shading shows improving or decreasing water quality, respectively.

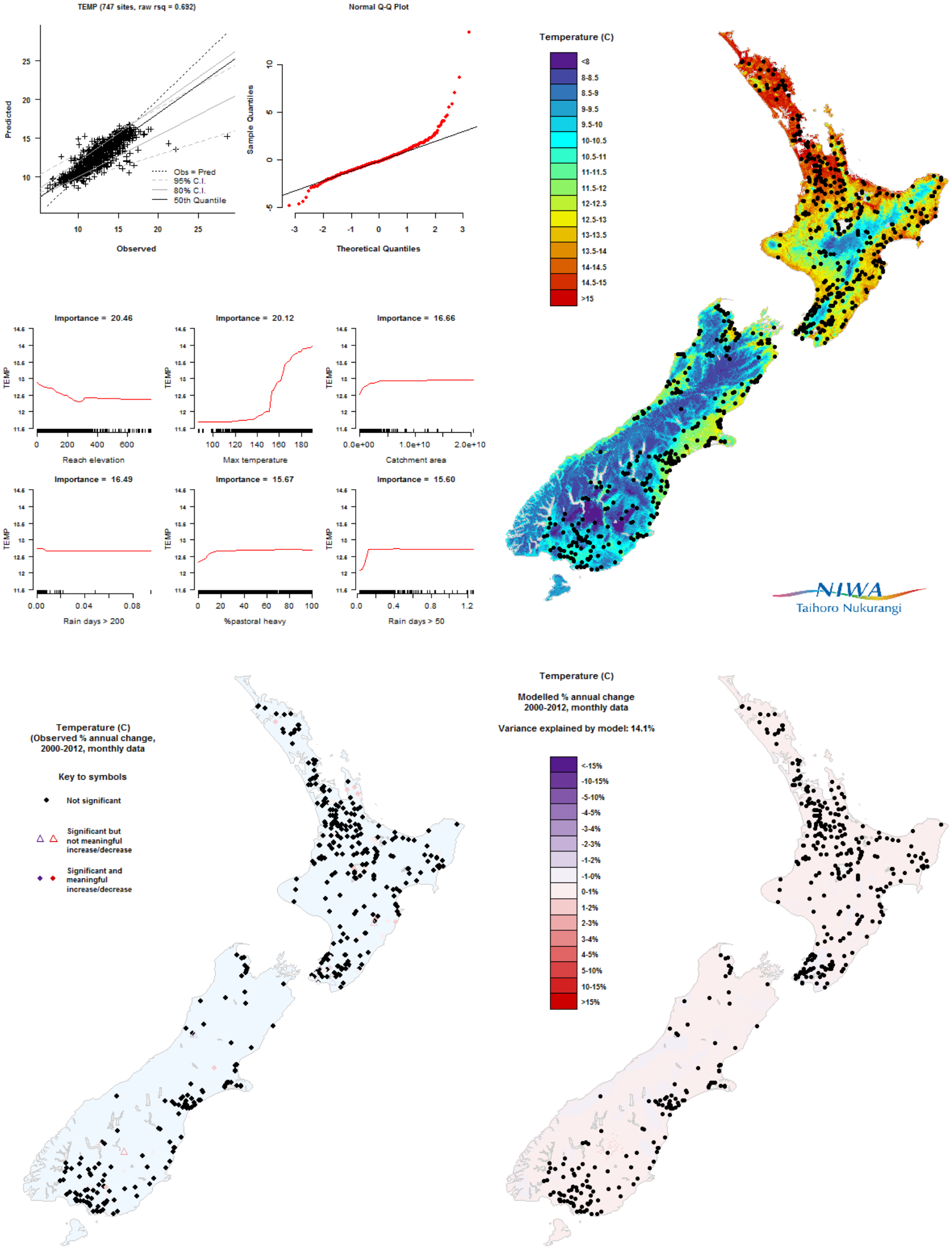


Figure A4: Modelled current state and trend data for median annual temperature (TEMP). Successive panels show diagnostic plots for the fitted random forest model representing current state (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites based on analysis of monthly data (lower left); and modelled monthly 2000-2010 trends for all NZReaches (lower right). For the trend maps, blue or red shading shows improving or decreasing water quality, respectively.

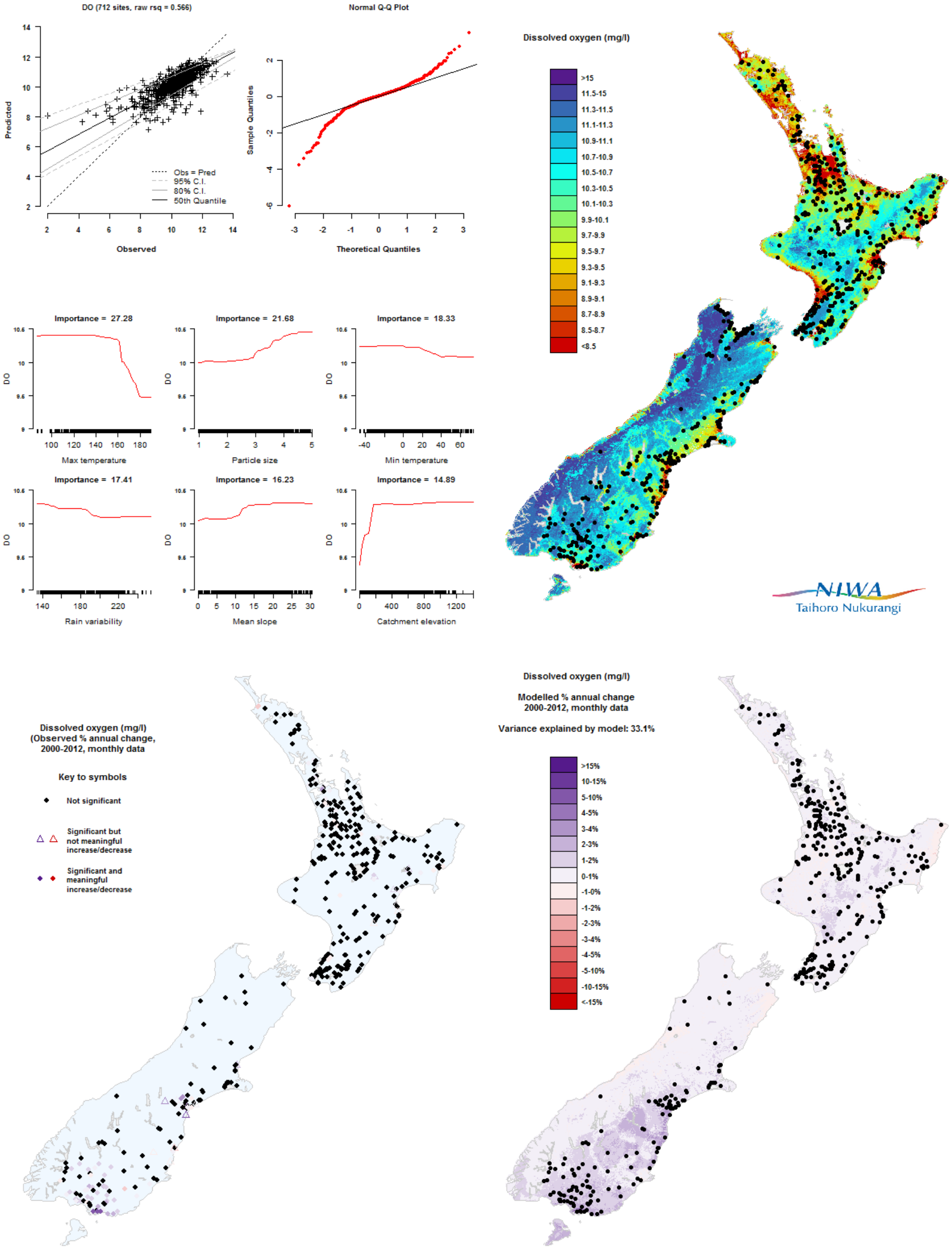
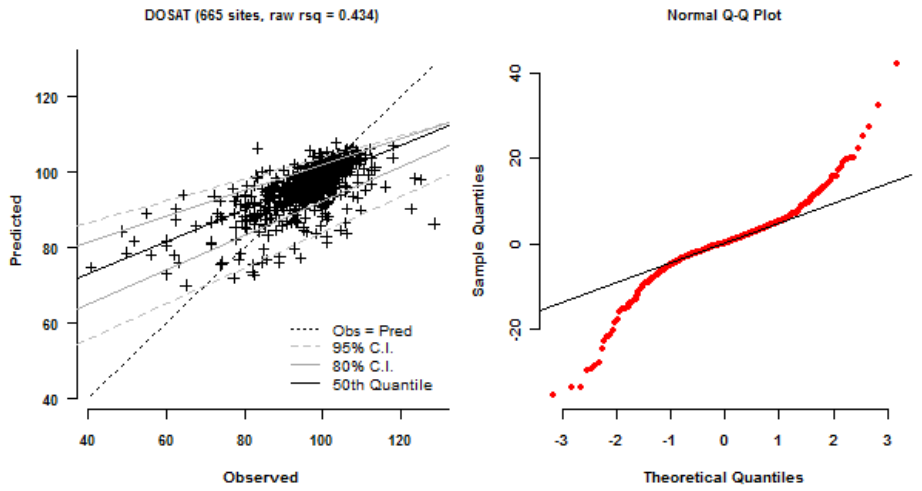
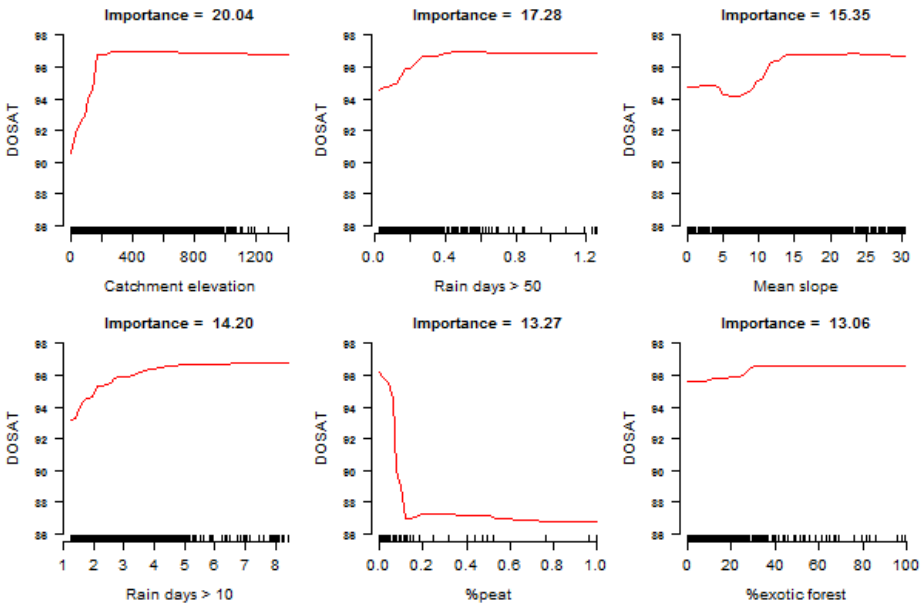
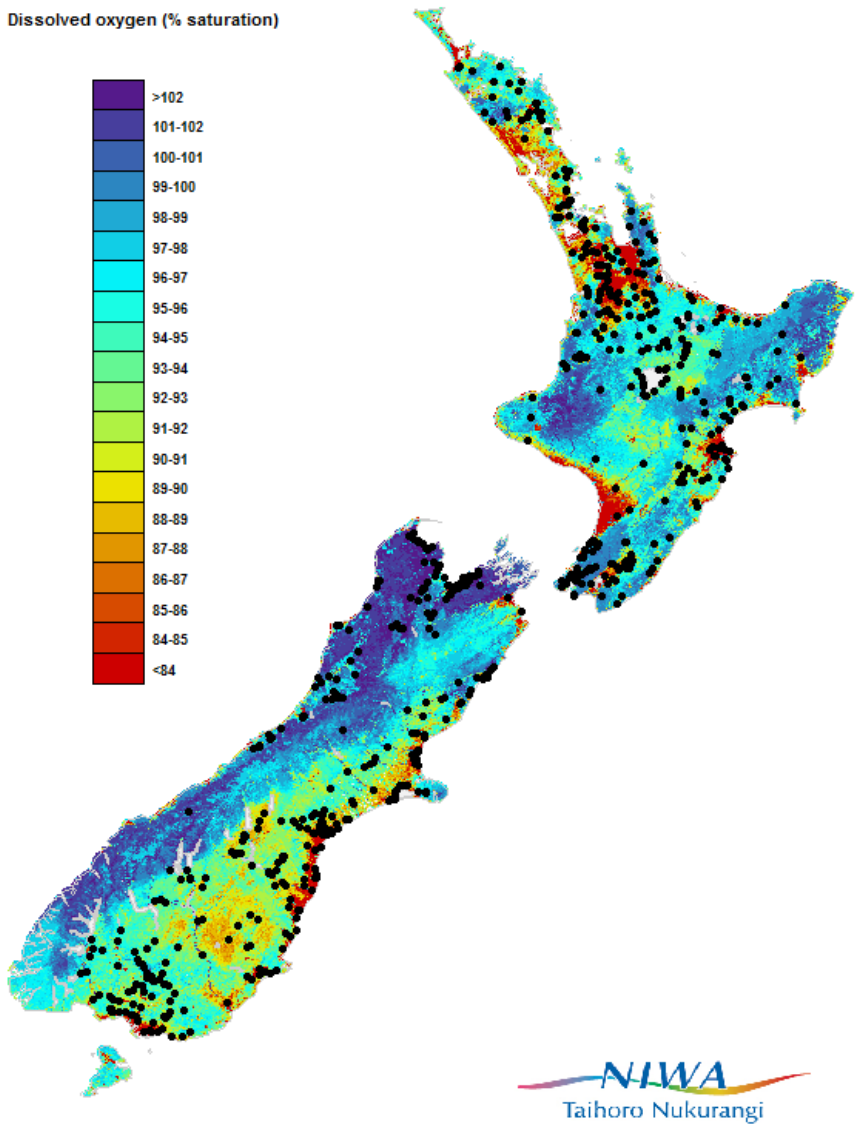
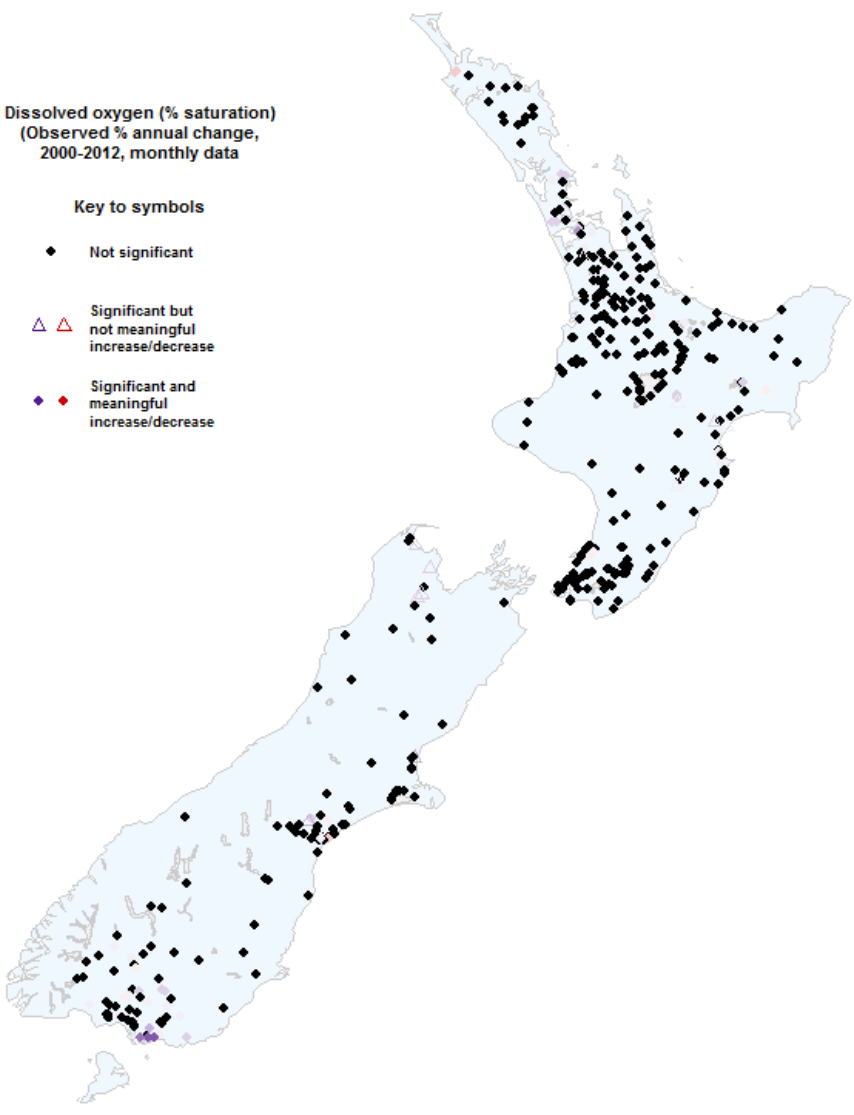


Figure A5: Modelled current state and trend data for dissolved oxygen (DO). Successive panels show diagnostic plots for the fitted random forest model representing current state (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites based on analysis of monthly data (lower left); and modelled monthly 2000-2010 trends for all NZReaches (lower right). For the trend maps, blue or red shading shows improving or decreasing water quality, respectively.









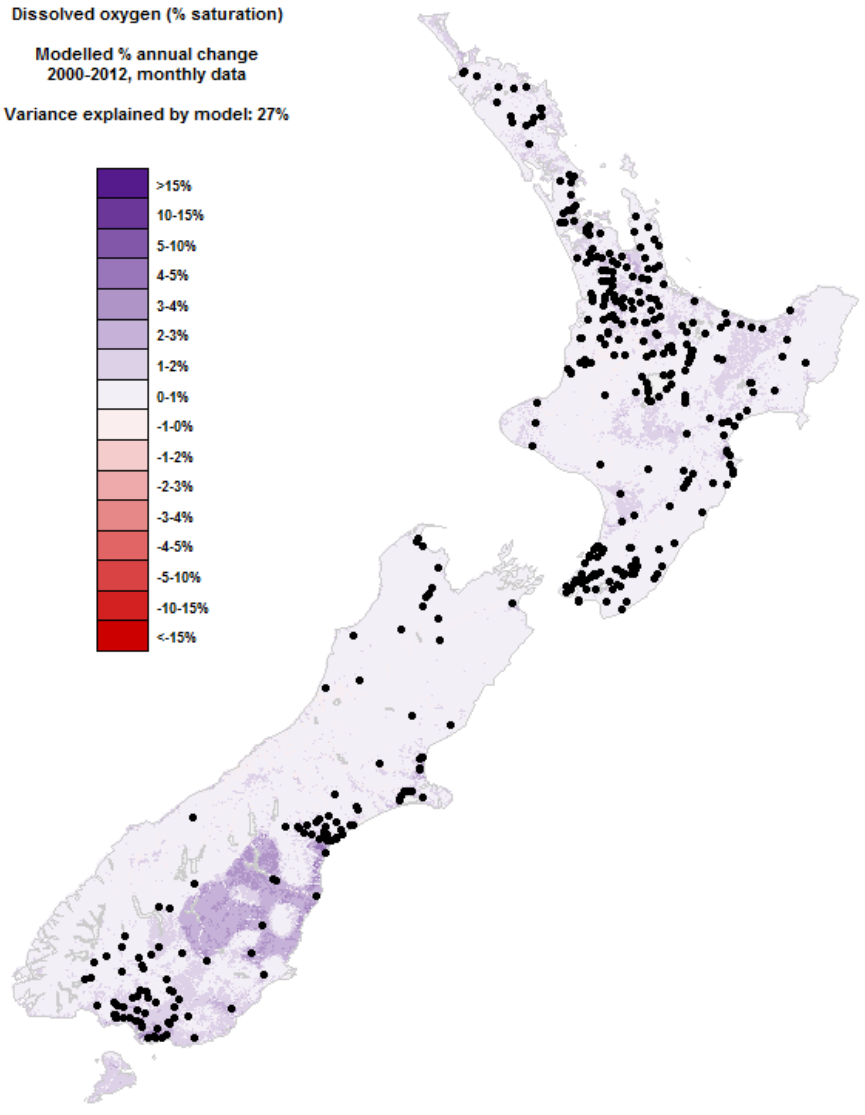


Figure A6: Modelled current state and trend data for dissolved oxygen % saturation. Successive panels show diagnostic plots for the fitted random forest model representing current state (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites based on analysis of monthly data (lower left); and modelled monthly 2000-2010 trends for all NZReaches (lower right). For the trend maps, blue or red shading shows improving or decreasing water quality, respectively.

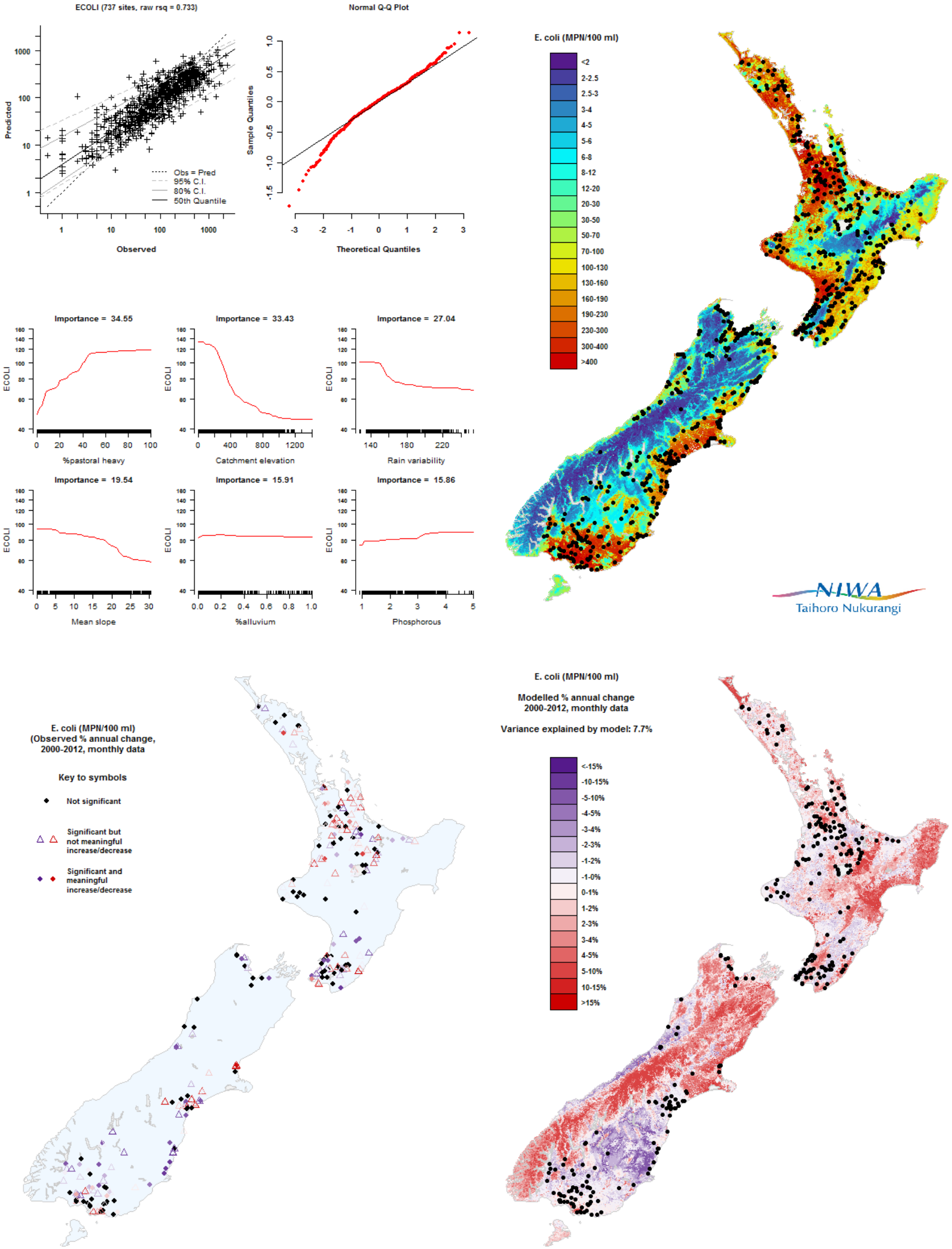
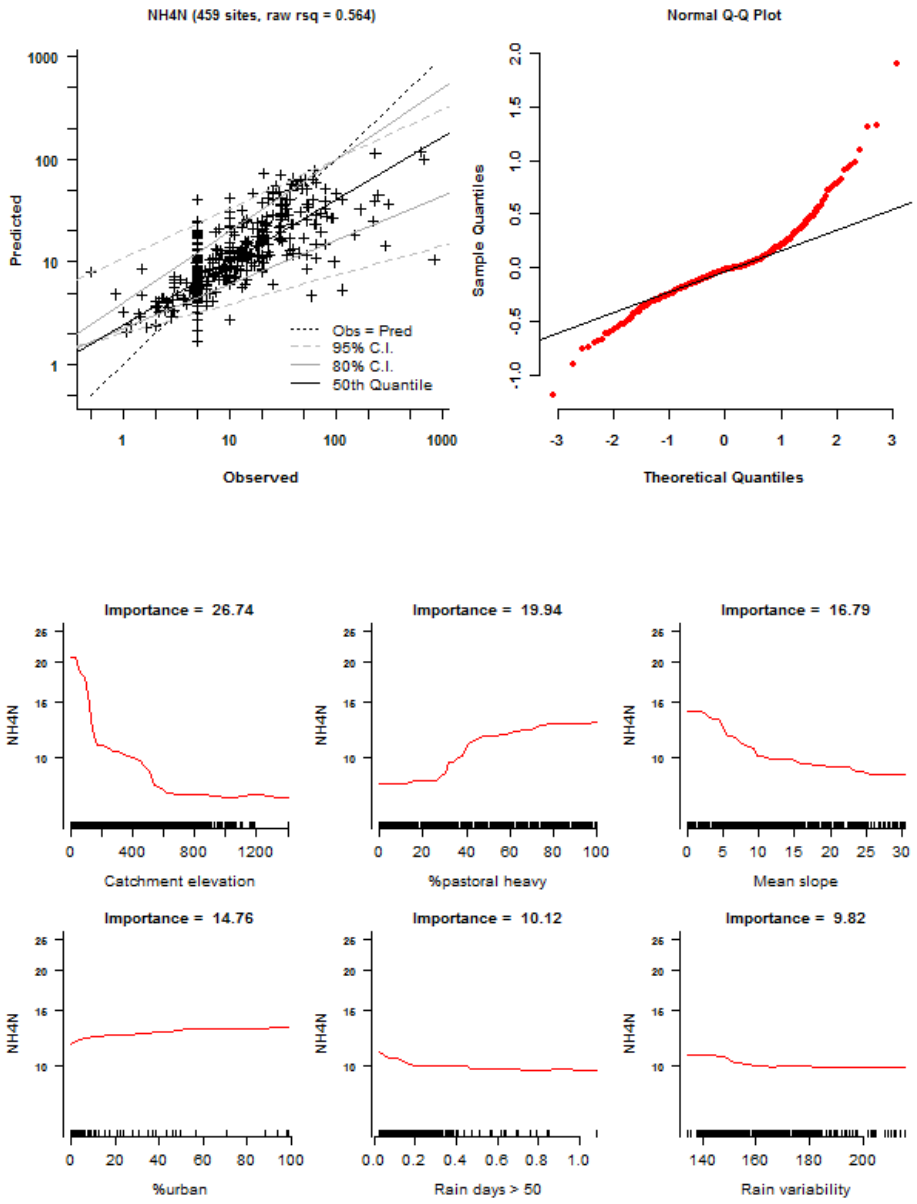
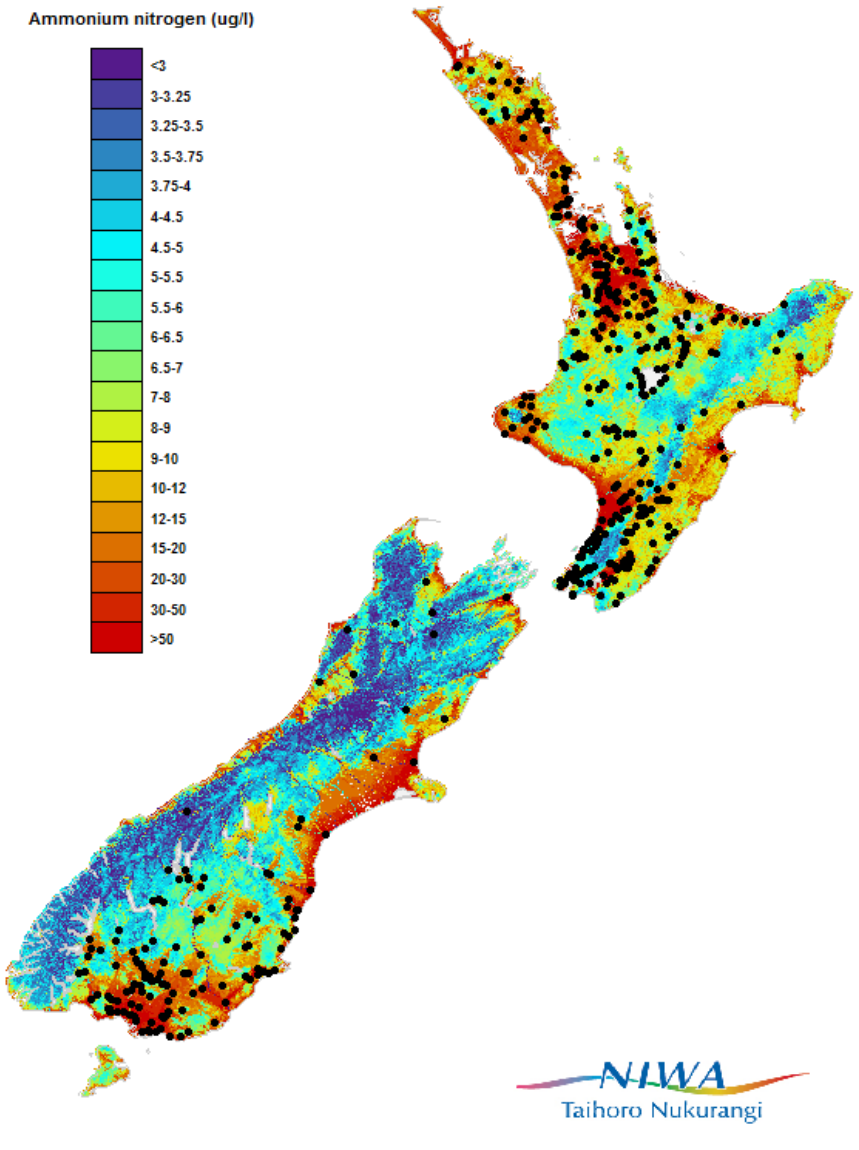
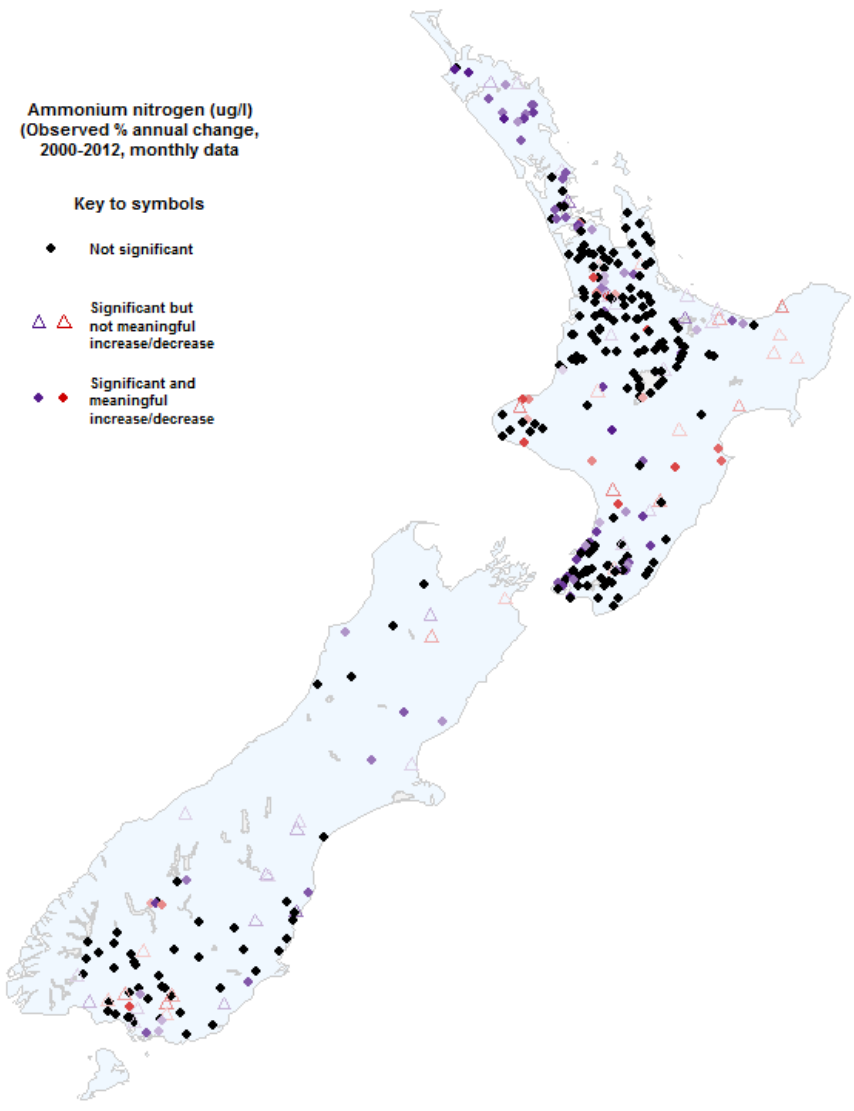


Figure A7: Modelled current state and trend data for *Escherichia coli* (ECOLI). Successive panels show diagnostic plots for the fitted random forest model representing current state (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites based on analysis of monthly data (lower left); and modelled monthly 2000-2010 trends for all NZReaches (lower right). For the trend maps, blue or red shading shows improving or decreasing water quality, respectively.







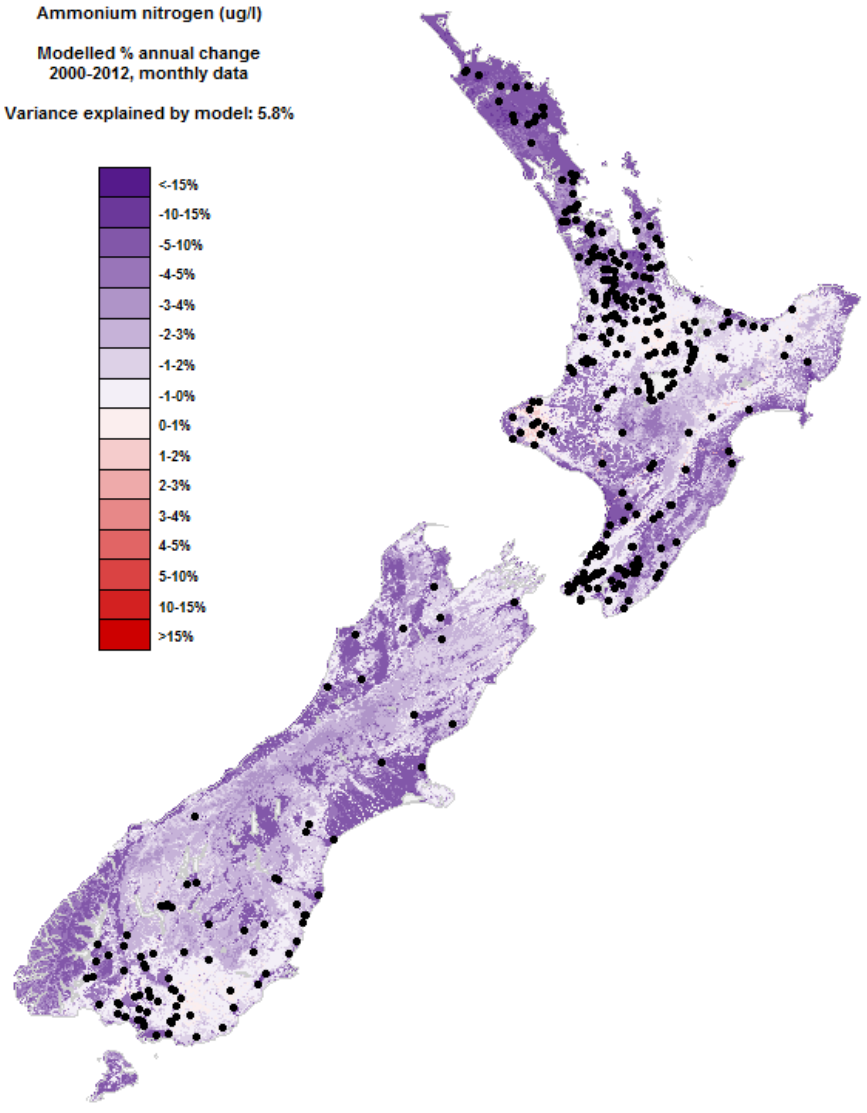


Figure A8: Modelled current state and trend data for ammonium nitrogen (NH4N). Successive panels show diagnostic plots for the fitted random forest model representing current state (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites based on analysis of monthly data (lower left); and modelled monthly 2000-2010 trends for all NZReaches (lower right). For the trend maps, blue or red shading shows improving or decreasing water quality, respectively.

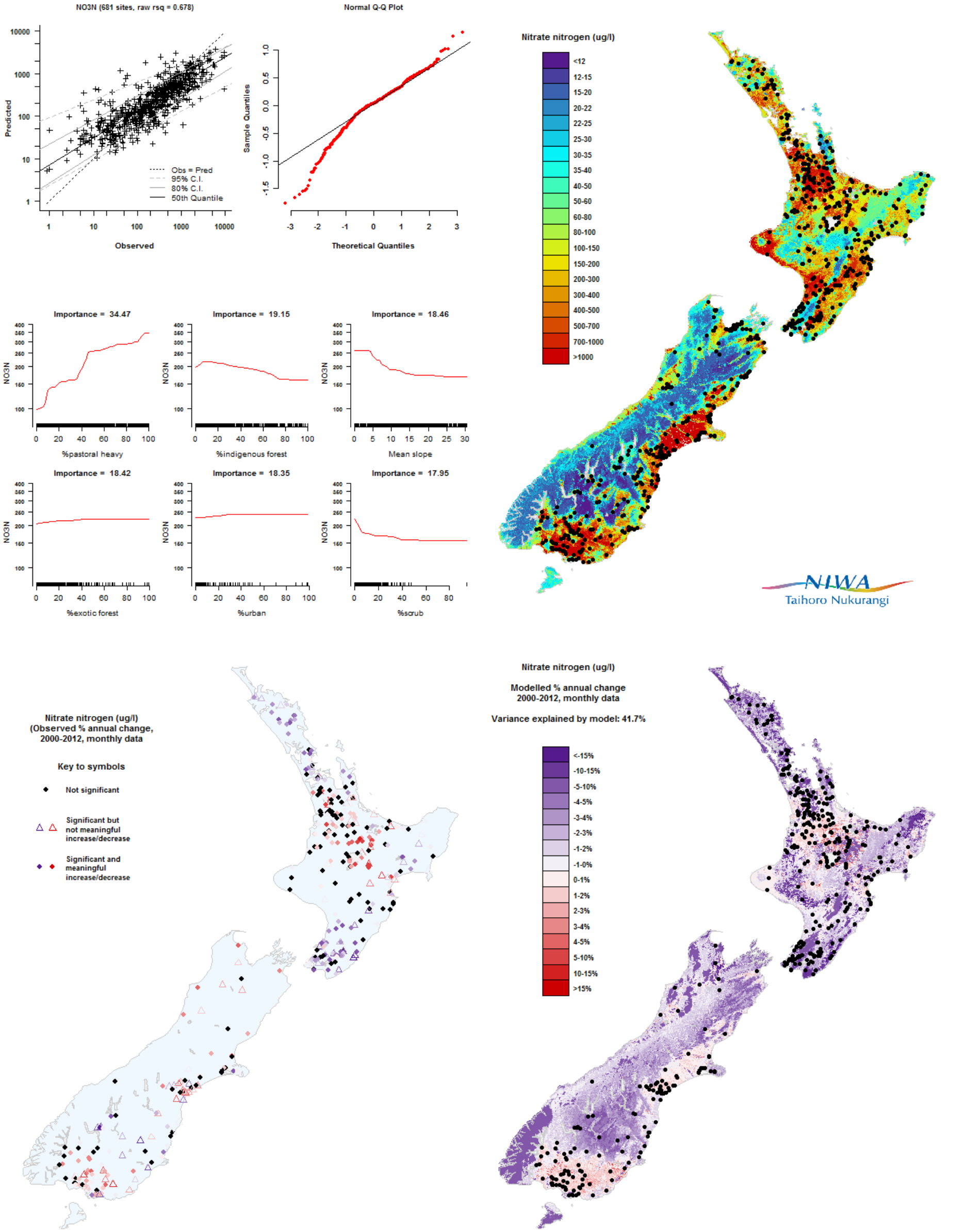
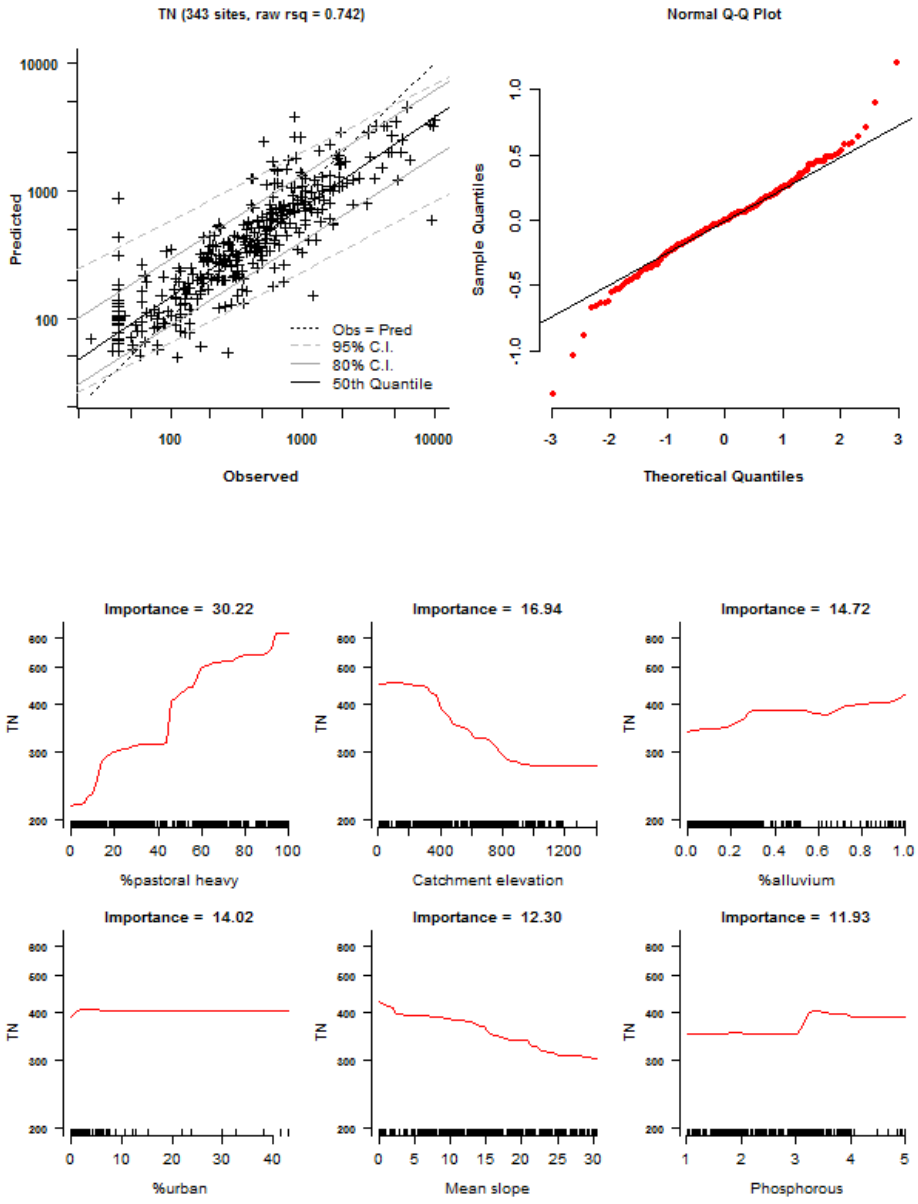
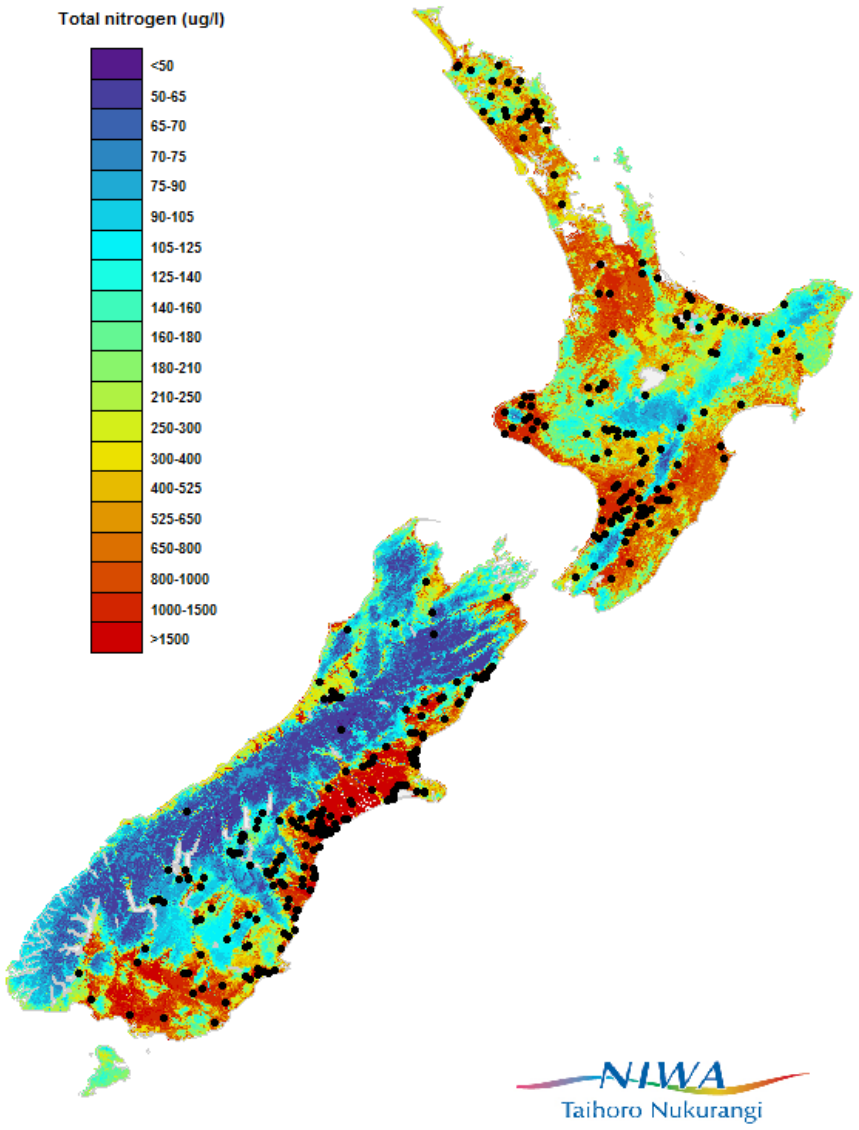
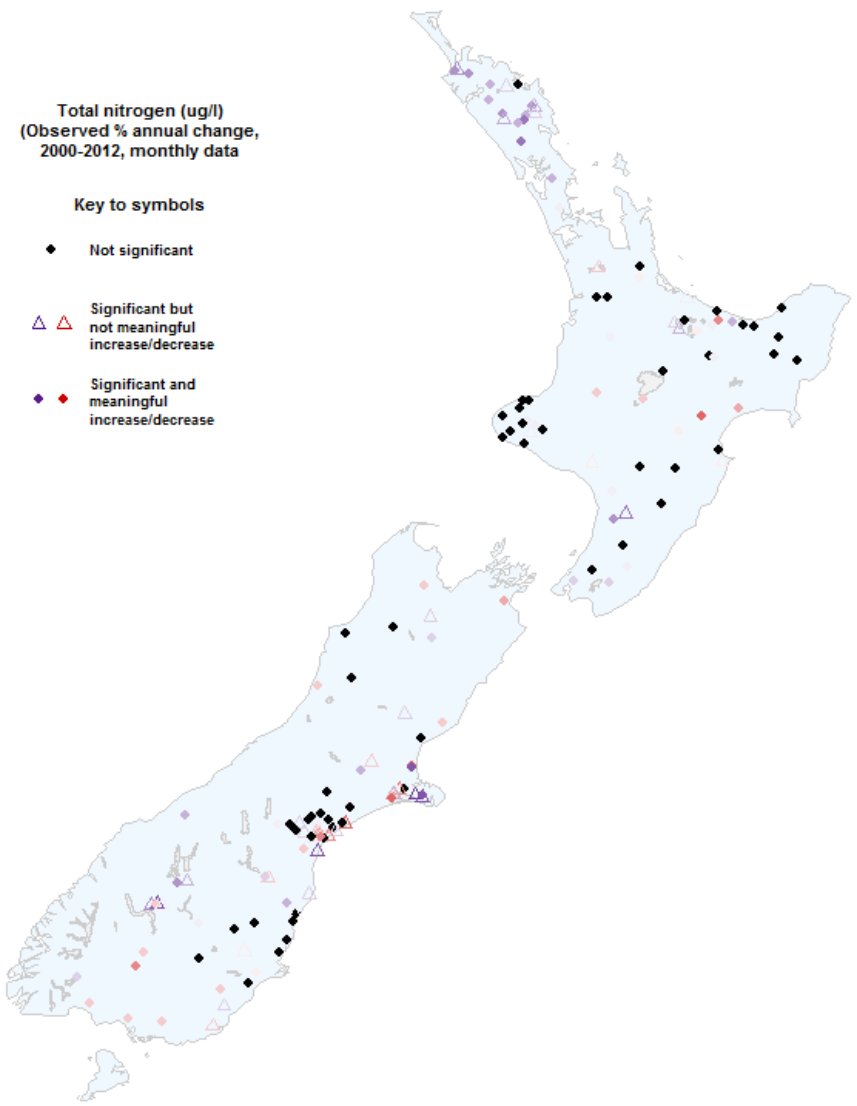


Figure A9: Modelled current state and trend data for nitrate nitrogen (NO3N). Successive panels show diagnostic plots for the fitted random forest model representing current state (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites based on analysis of monthly data (lower left); and modelled monthly 2000-2010 trends for all NZReaches (lower right). For the trend maps, blue or red shading shows improving or decreasing water quality, respectively.







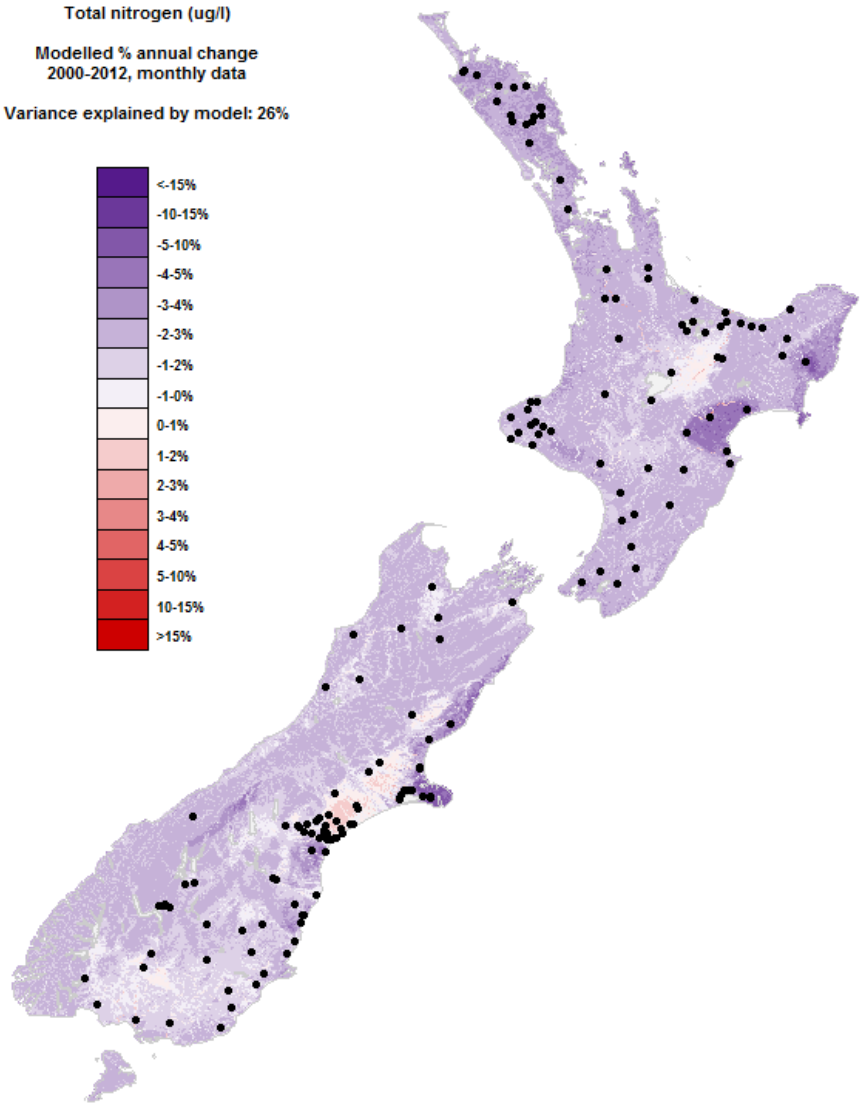


Figure A10: Modelled current state and trend data for total nitrogen (TN). Successive panels show diagnostic plots for the fitted random forest model representing current state (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites based on analysis of monthly data (lower left); and modelled monthly 2000-2010 trends for all NZReaches (lower right). For the trend maps, blue or red shading shows improving or decreasing water quality, respectively.

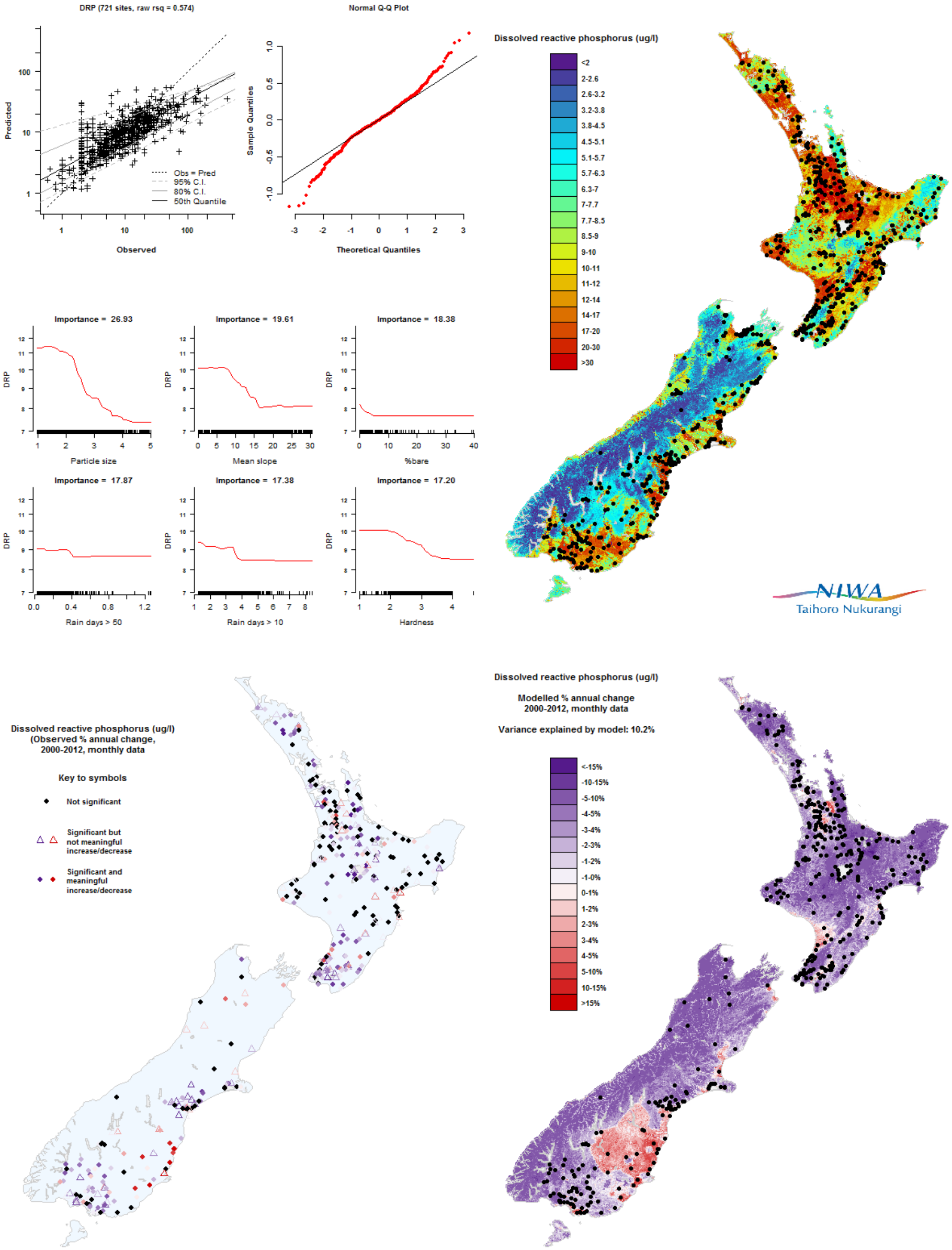
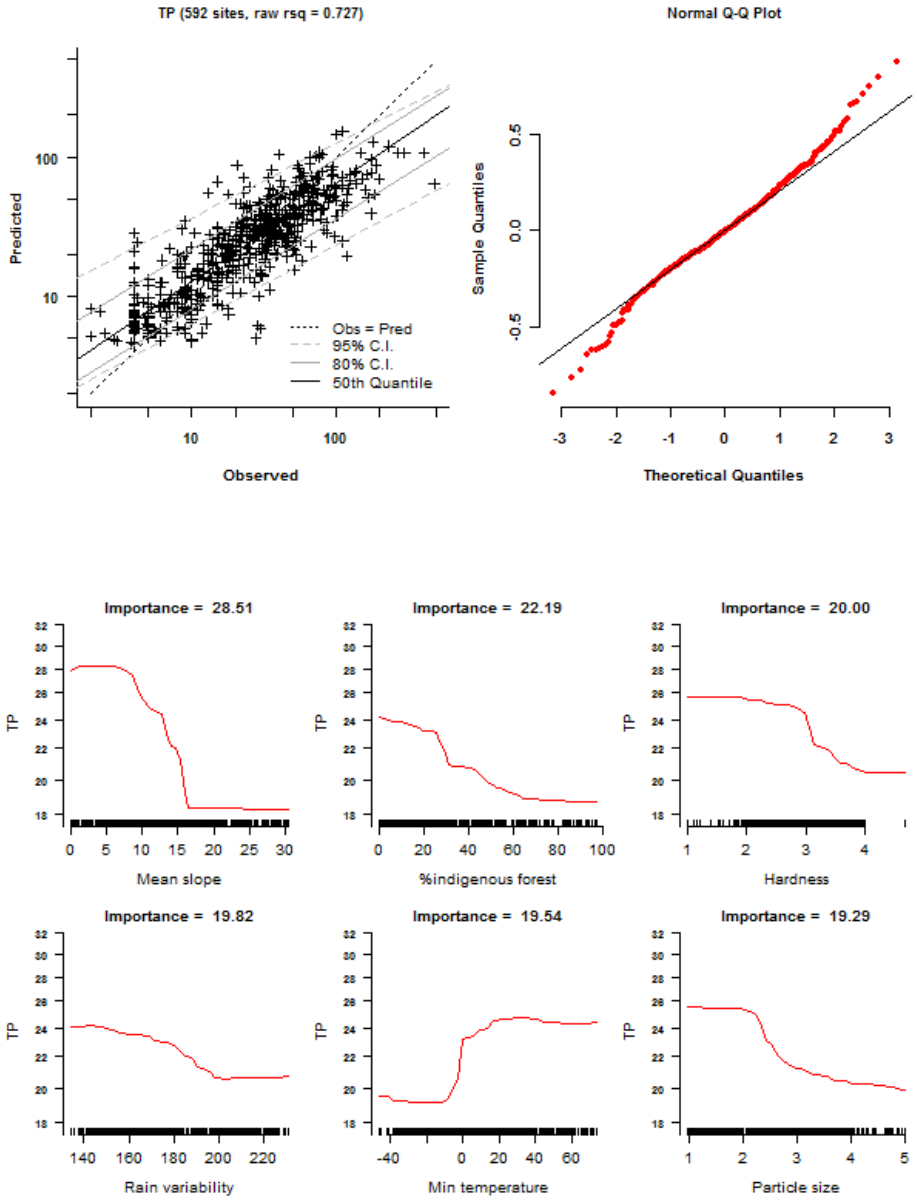
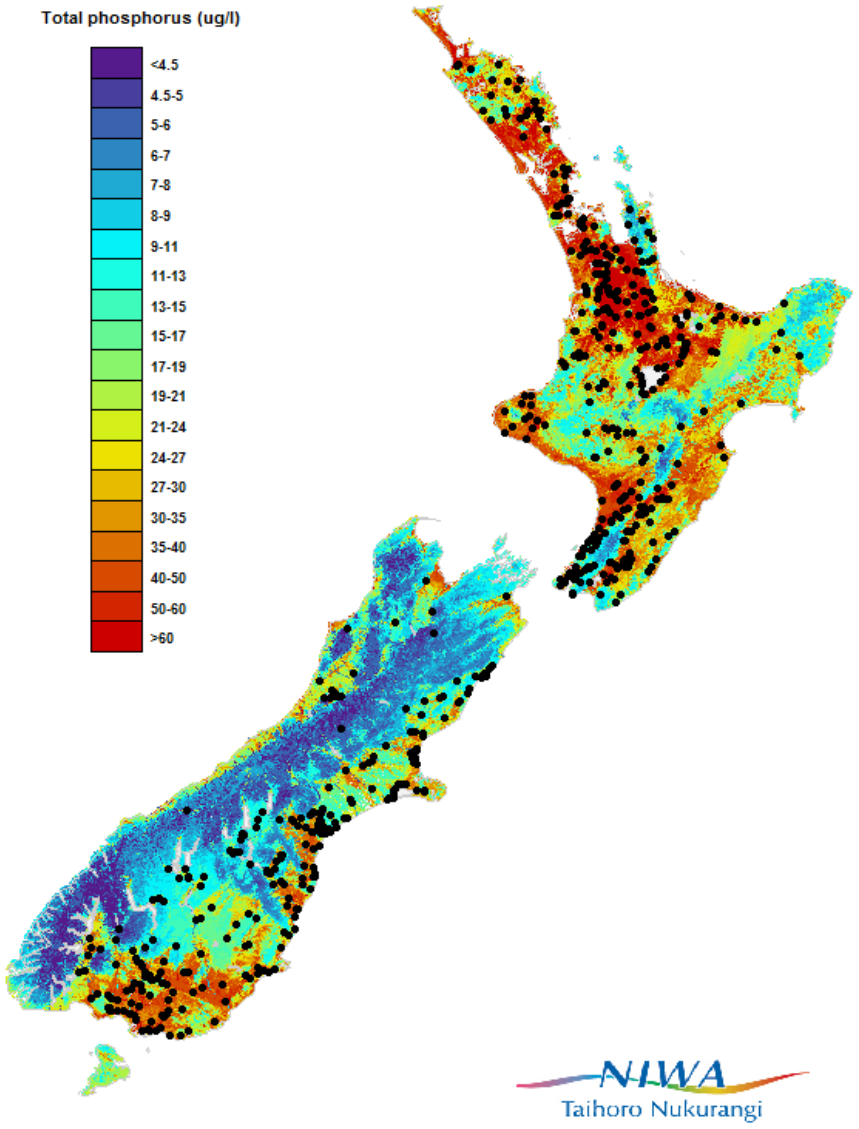
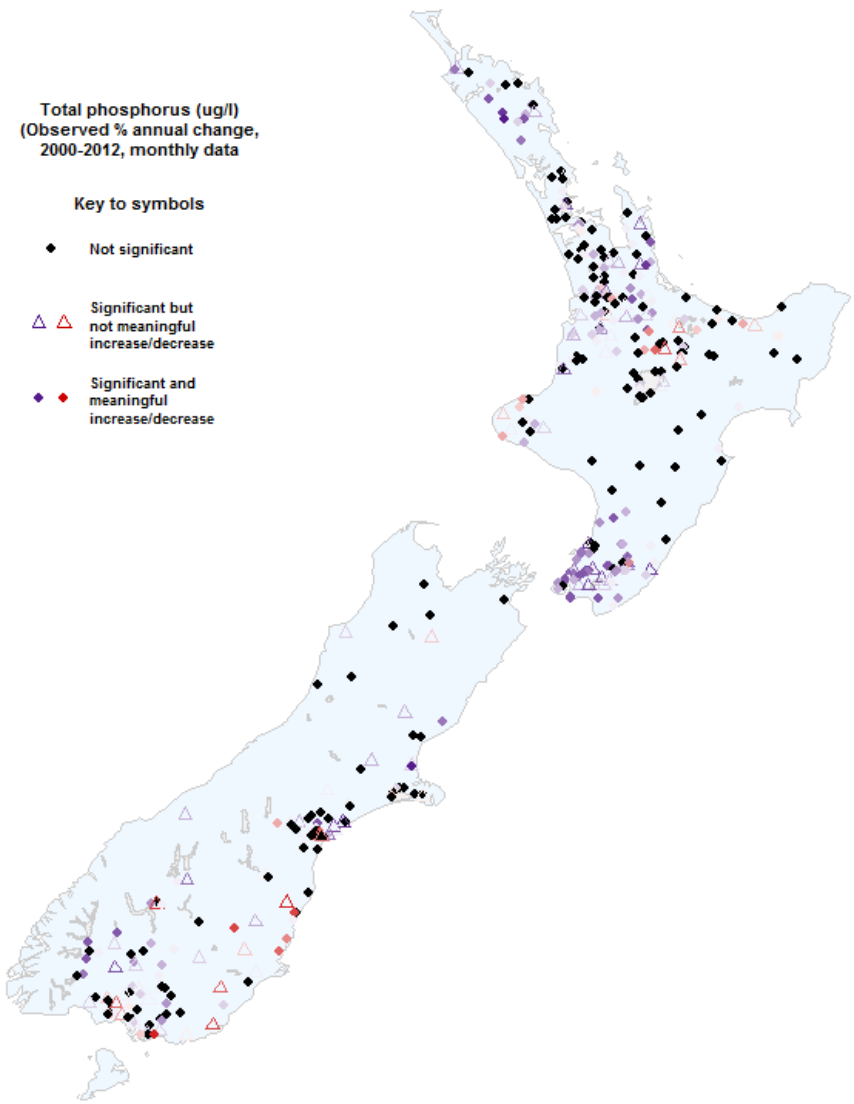


Figure A11: Modelled current state and trend data for dissolved reactive phosphorus (DRP). Successive panels show diagnostic plots for the fitted random forest model representing current state (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites based on analysis of monthly data (lower left); and modelled monthly 2000-2010 trends for all NZReaches (lower right). For the trend maps, blue or red shading shows improving or decreasing water quality, respectively.







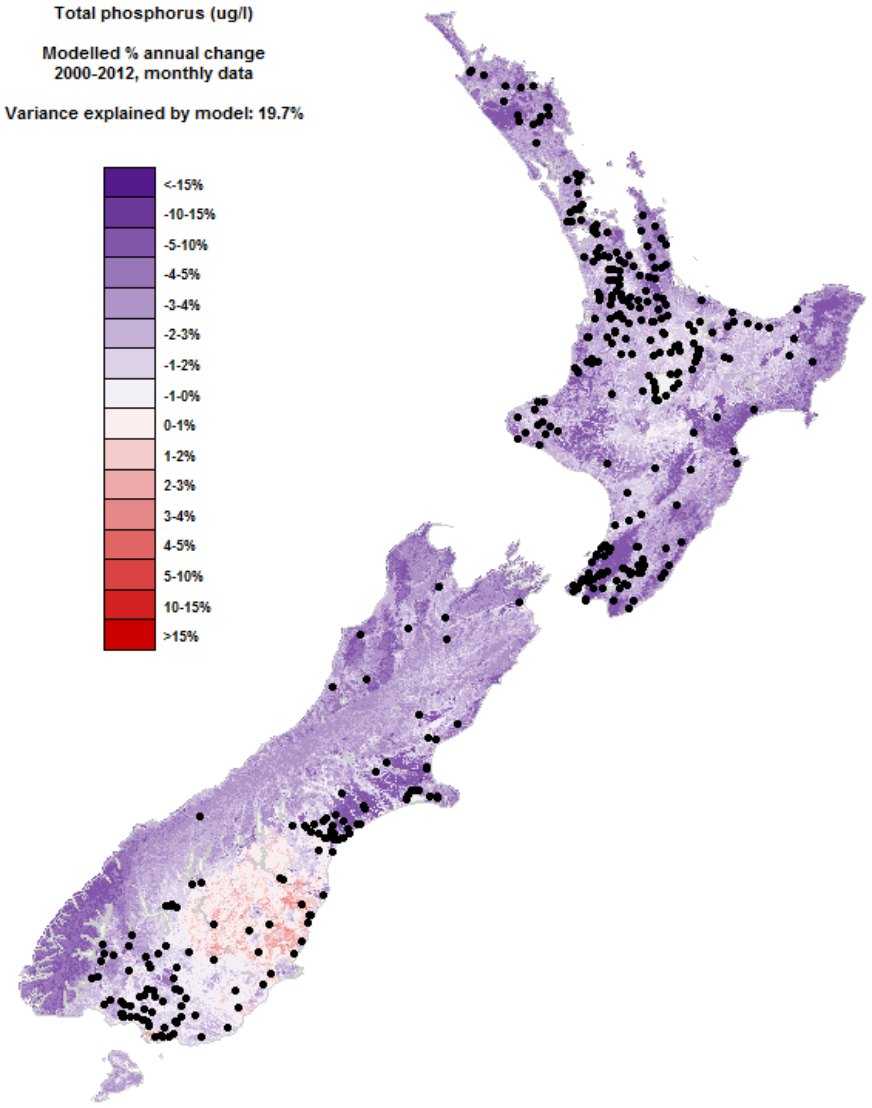


Figure A12: Modelled current state and trend data for total phosphorus (TP). Successive panels show diagnostic plots for the fitted random forest model representing current state (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites based on analysis of monthly data (lower left); and modelled monthly 2000-2010 trends for all NZReaches (lower right). For the trend maps, blue or red shading shows improving or decreasing water quality, respectively.

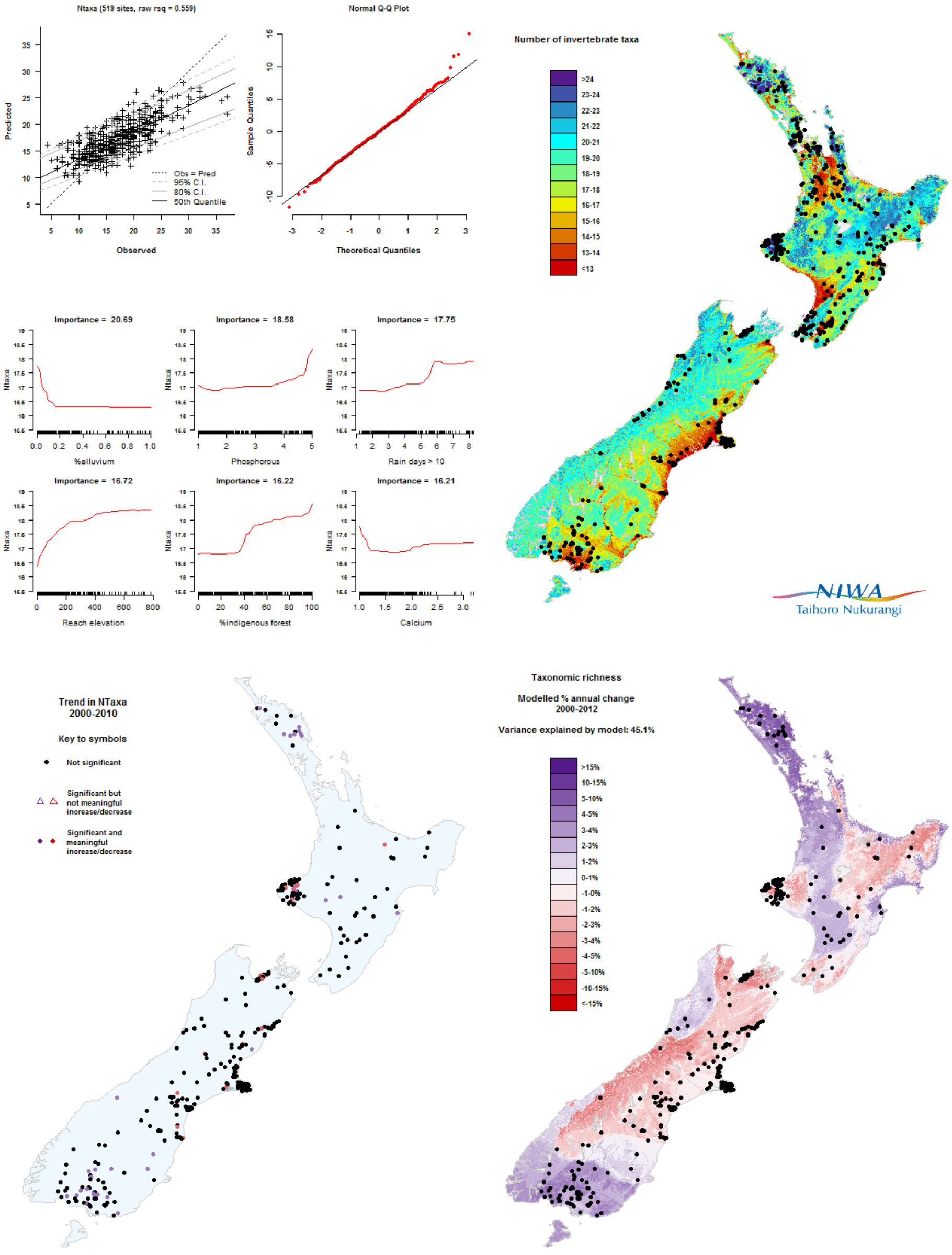


Figure B1: Modelled current state and trend data for taxonomic richness (NTaxa). Successive panels show diagnostic plots for the fitted random forest model (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites (lower left); and modelled 2000-2010 trends for all NZReaches (lower right). For trend maps, blue/red shading shows improving/decreasing index values, respectively.

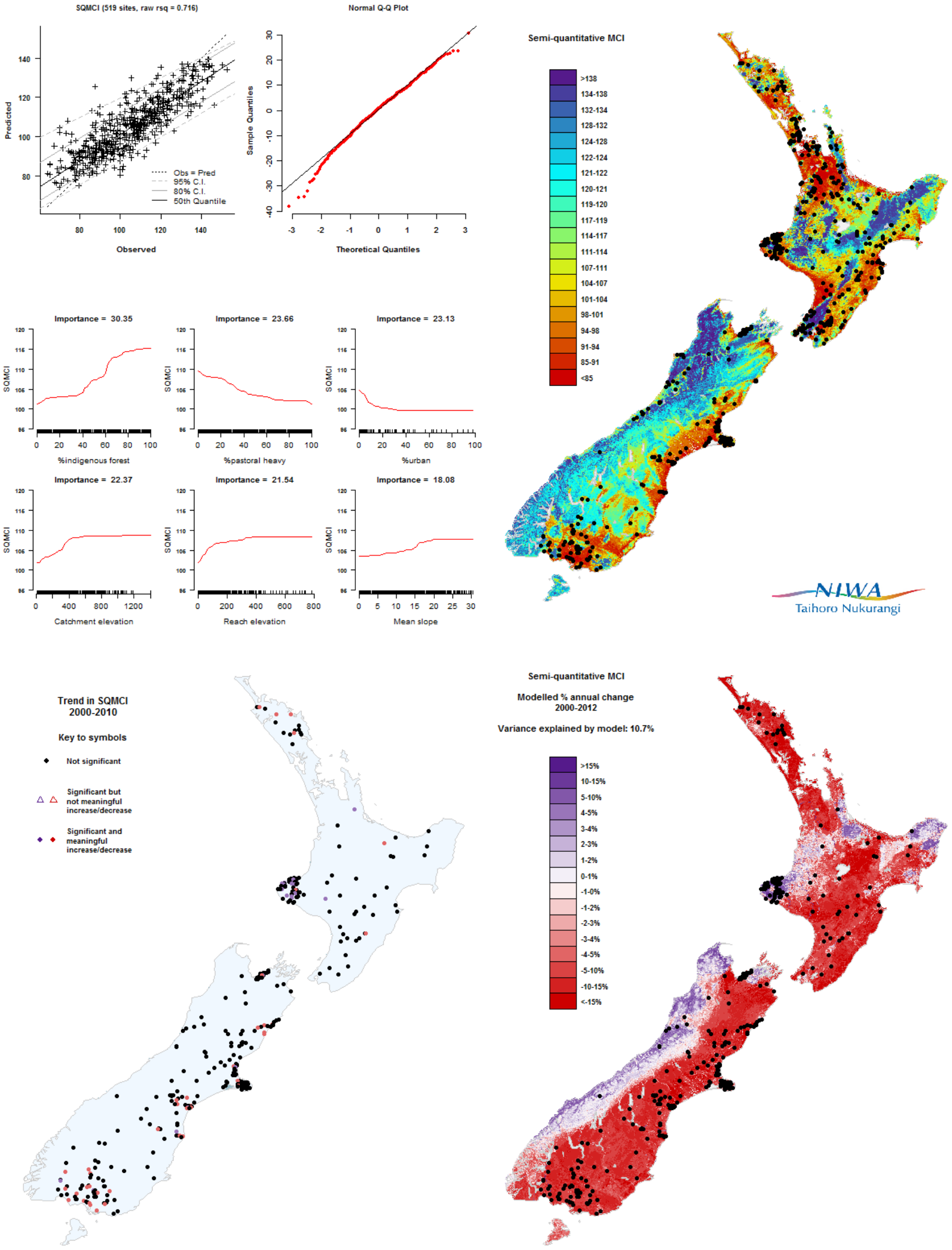
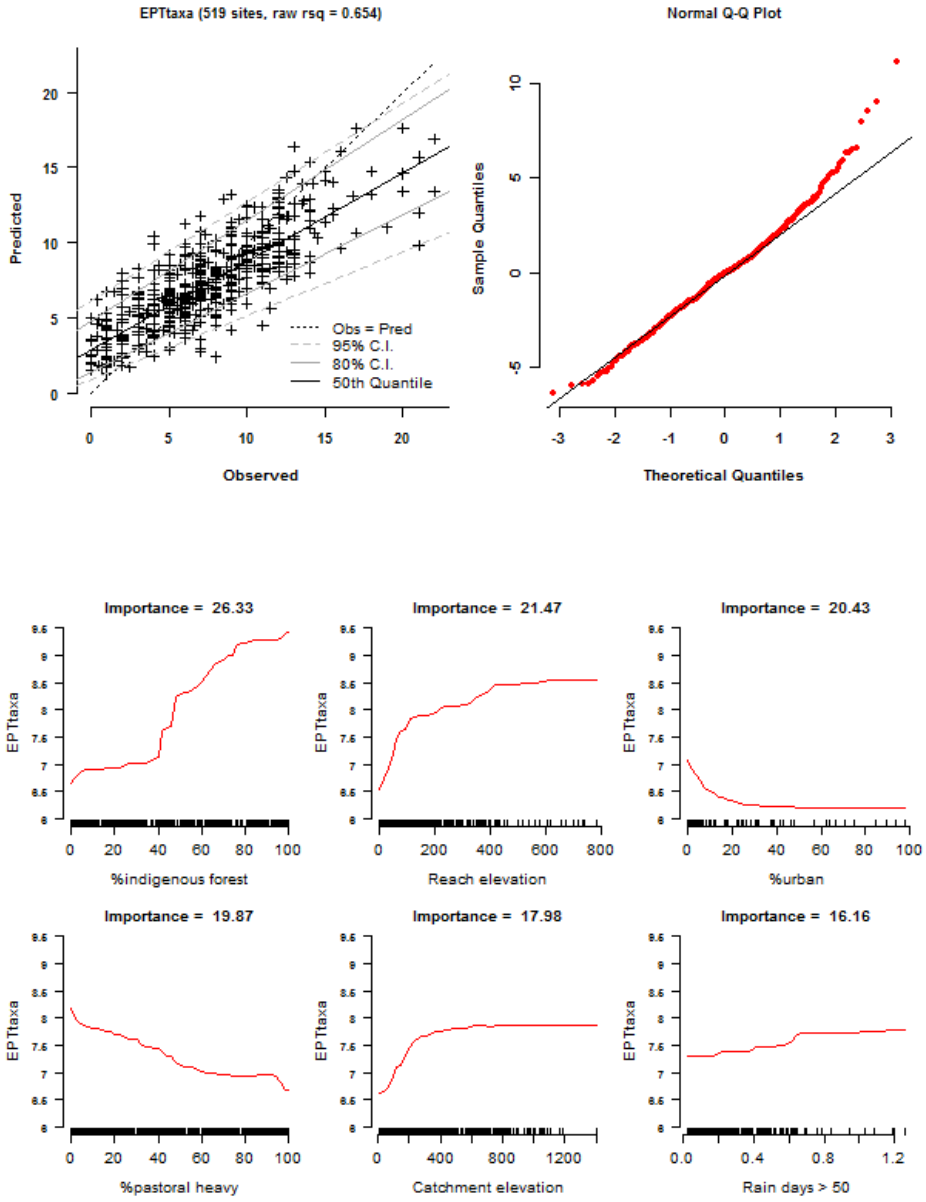
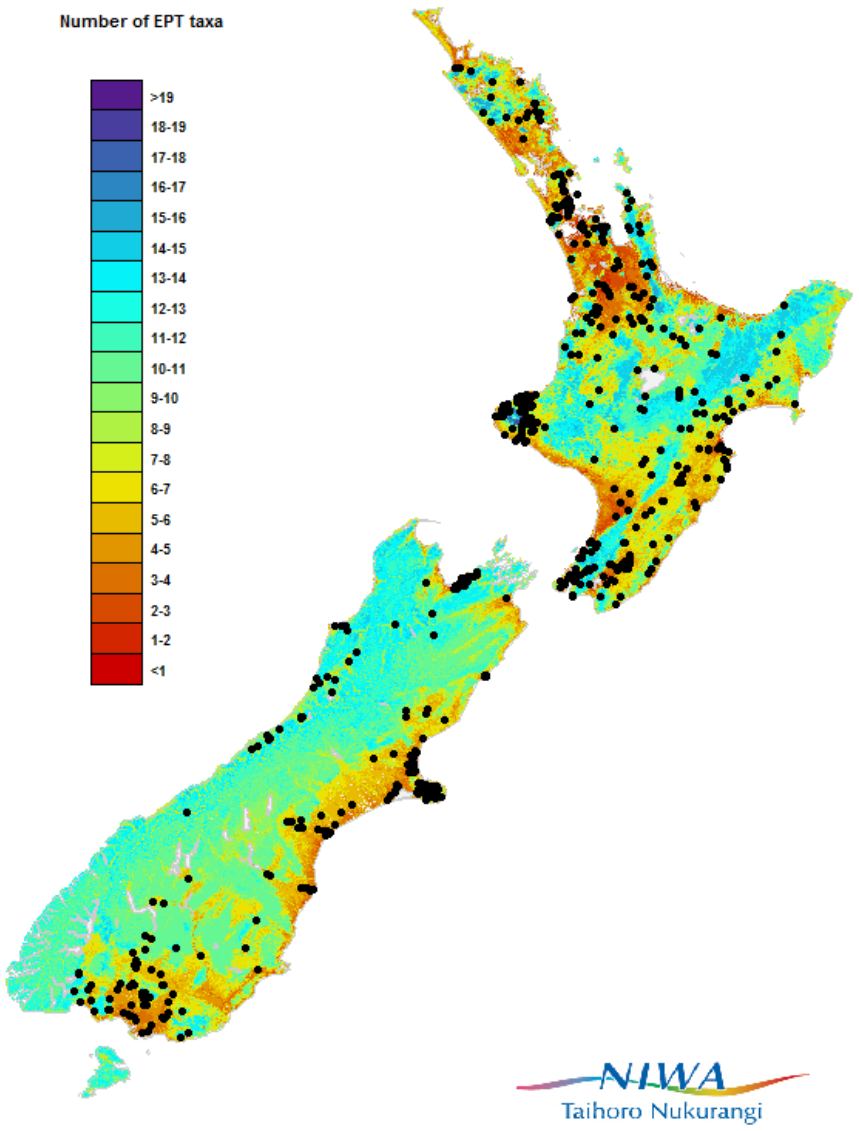
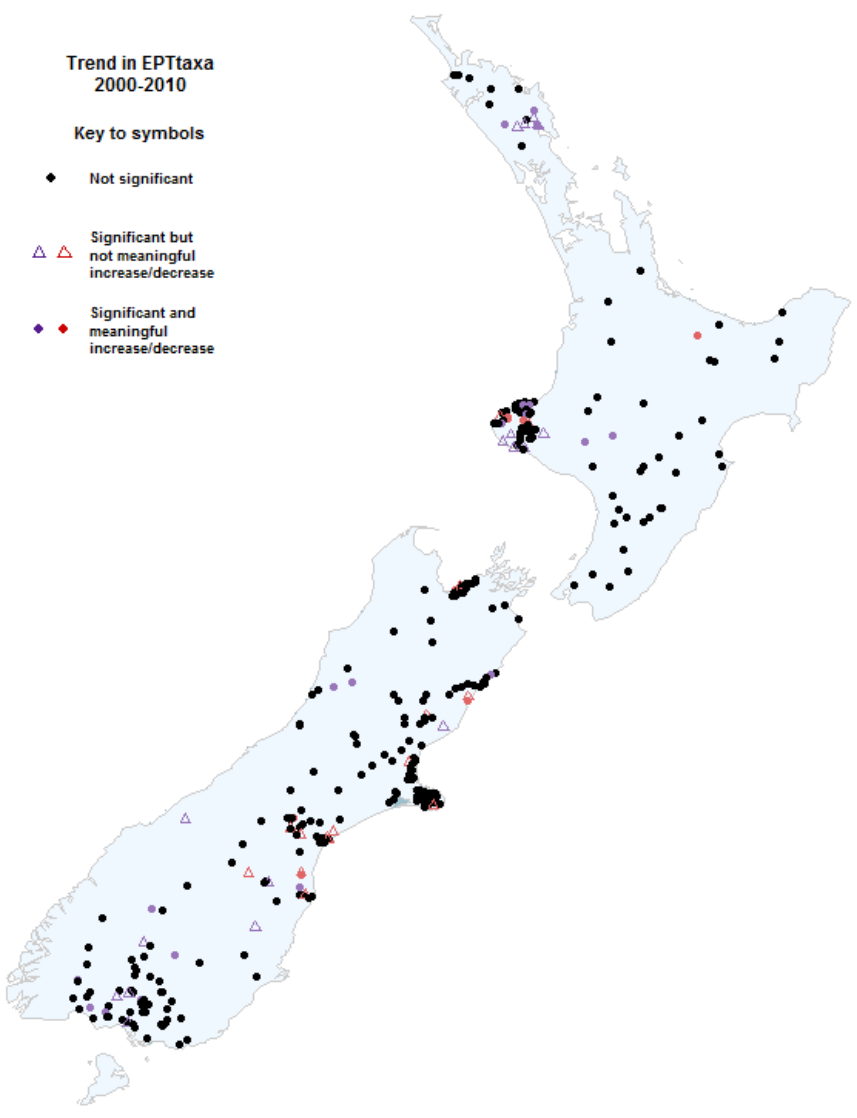


Figure B2: Modelled current state and trend data for semi-quantitative macroinvertebrate community index (SQMCI). Successive panels show diagnostic plots for the fitted random forest model (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites (lower left); and modelled 2000-2010 trends for all NZReaches (lower right). For trend maps, blue/red shading shows improving/decreasing index values, respectively.







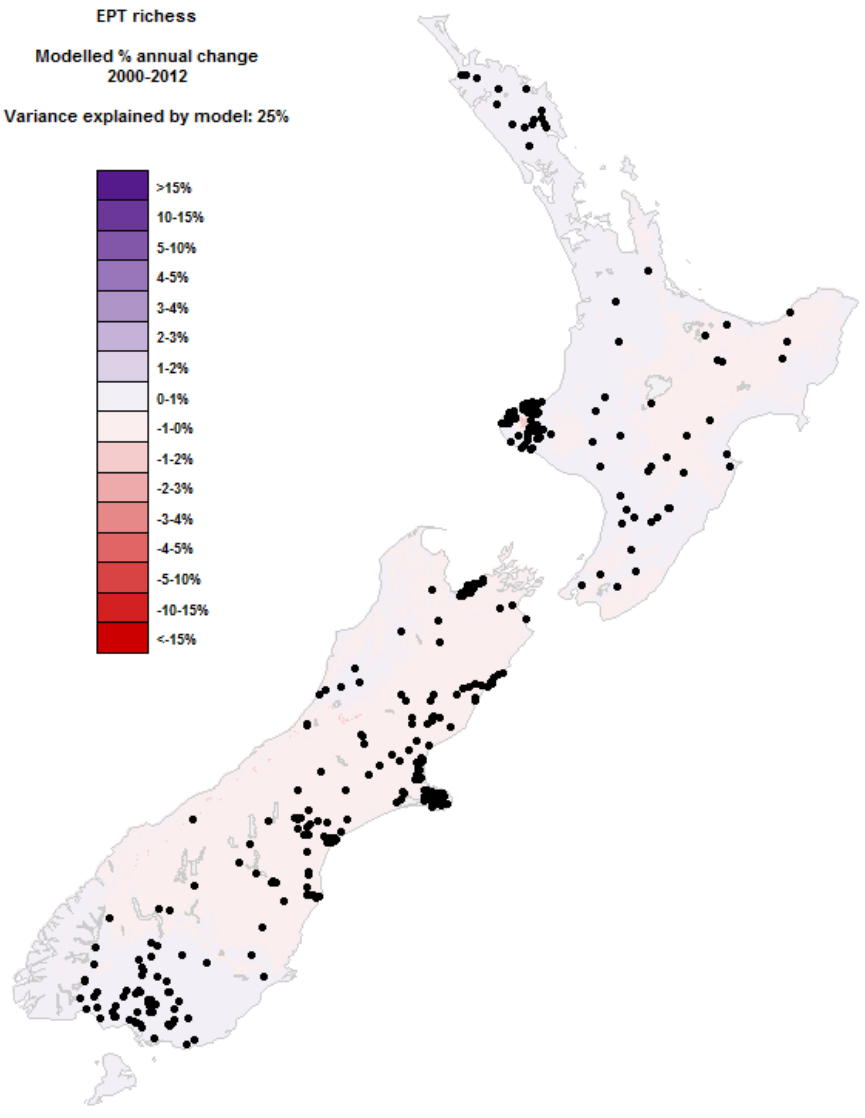
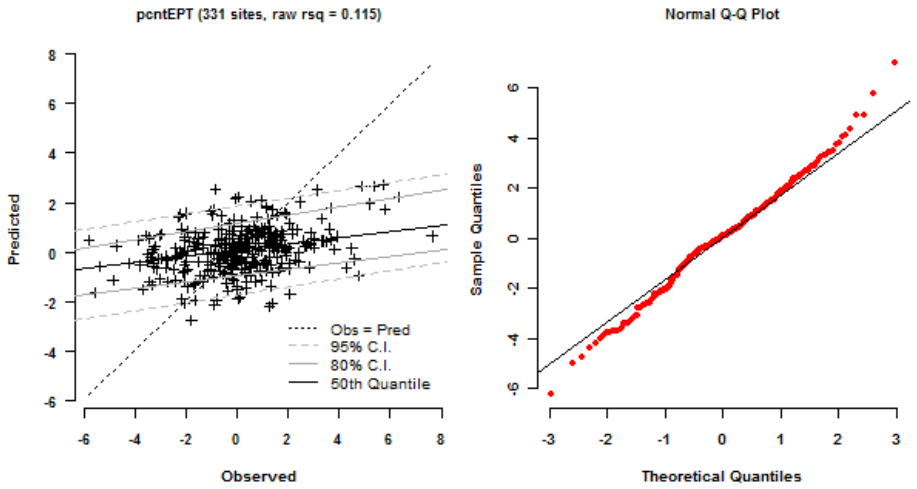
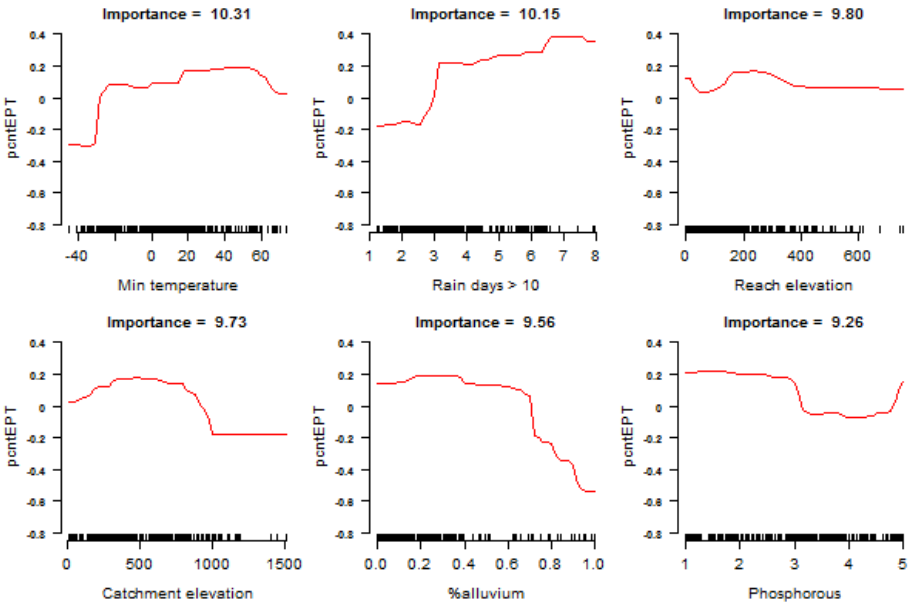
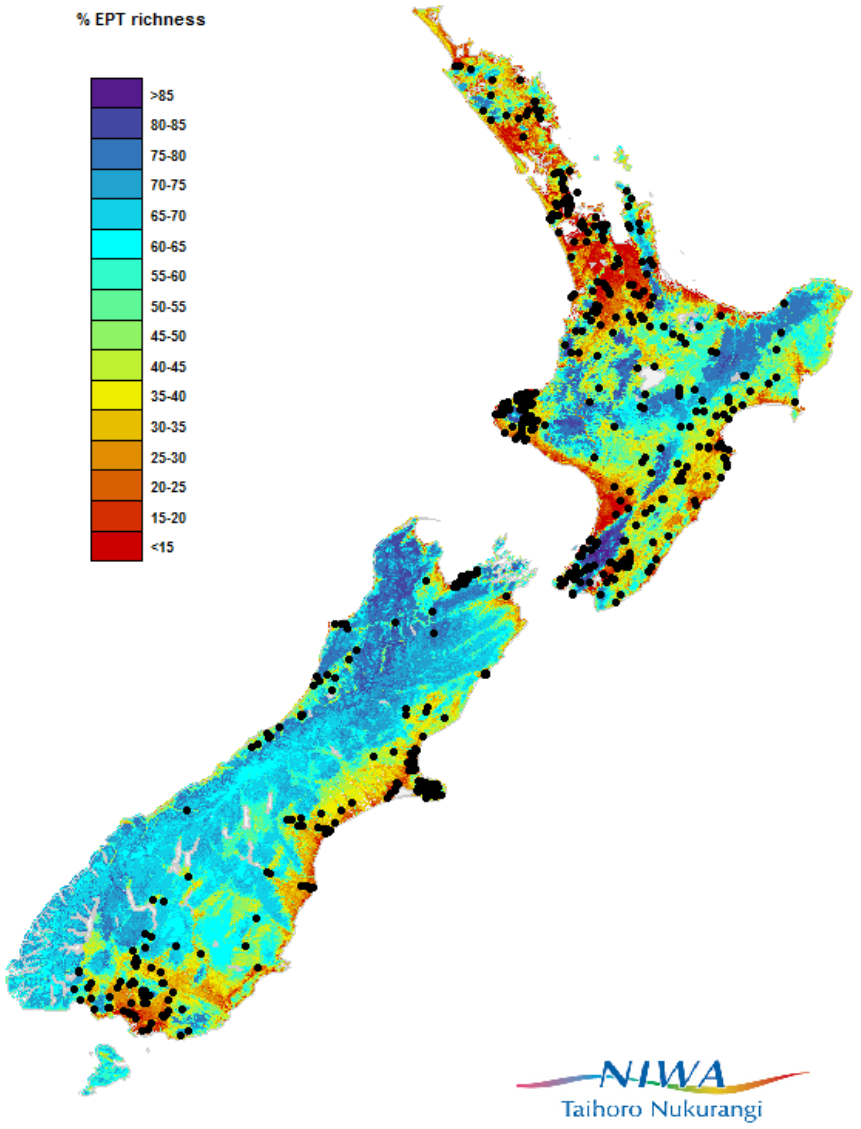
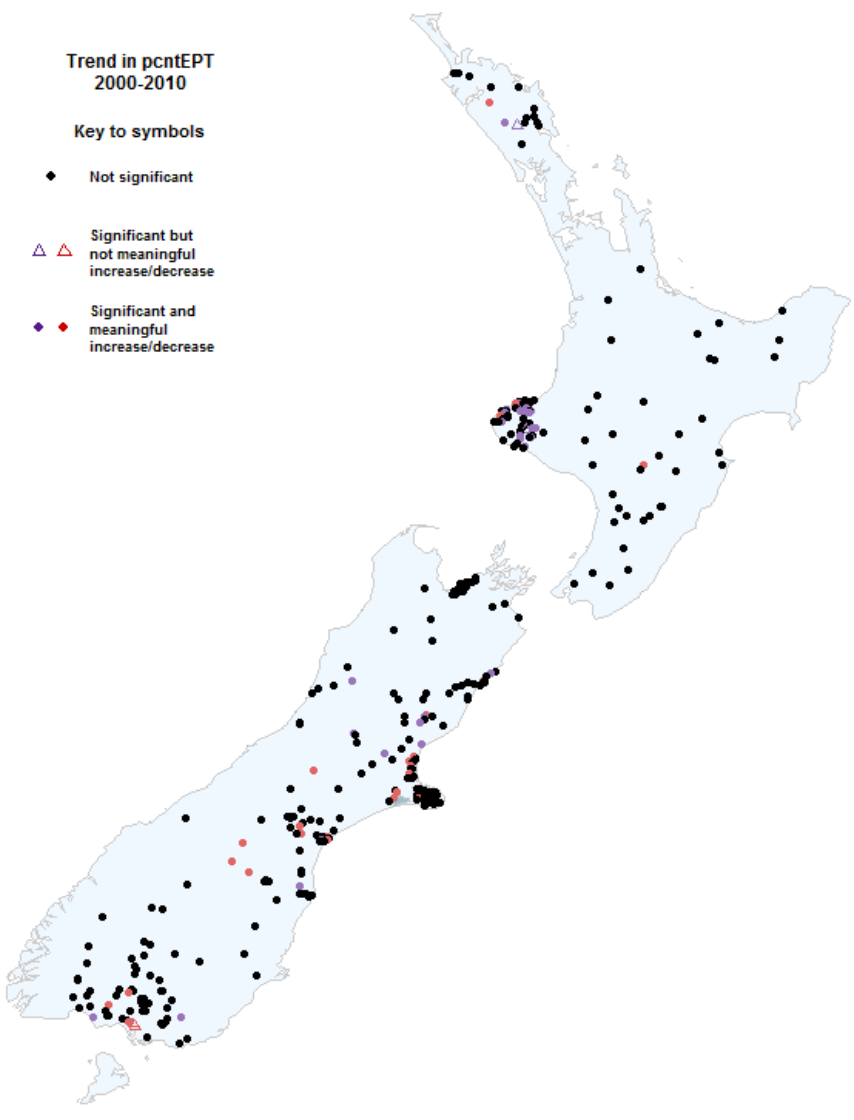


Figure B3: Modelled current state and trend data for EPT taxonomic richness (EPTtaxa). Successive panels show diagnostic plots for the fitted random forest model (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites (lower left); and modelled 2000-2010 trends for all NZReaches (lower right). For trend maps, blue/red shading shows improving/decreasing index values, respectively.









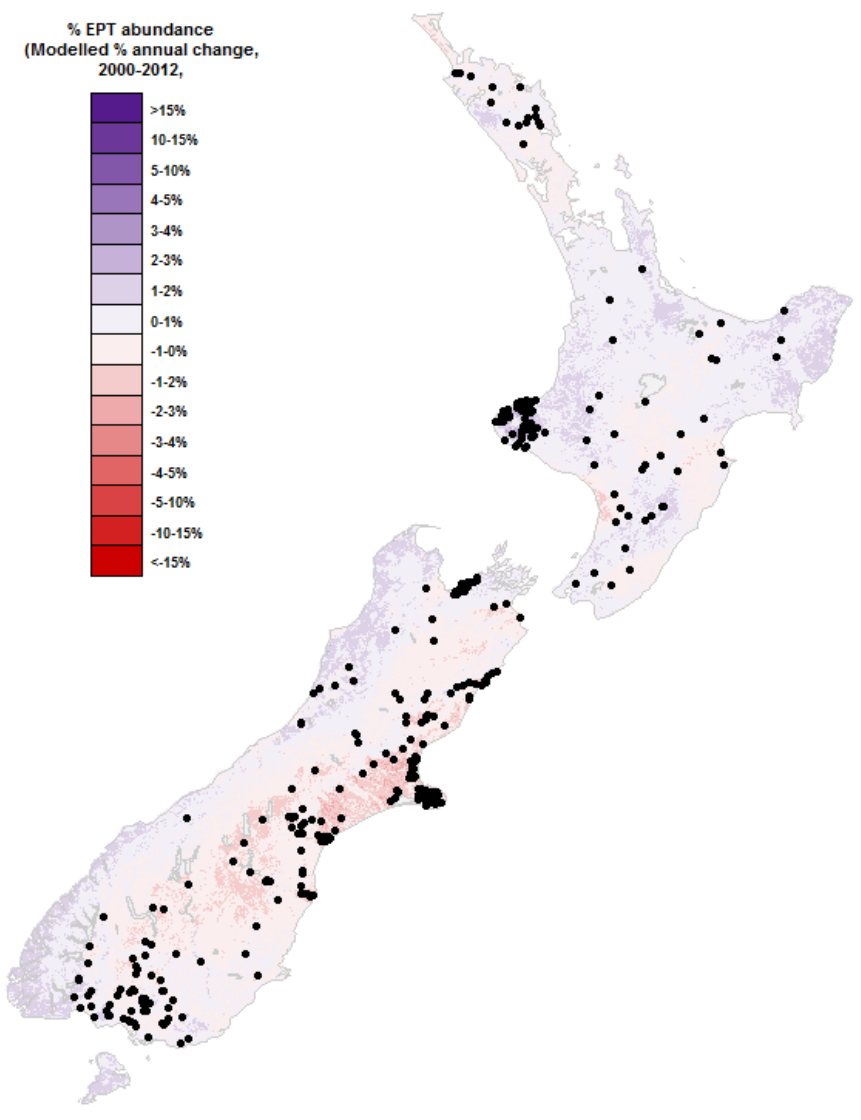


Figure B4: Modelled current state and trend data for percentage EPT abundance (pcntEPT). Successive panels show diagnostic plots for the fitted random forest model (top left); modelled current state for all NZReaches (top right); observed 2000-2010 trends for all sites (lower left); and modelled 2000-2010 trends for all NZReaches (lower right). For trend maps, blue/red shading shows improving/decreasing index values, respectively.