Taranaki water quality state spatial modelling

July 2023

Prepared By:

Caroline Fraser

For any information regarding this report please contact:

Caroline Fraser

Phone: 0220491779

Email: caroline@landwaterpeople.co.nz

LWP Ltd
PO Box 70
Lyttelton 8092
New Zealand

LWP Client Report Number:	2021-14
Report Date:	September 2021
LWP Project:	TRC State Spatial Modelling

Quality Assurance Statement

[Click here and type text]

Version	Reviewed By	
	Ton Snelder	
Taranaki State Spatial Modelling LWP Sep21 FINAL updated Jul23 (31/7/2023)		

Table of Contents

1 Introduction 5

2 Data 5

2.1 Water quality data 5

2.2 Predictor variable data 6

2.2.1 Catchment characteristics 6

2.2.2 Stocking density data 7

2.2.3 Summary of all predictor variables 7

3 Methods 8

3.1 Water quality state analyses 8

3.1.1 Grading of monitoring sites 8

3.1.2 Handling censored values 10

3.1.3 Time period for assessments and data requirements 10

3.1.4 pH Adjustment of Ammonia 11

3.1.5 Evaluation of compliance statistics 11

3.2 Spatial modelling of state and export coefficients 11

3.2.1 Random forest models 12

3.2.2 Model performance 13

3.2.3 Modelled relationships 14

3.2.4 Representativeness of sites used in RF models 14

3.2.5 Model predictions 15

3.2.6 Evaluating confidence intervals of spatial model predictions 15

4 Results 15

4.1 State 15

4.2 State spatial modelling 18

4.2.1 Model performance 18

4.2.2 Modelled relationships 20

4.2.3 Monitored site representativeness 21

4.2.4 Model predictions 22

5 Discussion 26

Acknowledgements 27

References 28

Appendix A: Comparison of water quality compliance statistics between start of time and 2017 30

Appendix B: Comparison of compliance statistics and NOF grades over different time periods 32

Appendix C: Continuous spatial model predictions 40

Figure 1: Maps showing NPS-FM NOF attribute state grades (excluding E. coli) for the 5-year period ending December (or June for macroinvertebrates) 2017. 17
Figure 2: Maps showing NPS-FM NOF attribute state grades the E.coli attribute states for the 5-year period ending December 2017. 18
Figure 2: Comparison of observed water quality compliance statistics versus values predicted by the RF models.....20
Figure 3: Relationships of predictors included in the 'reduced' random forest models with the water quality compliance statistics.....21
Figure 4 Probability-probability plots for the top 24 most important predictors used by the water quality compliance statistics spatial models describing the representativeness of the water quality monitoring sites used to fit the spatial models. 22
Figure 5: (a) Predicted NOF grades for selected water quality variables, for all segments of the regional network. 23
Figure 6: Comparison of water quality compliance statistics from the beginning of observation records with those ending in 2017. The black line is a 1:1 line. 30
Figure 8: Variation in NOF suspended fine sediment attribute grades for sites over time. 32
Figure 9: Variation in NOF Ammonia attribute grades for sites over time.....33
Figure 10: Variation in NOF DRP attribute grades for sites over time. 34
Figure 11: Variation in NOF E. coli attribute grades for sites over time. 35
Figure 12: Variation in NOF Nitrate attribute grades for sites over time. 36
Figure 13: Variation in NOF ASPM attribute grades for sites over time.....37
Figure 14: Variation in NOF MCI attribute grades for sites over time. 38
Figure 15: Variation in NOF QMCI attribute grades for sites over time. (Note this comparison is made using TRC SQMCI observations). 39
Figure 16: (a) Predicted water quality compliance statistics for selected water quality variables, for all segments of the regional network. 40

Table 1: Water quality variables and associated numbers of sites included in this study. 6
Table 2. Predictor variables used in spatial models. 7
Table 3: Details of the NOF attributes used to grade the state of the river monitoring sites. 9
Table 4: Performance criteria for statistics used in this study, ...14
Table 5: Summary of the number and percentage (in brackets) of sites assigned to state grades for the period ending December (or June for macroinvertebrates) 2017 for sites monitored within the Taranaki region. Total number of sites is the number of sites that met minimum data requirements outlined in section 3.1.3. 16
Table 6. Performance of the state statistics RF spatial models.....19

1 Introduction

This report describes an analysis of river water quality state in the Taranaki Region in two steps. First, the study evaluated water quality state at Taranaki Regional Council (TRC) river monitoring sites and graded each site into relevant attribute bands designated in Appendix 2A and 2B of the National Policy Statement – Freshwater Management (NZ Government, 2020). Second, data from river water quality sites were used to develop spatial models of river water quality state.

The primary purpose of the spatial modelling is to provide large-scale water quality assessments that are more representative of the true patterns of water quality than assessments based on aggregated data from individual monitoring sites. The latter approach can lead to conclusions about water quality patterns that are biased by the non-random locations of monitoring sites. Previous studies have shown that the aggregating river water quality monitoring data from sites nationally, without spatial modelling, leads to an over-representation of some types of catchment (e.g., catchments dominated by pastoral land cover) and under-represent other types of catchments (e.g., catchments dominated by native forest) (Snelder *et al.*, 2014; Whitehead, 2018). Spatial modelling of water quality state as a function of catchment and other characteristics reduces the problem of biased representation of the monitoring sites and produces predicted patterns of river water quality that can be used to inform decisions at unmonitored locations and can lead to other insights. For the spatial modelling, TRC monitoring sites were supplemented by monitoring data from neighbouring regions and the national river water quality network (NRWQN) to increase coverage of the range of environmental conditions in the region.

The results provide detailed data describing the grades assigned to river water quality sites in the region for a range of water quality measures. In addition, we describe the statistical performance of the spatial models, provide maps showing regional predictions of river water quality state for the water quality measures, and identify important relationships between water quality state and catchment conditions.

2 Data

2.1 Water quality data

River water quality data was used in this study to investigate spatial patterns in water quality state across the Taranaki region. We obtained water quality data timeseries representing physico-chemical, microbiological and biological variables from the TRC database. To supplement the spatial coverage provided by the TRC monitoring network, we also obtained water quality state timeseries data for monitoring sites in neighbouring regions (Manawatū-Whanganui and Waikato) and for sites within Taranaki that are monitored by NIWA, as part of the National River Water Quality Network (NRWQN) (Larned et al. 2018). Table 1 describes the variables and total numbers of sites by region.

Statistics such as median and 95th percentile values that define NPS-FM (2020) NOF attribute states were calculated for each site and water quality variable from the timeseries data as described in section 3.1.3. Note that the numbers of non-TRC sites reflect the availability of sites that complied with the data requirement rules used for calculating the statistics that are outlined in section 3.1.3.

Table 1: Water quality variables and associated numbers of sites included in this study. Where "MW" is Manawatū Whanganui and "Wk" is Waikato

Variable type	Variable name	Description	Units	Number of sites		
				TRC +NRWQN	MW	Wk
Physico-Chemical	CLAR	Black Disc Visibility	m	30	11	82
	DRP	Dissolved Reactive Phosphorous	mg l^{-1}	38	123	112
	NH_4-N	Ammoniacal Nitrogen (pH Adjusted)	mg l^{-1}	38	116	106
	NH_4-N_raw	Ammoniacal Nitrogen (raw)	mg l^{-1}	41	124	112
	NO_3-N	Nitrate	mg l^{-1}	32	124	112
	pH	Field pH	pH	38	NA	NA
	TN	Total Nitrogen	mg l^{-1}	32	124	112
	TP	Total Phosphorous	mg l^{-1}	30	124	112
Micro-biological	*E. coli*	*E. coli* cfu	cfu 100mL^{-1}	32	124	112
Biological	ASPM	Average score per metric	ASPM	85	0	0
	MCI	Macroinvertebrate Community Index	MCI	88	83	56
	SQMCI	Semi-Quantitative Macroinvertebrate Community Index	SQMCI	85	0	0

2.2 Predictor variable data

In this study, predictive spatial models of water quality state statistics were derived by combining monitoring site water quality statistics with predictors associated with the digital network to make predictions for unmonitored locations. The following sections describe the predictor variables used in these models.

2.2.1 Catchment characteristics

The spatial models were based on a digital drainage network that represents the region's streams and rivers and their associated catchments. We used the GIS-based digital drainage network, which underlies the River Environment Classification (REC; Snelder and Biggs, 2002). The digital network was derived from 1:50,000 scale contour maps and represents the rivers within the region as 16,627 segments bounded by upstream and downstream confluences, each of which is associated with a sub-catchment.

The digital drainage network is linked to a database describing a wide range of descriptors of the individual network segments and their upstream catchment characteristics (Wild *et al.*, 2005). We used catchment characteristics as predictors in the predictive spatial models (Table 2). Catchment topography was derived from a digital elevation model. Catchment climate characteristics were derived from climate station data as described by Wild *et al.* (2005). Catchment land cover descriptors were derived from the national Land Cover Database-3 (LCDB3) which differentiates 33 categories based on analysis of satellite imagery from 2008 (lris.scinfo.org.nz). Descriptions of catchment regolith are derived from the Land Resources Inventory (LRI) including interpretations of the LRI categories made by Leathwick et al. (2003). Descriptions of catchment hydrology were derived from national-scale hydrological modelling

(e.g., Booker and Snelder, 2012). The catchment characteristics considered in this study are summarised in Table 2.

2.2.2 Stocking density data

The catchment characteristics included five predictors that quantified the density of pastoral livestock in 2017 as indicators of land use intensity. These predictors were based on publicly available information describing the density of pastoral livestock[1]. These predictors improve the discrimination of catchment land use intensity compared to previous studies that have only had access to descriptions of the proportion of catchment occupied by different land cover categories (e.g., Whitehead, 2018). The densities of four livestock types (dairy, beef, sheep and deer) in each catchment were standardised using 'stock unit (SU) equivalents', which is a commonly used measure of metabolic demand by New Zealand's livestock (Parker, 1998). We express land use intensity as the total stock units divided by catchment area (i.e., SU ha^{-1}). We also use four additional predictors which describe the proportion of the stock units attributable to each of the four livestock types.

2.2.3 Summary of all predictor variables

Table 2. Predictor variables used in spatial models.

Predictor	Abbreviation	Description	Unit
Geography and topography	usArea	Catchment area	m^2
	usLake	Proportion of upstream catchment occupied by lakes	%
	usElev	Catchment mean elevation	m ASL
	usSlope	Catchment mean slope	degrees
	segAveElev	Segment mean elevation	degrees
Climate	usAvTWarm	Catchment averaged summer air temperature	degrees C x 10
	usAvTCold	Catchment averaged winter air temperature	degrees C x 10
	usAnRainVar	Catchment average coefficient of variation of annual rainfall	mm y^{-1}r
	usRainDays10	Catchment average frequency of rainfall > 10 mm	days $month^{-1}$
	usRainDays20	Catchment average frequency of rainfall > 20 mm	days $month^{-1}$
	usRainDays100	Catchment average frequency of rainfall > 100 mm	days $month^{-1}$
	segAveTCold	Segment mean minimum winter air temperature	degrees C x 10
Hydrology	MeanFlow	Estimated mean flow	m^3 s^{-1}
	nNeg	Mean number of days per year on which flow was less than that of the previous day	$Year^{-1}$
	MALF7	Mean annual 7-day low flow divided by the mean flow	Unitless
	FRE3	Mean number of events per year that exceeded three times the long-term median flow	$Year^{-1}$
	JulFlow	Mean daily flow for July divided by the mean daily flow	Unitless
	FloodFlow	Log10 mean annual 1-day maximum flow divided by the mean daily flow.	Unitless
Geology*	usHard	Catchment average induration or hardness value	Ordinal*
	usPhos	Catchment average phosphorous	Ordinal*

[1] https://statisticsnz.shinyapps.io/livestock_numbers/.

Predictor	Abbreviation	Description	Unit
	usParticleSize	Catchment average particle size	Ordinal*
	usCalcium	Catchment average calcium	
Land cover	usIntensiveAg	Proportion of catchment occupied by combination of high producing exotic grassland, short-rotation cropland, orchard, vineyard and other perennial crops (LCDB3 classes 40, 30, 33)	Proportion
	usIndigForest	Proportion of catchment occupied by indigenous forest (LCDB3 class 69)	Proportion
	usUrban	Proportion of catchment occupied by built-up area, urban parkland, surface mine, dump and transport infrastructure (LCDB3 classes 1,2,6,5)	Proportion
	usScrub	Proportion of catchment occupied by scrub and shrub land cover (LCDB3 classes 50, 51, 52, 54, 55, 56, 58)	Proportion
	usWetland	Proportion of catchment occupied by lake and pond, river and estuarine open water (LCDB3 classes 20, 21, 22)	Proportion
	usBare	Proportion of catchment occupied by bare ground (LCDB3 classes 10, 11, 12,13,14, 15)	Proportion
	usExoticForest	Proportion of catchment occupied by exotic forest (LCDB3 class 71)	Proportion
Stocking density data	SUTotal_2017	Stock unit density for all stock types in 2017 (i.e., total stock units)	SU ha^{-1}
	PropDairy_2017	Proportion of total stock unit density attributable to dairy cows in 2017	Proportion
	PropBeef_2017	Proportion of total stock unit density attributable to beef cows in 2017	Proportion
	PropSheep_2017	Proportion of total stock unit density attributable to sheep in 2017	Proportion
	PropDeer_2017	Proportion of total stock unit density attributable to deer in 2017	Proportion

3 Methods

3.1 Water quality state analyses

3.1.1 Grading of monitoring sites

The water quality state for river and lake monitoring sites is graded based on attributes and associated attribute state bands defined by the National Objectives Framework (NOF) of the National Policy Statement – Freshwater Management (NPS-FM) (Ministry for the Environment, 2020) (Table 3).

Each table of appendix 2 of the NPS-FM (2020) represents an **attribute** that must be used to define an objective that provides for a particular environmental **value**. For example, Appendix 2A, Table 6 defines the nitrate toxicity attribute, which is defined by nitrate-nitrogen concentrations that will ensure an acceptable level of support for "Ecosystem health (Water quality)" value. Objectives are defined by one or more **numeric attribute states** associated with each attribute. For example, for the nitrate-nitrogen attribute there are two numeric attribute states defined by the annual median and the 95th percentile concentrations.

For each numeric attribute, the NOF defines categorical numeric attribute states as four (or five) **attribute bands**, which are designated A to D (or A to E, in the case of the *E. coli* attribute). The attribute bands represent a graduated range of support for environmental values from high (A band) to low (D or E band). The ranges for numeric attribute states that define each attribute band are defined in Appendix 2 of the NPS-FM (2020). For most attributes, the D band represents a condition that is unacceptable (with the threshold between the C and the D band being referred to as "**bottom line"**) in any waterbody nationally. In the case of the Nitrate (toxicity) and Ammonia (toxicity) attributes in the 2020 NPS-FM, the C band is unacceptable, and for the DRP attribute, no bottom line is specified.

The primary aim of the attribute bands designated in the NPS-FM is as a basis for objective setting as part of the NOF process. The attribute bands are intended to be simple shorthand for communities and decision makers to discuss options and aspirations for acceptable water quality and to define objectives. Attribute bands avoid the need to discuss objectives in terms of technically complicated numeric attribute states and associated numeric ranges. Each band is associated with a narrative description of the outcomes for values that can be expected if that attribute band is chosen as the objective. However, it is also logical to use attribute bands to provide a grading of the current state of water quality; either as a starting point for objective setting or to track progress toward objectives.

A site can be **graded** for each attribute by assigning it to attribute bands (e.g., a site can be assigned to the A band for the Nitrate toxicity attribute). A site grading is done by using the numeric attribute state (e.g., annual median nitrate-nitrogen) as a **compliance statistic**. The value of the compliance statistic for a site is calculated from a record of the relevant water quality variable (e.g., the median value is calculated from the observed monthly nitrate-nitrogen concentrations). The site's compliance statistic is then compared against the numeric ranges associated with each attribute band and a grade assigned for the site (e.g., an annual median nitrate-nitrogen concentration of 1.3 mg/l would be graded as "B-band", because it lies in the range >1.0 to ≤2.4 mg/l). Note that for attributes with more than one numeric attribute state, we have provided a grade for each numeric attribute state (e.g., for the Nitrate (toxicity) attribute, grades are defined for both the median and 95th percentile concentrations).

Table 3 provides a summary of the NOF numeric attribute states calculated in this study. In addition to these NOF attributes, we have also calculated median states for Total Nitrogen (TN), Total Phosphorous (TP) and raw Ammoniacal Nitrogen (NH4N).

Table 3: Details of the NOF attributes used to grade the state of the river monitoring sites.

NPS-FM Reference – NOF Attribute	Numeric attribute state description	Units	Abbreviated name
A2A; Table 5 - Ammonia	Median concentration of Ammoniacal-N (pH adjusted)	mg l^{-1}	NOF.NH4N.Median
	95th percentile concentration of Ammoniacal-N (pH adjusted)	mg l^{-1}	NOF.NH4N.Q95
A2A; Table 6 - Nitrate	Median concentration of Nitrate	mg l^{-1}	NOF.NO3N.Median
	95th percentile concentration of Nitrate	mg l^{-1}	NOF.NO3N.Q95
A2A.; Table 8 - Suspended fine sediment	Median visual clarity	m	NOF.CLAR.Median
A2A; Table 9 - *Escherichia coli*	% exceedances over 260 cfu 100 mL^{-1}	%	NOF.ECOLI.260

NPS-FM Reference – NOF Attribute	Numeric attribute state description	Units	Abbreviated name
	% exceedances over 540 cfu 100 mL^{-1}	%	NOF.ECOLI.540
	Median concentration of *E. coli*	cfu 100 ml^{-1}	NOF.ECOLI.Median
	95th percentile concentration of *E. coli*	cfu 100 ml^{-1}	NOF.ECOLI.Q95
	Overall attribute state[1]	NA	NOF.ECOLI.Swim
A2B; Table 14 – Macroinvertebrates[2]	Median MCI score	-	NOF.MCI.Median
	Median ASPM score	-	NOF.ASPM.Median
	Median QMCI score[3]	-	NOF.QMCI.Median
A2B; Table 20 - DRP	Median concentration of DRP	mg l^{-1}	NOF.DRP.Median
	95th percentile concentration of DRP	mg l^{-1}	NOF.DRP.Q95

Notes:

(1) The overall attribute state is defined as the worst of the attribute state bands for the other 4 *E. coli* statistics.

(2) Following NPS-FM requirements Macroinvertebrate attributes are only calculated based on data collected in Dec-Mar.

(3) QMCI is not monitored in by TRC. TRC requested that their monitored SQMCI data was compared against the NPS-FM QMCI numeric attribute state.

3.1.2 Handling censored values

Censored values in the TRC water quality data were handled followed the methodology used by Larned et al (2018). Censored values were replaced by imputation for the purposes of calculating the compliance statistics. Left censored values (values below the detection limit(s)) were replaced with imputed values generated using ROS (Regression on Order Statistics; Helsel, 2012), following the procedure described in Larned *et al.* (2015). The ROS procedure produces estimated values for the censored data that are consistent with the distribution of the uncensored values and can accommodate multiple censoring limits. When there are insufficient non-censored data to evaluate a distribution from which to estimate values for the censored observations, censored values are replaced with half of their reported value.

Censored values above the detection limit were replaced with values estimated using a procedure based on “survival analysis” (Helsel, 2012). A parametric distribution is fitted to the uncensored observations and then values for the censored observations are estimated by randomly sampling values larger than the censored values from the distribution. The survival analysis requires a minimum number of observations for the distribution to be fitted; hence in the case that there were fewer than 24 observations, censored values above the detection limit were replaced with 1.1* the detection limit. The supplementary file outputs provide details about whether and how imputation was conducted for each site by criteria assessment.

3.1.3 Time period for assessments and data requirements

When grading sites based on NPS-FM attributes, it is general practice to define consistent time periods for all sites and to define the acceptable proportion of missing observations (i.e., data gaps) and how these are distributed across sample intervals so that site grades are assessed from comparable data. The time period, acceptable proportion of gaps and

representation of sample intervals by observations within the time period are commonly referred to as site inclusion or filtering rules (Larned *et al.*, 2018).

We chose time periods and filtering rules to be consistent with those used by Larned et al. (2018), in order to ensure that the state statistics calculated for the TRC sites were consistent with those calculated for the NRWQN and sites in the neighbouring regions. The grading assessments were made for the 5-year time period to the end of December 2017, with the exception that ASPM, MCI and SQMCI were evaluated for a 5-year time period to the end of June 2017 (aligning the assessment period with water-years). State observations were only included in the spatial models if they met the filtering requirements outlined in Larned et al. (2018): (1) for monthly monitored data, this required that 90% of months in the 5-year period had observations; (2) for macroinvertebrate observations, the requirement was that there was at least one observation in 4 of the 5 water years.

We also assessed the changes in water quality state over time for the monitored water quality sites within the Taranaki region. The outcomes of this analysis are described in detail in Appendix A and B. Briefly, for each site, we evaluated the compliance statistics associated with the numeric attribute states described in Table 3 and assigned grades for rolling 5-year period windows since the beginning of site records. It had initially been intended to develop separate spatial models that were representative of water quality state at the beginning of the region's monitoring record. However, there was a lack of donor sites from neighbouring regions, and limited variation in the water quality statistics for sites in Taranaki relative to the errors in the state spatial models. Therefore, this additional spatial modelling was not able to be pursued.

3.1.4 pH Adjustment of Ammonia

Ammonia is toxic to aquatic animals and is directly bioavailable. When in solution, ammonia occurs in two forms: the ammonium cation (NH_4^+) and unionised ammonia (NH_3); the relative proportions of the forms are strongly dependent on pH (and temperature). Unionised ammonia is significantly more toxic to fish than ammonium, hence the total ammonia toxicity increases with increasing pH (and/or temperature) (ANZECC, 2000). The NPS-FM attribute for ammonia requires a correction to account for pH. We applied a pH correction to NH_4-N to adjust values to equivalent pH 8 values, following the methodology outlined in Hickey (2014). For pH values outside the range of the correction relationship (pH 6-9), the maximum (pH<6) and minimum (pH>9) correction ratios were applied.

3.1.5 Evaluation of compliance statistics

For numeric attribute states specified as "Annual" (maximum, median, 95th percentile) in the NPS-FM (2020), we calculated the compliance statistics over the entire 5-year period used for the state assessment (i.e., 1 January 2013 to 31 December 2017, or 1 July 2012 to 30 June 2017).

3.2 Spatial modelling of state and export coefficients

We used statistical spatial modelling to predict state (e.g., NPS-FM compliance statistics) for all segments of the region's river network (section 3.2.1). The modelled predictions represent an estimate of state at unmonitored locations and can be used to make comparisons between locations.

3.2.1 Random forest models

We fitted a variety of water quality characteristics derived for each monitoring site (e.g., NPS-FM numeric attribute states) to a suite of predictor variables using random forest (RF) models (Breiman, 2001; Cutler *et al.*, 2007). An RF model is an ensemble of individual classification and regression trees (CART). In a regression context, CART partitions observations (in this case the individual water quality variables) into groups that minimise the sum of squares of the response (i.e., assembles groups that minimise differences between observations) based on a series of binary rules or splits that are constructed from the predictor variables. CART models have several desirable features including requiring no distributional assumptions and the ability to automatically fit non-linear relationships and high order interactions. However, single regression trees have the limitations of not searching for optimal tree structures, and of being sensitive to small changes in input data (Hastie *et al.*, 2001). RF models reduce these limitations by using an ensemble of trees (a forest) and making predictions based on the average of all trees. An important feature of RF models is that each tree is grown with a bootstrap sample of the fitting data (i.e., the observation dataset). In addition, a random subset of the predictor variables is made available at each node to define the split. Introducing these random components and then averaging over the forest increases prediction accuracy while retaining the desirable features of CART.

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

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

A partial dependence plot is a graphical representation of the marginal effect of a predictor variable on the response variable when the values of all other predictors are held constant at their mean values. The benefit of holding the other predictors constant is that the partial dependence plot effectively ignores their influence on the response variables. Partial

dependence plots do not perfectly represent the effects of each predictor variable, particularly if predictor variables are highly correlated or strongly interacting, but they do provide an approximation of the modelled predictor-response relationships that are useful for model interpretation (Cutler *et al.*, 2007)

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

Although RF models do not depend on distributional assumptions, transformation of the response variable to an approximately symmetric distribution can improve model performance. We investigated transformations (e.g. log10, sqrt or logit) of the modelled water quality (i.e., response) variables on the model performance. Where performance was improved, we made predictions using these models.

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

3.2.2 Model performance

Model performance was assessed by comparing observations with independent predictions (i.e., sites that were not used in fitting the model), which were obtained from the OOB observations. We summarised the model performance using five statistics; regression R^2, Nash-Sutcliffe efficiency (NSE), percent bias (PBIAS), the relative root mean square deviation (RSR) and the root mean square deviation (RMSD).

The regression R^2 value is the coefficient of determination derived from a regression of the observations against the predictions. The R^2 value indicates the proportion of the total variance explained by the model, but is not a complete description of model performance (Piñeiro *et al.*, 2008).

NSE indicates how closely the observations coincide with predictions (Nash and Sutcliffe, 1970). NSE values range from $-\infty$ to 1. A NSE of 1 corresponds to a perfect match between predictions and the observations. An NSE of 0 indicates the model is only as accurate as the mean of the observed data and values less than 0 indicate the model predictions are less accurate than using the mean of the observed data.

Bias measures the average tendency of the predicted values to be larger or smaller than the observed values. Optimal bias is zero, positive values indicate underestimation bias and negative values indicate overestimation bias (Piñeiro *et al.*, 2008). PBIAS is computed as the sum of the differences between the observations and predictions divided by the sum of the observations (Moriasi *et al.*, 2007).

RSR is a measure of the characteristic model uncertainty. It is estimated as the mean deviation of predicted values with respect to the observed values (the root mean square deviation), divided by the standard deviation of the observations (Moriasi *et al.*, 2007).

The normalization associated with PBIAS and RSR allowed the performance of models to be compared across all of the modelled water quality variables. Model predictions were evaluated to be very good, good, satisfactory or unsatisfactory, following the criteria proposed by Moriasi et al., 2007, outlined in Table 4.

Table 4: Performance criteria for statistics used in this study, from (Moriasi et al., 2007).

Performance Rating	RSR	NSE	PBIAS
Very good	RSR ≤ 0.50	NSE > 0.75	|PBIAS| <25
Good	0.50 < RSR ≤ 0.60	0.65 < NSE ≤ 0.75	25 ≤ |PBIAS| < 40
Satisfactory	0.60 < RSR ≤ 0.70	0.50 < NSE ≤ 0.65	40 ≤ |PBIAS| < 70
Unsatisfactory	RSR > 0.70	NSE ≤ 0.5	|PBIAS| ≥ 70

RMSD is a measure of the characteristic model statistical error or uncertainty. RMSD is the mean deviation of predicted values with respect to the observed values (distinct from the standard error of the regression model). We used RMSD to evaluate the confidence intervals of the predictions.

3.2.3 Modelled relationships

RF model importance measures were used to quantify the contribution of each predictor to the model prediction accuracy (Breiman, 2001; Cutler *et al.*, 2007). Partial dependence plots (PDPs) were used to describe the fitted predictor-response relationships (Cutler *et al.*, 2007).

We approximated the direction of the influence of each predictor by the sign of a linear regression fitted to the data representing the PDPs i.e., the regressor is the range in the predictor variable (the variable on the x-axis of the PDP) and the regressand is the corresponding marginal response (the variable on the y-axis of the PDP. There is a loss of information associated with representing the PDP as linear regression because PDPs can have non-linear shapes and describe non-monotonic responses. This loss of information was considered an acceptable trade-off with the simpler representation of the key modelled relationships. We reversed the sign of these slopes for variables for which increasing state indicated an improvement (this included the variables: MCI, CLAR). We used heat plots to graphically display the relative contributions and direction of influence of each of the predictors. In these plots, the intensity of the colour is a measure of the importance, and the direction of influence is indicated by the colour; red indicates that increasing values of the predictor corresponds to degrading state/load and green indicates that increasing values of the predictor correspond to improving state/load).

3.2.4 Representativeness of sites used in RF models

A graphical comparison was used to gauge how well all the monitoring sites used to fit the RF models represented environmental variation at the regional scale. Here, representativeness refers to the degree to which the distribution of the predictor variable over the monitored sites matches the distribution of the predictor variable over all segments of the digital river network in the region. Poor representativeness indicates reduced reliability of the model predictions

because some parts of the environmental conditions that are present in the region are not represented in the fitting data.

We made the comparison by assessing how closely the distributions of each predictor for the monitoring sites matched the distribution over all segments of the digital river network using probability-probability (P-P) plots. A P-P plot is a scatter plot of the cumulative frequency distributions (CFDs) of the two datasets. A CFD varies between 0 and 1 and the comparison line is the 45° line from (0,0) to (1,1). Probability-probability plots that are close to 1:1 line, indicate that the monitoring sites are a representative sample of the environmental conditions occurring across the whole region (i.e., over all segments of the river network). Biases in the representation of the whole region by the sites are associated with deviations from the 1:1 line (i.e., either above or below the 1:1 line). Inconsistent representation of the environmental conditions across the region by the sampling sites may also be associated with the probability-probability plot appearing as a 'S' curve (i.e., alternately above and below the 1:1 line). Note that representativeness of monitored sites is different from model bias, which is defined in Section 3.2.2.

3.2.5 Model predictions

Predictions are made with RF models by "running" new cases down every tree in the fitted forest and averaging the predictions made by each tree (Cutler et al., 2007). Some of the models in this study were fitted to log10- or square root transformed data and when the model predictions were back-transformed, we corrected for retransformation bias using the smearing estimate (Duan, 1983). The back-transformed predictions were used to produce regional maps depicting the variation in each modelled characteristic.

3.2.6 Evaluating confidence intervals of spatial model predictions

The 95% confidence intervals for values predicted by our spatial models of NPS-FM attribute states for individual segments can be obtained using the following equations. Equation 6 and 7 are used for calculating the intervals for the state estimates that were $\log_{10}$ of square root transformed prior to model fitting and the prediction uncertainty (RMSD) values have been reported in the $\log_{10}$ or square root transformed space.

$$95\%\ CI = 10^{[\log_{10}(x) \pm 1.96 \times RMSD]} \qquad \text{Equation 6}$$

$$95\%\ CI = (sqrt(x) \pm 1.96 \times RMSD)^2 \qquad \text{Equation 7}$$

where x is the estimated value in the original units, RMSD is the model error and 1.96 is the standard normal deviate or Z-score for probability ($0.025 \leq Z \geq 0.975$). The prediction confidence intervals for the $\log_{10}$-and square root transformed variables, when expressed in the original units of the variables, are asymmetric and their values vary in proportion to the predicted water quality value.

4 Results

4.1 State

Table 5 provides a summary of water quality grades for each NPS-FM attribute, demonstrating the number and percentage of sites that are classified in each NOF grade. Figure 1 provides maps for each attribute showing the sites coloured by their evaluated state grade. Predicted NOF compliance statistics and grades are provided in the supplementary file *TRC State with Time_v210916.xlsx*.

Table 5: Summary of the number and percentage (in brackets) of sites assigned to state grades for the period ending December (or June for macroinvertebrates) 2017 for sites monitored within the Taranaki region. Total number of sites is the number of sites that met minimum data requirements outlined in section 3.1.3.

NOF Attribute	**Total no. of sites**	**State grade**				
		A	**B**	**C**	**D**	**E**
NOF.ASPM.Median	74	11 (14.7%)	39 (52%)	18 (24%)	7 (9.3%)	NA
NOF.Clar	20	7 (35%)	3 (15%)	1 (5%)	9 (45%)	NA
NOF.DRP.Median	21	3 (14.3%)	2 (9.5%)	1 (4.8%)	15 (71.4%)	NA
NOF.DRP.Q95	12	5 (23.8%)	1 (4.8%)	5 (23.8%)	10 (47.6%)	NA
NOF.ECOLI.260	17	4 (23.5%)	1 (5.9%)	0 (0%)	5 (29.4%)	7 (41.2%)
NOF.ECOLI.540	17	1 (5.9%)	1 (5.9%)	3 (17.6%)	5 (29.4%)	7 (41.2%)
NOF.ECOLI.Median	17	5 (29.4%)	NA	NA	5 (29.4%)	7 (41.2%)
NOF.ECOLI.Q95	17	2 (11.8%)	0 (0%)	0 (0%)	15 (88.2%)	NA
NOF.ECOLI.Swim	17	4 (23.5%)	1 (5.9%)	0 (0%)	5 (29.4%)	7 (41.2%)
NOF.MCI.Median	74	12 (15.8%)	17 (22.4%)	40 (52.6%)	7 (9.2%)	NA
NOF.NH4N.Q95	19	5 (26.3%)	12 (63.2%)	2 (5.3%)	0 (0%)	NA
NOF.NH4N.Median	19	13 (68.4%)	6 (31.6%)	0 (0%)	0 (0%)	NA
NOF.NO3N.Median	18	10 (55.6%)	7 (38.9%)	1 (5.6%)	0 (0%)	NA
NOF.NO3N.Q95	18	10 (55.6%)	7 (38.9%)	1 (5.6%)	0 (0%)	NA
NOF.QMCI.Median	74	30 (40%)	10 (13.3%)	16 (21.3%)	19 (25.3%)	NA

Figure 1: Maps showing NPS-FM NOF attribute state grades (excluding E. coli) for the 5-year period ending December (or June for macroinvertebrates) 2017.

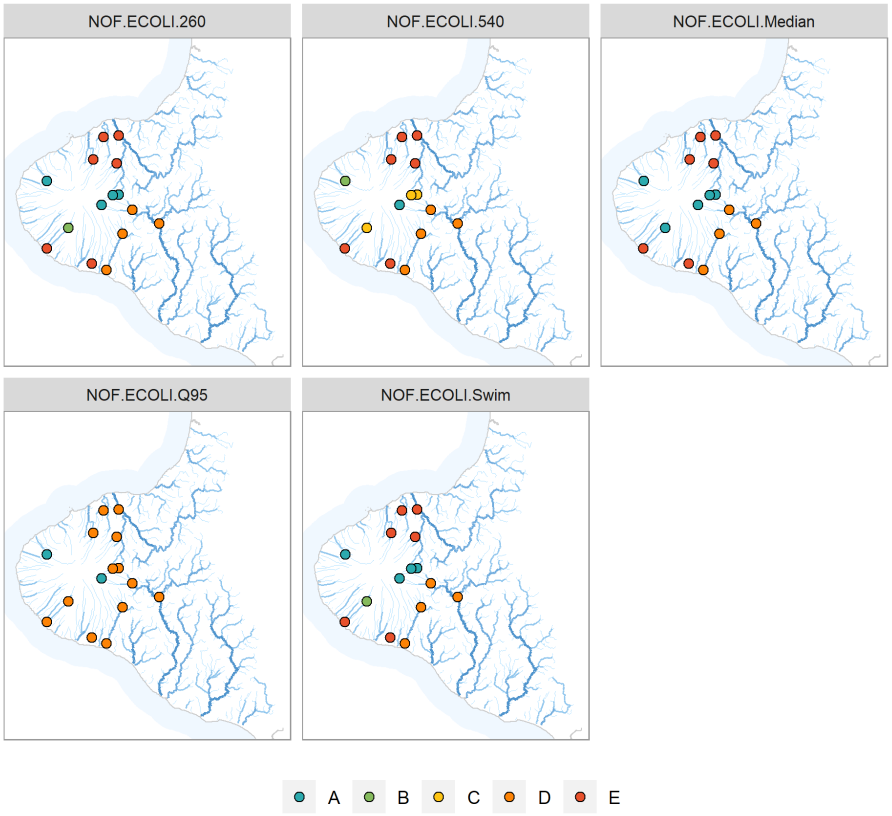

Figure 2: Maps showing NPS-FM NOF attribute state grades the E.coli attribute states for the 5-year period ending December 2017.

4.2 State spatial modelling

4.2.1 Model performance

We considered two alternative transformations of the data (as well as untransformed data) in order to optimise the performance of the spatial models of state statistics. Generally, the physico-chemical variables yielded best model performance when the response was $\log_{10}$ transformed (as they are generally strongly right skewed), whereas variables in the units of % or proportion (e.g., G260, G540) performed best with a square root transformation. Variables with approximately normal or uniform distributions (e.g., MCI, QMCI, ASPM) showed little to no improvement following variable transformation. The transformations used for each variable are listed along with model performance measures (in transformed units) in Table 6.

Table 6. Performance of the state statistics RF spatial models. Performance was determined using independent predictions (i.e., sites that were not used in fitting the models) generated from the out-of-bag observations. N=Total number of sites used to fit te model, N_T= Number of sites from Taranaki used, R^2 = coefficient of determination of observation versus predictions, NSE = Nash-Sutcliffe efficiency, PBIAS = percent bias, RSR = relative root mean square error, RMSD = root mean square deviation. RMSD units are the transformed original units.

Attribute Name	N	N_T	R^2	NSE	PBIAS	RSR	RMSD	Trans-formation
NOF.Clar	113	20	0.59	0.57	8.39	0.65	0.21	Log10
NOF.DRP.Median	256	21	0.43	0.43	0.11	0.75	0.33	Log10
NOF.DRP.Q95	256	21	0.39	0.39	1.26	0.78	0.41	Log10
NOF.ECOLI.260	253	17	0.70	0.70	-0.36	0.55	0.13	Sqrt
NOF.ECOLI.540	253	17	0.67	0.67	-0.56	0.57	0.12	Sqrt
NOF.ECOLI.Median	253	17	0.67	0.67	-0.11	0.58	0.32	Log10
NOF.ECOLI.Q95	253	17	0.64	0.64	-0.25	0.60	0.37	Log10
NOF.NH4N.Q95	241	19	0.43	0.42	0.59	0.76	0.54	Log10
NOF.NH4N.Median	241	19	0.25	0.23	0.54	0.88	0.51	Log10
NOF.NO3N.Median	254	18	0.66	0.66	-0.92	0.58	0.38	Log10
NOF.NO3N.Q95	254	18	0.77	0.77	-1.54	0.48	0.24	Log10
NOF.MCI.Median	204	74	0.74	0.74	-0.08	0.51	9.59	None
NOF.ASPM.Median	74	74	0.65	0.64	-0.16	0.60	0.07	None
NOF.QMCI.Median	74	74	0.52	0.52	-0.62	0.69	1.05	None
NH4N.raw.Median	258	22	0.23	0.20	0.93	0.90	0.49	Log10
TN.Median	256	20	0.74	0.74	1.19	0.51	0.22	Log10
TP.Median	256	20	0.65	0.65	0.30	0.59	0.25	Log10

The RF model for the 95th percentile of NO3N, had very good performance as indicated by the following statistics: NSE > 0.75, RSR < 0.5 (Table 6). The RF models of E. coli (G260, G540, median), MCI, ASPM, NO3N (median), TN and TP had good performance as indicated by the following statistics: NSE > 0.65, RSR < 0.6 (Table 6). The RF models of Clarity, E. coli (95th percentile), and QMCI had satisfactory performance as indicated by the following statistics: NSE > 0.65, RSR < 0.6 (Table 6). The models for DRP (median and 95th percentile) and NH4N (adjusted median and annual maximum, and raw median) had poorer performance, with NSE values of 0.43,0.39, 0.25, 0.43 and 0.23, respectively. Most models had very low bias; the largest bias was 8.4% for Clarity. RMSD values provide an indication of the magnitude of the characteristic error in the transformed units of each variable. Scatter plots of predicted versus observed water quality compliance statistics indicating the model performance are shown in Figure 3.

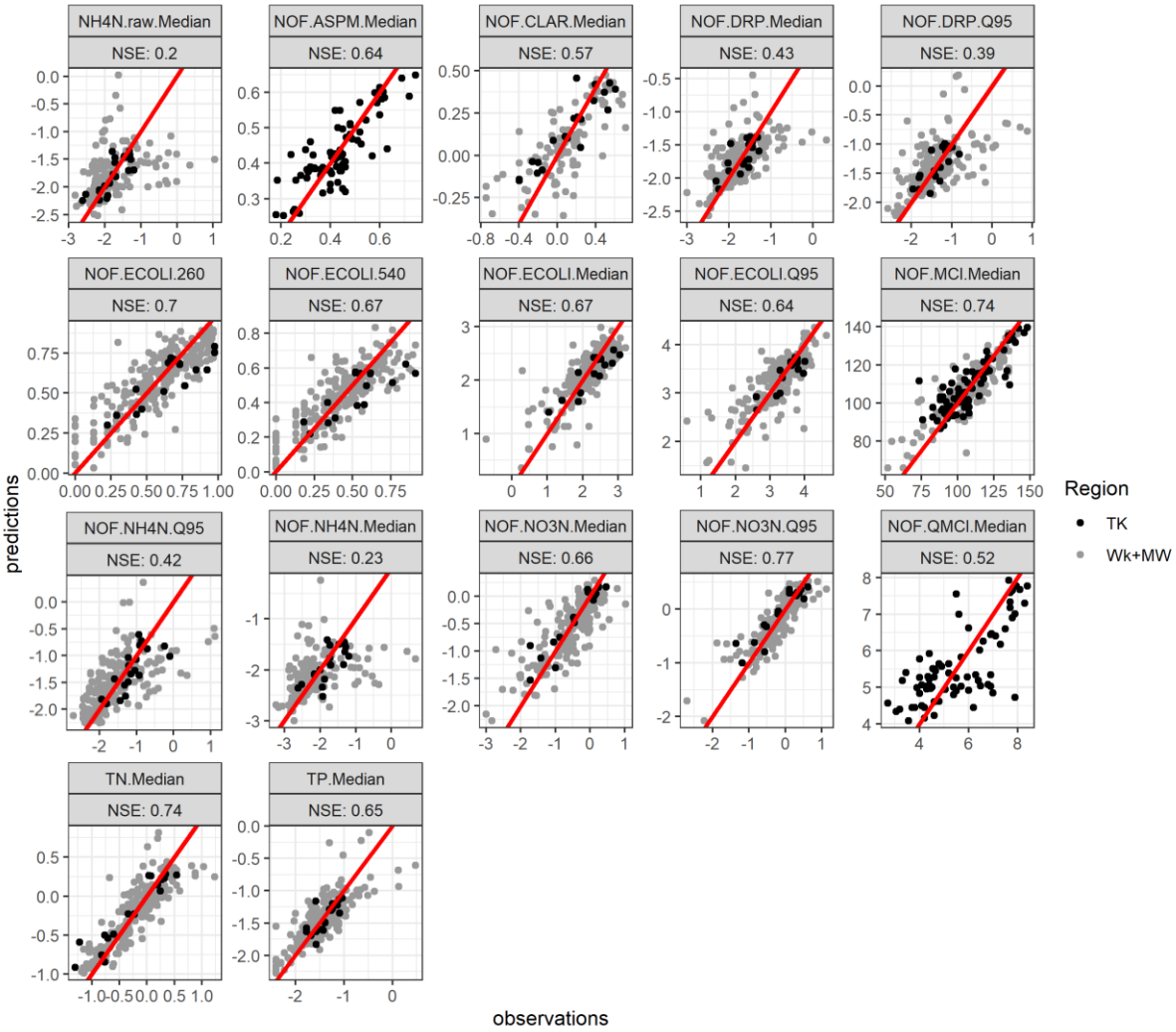

Figure 3: Comparison of observed water quality compliance statistics versus values predicted by the RF models. Points in black for sites within the Taranaki region. Points shown in grey are for sites in neighbouring regions that were also used to train the RF models. Note that the observed values are plotted on the Y-axis and predicted values on the X-axis, following (Piñeiro et al.*, 2008). The solid red line is one-to-one. Units for the variables are the transformed values (as per Table 6) of the original units.*

4.2.2 Modelled relationships

Figure 4 illustrates the relative importance and the direction of the fitted relationships between the water quality compliance statistics and the model predictors for each model.

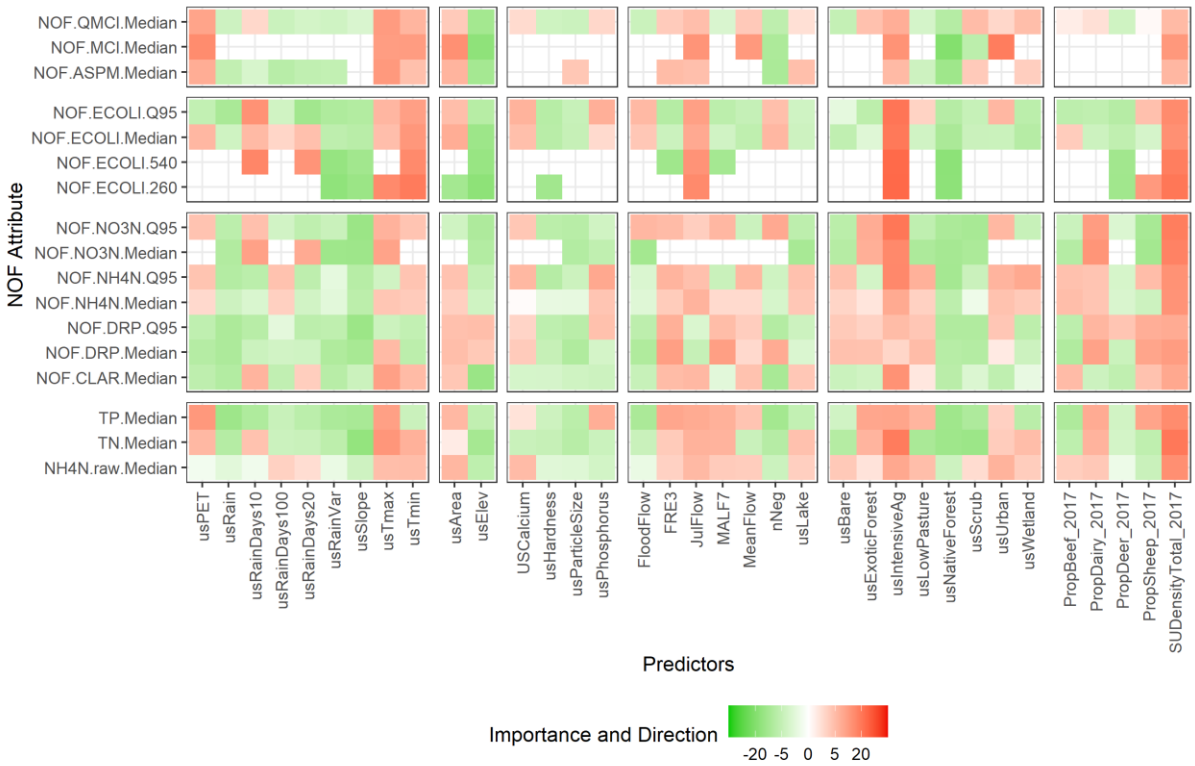

Figure 4: Relationships of predictors included in the 'reduced' random forest models with the water quality compliance statistics. Colours indicate the importance and direction of influence of each predictor on the modelled state statistics. Red indicates increasing predictor magnitudes are associated with increasing values of the state statistics, whereas green indicates increasing predictor magnitudes are associated with decreasing values. Blank cells indicate that the predictor was not included in the "'reduced'" random forest model.

4.2.3 Monitored site representativeness

The representativeness of the monitoring sites used in fitting the RF models (both from TRC and neighbouring regions) of the environmental gradients defined by the 24 most important predictor variables were inconsistent (Figure 5). The monitoring sites were generally biased towards higher values of many predictors as indicated by the probability-probability plot line lying above the red 1:1 line in Figure 5 (e.g., FRE3, *usElev, PropDeer_2017, PropDairy_2017, usUrban, usScrub*). This indicates that the monitoring sites generally overestimate catchments with: a high relative contribution to stocking units from deer and dairy cows, flashier flows, higher mean elevations and the presence of urban areas and scrub. The monitoring sites were biased towards lower values of some predictors as indicated by the probability-probability plot line lying below the red 1:1 line in Figure 5 (e.g., *usSlope, usRainDays10, usParticleSize*). This indicates the sites generally under-represent rivers with catchments with steeper slopes, higher frequency of rainfall events greater than 10mm, and catchment geology comprising larger particle sizes. The monitoring sites were biased towards median values of some predictors as indicated by the probability-probability plot line forming a flat S-curve, relative to the red 1:1 line in Figure 5 (e.g., *SUDensityTotal_2017, usNativeForest, usIntensiveAg)*. This indicates the sites generally under-represent rivers with catchments with very high or very low stocking density and native forest coverage. The monitoring sites were biased towards upper and lower values of some predictors as indicated

by the probability-probability plot line forming a steep S-curve, relative to the red 1:1 line in Figure 5 (e.g., *JulFlow, usTmax).* This indicates the sites generally over-represent rivers with catchments with very high or very low relative winter flows and average maximum temperatures.

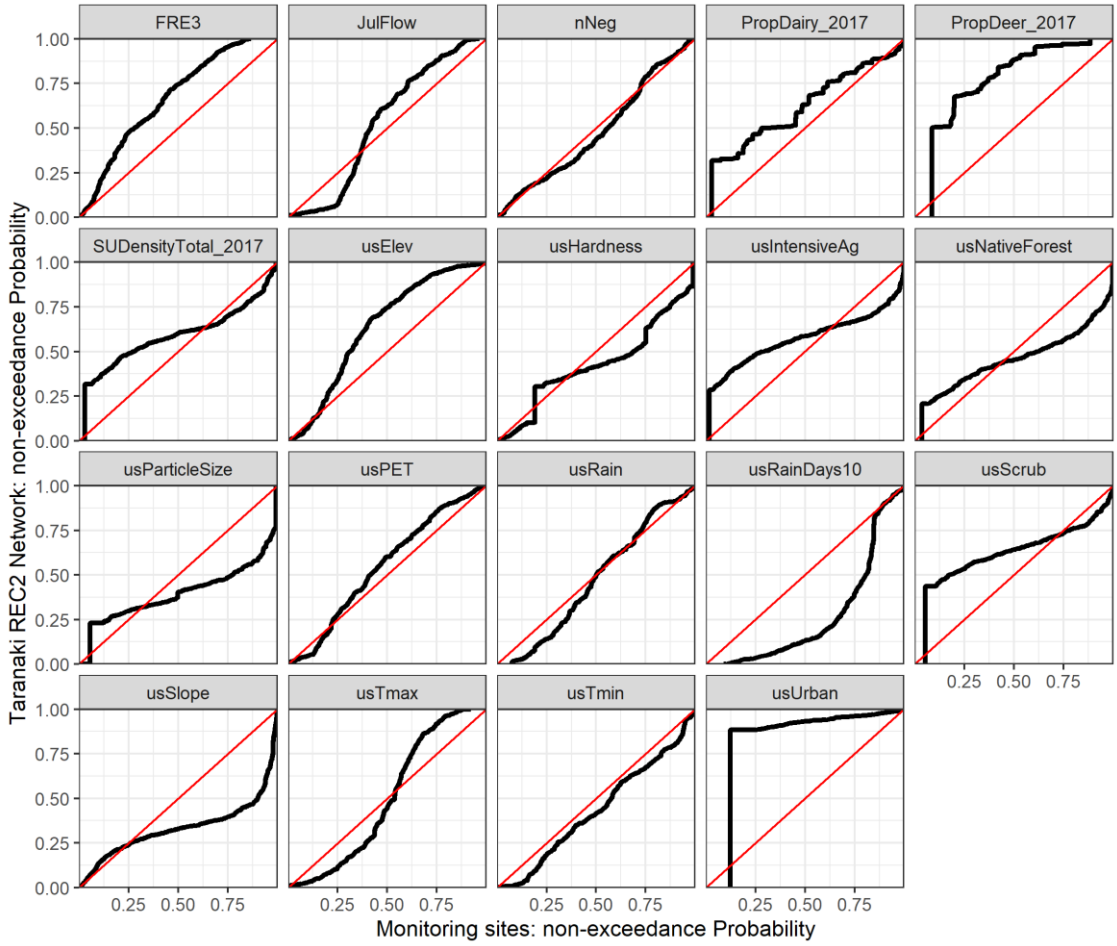

Figure 5 Probability-probability plots for the top 24 most important predictors used by the water quality compliance statistics spatial models describing the representativeness of the water quality monitoring sites used to fit the spatial models.

4.2.4 Model predictions

Figure 6 (a-d) shows maps of NOF grades evaluated from the spatial model predictions. Maps of the continuous water quality compliance statistics spatial model predictions are provided in Appendix C. There were some patterns in NOF grades that were consistent across all model predictions. For example, water quality tended to be least degraded in the eastern headwaters, in the northern part of the region and on Mount Taranaki. The most degraded areas were typically along the coastal areas (particularly on the western and southern coasts), as well as the low-lying areas around Stratford. Supplementary files with the estimated water quality compliance statistics and their 95% confidence intervals for all REC2 reaches in the Taranaki region are provided in TRCWQ_PredictionsDF_REC2_for2017__210826.csv

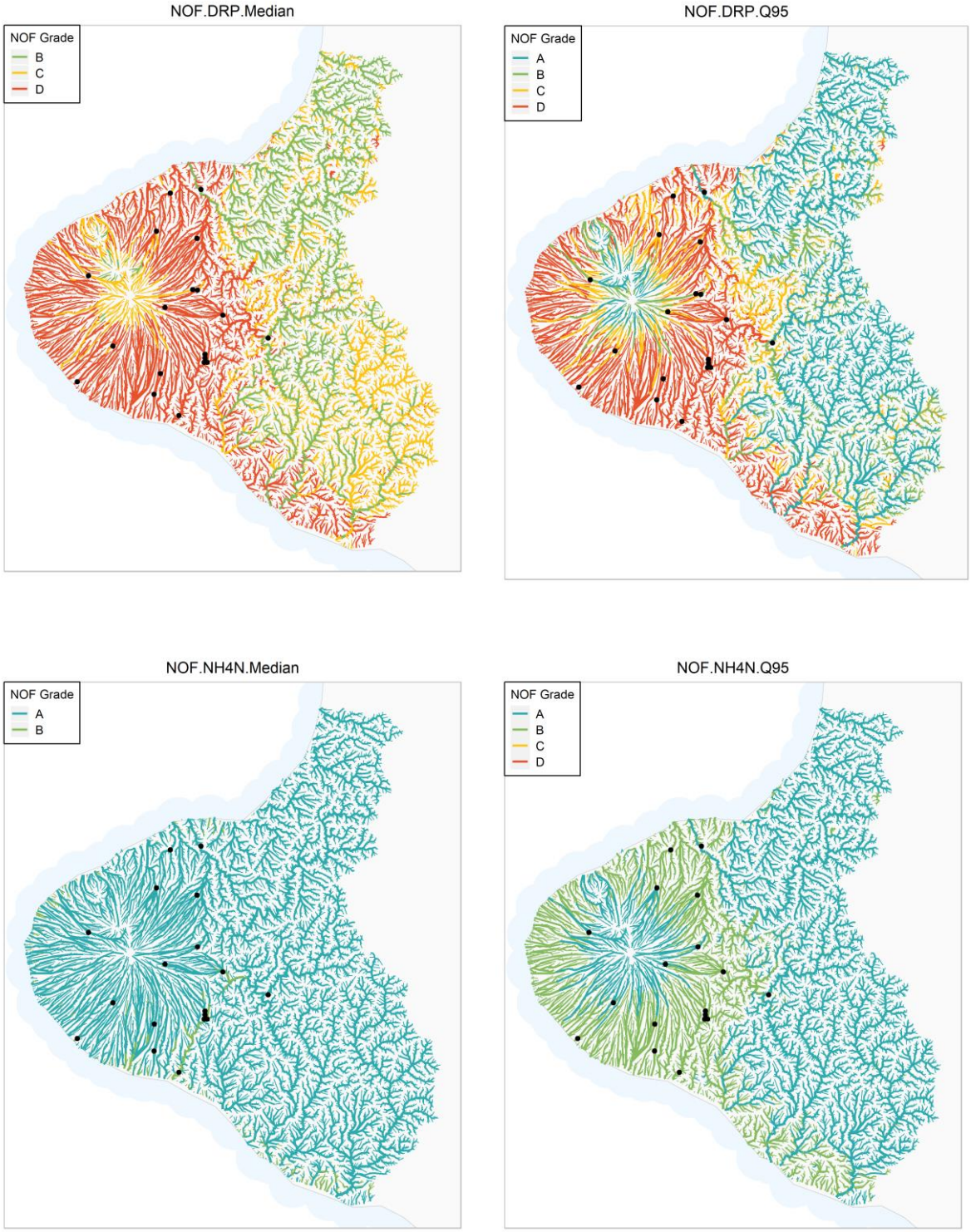

Figure 6: (a) Predicted NOF grades for selected water quality variables, for all segments of the regional network. Black dots indicate TRC sites used in model fitting.

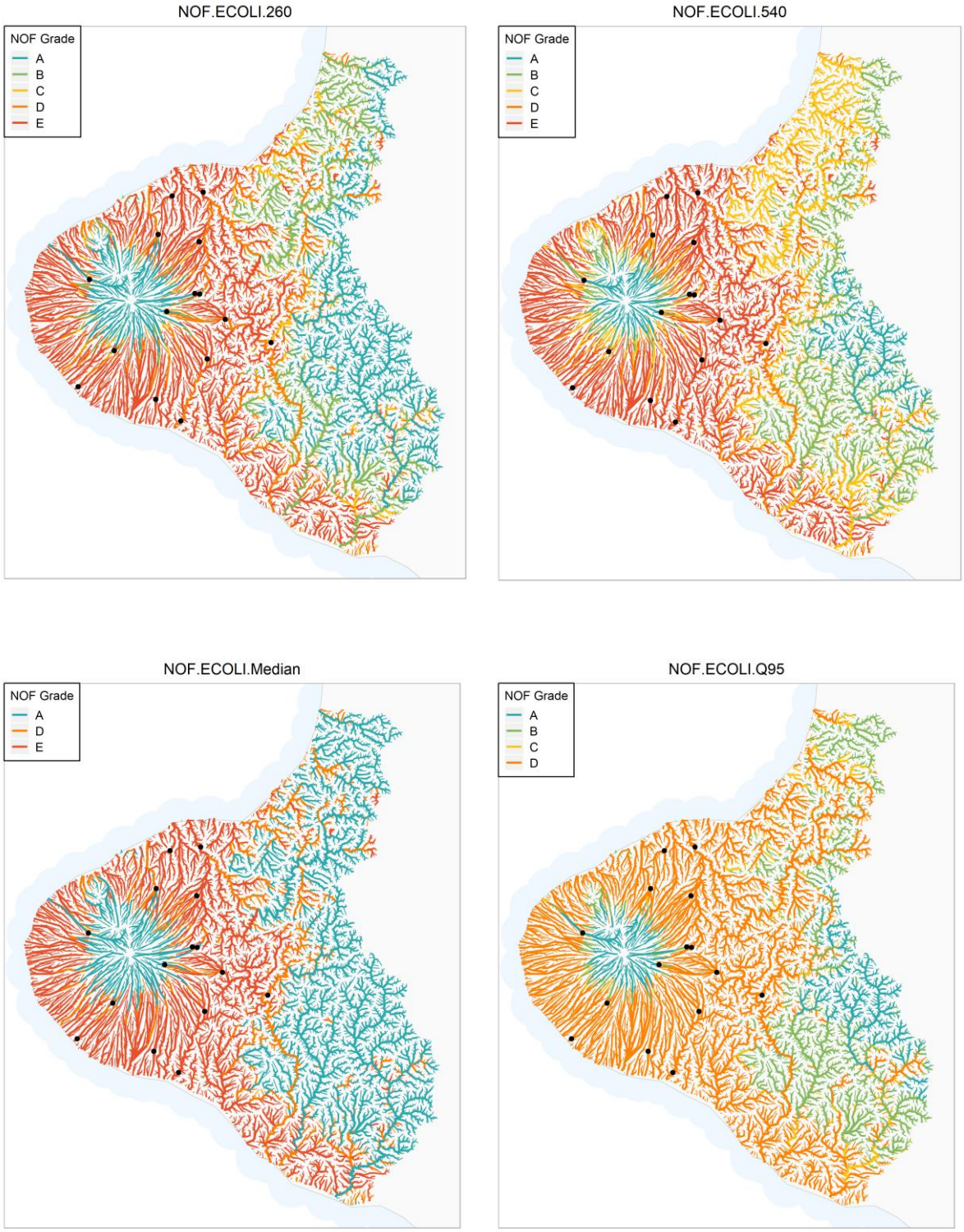

Figure 6: (b) Predicted NOF grades for selected water quality variables, for all segments of the regional network. Black dots indicate TRC sites used in model fitting.

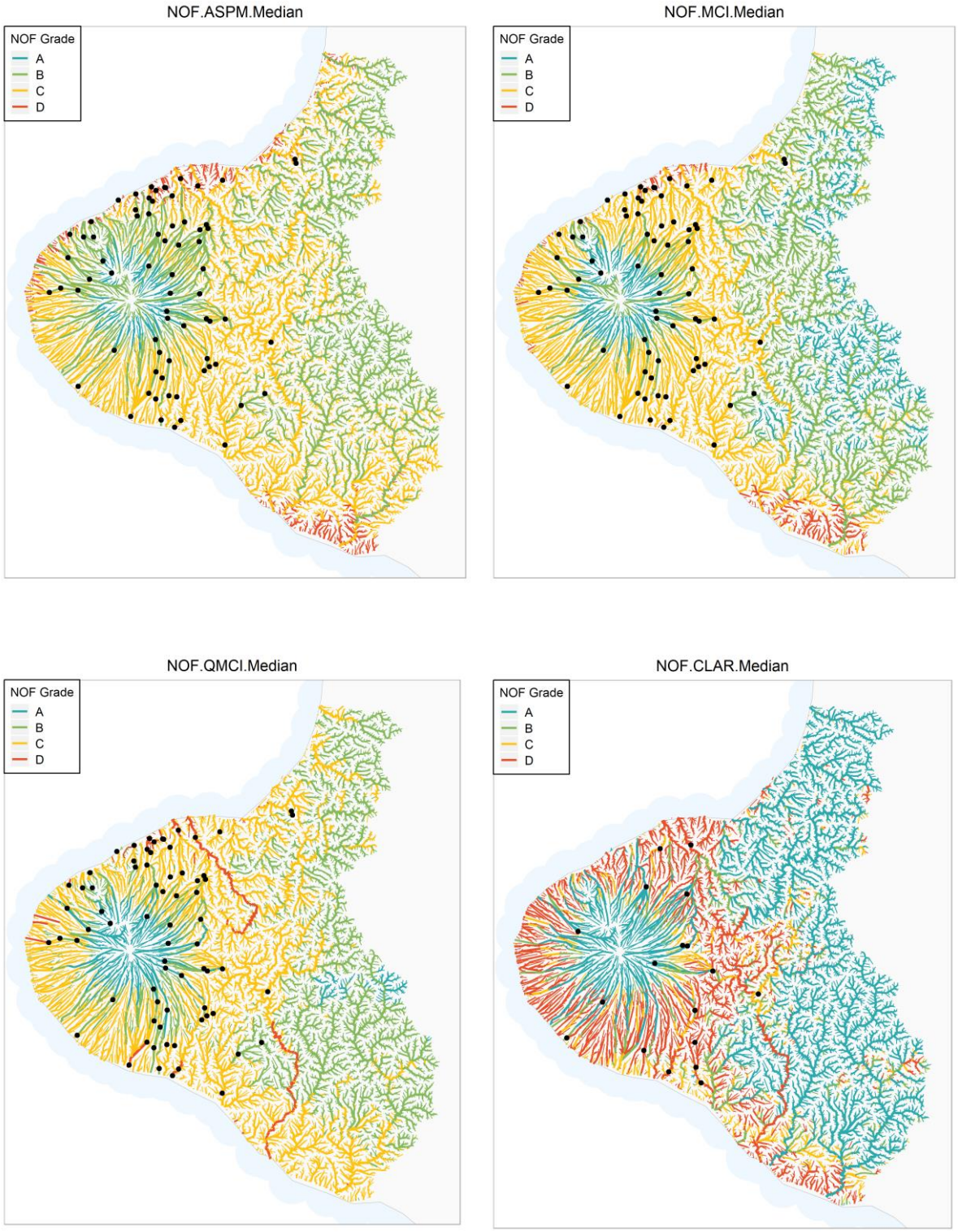

Figure 6: (c) Predicted NOF grades for selected water quality variables, for all segments of the regional network. Black dots indicate TRC sites used in model fitting.

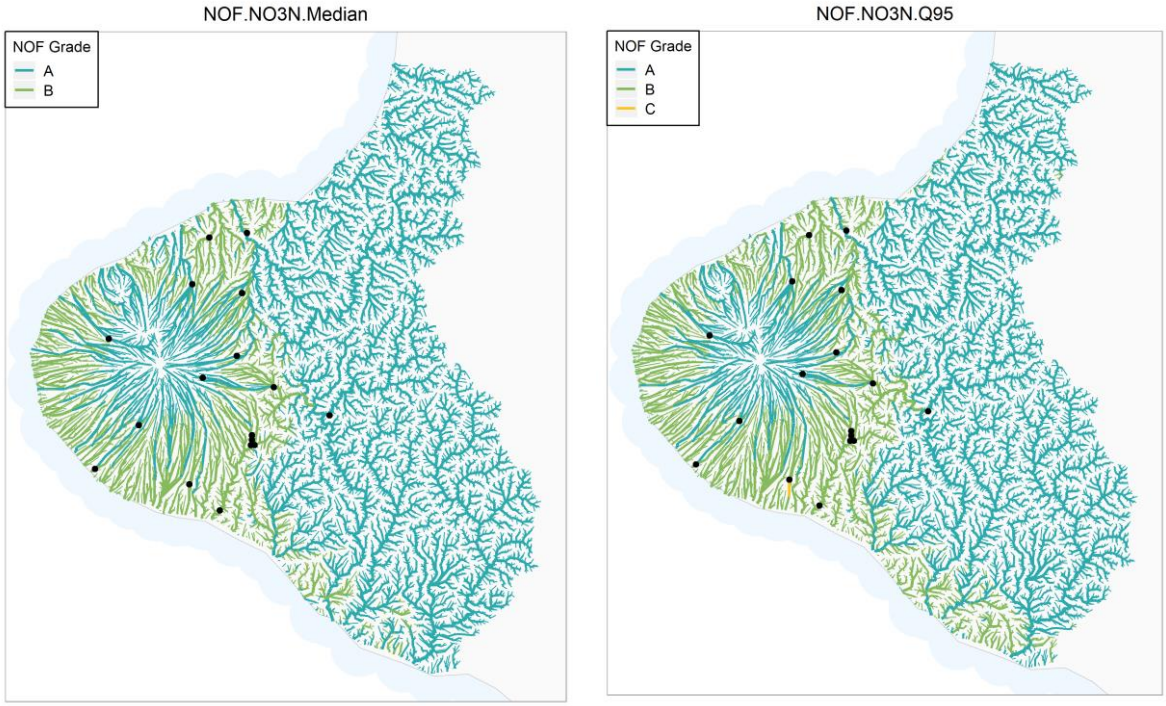

Figure 6: (d) Predicted NOF grades for selected water quality variables, for all segments of the regional network. Black dots indicate TRC sites used in model fitting.

5 Discussion

Our spatial models represent broad scale patterns in water quality (as NPS-FM attribute state statistics) based on catchment characteristics as predictor variables. The diversity of important predictor variables in the models indicates that a complex mixture of natural and anthropogenic processes (e.g., geochemical reactions, atmospheric deposition, anthropogenic nutrient input, geomorphic processes, microbial activity) influence water quality outcomes. The differences in the performance of the RF models among water quality variables (Table 6) may reflect differences in the biophysical processes that control those variables. Some biophysical processes may be poorly represented by the catchment-averaged spatial predictor variables. For example, concentrations of dissolved and total nitrogen and phosphorus in rivers are influenced to differing degrees by adsorption-desorption processes, deposition and suspension, and biological assimilation, transformation and removal; these mechanisms are not explicitly represented in the RF models. The absence of predictors that account for these and other processes means that some level of unexplained variation is inevitable.

Predictions made for individual locations are uncertain, and these uncertainties are quantified by the model RMSD values (Table 6). However, the bias of the spatial models for each contaminant was low (Table 6). This indicates that the predicted patterns reflect broad scale relative differences in water quality state between locations.

Acknowledgements

We thank Regan Phipps of Taranaki Regional Council for assistance with defining the aims of this study, providing the input datasets, and reviewing the draft report. Thanks also to Lizzie Ingham and Jeremey Wilkinson for feedback on the draft report.

References

ANZECC, A., 2000. Australian and New Zealand Guidelines for Fresh and Marine Water Quality. Australian and New Zealand Environment and Conservation Council and Agriculture and Resource Management Council of Australia and New Zealand, Canberra:1–103.

Booker, D.J. and T.H. Snelder, 2012. Comparing Methods for Estimating Flow Duration Curves at Ungauged Sites. Journal of Hydrology.

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

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

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

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

Helsel, D.R., 2012. Reporting Limits. Statistics for Censored Environmental Data Using Minitab and R. John Wiley & Sons, pp. 22–36.

Hickey, C., 2014. Derivation of Indicative Ammoniacal Nitrogen Guidelines for the National Objectives Framework. Memo prepared for Ms Vera Power, Ministry for the Environment, by NIWA.

Larned, S., T. Snelder, M. Unwin, G. McBride, P. Verburg, and H. McMillan, 2015. Analysis of Water Quality in New Zealand Lakes and Rivers. Prepared for the Ministry for the Environment. Wellington: Ministry for the Environment.

Larned, S., A. Whitehead, C.E. Fraser, T. Snelder, and J. Yang, 2018. Water Quality State and Trends in New Zealand Rivers. Analyses of National-Scale Data Ending in 2017. prepared for Ministry for the Environment, NIWA.

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

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

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

NZ Government, 2020. National Policy Statement for Freshwater Management 2020.

Parker, W.J., 1998. Standardisation between Livestock Classes: The Use and Misuse of the Stock Unit System. Proceedings of the Conference New Zealand Grassland Association., pp. 243–248.

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

Snelder, T.H. and B.J.F. Biggs, 2002. Multi-Scale River Environment Classification for Water Resources Management. Journal of the American Water Resources Association 38:1225–1240.

Snelder, T., D.. J. Booker, M. Unwin, and S.A. Wood, 2014. State and Trends of River Water Quality in the Manawatū River Catchment. Aqualinc Research Ltd.

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

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

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

Appendix A: Comparison of water quality compliance statistics between start of time and 2017

The NPS-FM requires a 'baseline state' to be defined as either the state in September 2017, or the state at the beginning of the site observation record. We calculated the water quality compliance statistics at the start of each observation record (the 'start state') as the water compliance statistic for the first 5-year period that complied with our data requirement filtering rules. The start states are compared with the 2017 states (hereafter 2017 state) in Figure 7.

For clarity and macroinvertebrate attributes (ASPM, MCI, QMCI), points lying above the 1:1 line indicate that start state was better than the 2017 state. For other variables, points lying below the 1:1 line indicate start state was better than the 2017 state (i.e., that water quality state has degraded over the time period).

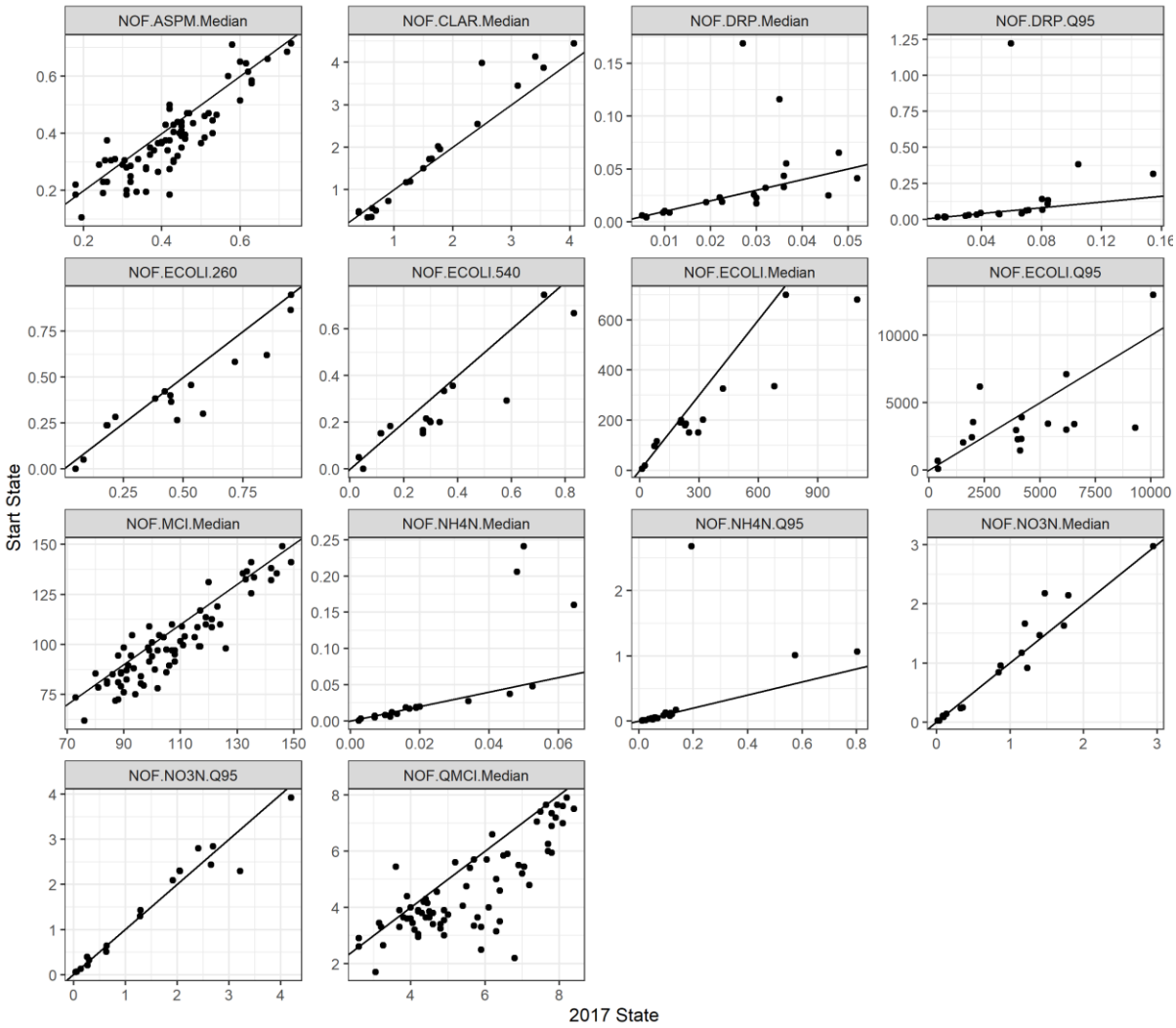

Figure 7: Comparison of water quality compliance statistics from the beginning of observation records with those ending in 2017. The black line is a 1:1 line.

Macroinvertebrate state at most sites was better in 2017 compared with the start state, and those that indicate some degradation, are only worse by a small amount (generally not enough to change the site's NOF grade). The *E. coli* compliance statistics show relatively consistent degradation across sites compared to the start state. In general, the differences between the

start state and the 2017 state were small in comparison to the variability in the water quality compliance statistics across sites.

In general, these differences between start states and 2017 state were small, or of similar magnitude to the RMSD of the state spatial models (e.g., Figure 3 and Table 6). As such, we concluded that assessment of the differences between spatial models based on the 2017 state and some earlier time period would not yield statistically significant results. We recommend, that where observed start states for monitoring sites indicate a higher water quality state than 2017, then the start state be used as a baseline, but for any assessment that uses the modelled compliance statistics to define state, the modelled (2017) predicted state is used as the baseline.

Appendix B: Comparison of compliance statistics and NOF grades over different time periods

For river water quality monitoring sites in the Taranaki region, we calculated compliance statistics for all 5-year periods within the records that complied with the data requirement rules outlined in section 3.1.3. Compliance statistics and NOF grades are provided in tabular form in the supplementary file: *TRC State with Time_v210826.xlsx.* Summaries of the variation in NOF grades for each site and NOF numeric attribute state are shown in Figure 8 to Figure 13.

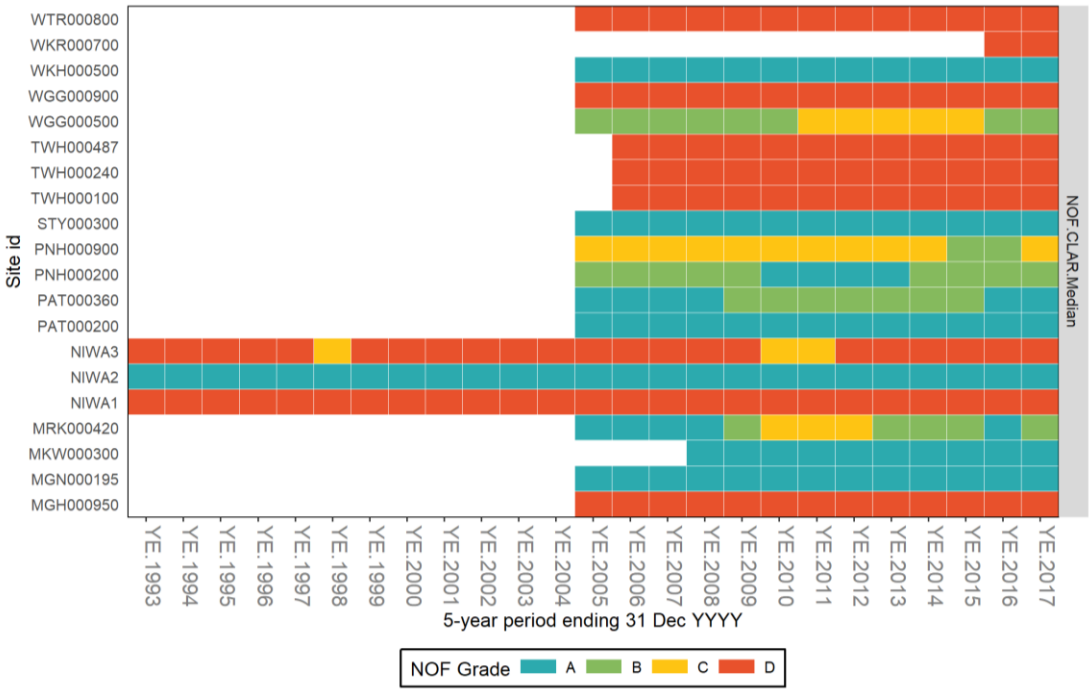

Figure 8: Variation in NOF suspended fine sediment attribute grades for sites over time.

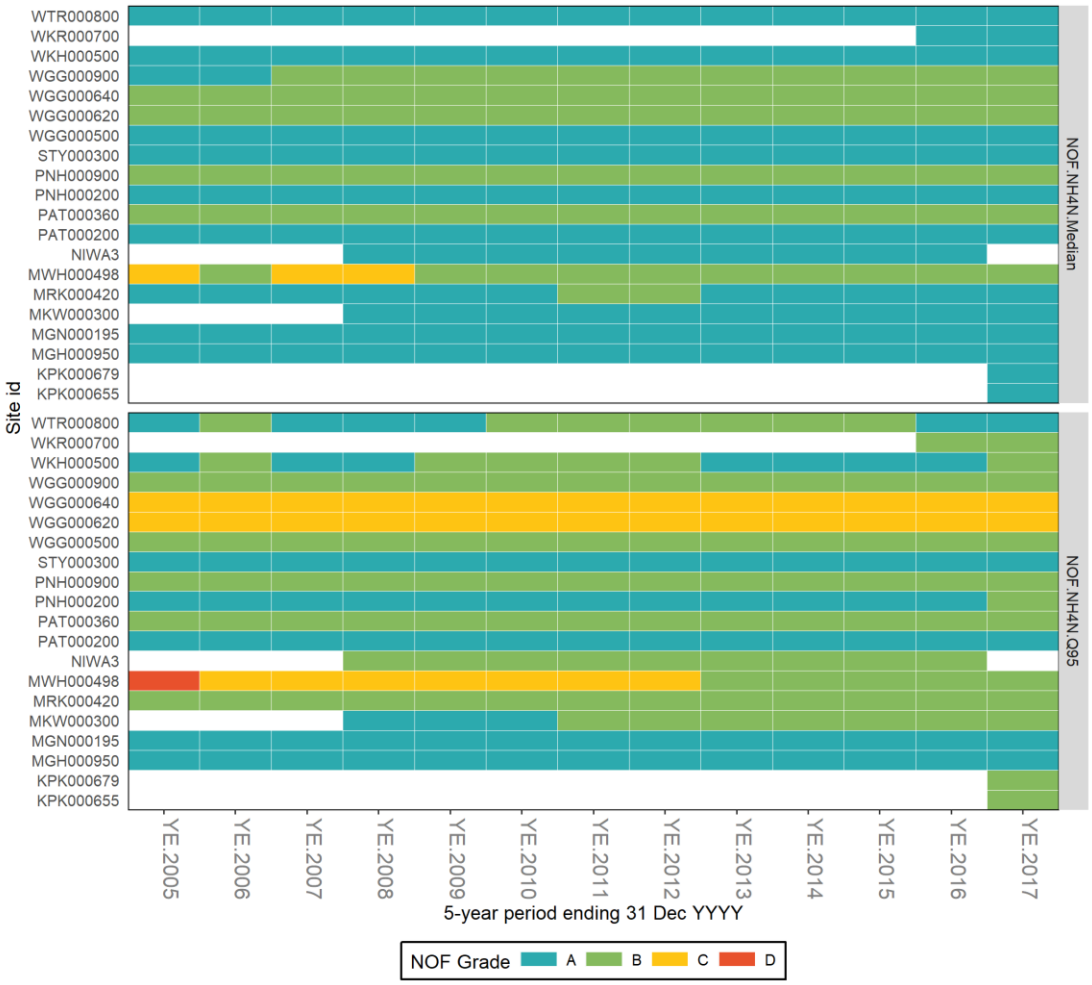

Figure 9: Variation in NOF Ammonia attribute grades for sites over time.

Figure 10: Variation in NOF DRP attribute grades for sites over time.

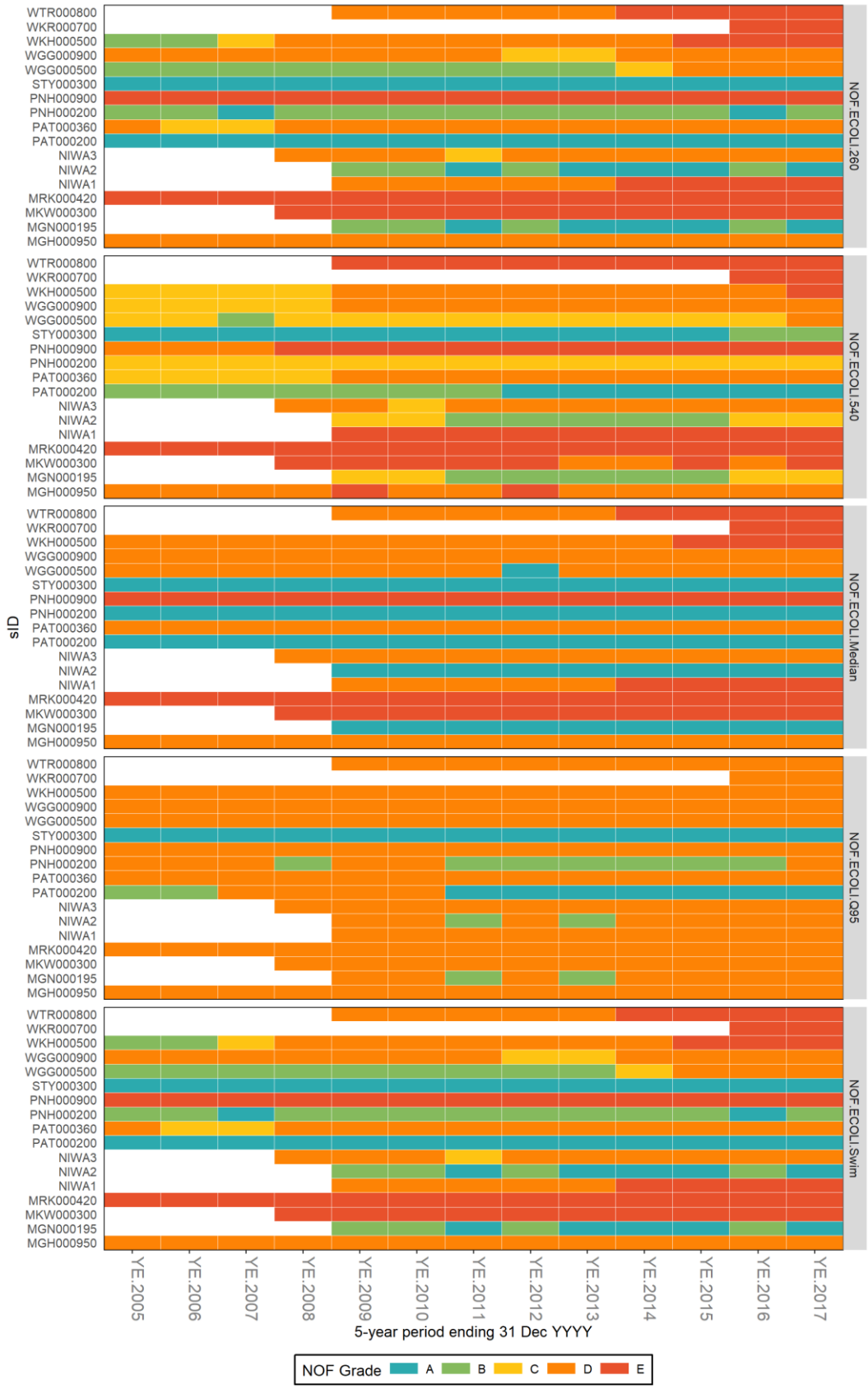

Figure 11: Variation in NOF E. coli attribute grades for sites over time.

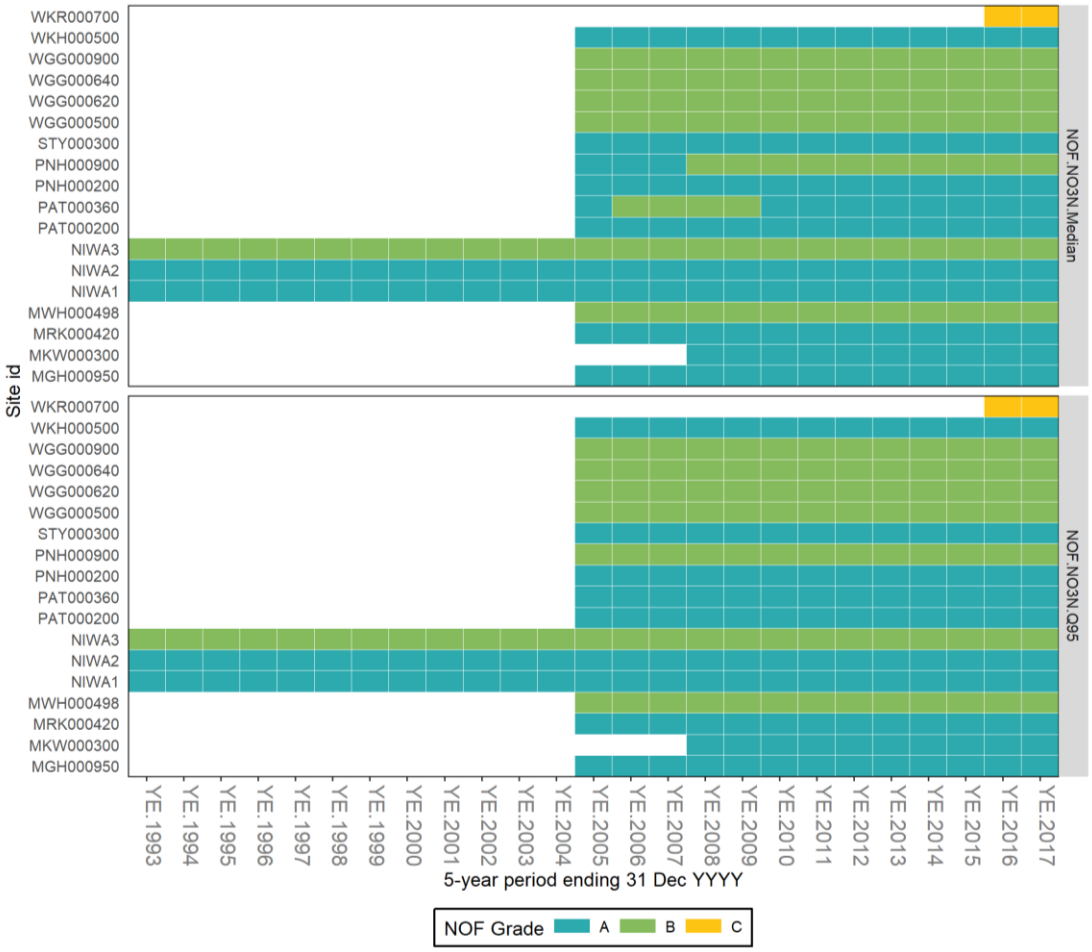

Figure 12: Variation in NOF Nitrate attribute grades for sites over time.

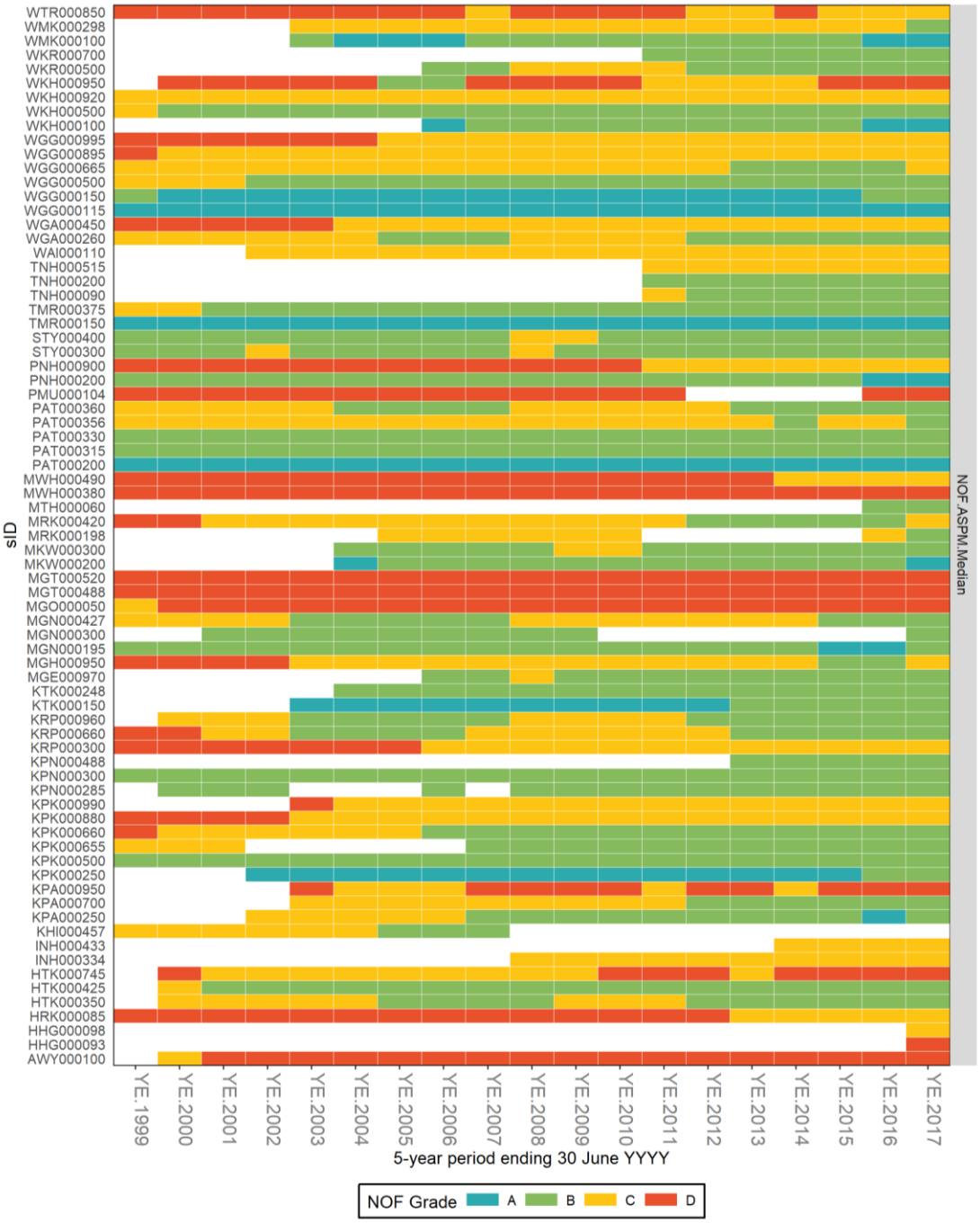

Figure 13: Variation in NOF ASPM attribute grades for sites over time.

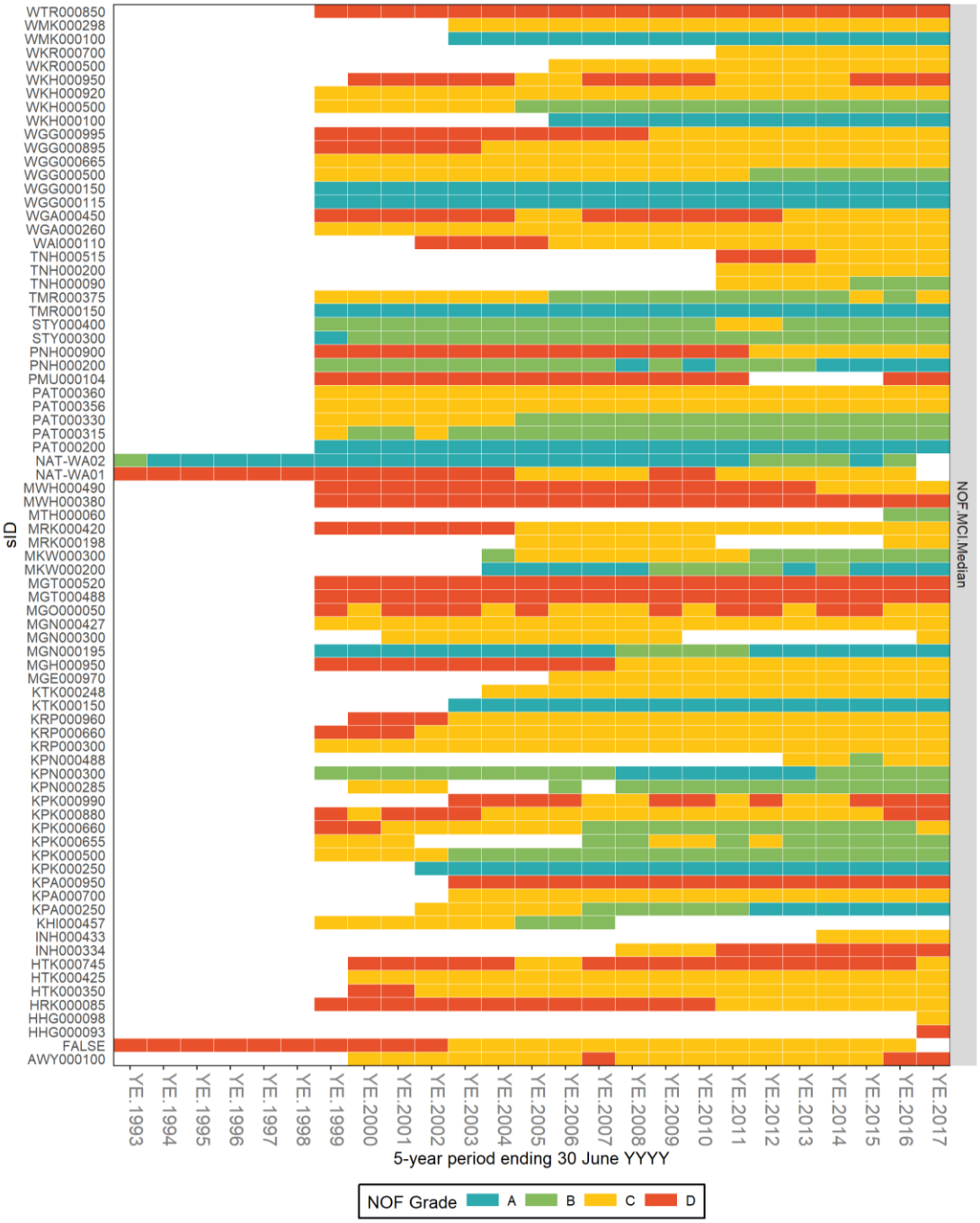

Figure 14: Variation in NOF MCI attribute grades for sites over time.

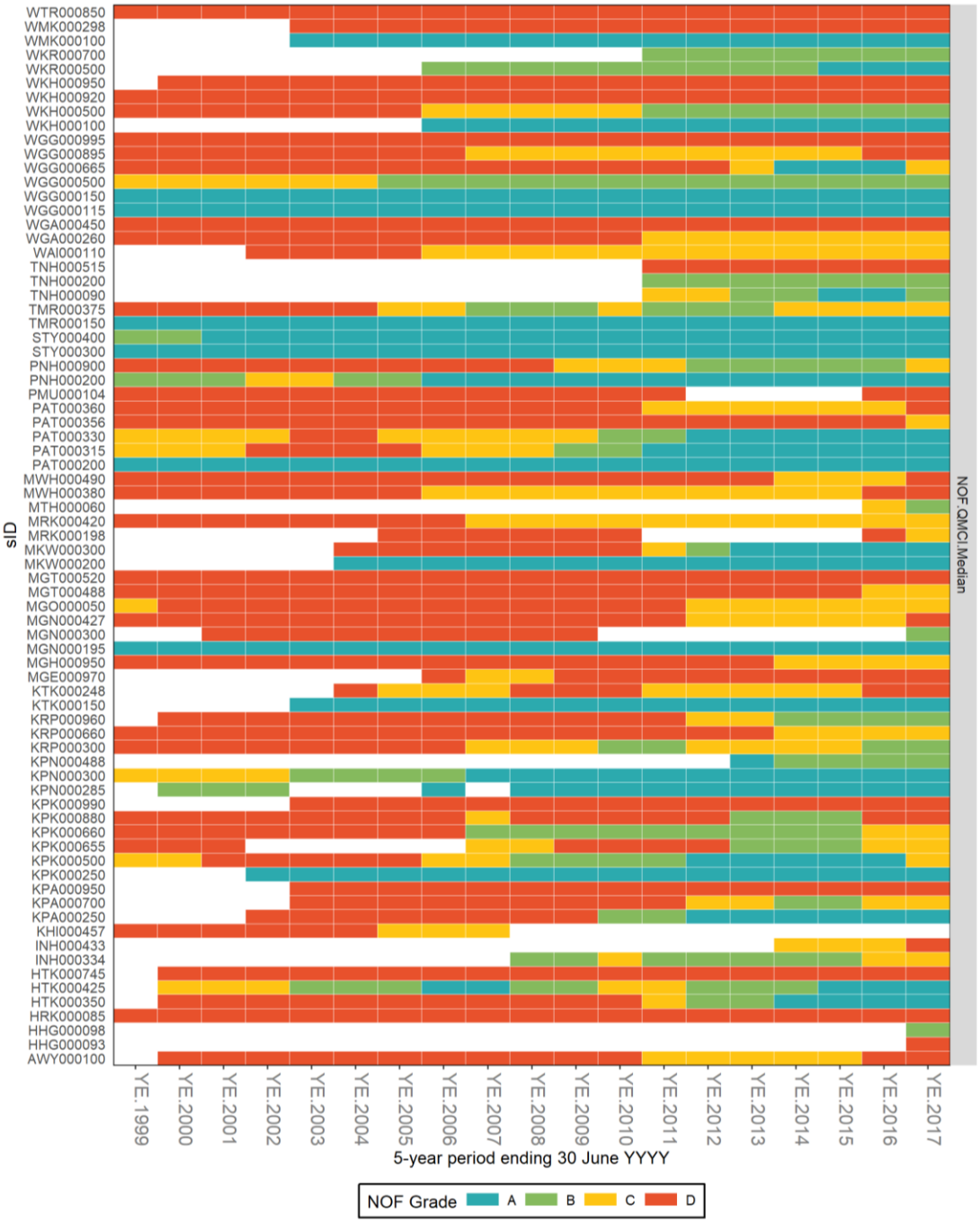

Figure 15: Variation in NOF QMCI attribute grades for sites over time. (Note this comparison is made using TRC SQMCI observations).

Appendix C: Continuous spatial model predictions

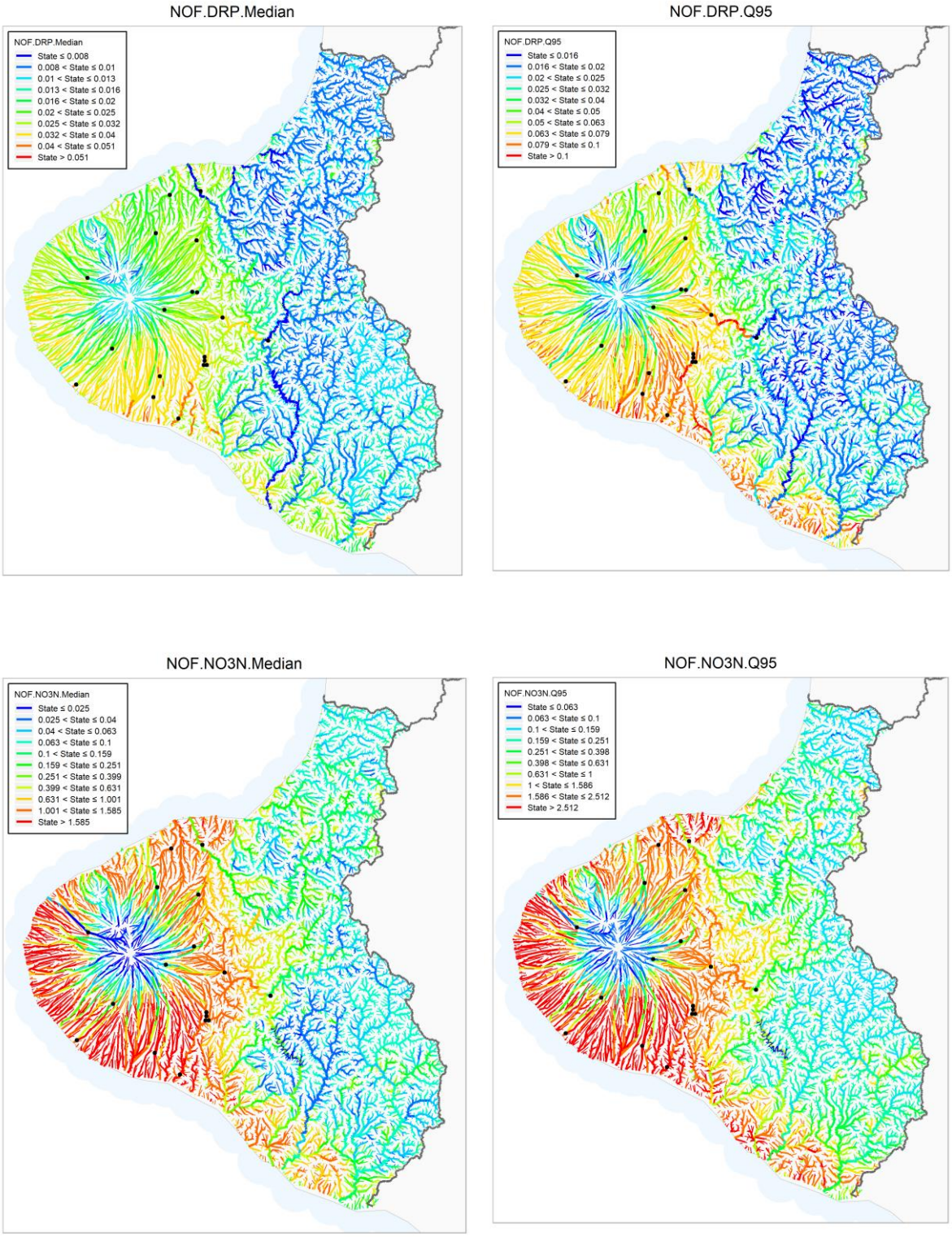

Figure 16: (a) Predicted water quality compliance statistics for selected water quality variables, for all segments of the regional network.

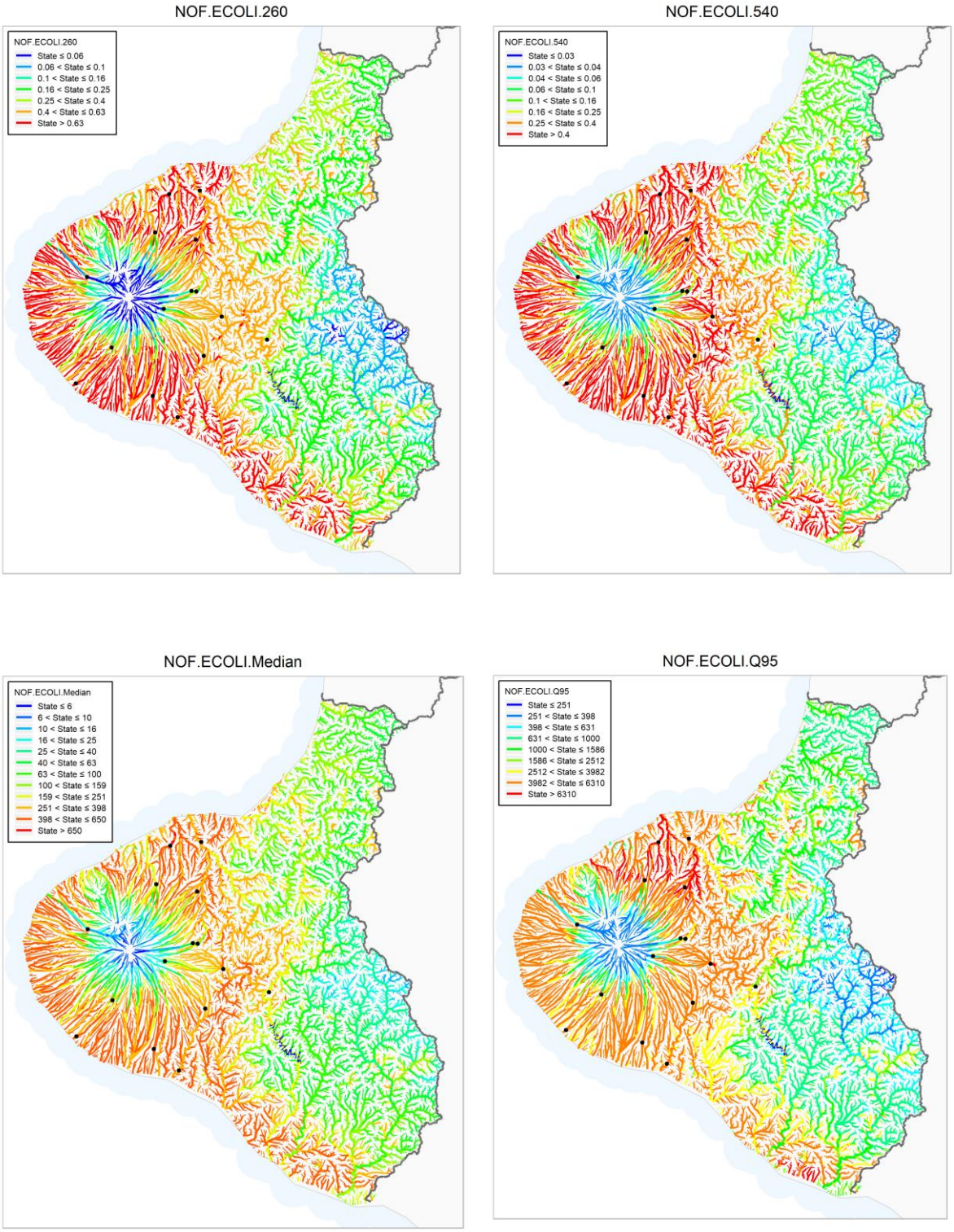

Figure 16 (b): Predicted water quality compliance statistics for selected water quality variables, for all segments of the regional network.

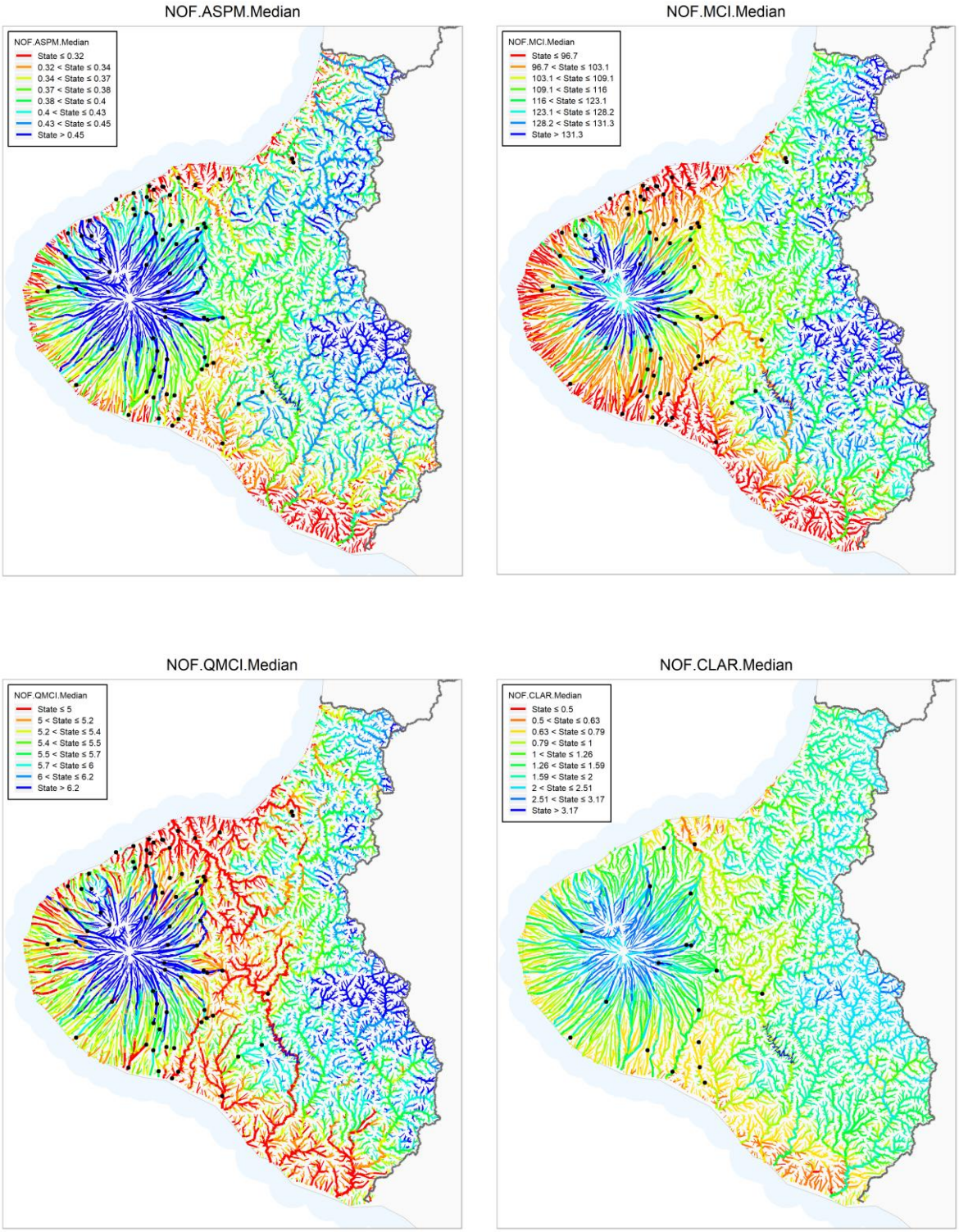

Figure 16 (c): Predicted water quality compliance statistics for selected water quality variables, for all segments of the regional network.

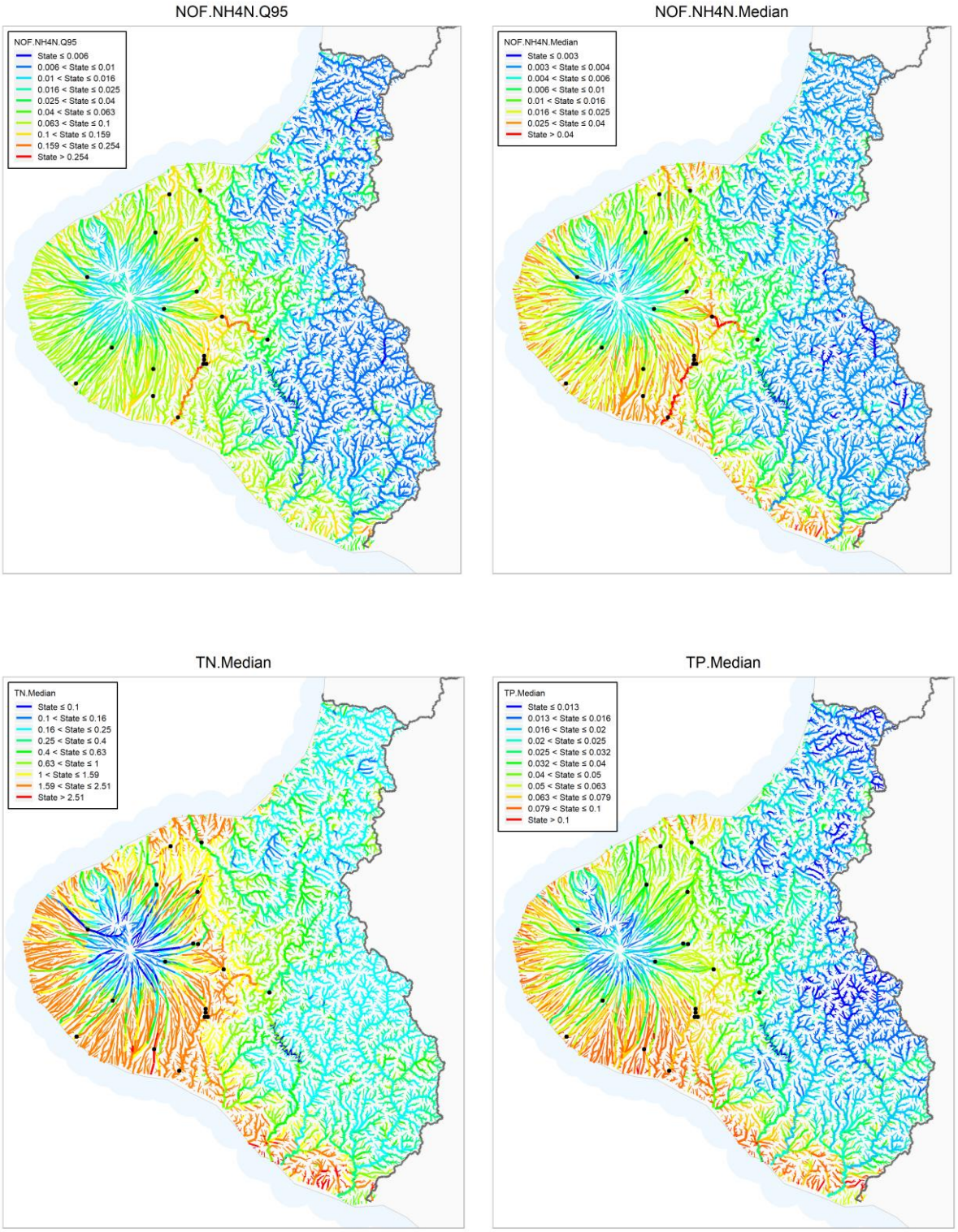

Figure 16 (d): Predicted water quality compliance statistics for selected water quality variables, for all segments of the regional network.