Relationships between instream concentrations and river nutrient loads

March 2020

Prepared By:

Caroline Fraser and Ton Snelder

For any information regarding this report please contact:

Ton Snelder

Phone: 027 575 8888

Email: ton@lwp.nz

LWP Ltd

PO Box 70

Lyttelton 8092

New Zealand

LWP Client Report Number:	LWP Client Report 2019-09
Report Date:	March 2020
LWP Project:	2019-09

Quality Assurance Statement

[Click here and type text]

Version	**Reviewed By**	
	[Name]	[Signature]

Table of Contents

TABLE OF FIGURES

TABLE OF TABLES

Executive Summary

The National Policy Statement for Freshwater Management (NPS-FM) (Ministry for Environment, 2017) directs regional councils to develop regional plans for managing freshwater resources. Plans must contain freshwater objectives, policies and limits. Objectives must express numerically (where practicable) the desired environmental state of water bodies[1]. Polices must describe how the objectives will be achieved and limits must describe restrictions to resource-using activities that will allow the objectives to be achieved.

Water quality objectives for rivers are generally associated directly or indirectly with contaminant concentration criteria. Therefore, the development of appropriate objectives and limits involves catchment analysis to determine the relationship between contaminant discharging activities and instream concentrations. However, concentrations are a complex outcome of mixing of multiple sources of contaminants whose individual contributions are temporally variable. It is therefore generally more tractable to base catchment analyses on loads (i.e., a mass over a period of time such as kg year^{-1}). A challenge then remains of how to relate concentration targets to instream catchment loads.

The purpose of the current project was to develop functional relationships between instream loads of the nutrients nitrogen and phosphorus and their corresponding instream concentrations. The objective was to define these relationships in such a way that they have national coverage and can be applied to any site for which an estimate of either instream load or concentration exists. The modelled relationships were to apply to nutrient variables that are relevant to NPS-FM objectives: total nitrogen (TN), total phosphorus (TP), nitrate-nitrogen (NO_3N) and dissolved reactive phosphorus (DRP). The models should also be invertible such that river loads can be related to concentrations and vice-versa, and model uncertainties should be able to be estimated.

We reviewed a range of alternative modelling approaches that have been employed to represent this linkage, ranging from complex physically based models through to simple statistical models. We found that the most complex models were too resource intensive for most applications, and even existing studies done within New Zealand were too limited to allow for generalisation to the national scale. Simple national and regional scale approaches to load-concentration relationships are required to support catchment management, which can improve on the pragmatic assumptions that have been made to date in data-scarce catchments. The approach of Oehler and Elliot (2011) and applied in the CLUES model is an approach to defining load-concentration relationships (for TN and TP). The approach of Oehler and Elliot (2011) was extended by this study to increase the range of nutrient variables covered and to increase the representation of New Zealand's streams and rivers.

The approach of Oehler and Elliot (2011) defines for any point on a river a value R, which is the ratio of median concentration to load (expressed as load normalised by mean flow). Because nutrient concentration criteria are not only defined by median concentrations, but also sometimes by means or 95th percentiles, we extended R to cover three concentration statistics: R_{median}, R_{mean} and R_{p95}.

We collated flow and concentration data for ~600 sites and four nutrient variables (TN, TP, DRP and NO3N) and calculated loads, concentration statistics (mean, median and 95th percentiles) and subsequently R (R_{median}, R_{mean} and R_{p95}). We used these observed values of R to develop three alternative modelling approaches to predict R at any location across the

[1] Policy CA2 NPS-FM

country. Therefore, we present 3 (model approaches) x 4 (nutrient variables) x 3 ($R_{statistic}$) = 36 models within this report.

Our alternative models for estimating R represent a range of complexity and input data requirements. The simplest model, the 'global model' is simply based on the median observed values of R. The 'catchment characteristics model' is derived using a random forest model, with predictors describing catchment characteristics that have national coverage. The predicted values of R can be read from a lookup table for any river segment in New Zealand. Finally, the 'distributional characteristics model' relies on inputs based on paired observations of concentrations and flows at a site. As the input requirements increased, so did the confidence in the estimated values of R, which is reflected in reducing prediction uncertainty for the estimated values of R.

We provide a worked example that is linked to an NPS-FM objective for periphyton biomass. The example demonstrates how each of the model approaches can be used predict R and how this is then used to determine catchment load targets that are constituent with the instream concentration criteria associated with the periphyton biomass objective.

1 Introduction

Under the National Policy Statement – Freshwater Management (NPS-FM) regional councils are required to set numeric objectives and limits in regional plans. Many objectives are either defined by, or linked to, concentrations of nitrogen and phosphorus (hereafter; nutrients). For example, objectives for toxicity must be defined in terms of nitrate concentrations and objectives for river periphyton biomass must be linked to associated dissolved inorganic nitrogen (DIN) and dissolved reactive phosphorus (DRP) concentrations (a mass per volume of water such as mg L^{-1}).

Limits are the management actions that are applied to resource-using activities in a catchment in order to achieve the objectives. For objectives associated with nutrients, appropriate actions include restricting the intensity of land use on individual land parcels and restricting the amount of nutrient that can be discharged from point sources within a catchment. Part of the development of appropriate objectives and limits involves catchment analysis to determine the relationship between contaminant discharging activities and instream concentrations. However, concentrations are a complex outcome of mixing of multiple sources of contaminants whose individual contributions are temporally variable. It is therefore generally more tractable to base catchment analyses on loads (i.e., a mass over a period of time such as kg $year^{-1}$). Basing catchment analyses on loads is also consistent with the use of budgeting models such as OVERSEER, which estimate the contributions of nitrogen and phosphorus from individual land parcels in terms of annual loads.

Although loads are a more appropriate characteristic than concentrations for catchment analyses, loads at a site ultimately need to be reconciled with objectives that are expressed in terms of concentration criteria. The concentrations that define, or are linked to, freshwater objectives are generally a statistic representing something about the distribution of concentrations in the receiving environment. Most commonly the statistic represents the central tendency of the distribution (i.e., the median or the mean) but it can also represent the extremes (e.g., the 95^{th} percentile concentration). There is therefore a need to be able to convert a contaminant concentration at a site to an associated load.

Sometimes it is assumed that a target load can be calculated as the concentration criteria multiplied by the mean flow, or alternatively the concentration is the load divided by the mean flow (e.g., Norton and Kelly; Roygard and McArthur, 2008). However, the dynamics of catchment contaminant concentration dynamics means that this is generally an approximation, which may not be appropriate (Oehler and Elliott, 2011). Catchment analyses to support objective and limit setting would therefore benefit from models that convert between loads and concentrations.

The aim of this study was to develop model-based tools to convert instream nutrient concentrations to instream nutrient loads, and vice-versa. The purpose of these tools is to enable the estimation of a nutrient load that is consistent with a nutrient concentration criterion or to establish the nutrient concentration that results from a given load at any site in New Zealand.

2 Background

2.1 Catchment water quality modelling approaches

Land-water-nutrient systems are complex and highly variable in space and time (Anastasiadas *et al.*, 2013). There are therefore considerable challenges in evaluating the management actions that are necessary to achieve targets. Generally, modelling is used to explore the outcomes of alternative land and water management scenarios. A series of scenarios is generally used to identify a set of interventions that will achieve desired outcomes. Most models that are used to represent this problem, can broadly be classed as source-pathway-receptor models (Holdgate, 1980).

Figure 1: The source-pathway-receptor framework, with examples relevant to catchment water quality management

The source describes the how and where the nutrients enter the system, e.g., point sources (such as wastewater treatment plant or industrial discharges), or diffuse sources (e.g., losses from agricultural or forestry land). The pathway describes how nutrients are transported from the source to the receptors. The transportation is generally via hydrological flow paths, such as infiltration, overland runoff, groundwater movement and conveyance by rivers. The pathway also accounts for transformation, uptake and exchanges of nutrients as they travel from source to receptor. The receptors are the agents that respond to the environmental state that results from the transported nutrients. In a river water quality example, the receptor is generally described in terms of a characteristic of the aquatic ecosystem (e.g., the community of fish or macroinvertebrate species or the trophic state) or a characteristic of relevance to humans (e.g., suitability of water for drinking or recreation). In most examples of land-water-nutrient models that are concerned with rivers as the receptor, the environmental state is expressed as an instream nutrient concentration and the objective is expressed in terms of the receptor's response to that concentration. For example, objectives for maximal periphyton biomass express the trophic response of rives to instream nutrient concentrations.

Variants of source-pathway-receptor models are related to the representation of the model components (e.g., each of the sources, pathways and receptors) and their linkages (i.e., transfers, interactions, and coupling), as well as the spatial and temporal discretisation. In the literature review that follows, we have explored alternative models and suites of models used to relate source loads to instream nutrient concentrations (and vice-versa), and their suitability under different application requirements (e.g., data availability, acceptable uncertainties and assumptions).

2.1.1 Statistical Models

Statistical modelling to directly estimate river water quality (e.g., through regression or machine learning techniques) have been successfully used to evaluate water quality in unmonitored locations (e.g., Snelder *et al.*, 2017; Unwin *et al.*, 2010). In order to use this class of models to make estimates of changes in water quality, it is necessary to make assumptions that trade space-for-time, that is, it is assumed that that the spatial relationships between water quality and drivers is similar to what would be seen between water quality and drivers over time at a single location (Singh *et al.*, 2011). The predictors in these models are generally not refined enough to be able to represent changes in land use (including land use intensity), let alone changes in land management or mitigations (hereafter referred to as management actions). Further, the appropriateness of a trading-space-for-time substitution, is highly dependent on large training datasets representative of the domain of application. Hence, the suitability of statistical modelling in evaluating management actions as part of scenario analysis is limited, and statistical modelling approaches are not discussed further in this review.

2.1.2 Distributed physically-based models

The most detailed approach to evaluating the complex relationships between source loads and instream concentration is to use distributed, daily (or sub-daily) time stepping, physically based models (herein abbreviated to DPBM). Examples of this class of model include INCA (Wade *et al.*, 2001), SWAT (Gassman *et al.*, 2007) and MIKE SHE (Graham and Butts, 2005). The theoretical advantage for these models is that all phases/components of the model are represented at a high frequency (e.g., daily, or sub-daily), there is a high spatial discretisation and the model components are represented based on assumptions about the physical processes that they represent. These models can represent processes at high spatial resolution and they are mass conservative; hence they can be used to predict concentrations throughout the model domain. However, there are significant challenges with the calibration, validation and uncertainty assessment of these type of models (Beven, 1993).

DPBMs are generally used in case study specific locations, where large amounts of data are available for parameterisation and validation, and when considerable financial resources and technical expertise can be invested for the modelling process. This class of models is generally not suitable for regional analysis, or for studies in data-poor catchments. Further, because their application even within one catchment is expensive (in terms of time, money and data), the ability to extrapolate outcomes from a small number of specific case studies that can afford this level of modelling, is limited. Applications of MIKE-SHE and SWAT in the New Zealand context have encountered difficulties with calibration and validation of the water quality predictions (Durney *et al.*, 2016; Fenemor, 2013).

Where data and time do not allow the development of DPBMs, alternatives are employed that compromise in terms of spatial and temporal resolution and process representation, for all, or components of water quality catchment models. Anastasiadas et al. (2013) provide a useful summary of water quality models (for predicting loads) used within the New Zealand context, categorising them based on the processes and time and space scales that they represent. In New Zealand, agricultural nutrient sources are commonly represented using the OVERSEER model. Overseer provides farm-scale nutrient budgeting and loss estimation of nitrogen and phosphorus on an annual steady state basis (Roberts and Watkins, 2014; Shepherd and Wheeler, 2013; Shepherd *et al.*, 2013; Wheeler *et al.*, 2014). OVERSEER has been developed within New Zealand, specifically to represent local climate, soil, and farming systems and practices. Farmers can use OVERSEER to help with effective on-farm nutrient management

(to optimise profitability and production) and many regional councils are now moving towards requiring farms to develop nutrient budgets using OVERSEER for regulatory purposes and/or using OVERSEER to support plan development (Baker-Galloway, 2013; Upton, 2018). As such, OVERSEER is increasingly being used to estimate farm scale nutrient leaching for use within many water quality management applications.

In a recent application with the Ruamahanga catchment (Greater Wellington), OVERSEER was used to estimate steady state nutrient losses (as average loads per year) and these were coupled with daily time stepping groundwater and surface water models (Blyth, 2018; Blyth *et al.*, 2018). This was achieved by determining effective concentrations for baseflows and runoff flows (based on surface and groundwater flows from hydrological models and the OVERSEER annual loads, to ensure mass balance). This allowed the annual steady state load to be distributed across time, and hence produce daily timeseries of instream nutrient concentrations. This approach assumes that variability in concentrations with time is only dependent on hydrology, and is not due to intra-annual variability in on-farm nutrient inputs/outputs. The source concentrations were further adjusted during the calibration process. The predicted instream nutrient concentrations were compared with nutrient targets to assess whether water quality objectives would be achieved under different scenarios. While the model was found to perform well in terms of predictions of instream concentrations following calibration (Blyth *et al.*, 2018), the collaborative modelling process (surface water, groundwater, nutrients etc) was a multi-year project and resource and data intensive. We were unable to assess the accuracy of the predicted relationships between source load reduction and corresponding reductions in instream concentrations based on the outputs as presented in (Blyth, 2018; Blyth *et al.*, 2018), as this was not part of their modelling objective and was not documented.

2.1.3 Hybrid models

Another class of catchment models operate on an average annual load basis. In this type of approach explicit functions are required to relate the average annual loads to the instream concentration statistic that is of interest. The CLUES model is a widely used spatially distributed model for estimating annual steady state loads (Elliott *et al.*, 2016), and has been used across New Zealand to evaluate catchment loads of TN and TP (e.g., Palliser *et al.*, 2015; Semadeni-Davies *et al.*, 2015; Semadeni-Davies and Sunil Kachhara, 2017). CLUES comprises a series of component models, OVERSEER and SPASMO for modelling nutrient *sources* and SPARROW for transmission and instream processes (*pathway*). The linked models generate load estimates. The loads are then converted into instream median concentrations (indicators for *receptors*) using explicit functions derived from regressions on catchment characteristics (described in Oehler and Elliott, 2011).

2.1.4 Pragmatic solutions

In many examples in New Zealand catchment management, pragmatic assumptions are made regarding the changes in load required to meet target nutrient concentrations. For example, a target load for the catchment may be estimated by taking the target concentration (typically expressed as a median concentration) and multiplying by a mean average flow .e.g., (Norton and Kelly; Roygard and McArthur, 2008; Scott, 2013). Comparing this target load to the "observed" load then suggests a target nutrient load reduction. If attenuation and other nutrient losses during transmission are represented by a linear transfer function, it follows that the required percentage reduction in instream load can be achieved by an equivalent reduction in source loads. The required source load reductions must take into account the load

contributions from non-manageable sources such as loads from naturally vegetated areas and atmospheric deposition, which cannot be reduced.

While practical, the above approach to calculating a target loads is likely to produce erroneous results because the relationship between concentration and load is complicated due to catchment contaminant concentration dynamics (Oehler and Elliott, 2011). For example, estimating the load required to achieve a median concentration criteria as the criteria multiplied by the mean flow will often under-estimate the required load and lead to an overestimate in the required source load reduction.

2.1.5 Previous approaches to spatial modelling of load-concentration relationships

There is limited literature concerning approaches to develop explicit functions to describe the relationship between river loads and instream concentrations; studies either adopt a catchment by catchment based approach (i.e., distributed integrated catchment modelling) to directly estimate instream concentrations, or alternatively, particularly in data-sparse or low pressure catchments, work with the very simplified representations. The most significant piece of research, particularly in a New Zealand context, examining these relationships is the work by Oehler and Elliott, (2011), as applied in the CLUES model. In the following section we will explore the detail of that study.

The regression relationships derived by Oehler and Elliott, (2011), describe a scalar (R) to convert yearly flow-weighted concentrations (C_{fw} = mean annual load/mean annual flow) to median concentration:

$$C_{median} = C_{fw} \times R \qquad \text{(Equation 1)}$$

Based on 72 sites with 20 years of monthly monitoring data from the National River Water Quality monitoring Network (NRWQN) (NIWA, 2009), Oehler and Elliott (2011), showed that R is spatially variable and varies with nutrient species (Figure 2). Oehler and Elliott (2011) developed models describing the spatial variation in R for TN and TP across New Zealand, based on a national database of catchment characteristics (e.g., REC, LCDB2, FWENZ). The regression relationships were modelled using a machine learning approach (Boosted Regression Trees; BRT). For comparative purposes, Oehler and Elliot (2011) also developed BRT models of C_{median} with loads as one of the predictors ("direct" models). The "direct" models generally yielded a poorer model fit compared to the estimates to C_{median} following equation (1) with the BRT models of R. The authors also noted that the "direct" models were less useful for generalisation and application to scenario analysis, as they were dependent on land cover.

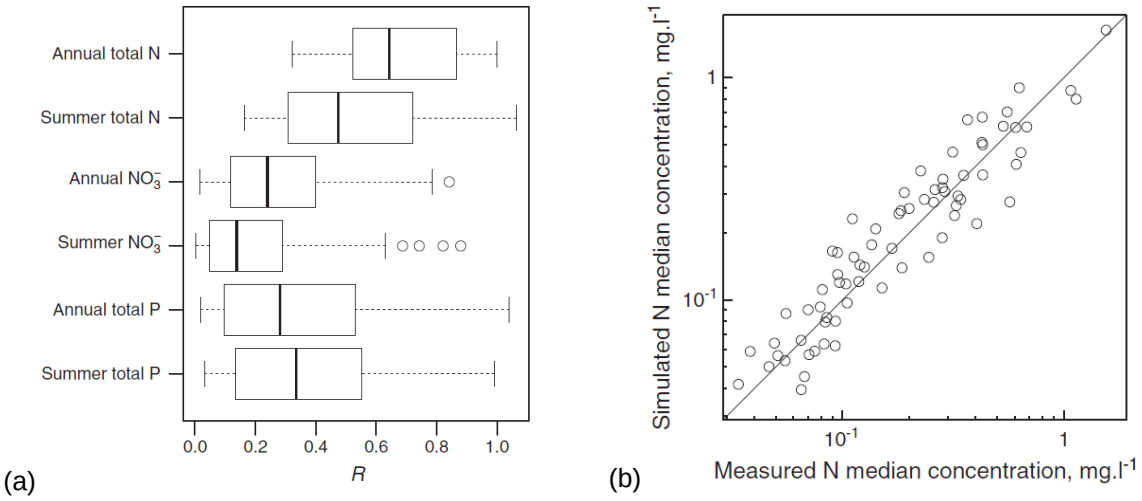

Figure 2: (a) Boxplot of measured R. The boxes show quartiles and the whiskers show 1.5 times the inter-quartile range, and circles are outlier candidates. (b) Cross-validation performance of the Ratio Model for summer TN C_{median}. (from Oehler and Elliott, 2011)

Oehler and Elliott, (2011) found that the scalar, R, was mostly less than one (i.e., flow weighted concentrations were greater than median concentrations, Figure 2). Hydrological predictions (e.g., variables describing mean flows, flood flows and yields) explained around 1/3 of the variability in R. Catchment characteristics relating to sediment sources and transport were also important. Model performance (in terms of estimates of concentration) was good for TN (Figure 2**Error! Reference source not found.**, $r^2 = 0.78$), but slightly poorer for TP ($r^2 = 0.56$). Characteristics relating to land use were not included in the model, which simplifies the use of the relationships in scenario analysis, if it is assumed that R does not change under land management/land use change scenarios. This approach provides a simple approximation of the relationship between loads and concentrations, particularly for catchments with limited water quality monitoring data.

2.1.6 Summary

The purpose of the current project is to develop functional relationships between river loads and instream concentrations. The objective is to define these relationships in such a way that they have national coverage and can be applied to any site for which an estimate of either instream loads or concentrations exists. In addition, the relationships should be structured so that they can be used to investigate various catchment scenarios (e.g., the implementation of mitigations). The model's relationships should cover constituent water quality species relevant to NPS-FM objectives: TN, TP, DRP and NO_3N. The models should also be invertible such that river loads can be related to concentrations and vice-versa, and model uncertainties should be able to be estimated.

With these objectives in mind, we concluded that investigating approaches using DBPMs was not appropriate for our application given that (1) these models are extremely resource intensive even for a single site/catchment application, let alone the number that would be required to be able to provide sufficient information to generalise the results nationally and (2) the performance of this class of models for water quality is often quite poor, primarily due to challenges associated with calibration. We also consider that collating outputs from various sources (e.g., detailed modelling studies from across the country) to examine load-concentration relationships would also be hampered by the differentiating between differences

related to difference in modelling approaches and time periods, and those due to differences between the load-concentration relationships.

Simple national and regional scale approaches to load-concentration relationships are required to support catchment management, which can improve on the pragmatic assumptions that have been made to date in data-scarce catchments. The approach of Oehler and Elliott (2011) and applied in the CLUES model is a more rigorous approach to defining load-concentration relationships, however it can be improved in terms of constituent species covered, and the representation of New Zealand's stream and rivers.

2.2 Estimating loads given a concentration target

The method we have employed in this study follows a similar approach to Oehler and Elliott, (2011), but builds on that earlier work by: (1) incorporating more data; (2) incorporating updated knowledge about load estimation procedures; (3) including models for both NO_3N and DRP; and (4) calculating R values that also relate to the mean and 95th percentile concentrations.

In practice, we have a concentration target specified as a median, mean or 95th percentile. We want to know the load that corresponds to the target. The underlying assumption is that there is a relationship between a concentration statistic and load at a site that is constant such that a change in the load will result in a change in the concertation as presented in equation (1) and generalised below in equation 1a

$$R_{Statistic} = \frac{C_{Statistic}}{C_{FW}} \qquad \text{Equation 1a}$$

where $C_{FW} = {}^{L}/_{Q}$ is the flow weighted concentration, L is the mean annual load, Q is the mean flow, and $C_{Statistic}$ can be the concentration target defined by a median, mean or 95th percentile concentration. Then at a site with a defined concentration target, the load can be estimated if the associated value $R_{Statistic}$ is known as follows:

$$Load = \frac{C_{Statistic} \times Q}{R_{Statistic}} \qquad \text{Equation 2}$$

Conversely, at a site with a modelled load, the associated instream concentrations can be estimated as follows:

$$C_{Statistic} = \frac{Load \times R_{Statistic}}{Q} \qquad \text{Equation 3}$$

Our objective was to estimate the value of $R_{Statistic}$ at any location. This study has developed three methods to estimate the value of $R_{Statistic}$.which we refer to as the global model, the catchment characteristic model and the distributional characteristic model.

The following sections detail the data and methods used to determine the observed $R_{Statistic}$values at any site and the development of the three types of models that can be used to predict $R_{Statistic}$.

3 Data

3.1 Water quality data

This study was based on a national water quality dataset (1000+ sites) compiled for recent national environmental monitoring (Larned et al., 2018). The water quality data consisted of

measurements from river monitoring sites associated with regional council state of environment (SOE) monitoring networks and the National River Water Quality Network (NRWQN). Figure 3 provides a summary of the start years, end year and the total number of observation years for the complete dataset, for the variables included in this study. The total number of sites used in our study is slightly smaller than represented by Figure 3 due to filtering to meet minimum data requirements. However, the data represents a much larger number of sites for developing R values than that used by Oehler and Elliot (2011), whose relationships were derived based on the 72 NRWQN sites.

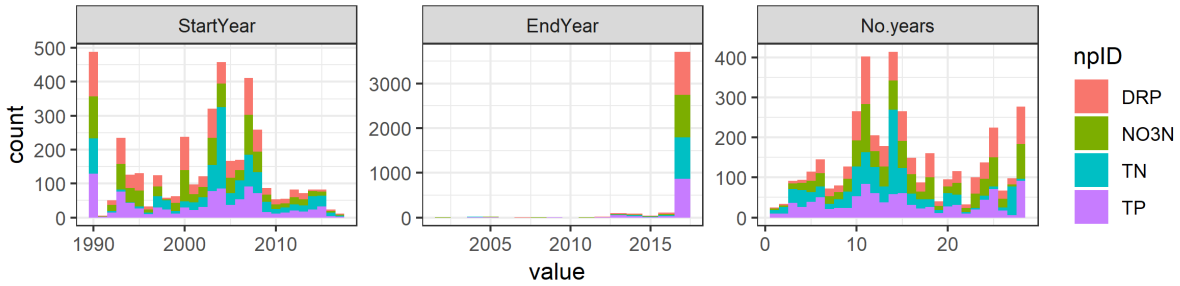

Figure 3: Summary of start years, end years and total number of observation years for the complete water quality data set

3.2 Flow data

The calculation of loads requires that water quality observations are associated with the flow at the time of sampling, and preferably that there is also a continuous timeseries of daily flow at the site. In this study, flow estimates for each monitoring site were based on measured or modelled daily mean flow. For monitoring sites with flow recorders on the same reach, daily mean flows were calculated from measured flow (approximately 32% of sites had observed flow data). However, most river monitoring sites are not associated with a flow recorder, and daily mean flows for these sites were estimated by hydrological modelling. We used predicted flows from the Larned et al. (2018), which were derived from the TopNet hydrological model, corrected using flow-duration curves, which were in turn estimated with random forest models (Booker and Snelder, 2012a; Booker and Woods, 2014). TopNet is a spatially distributed time-stepping model that combines water-balance models with a kinematic wave channel-routing algorithm (McMillan *et al.*, 2013).

3.3 Spatial framework

The spatial framework for the analysis was a GIS-based digital drainage network comprising rivers and catchment boundaries that is the basis for the River Environment Classification (REC; Snelder and Biggs, 2002). We used the digital drainage network from version 2 of the REC (REC2). The digital network was derived from 1:50,000 scale contour maps; it represents New Zealand's rivers as 590,000 segments (delineated by upstream and downstream confluences), each of which is associated with a sub-catchment. All observation sites were associated with a "nzsegment" value, which is a unique identifier for the segments of the REC2 digital river network.

3.4 Site metadata

Following the approach of Oelher and Elliot (2011) we postulated that variation in the relationship between the load and the concentration across sites might be at least partly explained by differences in catchment characteristics. To explore this hypothesis, we

compiled a dataset of site attributes for all monitoring sites in the study. We categorised these as “catchment characteristics” and “land use intensity data”. These data represent updates to the explanatory variables used by Oehler and Elliot (2011). Firstly, we use updated land cover data (we use LCDB3 rather than LCDB2) and secondly we incorporated land use intensity data based on animal stock density data, which were not available until recently.

3.4.1 Catchment characteristics

The digital drainage network is linked to a database describing a wide range of descriptors of the individual network segments (Wild *et al.*, 2005). We used several catchment characteristics as predictors in our models (Table 1). 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, 2012b).

Table 1: Table of catchment characteristic predictors 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	$Year^{-1}$
	MAHF30	Mean annual 30-day high flow divided by the mean flow	Unitless
	Lcv	L-moments coefficient of variation	Unitless
	Reversal	Number of negative and positive changes in water conditions from one day to the next	$Year^{-1}$
	MALF30	Mean annual 30-day low flow divided by the mean flow	Unitless
	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*
	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

3.4.2 Land use intensity data

Previous spatial modelling has used the proportion of catchment occupied by pasture as a predictor of water quality (e.g., Snelder *et al.*, 2018a; Unwin *et al.*, 2010). This is justified by the strong relationship between the proportion of catchment occupied by pasture and observed variation in water quality between monitoring sites (Julian *et al.*, 2017; Larned *et al.*, 2016). The area of pasture has not varied substantially in most catchments in New Zealand over the past 30 years (Julian *et al.*, 2017). However, the density and types of livestock on pasture land over much of the country currently vary considerably and have changed in that period (MFE and StatsNZ, 2019). Therefore, predictions of current water quality are likely to be better informed by indicators of land use intensity, (e.g., measuring inputs such as fertilizer, cultivation frequency), carrying rates (i.e., numbers of livestock) or production such as crop yields, milk solids.

Data describing land use inputs (e.g., fertilizer, cultivation frequency) and production (e.g., crop yields, milk solids) at the extent of the entire country are not available in New Zealand. However, data describing stocking rates (i.e., numbers of livestock), derived from the annual agricultural production census (APC), have recently been made available by Statistics New Zealand at a level of resolution that is appropriate to the analyses undertaken by this study. We used the APC data to derive indicators of the intensity of pastoral agriculture by combining spatial data describing the type of land cover (i.e., grazed grassland and plantation forest, from LUCAS; Newsome et al. 2018) with data describing the number and types of livestock. Details of the procedure used to derive these statistics are provided in Appendix A.

The resulting dataset described the stock unit density in total (i.e., all stock types) and by stock type for the catchment of every segment of the river network and the proportion catchment occupied by plantation forest for the latest version of LUCAS.

Table 2. Land use intensity predictor variables used in spatial models.

Land use intensity predictor	Description	Unit
SUTotal	Stock unit density for all stock types in 2017 (i.e., total stock units)	SU ha^{-1}
SUDairy	Proportion of total stock unit density attributable to dairy cows in 2017	Unitless
SUBeef	Proportion of total stock unit density attributable to beef cows in 2017	Unitless
SUSheep	Proportion of total stock unit density attributable to sheep in 2017	Unitless
SUDeer	Proportion of total stock unit density attributable to deer in 2017	Unitless
pForest	Proportion of upstream catchment area in forestry	Unitless

4 Methods

4.1 Overview

The methodology that follows describes the various steps conducted to take the raw input data described in the previous section through to the development of three alternative models of $R_{statistic}$. These steps are demonstrated schematically in Figure 4.

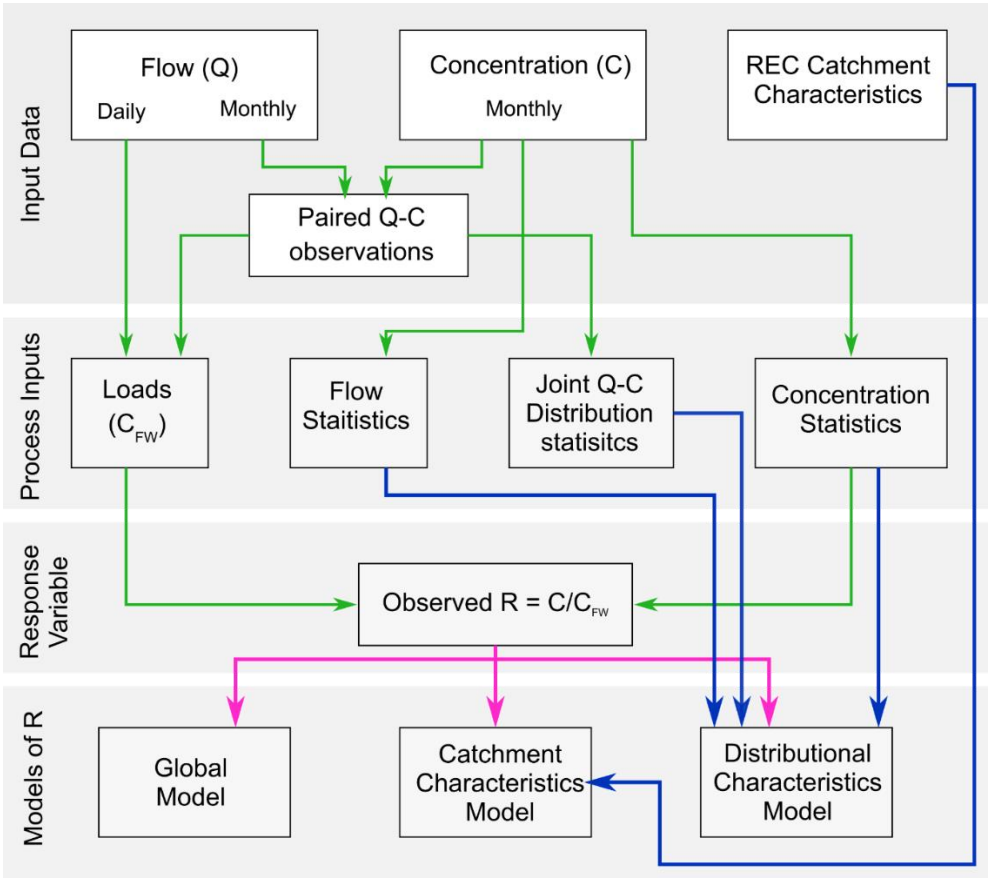

Figure 4: Process diagram indicating the methodological steps required to go from raw input data through to development of alternative models or R. Blue arrows indicate inputs required for prediction in the models of R. Pink arrows indicate inputs only required for the development of the models of R. Green arrows indicate other process steps performed in this study.

4.2 Concentration statistics

In order to evaluate $R_{statistic}$ at any given site, it was first necessary to evaluate the concentration statistic of interest (mean, median and 95th percentile) from the observed water quality data. We followed procedures that are consistent with those used in national environmental reporting of river water quality state (e.g. Larned *et al.*, 2015, 2018). These methods are outlined below.

4.2.1 Time period and filtering rules for analysis of concentration statistics

The statistical robustness with which concentration statistics can be determined depends on the variability in the measurements between sampling occasions and the number of

observations. As a general rule, there are diminishing returns on increasing sample size with respect to confidence for sample sizes greater than 30 (McBride, 2005).

In this study, a period of five years represented a reasonable trade-off for most of the targets because it yielded a sample size that was 30 or more for many sites and nutrient variable combinations, and typically 60 samples based on monthly sampling. We used a five-year period for the analysis of concentration statistics (Jan 2013 – Dec 2017) as it is consistent with the periods used in previous national water-quality state analyses (e.g., Larned et al. 2015, Larned et al. 2019). Because water quality data tends to be seasonal, it is also important that each season is well-represented over the period of record. In New Zealand, monthly monitoring is generally undertaken at SOE sites and seasons are therefore defined by months. We therefore applied a rule that restricted site × nutrient variable combinations in the concentration statistic analyses to those with measurements for at least 90% of the sampling intervals in that period (at least 54 of 60 months). Site × nutrient variable combinations that did not comply with these rules were excluded from the analysis.

4.2.2 Censored values in concentration observations

Censored values were replaced by imputation for the purposes of calculating the concentration 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 it can accommodate multiple censoring limits. 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 where fewer than 24 total observations existed, censored values above the detection limit were replaced with 1.1 times the detection limit. Sites with greater than 15% of observations that were censored were excluded from the study. This restriction reduced the influence on the imputed values on the concentration statistics, and also reduce the influence on censored values in the fitting of rating curves for the load calculations (see section 4.3).

4.2.3 Calculation of concentration statistics and confidence intervals

For each monitoring site and nutrient variable, we calculated the concentration statistic using the mean and percentiles (50th and 95th) derived from the distribution of observed values for the period 2013 to 2017 (inclusive). All percentiles were calculated using the Hazen method.[2]

As water quality concentrations tend to be log-normally distributed, we have calculated the 95% confidence interval of the mean using the Cox method for log-normally distributed data (Olsson, 2005), given by:

$$CI = \bar{C} + \frac{S^2}{2} \pm z\sqrt{\frac{S^2}{n} + \frac{S^4}{2(n-1)}} \qquad \text{(Equation 4)}$$

[2] (http://www.mfe.govt.nz/publications/water/microbiological-quality-jun03/hazen-calculator.html) Note that there are many possible ways to calculate percentiles. The Hazen method produces middle-of-the-road results (McBride 2005).

The confidence interval of the percentiles was evaluated using the R package jmuOutlier, and the function quantileCI, which produces exact confidence intervals on quantiles corresponding to the stated probabilities, based on the binomial test.

4.3 Load calculations

The second component required to calculate $R_{statistic}$ is an estimate of the flow weighted concentration (C_{fw}), which is the load divided by the mean flow. Mean flows were estimated from the data described in section 3.2, preferentially using observed flows, and when these were not available deferring to modelled flows. The following section describes the methodology employed to estimate at site loads.

We used rating curve methods to calculate loads that comprised two steps: (1) the generation of a series of flow and concentration pairs representing 'unit loads' and (2) the summation of the unit loads over time to obtain the total load. In practice step 1 precedes step two but in the explanation that follows, we describe step 2 first.

If flow and concentration observations were available for each day, the load would be the summation of the daily flows multiplied by their corresponding concentrations:

$$L = \frac{K}{N} \sum_{j=1}^{N} C_j Q_j \qquad \text{(Equation 5)}$$

where L: mean annual load expressed as an annual load (kg yr^{-1}), ha, K: units conversion factor (31.6 kg s mg^{-1} yr^{-1}), C_j: contaminant concentration for each day in period of record (mg m^{-3}), Q_j: daily mean flow for each day in period of record (m^3 s^{-1}), and N: number of days in period of record. Censored values are included in the assessment at their face value.

In this summation, the individual products represent unit loads. Because concentration data are generally only available for infrequent days (i.e., generally in this study, monthly observations), unit loads can be only be calculated for these days. However, flow is generally observed continuously, or the distribution of flows can be estimated for locations without continuous flow data, and there are often relationships between concentration and flow, time and/or season. Rating curves exploit these relationships by deriving a relationship between the sampled nutrient concentrations (c_i) and simultaneous observations of flow (q_i). Depending on the approach, relationships between concentration and time and season may be included in the rating curve. This rating curve is then used to generate a series of flow and concentration pairs (i.e., to represent Q_j *and* C_j in equation 5) for each day of the entire sampling period (i.e., step 1 of the calculation method; Cohn *et al.*, 1989). The estimated flow and concentration pairs are then multiplied to estimate unit loads, and these are then summed to estimate mean annual loads (i.e., step 2 of the calculation method; Equation 5).

There are a variety of rating curve calculation methods that combine infrequently observed concentration data and flow information to estimate daily contaminant concentrations at a site. These concentrations can then be used in equation 5 to calculate site contaminant loads. Identifying the most appropriate rating curve calculation method and approximations to use when daily concentration are not available requires careful inspection of the available data for any given nutrient variable and site.

In this study, loads were calculated by (1) identifying the best rating curve method for each site (through manual inspection of all possible rating curves for each site), (2) calculating loads using daily flow time series (either observed or modelled, as described in section 3.2.

We used all available flow-concentration observations at each site in order to characterize the rating curves and made load estimates for a prediction year of 2015 (the middle of the time period that concentration statistics were calculated over). Setting temporal terms to the middle date of the state time period (for those models that use time variable components), allows trends and seasonality to be accounted for.

Sites were excluded from the analysis where they did not meet the following load calculation criteria:

1. Observations in at least 8 years in the most recent 10 years;
2. 60 total observations;
3. At least 80% of all quarters in the most recent 10 years.

Table 3 provides a summary of the total number of sites available for this analysis, and the numbers that met the load (described above) and state (described in section 4.2.1) filtering requirements.

Table 3: Summary of number of sites used in this study following application of filtering rules.

Nutrient Variable	Total number of sites with Q and C data	Sites that meet load filtering rules	Sites that meet state filtering rules	Sites that meet all filtering rules
DRP	1033	782	740	520
NO3N	1034	762	743	641
TN	996	679	678	584
TP	992	673	679	553

We expressed all nutrient loads as flow weighted concentrations, C_{FW} (i.e. mg L^{-1}) by dividing the annual load by the mean flow (m^3s^{-1}) and applying appropriate units conversion. Details of the alternative load calculation methods are provided in the following sections.

4.3.1 Load calculation methods

4.3.1.1 L7 model

Two regression model approaches to defining rating curves of (Cohn *et al.*, 1989, 1992) and (Cohn, 2005) are commonly used to calculated loads. The regression models relate the log of concentration to the sum of three explanatory variables: discharge, time, and season. The L7 model is based on seven fitted parameters given by:

$$ln(\widehat{C}_i) = \beta_1 + \beta_2\left[ln(q_i) - \overline{(ln(q))}\right] + \beta_3\left[ln(q_i) - \overline{(ln(q))}\right]^2 + \beta_4(t_i - \bar{T}) + \beta_5(t_i - \bar{T})^2 + \beta_6 sin(2\pi t_i) + \beta_7 cos(2\pi t_i) \quad \text{(Equation 6)}$$

where, *i* is the index for the concentration observations, $\beta_{1,2,..7}$: regression coefficients, t_i: time in decimal years, $\bar{T}$: mean value of time in decimal years, $\overline{(ln(q))}$ mean of the natural log of discharge on the sampled days, and $\widehat{C}_i$: is the estimated i^{th} concentration.

The coefficients are estimated from the sample data by linear regression, and when the resulting fitted model is significant ($p < 0.05$), it is then used to estimate the concentration on

each day in the sample period, $ln(\widehat{C_j})$. The resulting estimates of $ln(\widehat{C_j})$ are back-transformed (by exponentiation) to concentration units. Because the models are fitted to the log transformed concentrations the back-transformed predictions were corrected for retransformation bias. We used the smearing estimate (Duan, 1983) as a correction factor (S):

$$S = \frac{1}{n}\sum_{i=1}^{n} e^{\hat{\varepsilon}_i} \quad \text{(Equation 7)}$$

where, $\hat{\varepsilon}$ are the residuals of the regression models, and *n* is the number of flow-concentration observations. The smearing estimate assumes that the residuals are homoscedastic and therefore the correction factor is applicable over the full range of the predictions.

The average annual load is then calculated by combining the flow and estimated concentration time series:

$$L = \frac{KS}{N}\sum_{j=1}^{N} \hat{C}_j Q_j \quad \text{(Equation 5a)}$$

If the fitted model is not significant, $\widehat{C_j}$ is replaced by the mean concentration and S is unity.

To provide an estimate of the load at a specific date, (i.e. t^{est} = 1/3/2004) a transformation is performed so that the year components of all dates (t_j) are shifted such that all transformed dates lie within a one-year period centred on the proposed observation date (i.e. *Y*=1/9/2003 to 31/8/2004). For example, flow at time *t*=13/6/2007 would have a new date of *Y* =13/6/2004, and a flow at time *t*=12/11/1998 would have a new date of *Y*=12/11/2003.

$$ln\left(\widehat{C_j^Y}\right) = \beta_1 + \beta_2\left[ln(q_j) - \overline{(ln(q))}\right] + \beta_3\left[ln(q_j) - \overline{(ln(q))}\right]^2 + \beta_4(Y_j - \bar{T}) + \beta_5(Y_j - \bar{T})^2 + \beta_6 sin(2\pi Y_j) + \beta_7 cos(2\pi Y_j) \quad \text{(Equation 6a)}$$

where $\widehat{C_j^Y}$ is the estimated *j*th concentration for the estimation year, and Y_j is the transformed date of the *i*th observation, and all other variables are as per equation 6. We use this approach to estimate loads for the analysis that are representative of the middle of the state time period (i.e. the full calendar year of 2015). The regression coefficients ($\beta_{1,2,..7}$) are those derived from fitting equation 6 to the observation dataset. It follows that the estimated load for the year of interest can be calculated by:

$$L^Y = \frac{KS}{N}\sum_{j=1}^{N} \hat{C}_j^Y Q_j \quad \text{(Equation 5b)}$$

4.3.1.2 L5 Model

The L5 model is the same as L7 model except that two quadratic terms are eliminated:

$$ln(\widehat{C_i}) = \beta_1 + \beta_2(ln(q_i)) + \beta_3(t_i) + \beta_4 sin(2\pi t_i) + \beta_5 cos(2\pi t_i) \quad \text{(Equation 8)}$$

The five parameters are estimated, and loads are calculated in the same manner as the L7 model. Following the approach outlined for the L7 model, the L5 model can be adjusted when used for prediction to provide estimates for a selected load estimation date:

$$ln\left(\widehat{C_j^Y}\right) = \beta_1 + \beta_2[ln(q_j)] + +\beta_4(Y_j - \bar{T}) + +\beta_6 sin(2\pi Y_j) + \beta_7 cos(2\pi Y_j) \quad \text{(Equation 8a)}$$

4.3.1.3 Flow stratification

Roygard *et al.* (2012) employed a flow stratification approach to defining rating curves. This approach is based on a non-parametric rating curve, which is defined by evaluating the mean concentration within equal increments of the flow probability distribution (flow 'bins'). In their application, Roygard *et al.* (2012) employed ten equal time-based categories (flow decile bins), defined using flow distribution statistics and then calculated mean concentrations within each bin. This non-parametric rating curve can then be used to estimate nutrient concentrations, $\hat{C}$, for all days with flow observations. At step 2, the load is calculated following equation (5a), providing an estimate of average annual load over the observation time period.

$$L = \frac{K}{N} \sum_{j=1}^{N} \hat{C}_j Q_j \quad \text{````} \quad \text{(Equation 5c)}$$

where $\widehat{C_j}$ is calculated mean concentration associated with the flow quantile bin of the flow Q_j., and all other variables are as per equation 5.

4.3.1.4 Flow stratification with trend

We have also included a modified version of the flow stratification method to account for trends in water quality. This is useful in the instance that loads are required to be estimate for a particular point in time, rather than as an average over the complete observation period, particularly when there are strong trend evident. We detrended the observation data by fitting equation 9 to the concentration time series.

$$ln(\widehat{C_i}) = \beta_1 + \beta_2(t_i) \quad \text{(Equation 9)}$$

The form of equation 9 follows that used to represent trends in the L5 model. We then use the concentration residuals to develop a non-parametric rating curve following the method described in 4.3.1.3. $\widehat{C_j}$ is calculated as the mean residual concentration associated with the flow quantile bin of the flow Q_j., plus the predicted value of concentration from equation 6 at time T_j.which is multiplied by the smearing coefficient to account for the log transformation of equation 9.

4.3.2 Precision of load estimates

The statistical precision of a sample statistic, in this study the mean annual load, is the amount by which it can be expected to fluctuate from the population parameter it is estimating due to sample error. In this study, the precision represents the repeatability of the estimated load if it was re-estimated using the same method under the same conditions. Precision is characterised by the standard deviation of the sample statistic, commonly referred to as the standard error. We evaluated the standard error of each load estimate by bootstrap resampling (Efron, 1981). For each load estimate we constructed 100 resamples of the concentration data (of equal size to the observed dataset), each of which was obtained by random sampling with replacement from the original dataset. Using each of these datasets, we recalculated the site load and estimated the 95% confidence intervals, using the boot r package. We represent precision in the results as the 95% confidence interval range, standardised by the load estimate (i.e., represented as a proportion).

4.3.3 Selection of best load estimation methodology

Loads were calculated for all site and nutrient variable combinations using each of the four load estimation methods. All load estimates were converted to flow weighted concentrations (by dividing by the long term mean flow) and were then combined with observed median

concentrations to calculate R_{median} as the ratio of the C_{median} to C_{FW} (Equation 1a) for each nutrient variable, site and method. We evaluated the performance of each rating curve method for predicting observed concentrations, using a range of model performance measures (details are in section 4.6). We identified site and nutrient variable combinations that had any of:

1. R_{median} values <0.1 or R_{medain} values >1.2;
2. Large C_{fw} values (values varied by nutrient variable);
3. Large differences in the loads calculated using different methods.

For these site and nutrient variable combinations (approximately 10-20% of sites for each nutrient variable), we manually inspected diagnostic plots (e.g. C-Q plots, C-T plots, comparisons of sampled flow distributions relative to observed flow distributions) We used expert judgement to select the most appropriate load estimation methodology for each site and variable that were outside of the three criteria outlined above. As well as selecting from one of the four rating curve methods described above, we also allowed sites to be discarded at this stage if no method appeared to satisfactorily describe the observed behaviour. This process also suggested that, for the manually inspected sites, the selection of the model with the lowest RMSD (in terms of performance in predicting observed concentrations) was the criteria most consistent with the outcomes of the expert judgement. As such, for the remainder of the site and nutrient variable combinations that were not flagged by the abovementioned criteria (and were therefore not individually inspected), the most appropriate load estimation method was selected as the rating curve method that yielded the lowest RMSD.

4.4 Observation of R at monitoring sites

We calculated R at each monitoring site as the ratio of $C_{statistic}$ to C_{fw} (Equation 1a). We combined the uncertainties of $C_{statistic}$ and C_{fw} to estimate the uncertainty of the observed R. We estimated the confidence intervals for R (as proportions), by adding the confidence intervals as proportions for each of the concentration statistic and the flow weighted concentration. Because the confidence intervals were estimated non-parametrically they are not necessarily symmetric, therefore the upper and lower confidence intervals were calculated independently, (and converted back into absolute values rather than proportions) using the following equations:

$$R_{Statistic\ LCI} = R_{Statistic}\left(1 - \left(\frac{C_{Statistic} - C_{Statistic\ LCI}}{C_{Statistic}} + \frac{C_{FW\ UCI} - C_{FW}}{C_{FW}}\right)\right) \quad \text{(Equation 10)}$$

$$R_{Statisitc\ UCI} = R_{Statistic}\left(1 + \left(\frac{C_{Statistic\ UCI} - C_{Statistic}}{C_{Statistic}} + \frac{C_{FW} - C_{FW\ UCI}}{C_{FW}}\right)\right) \quad \text{(Equation 11)}$$

The confidence intervals for $R_{statistic}$ are primarily used for plotting purposes and to provide a qualitative evaluation of the reliability of the estimates of R, which has implications for subsequent reliability of the models developed for R. This approach is likely to provide pessimistic uncertainty ranges, as we might expect the $C_{statistic}$ and C_{fw} errors to covary.

4.5 Predictions of R

We modelled and made predictions of R using three alternative approaches: a global model, a catchment characteristics model and a distributional characteristics model. These models represent a spectrum of complexity and user data requirements (as summarised in Table 4).

Table 4: User data requirements for alternative models of R

Model Name	Data Requirements for prediction of R
Global	None
Catchment characteristics	Database of R (predictions made by this study); available for all REC2 segments across NZ
Distributional characteristics	Paired observations of concentration (C) and discharge (Q) that meet filtering requirements set out in section 4.3 and 4.2. At least 10 years of continuous flow data (or appropriate estimates of the required flow distributional parameters from a reliable source).

4.5.1 Global model

The first model is the simplest and assumes that $R_{statistic}$ is a global constant for each nutrient variable. $R_{statistic}$ is estimated as the median of the observed values of $R_{statistic}$. To provide estimates of model error that are more representative of expected errors for sites not included in fitting the data, we performed a 10-fold cross-validation procedure, whereby for each fold, 10% of the data were randomly excluded and the estimate of $R_{statistic}$ at the excluded sites was the median of $R_{statistic}$ for the remaining 90% of sites.

4.5.2 Catchment characteristics model

We fitted $R_{statistic}$ for each nutrient variable derived for each monitoring site to a suite of predictor variables using random forest (RF) models (Breiman, 2001; Cutler *et al.*, 2007). These predictors represented values that were available for all segments of the digital river network and did not include any predictors that would rely on more detailed information about the specific flow or concentration relationships at a site. Details of RF models and how they were applied in this study are provided by (Snelder *et al.*, 2018; Whitehead, 2018).

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 MSE is within one standard error of the best model (where the best model is the model with the lowest cross validated MSE) (Breiman *et al.*, 1984). The standard error of MSE is the standard deviation of MSE values calculated for each of the hold-out observations (i.e., the cross validation folds). Importance levels for predictor variables were not recalculated at each reduction step to avoid over-fitting (Svetnik *et al.*, 2004).

All calculations were performed in the R statistical computing environment (R Development Core Team 2009) using the *randomForest* package and other specialised packages. (Here R refers to the software, rather than the ratio R).

4.5.2.1 Modelled relationships

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

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

We approximated the direction of the influence of each predictor by the sign of a linear regression through the partial dependence plots. 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 concentration statistic/load and green indicates that increasing values of the predictor correspond to improving concentration statistic/load).

4.5.2.2 Model predictions

Predictions are made with RF models by “running” new cases (i.e., sites that were not part of the training data) down every tree in the fitted forest and averaging the predictions made by each tree (Cutler *et al.*, 2007). The output is a database of predictions of R associated with each individual segment of the digital river network.

4.5.3 Distributional characteristics model

The distributional characteristics model is a linear regression model for which the predictor variables are various statistics derived from the observations of flow and concentration (the distributional characteristics), including statistics for the flow, concentration and joint flow-concentration distribution Table 5. In most cases where there is sufficient data to estimate R using the distributional characteristics model, there will also be sufficient data to estimate R directly (i.e. by calculating the loads and then dividing the observed concentrations by the load). However, there may be occasions where loads cannot be estimated due to insufficient data, and in these cases the distributional characteristics model may be useful. More

importantly, we considered the distributional characteristics model might provide some insights into the mechanisms that cause spatial variation in R.

The concentration and joint concentration-flow characteristics only use the data from the 5 year period over which the concentration statistics were calculated, and flow quantiles used to characterise the join C:Q distribution are derived from the paired C:Q observations (not the continuous flow record). The characteristics of the flow distribution were derived from the complete flow timeseries. We considered the use of other statistics but found that these were largely highly correlated with the chosen predictors. Table 5 provides a summary of the predictors used in the distributional characteristics model.

Table 5: Summary of predictors used in the distributional characteristics model

Predictor Name	Description	Units
WQ_sd	Standard deviation of the concentration observations	mg L^{-1}
WQ_cv	Coefficient of variation of the concentration observations	-
CQ9050	Ratio of the median concentration associated with flows between the 80:100th flow quantiles ("90") to the median concentration associated with flows between the 40:60th flow quantiles ("50").	-
CQ9010	Ratio of the median concentration associated with flows between the 80:100th flow quantiles ("90") to the median concentration associated with flows between the 0:20th flow quantiles ("10").	-
CT_Slope_Std	Coefficient β_2 from the regression: $\ln(C) \sim \beta_1 + \beta_2(T)$, standardised by dividing by the mean concentration	C – mg L^{-1} T - days
Q_cv	Coefficient of variation of the continuous flow record	-
MeanFlow	Mean Flow	m^3s^{-1}

After visual inspection of scatter plots of R versus the predictor variables, we chose to $\log_{10}$ transform the predictors to linearise these relationships. As CT_Slope_Std includes both positive and negative values, this variable was transformed by taking the square root of the absolute value, then assigning the sign of the original value (this is represented as the function F(x) in the equation below).

The distributional characteristics model takes the form of:

$$R_{Statistic} = \beta_1 + \beta_2 \log_{10}(WQ_{sd}) + \beta_3 \log_{10}(WQ_{cv}) + \beta_4 \log_{10}(CQ9050) + \beta_5 \log_{10}(CQ9010) + \beta_6 \mathrm{F}\left(CT_{Slope_{Std}}\right) + \beta_7 \log_{10}(Q_{cv}) + \beta_8 \log_{10}(MeanFlow) \quad \text{(Equation 12)}$$

Where β_1, $\beta_2 \ldots$, β_8 are the fitted coefficients. The coefficients of equation 12 were determined through least squares regression employing a stepwise model building process to fit the most parsimonious model starting with all the explanatory variables. The Akaike information criterion (AIC; Akaike 1973) was used to apply a penalised log-likelihood method to evaluate

the trade-off between the degrees of freedom and fit of the model as explanatory variables were added or removed (Crawley, 2002). The procedure identified the preferred model as that with the lowest AIC value. We performed a 10-fold cross validation procedure to provide more representative estimates of model error for predictions at new sites (i.e., sites that were not part of the training data).

4.5.3.1 Modelled Relationships

We repeated the regression with all predictor variables standardised such that they have means of zero and standard deviations of one. In doing this, the magnitude and sign of the regression coefficients can be used to interpret the relationships between the response (R for each nutrient variable) and each of the predictor variables (Saltelli *et al.*, 2008).

4.6 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 either from the 10-fold cross validation process for the linear regression models, or from the OOB observations for the Random Forest models. We summarised the model performance using seven statistics; regression R^2; Nash-Sutcliffe efficiency (NSE); percent bias (PBIAS) the relative root mean square deviation (RSR); the root mean square deviation (RMSD); the median absolute error (MAE); and the relative median absolute error (RMAE).

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 how much of the variation in the observed values is explained by the variation in the predicted values, 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 square root of the mean squared 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.*, 2015).

The normalization associated with PBIAS and RSR allowed the performance of models to be compared across all nutrient variables. Model predictions were evaluated to be very good, good, satisfactory or unsatisfactory based on the criteria (for nutrient modelling) proposed by Moriasi et al.,(2015), outlined in Table 6.

Table 6: Performance ratings for statistics used in this study, from (Moriasi et al.*, 2015).*

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

RMSD is a measure of the characteristic model statistical error or uncertainty. RMSD is the square root of the mean squared 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.

MAE is the median absolute error. It is similar to the RMSD, but is more robust to outliers, so potentially can provide a more appropriate estimation of the typical model errors, if the number of outliers are small. RMAE is the normalised median absolute error: MAE divided by the standard deviation of the observations.

4.7 Evaluating differences in model performance

In order to evaluate the difference in the model performance we tabulated the RMSD and RMAE for each combination of model, statistic and nutrient variable. To determine the statistical significance of these differences we calculated the residuals for each model and then compared the variances of the residual distributions between all pairs of models using the F-test (which is a test for equality of variances).

5 Results

5.1 Concentrations

The distributions of the three site concentration statistics (mean, median and 95th percentile) were approximately log-normally distributed (Figure 5). Maps of the concentration statistics are provided in Appendix B.

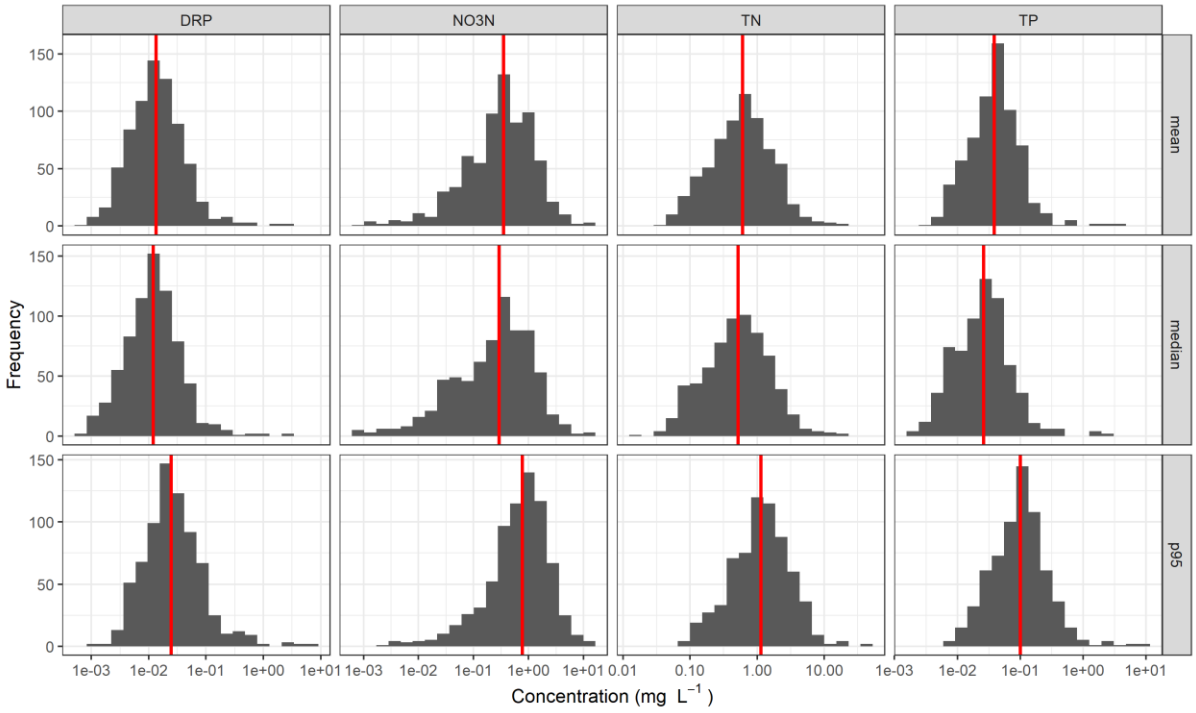

Figure 5: Histograms of observed concentration statistics (mean, median and 95th percentile) derived for the monitoring sites, by nutrient variable. Note, x-axis is a log scale. The red lines indicate the median of the site statistics.

5.2 Estimated loads

Table 6 shows a summary of the number of sites for which each of the alternative load calculation methods described in the methods was selected as most appropriate. In general, the L5 model was most frequently selected. However, for TP, the flow stratification method with trends (FST) method was judged to be most appropriate more often than the L5 method. The reflects the tendency for TP flow-concentration observations to be erratic compared to the patterns observed for the nitrogen species. A small (1-3%) number of sites were excluded from the analysis at this stage because either none of the methods could adequately represent the observations, or, upon inspection the load estimates were considered to be unreliable.

Figure 6: Table summarising number and percentages (in brackets) of sites for which each of the alternative load calculation methods were used. "No model" indicates sites that were not assigned a rating curve methodology and were discarded from the analysis.

Nutrient variable	L7	L5	FS	FST	No model
DRP	124 (17%)	234 (32%)	26 (4%)	345 (47%)	6 (1%)
NO3N	31 (4%)	250 (34%)	59 (8%)	387 (53%)	7 (1%)
TN	37 (6%)	234 (35%)	33 (5%)	361 (54%)	4 (1%)
TP	63 (9%)	318 (47%)	41 (6%)	243 (36%)	5 (1%)

Figure 7 shows histograms of the calculated loads for all sites. Loads are demonstrated as both flow weighted concentrations (C_{FW} mg L^{-1}) as well as export coefficients (loads standardised by catchment area, kg ha^{-1} yr^{-1}). All subsequent analyses and results represent loads as flow weighted concentrations. However, we have included export coefficients in Figure 7 (as well as a comparison of the relationship between flow C_{FW} and the export coefficients), as a comparison point for readers who may not be familiar with typical values for flow weighted concentrations. Loads were approximately log-normally distributed. Maps of calculated loads are provided in Appendix C.

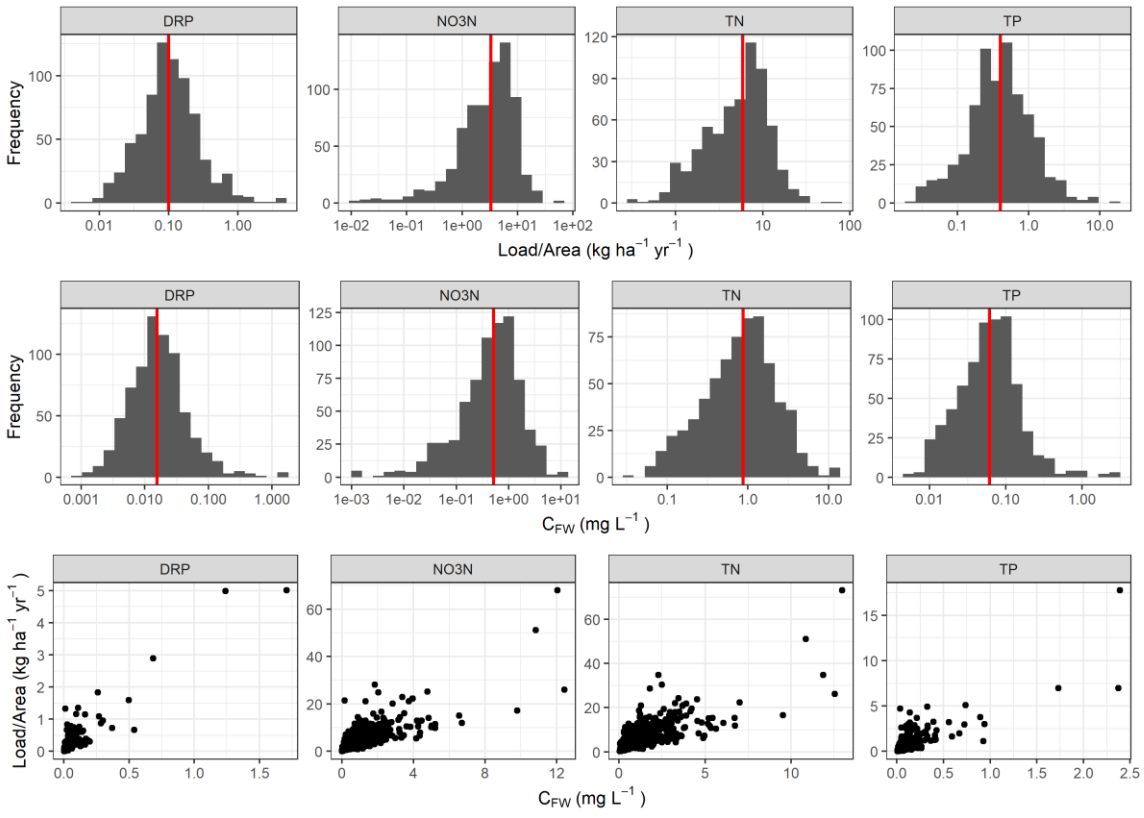

Figure 7: Histograms of loads as load per unit area and flow weighted concentrations. The third row shows a comparison of the loads of the two alternative units.

5.3 Observed R

The majority of observed R values at monitoring sites were less than one when the concentration statistic was either the median or the mean for all nutrient variables (Figure 8). This indicates that loads would generally be under-estimated if the simple assumption was made that these can be calculated by multiplying the statistic (i.e., mean or median concentration) by the mean flow. When the statistic type was the 95th percentile, R values had the reverse pattern (i.e., most observed R values were greater than one).

The uncertainty of the observed R values were least for R_{median} and greatest for R_{p95} (Figure 9). As a measure of relative uncertainty, we compared the median of the difference between the upper and lower confidence intervals at each site against the interquartile range (ICR) of the observed values of R (Table 7). The typical uncertainties were 1.3,1.7 and 3.1 x ICR for R_{median}, R_{mean} and R_{p95}. In other words, the average uncertainty at a site was larger than the average difference in R between sites for all statistics. The generally higher uncertainties for R_{mean} and R_{p95} compared to R_{median} are associated with the large uncertainty of the estimates of these statistics from five years of monthly data.

Site values of R were spatially variable for all nutrient variables (Figure 10, Figure 11, Figure 12). Some spatial patterns were evident in the mapped R values. For example, sites with high values of R for all nutrient variables were clustered in the central North Island and tended to occur on mainstems of larger rivers. Sites with low values of R for all nutrient variables tended to occur on the eastern coasts of both the North and South Islands.

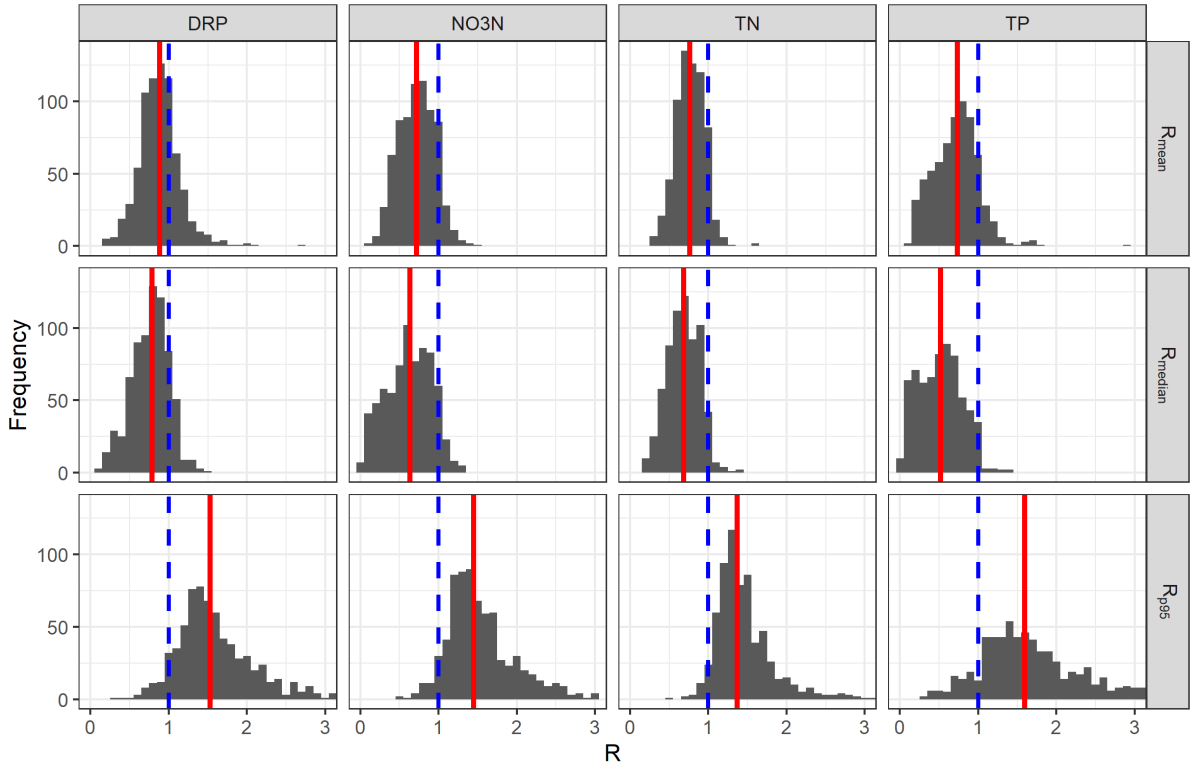

Figure 8: Histograms of observed R values (mean, median and p95) at the monitoring sites. Site median values of R are indicated by the red line. The blue dashed line indicates an R value of 1 to provide a reference point: values of R greater than one indicate that the concentration statistic is greater than the flow weighted concentration.

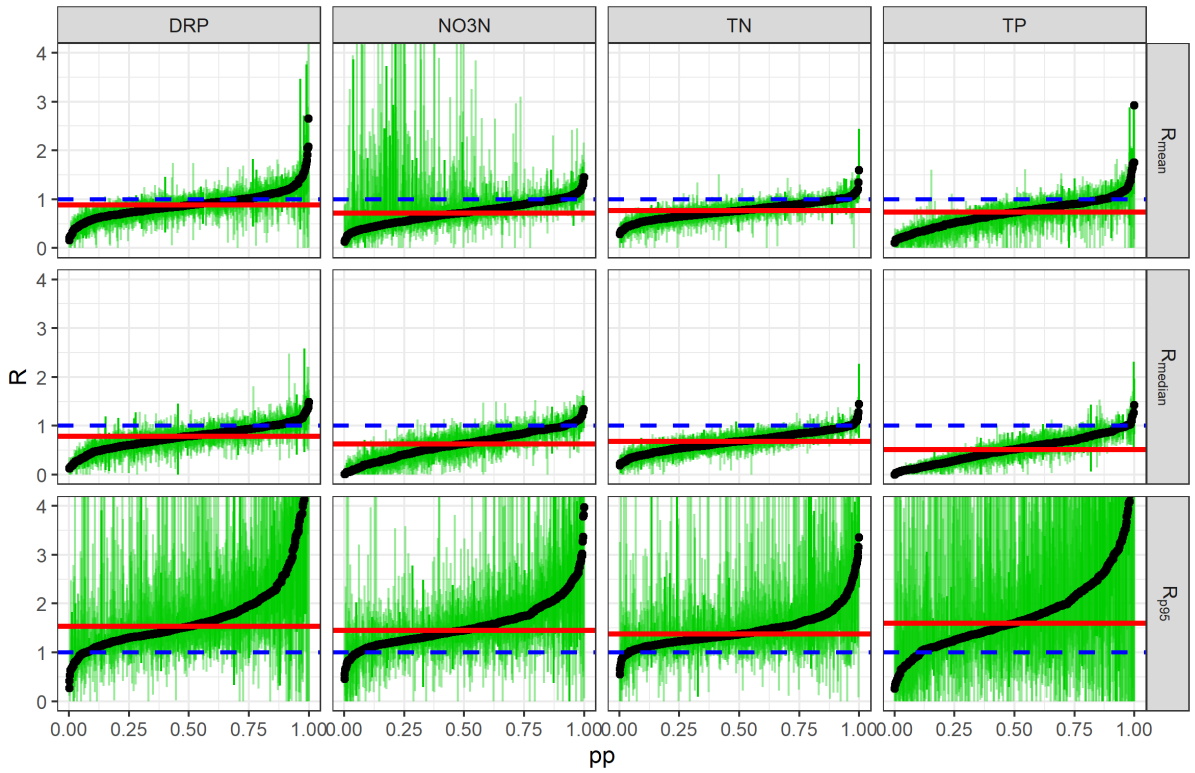

Figure 9: Cumulative distribution plots of R , including 95% confidence intervals (green lines). Red lines indicate the median value of R and blue lines are at unity, for comparison.

Table 7 Summary of relative uncertainties, as the ratio of the median difference between the upper and lower confidence intervals, standardised by the interquartile range:

Nutrient variable	R_{median}	R_{mean}	R_{p95}
DRP	1.26	1.18	1.78
NO3N	1.19	2.08	1.90
TN	1.03	1.05	2.50
TP	0.92	1.28	2.40

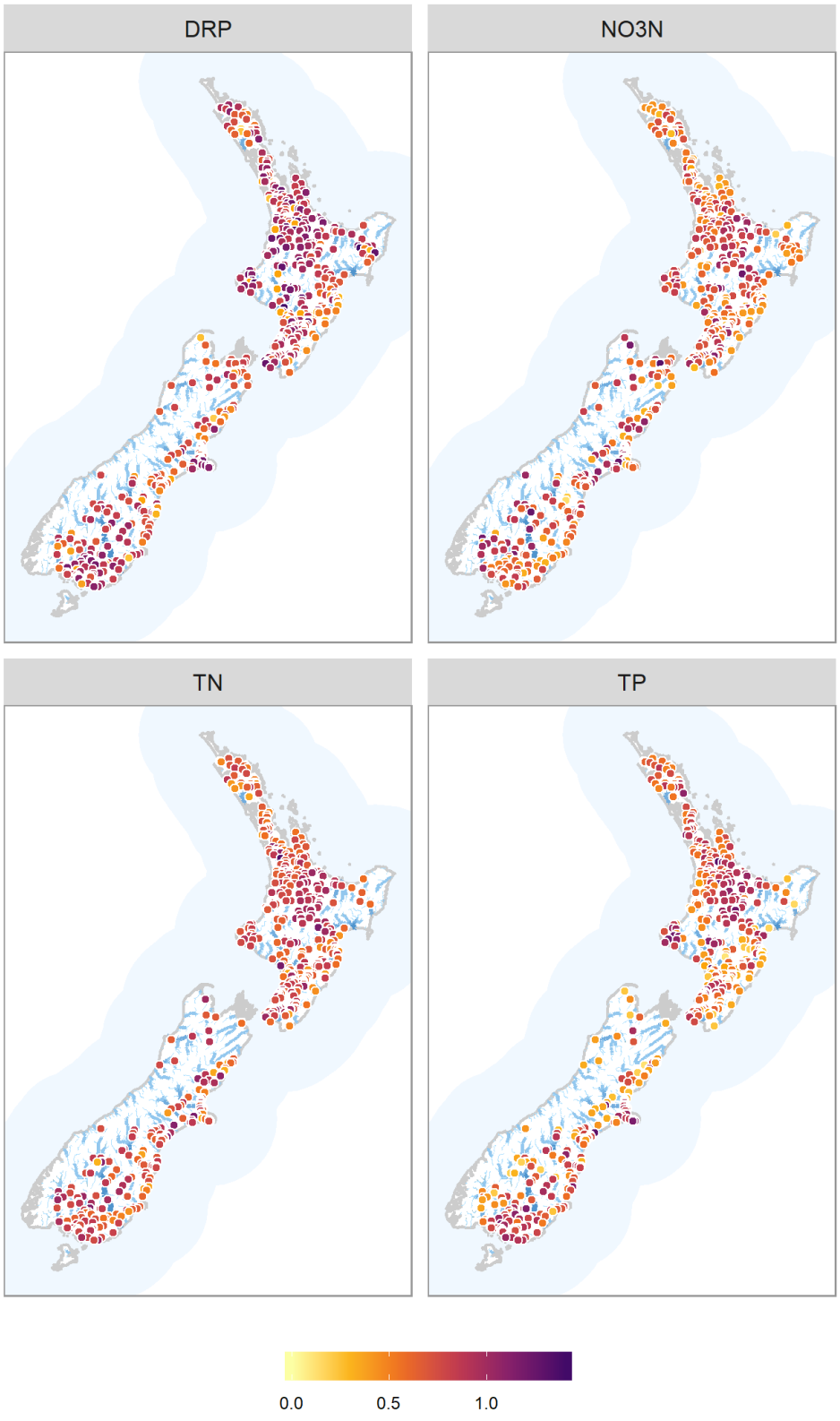

Figure 10: Maps showing observed values of R_{mean} at monitoring sites.

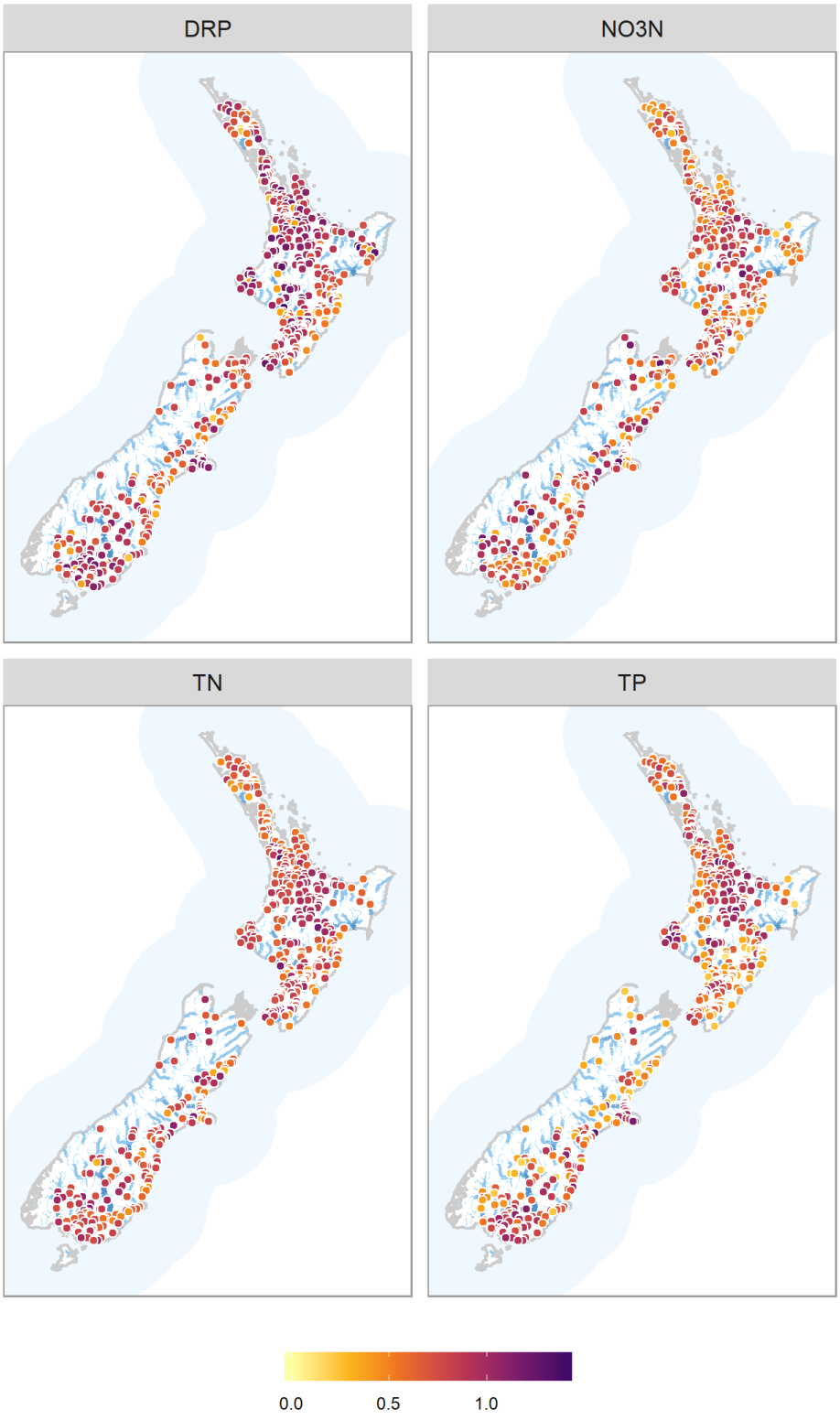

Figure 11: Maps showing observed values of R_{median} at monitoring sites.

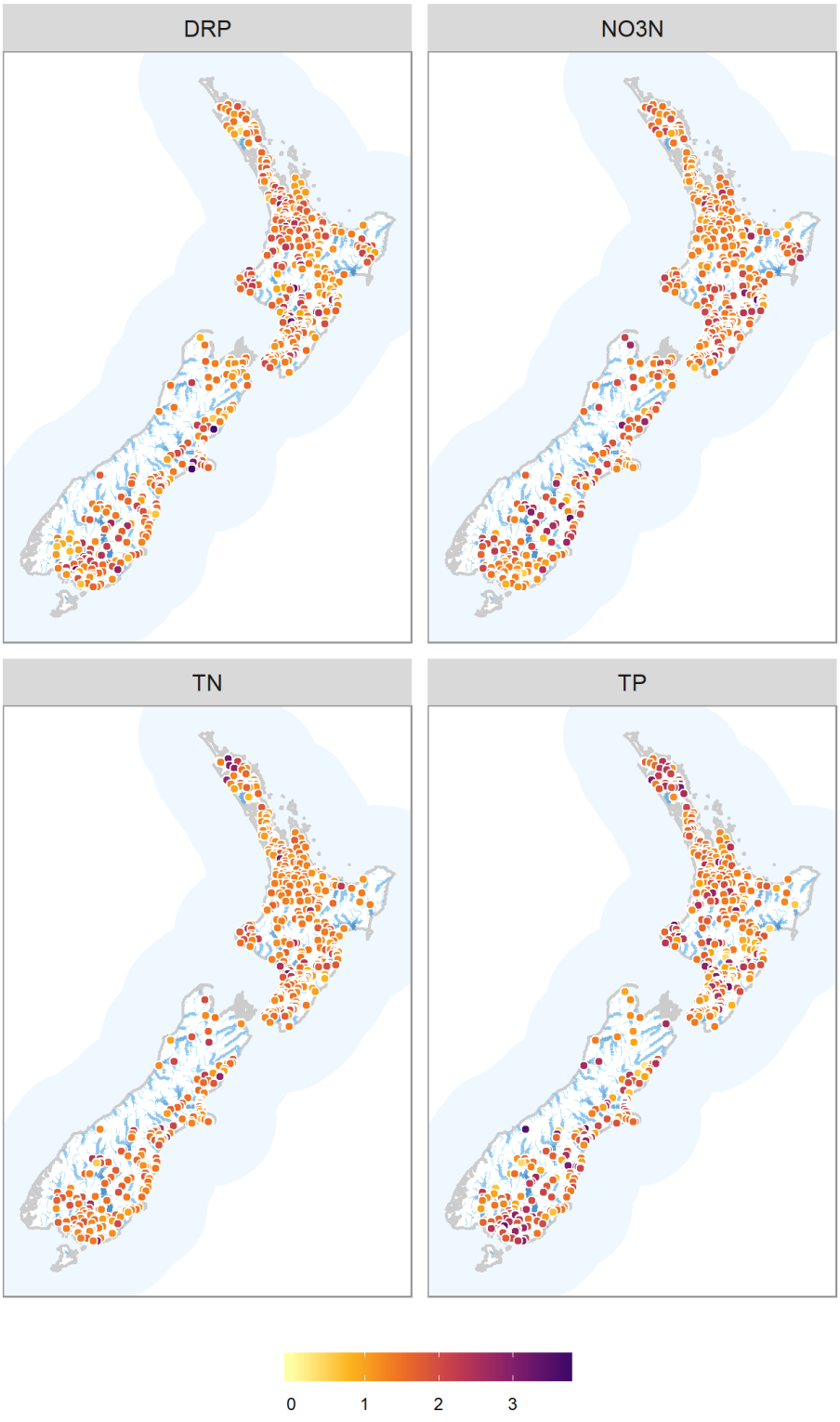

Figure 12: Maps showing observed values of R_{p95} at monitoring sites. (Note, the colour scale has been stopped at 4, but there are a small number of sites with values greater than this).

5.4 Global model

The global models for R (R_{mean}, R_{median} and R_{p95}) are simply the median of the observed site R values, as demonstrated by red lines in in Figure 8. Performance statistics for these models are provided in Table 8. Note that some performance statistics are excluded from Table 8 (R^2, NSE and RSR), because these values tend to towards, 0, 0 and 1 as is the case for a model that is based on the mean and that has normally distributed residuals. Because our median estimates were not strongly dissimilar from the mean values, and the distributions are not highly skewed, there was little variance in these performance measures, and they all approached the previously described limits.

Table 8. Performance of the global models of R (for mean, median or p95). Performance was determined using independent predictions (i.e., sites that were not used in fitting the models) generated from the cross-validated predictions of R. PBIAS = percent bias, RMSD = root mean square deviation, RMAE = relative median absolute error, MAE = median absolute error.

R	Nutrient variable	PBIAS	RMSD	RMAE	MAE
R_{mean}	DRP	1.17	0.30	0.49	0.15
	NO3N	0.28	0.23	0.76	0.17
	TN	0.44	0.18	0.70	0.13
	TP	-1.67	0.30	0.63	0.19
R_{median}	DRP	-2.76	0.24	0.68	0.16
	NO3N	-2.13	0.29	0.76	0.22
	TN	-0.36	0.20	0.73	0.15
	TP	-0.30	0.28	0.77	0.21
R_{p95}	DRP	14.86	1.31	0.23	0.29
	NO3N	6.45	0.47	0.52	0.24
	TN	6.15	0.39	0.46	0.18
	TP	10.28	0.86	0.47	0.40

5.5 Catchment characteristics model

5.5.1 Model performance

The performance of the catchment characteristics models of R (mean, median and p95) are summarised in Figure 13 and Table 9. According to the performance criteria outlined in section 4.6, only the model for TP R_{median} had a satisfactory performance; all other models were classed as "unsatisfactory". Nevertheless, the models were generally unbiased, and in terms of characteristic errors (RMSD), the catchment characteristics model was approximately 20% better than the global models for R_{mean} and R_{median} and on average 5% better than the global model for R_{p95}.

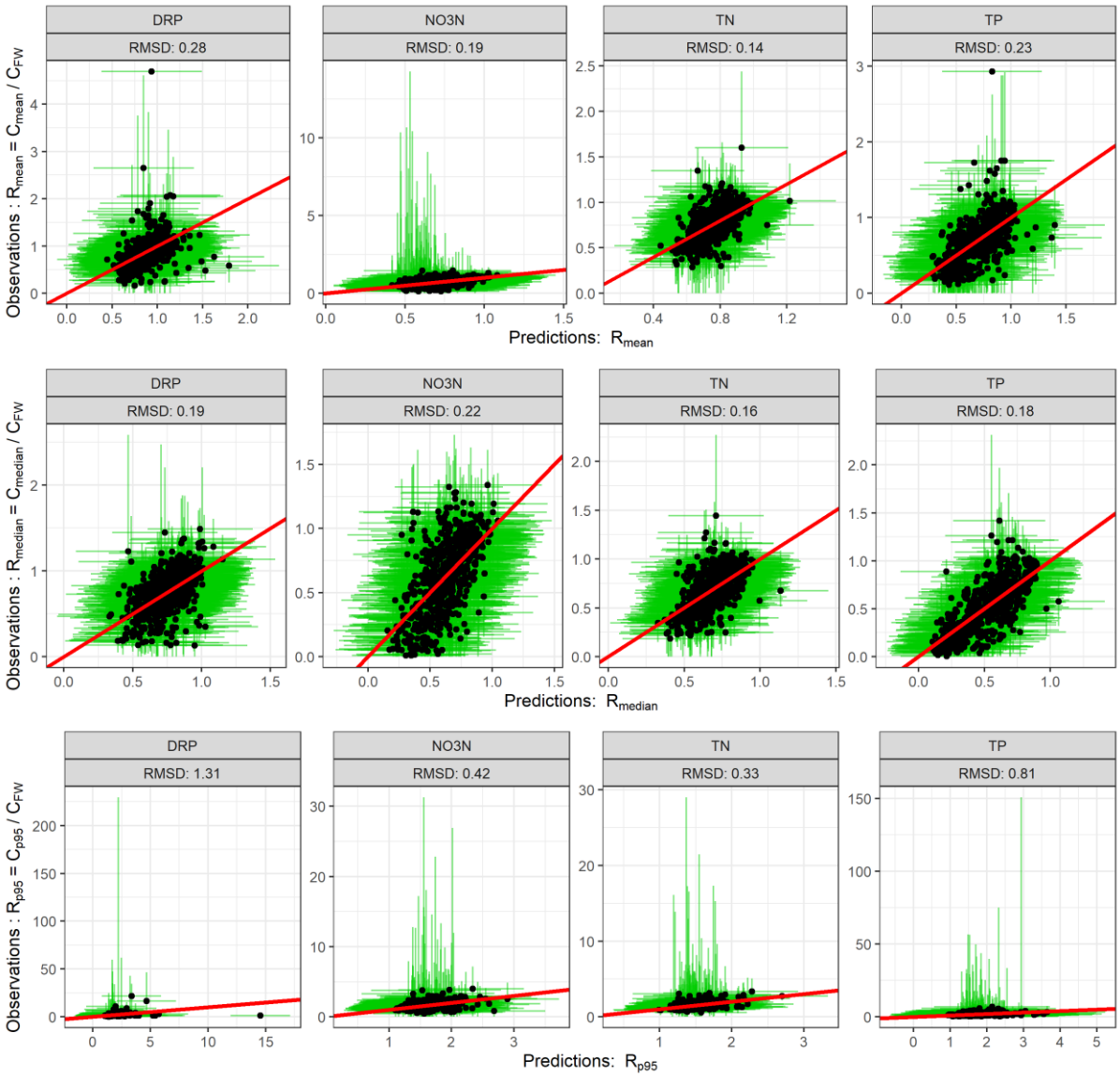

Figure 13: Scatter plots of the predicted R (mean, medians and p95) from the catchment characteristics model versus observed R (C/C_{FW}). Red line shows 1:1 line as a comparison point. Green lines indicate the 95% confidence intervals of both observations and predictions.

Table 9. Performance of the catchment characteristics model of R (mean, median or p95). Performance was determined using independent predictions (i.e., sites that were not used in fitting the models) generated from the out-of-bag observations of R. 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, RMAE = relative median absolute error, MAE = median absolute error.

	Nutrient variable	R^2	NSE	PBIAS	RSR	RMSD	RMAE	MAE
R_{mean}	DRP	0.14	0.12	-0.67	0.94	0.283	0.39	0.12
	NO3N	0.33	0.33	-0.03	0.82	0.188	0.48	0.11
	TN	0.35	0.35	-0.02	0.81	0.143	0.44	0.08
	TP	0.37	0.37	-0.67	0.79	0.232	0.39	0.12
R_{median}	DRP	0.32	0.32	0.47	0.82	0.195	0.49	0.12
	NO3N	0.40	0.40	-0.16	0.77	0.222	0.51	0.15
	TN	0.37	0.36	0.00	0.80	0.161	0.47	0.10
	TP	0.58	0.58	-0.47	0.65	0.178	0.39	0.11
R_{p95}	DRP	0.06	-0.03	-3.03	1.01	1.315	0.26	0.33
	NO3N	0.18	0.17	-1.00	0.91	0.419	0.47	0.21
	TN	0.22	0.22	-0.28	0.88	0.326	0.47	0.18
	TP	0.10	0.08	-2.03	0.96	0.806	0.55	0.46

5.5.2 Modelled relationships

Figure 14 demonstrates the relative importance and direction of influence of predictors on the modelled response (i.e., $R_{statistic}$) fitted by the RF models. For R_{mean} and R_{median} hydrological characteristics were important, with increasing R values associated with increasing baseflow contributions (MALF7 and MALF30) and decreasing R values with higher degrees of hydrological “flashiness” (i.e., FRE3 and JulFloodFlow).

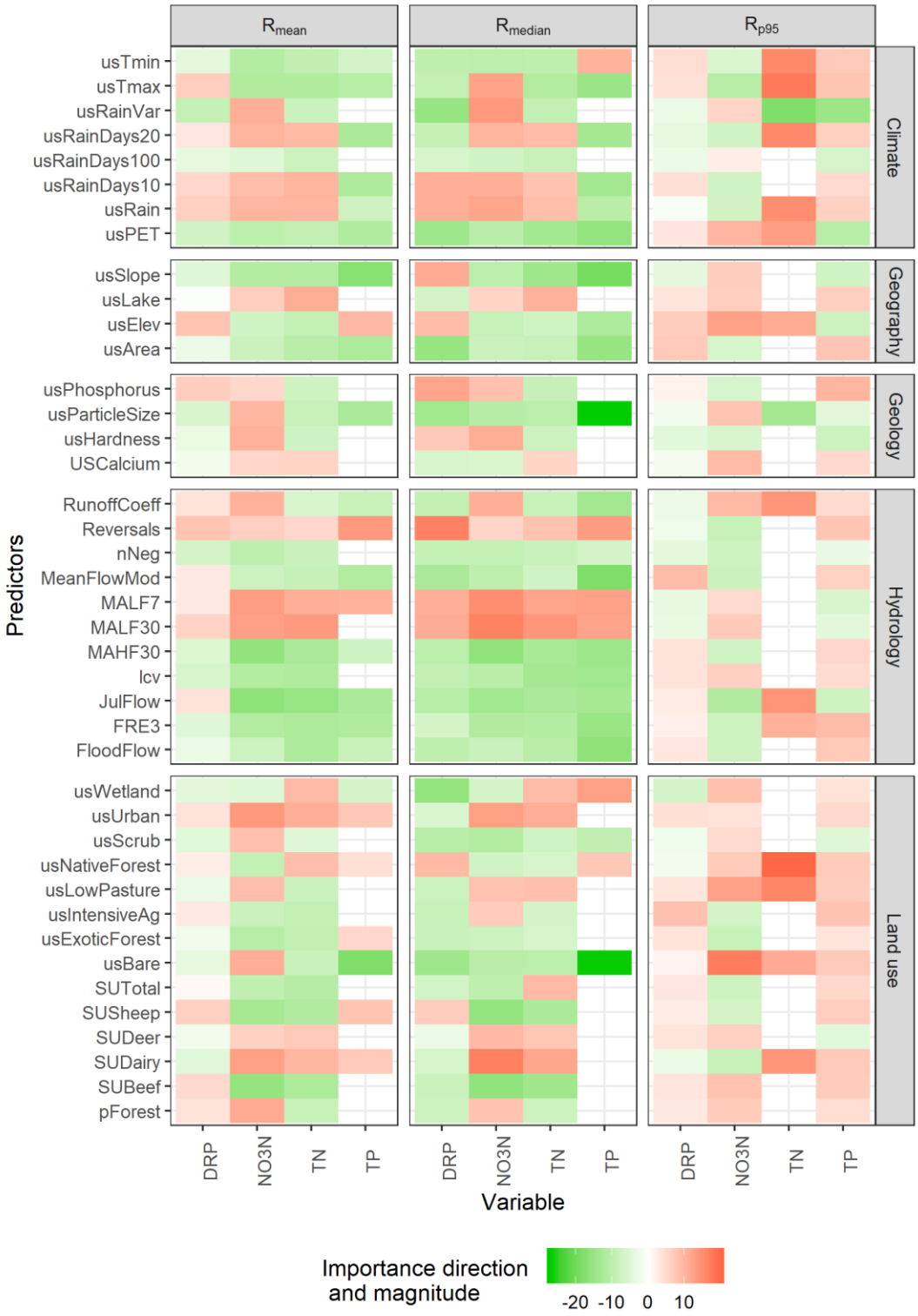

Figure 14: Importance of predictors included in the 'reduced' random forest $R_{statistic}$ models. The colours indicate the importance and direction of influence of each predictor on the modelled $R_{statistic}$ predictions. Red indicates increasing predictor magnitudes are associated with increasing $R_{statistic}$, whereas green indicates increasing predictor magnitudes are associated with decreasing $R_{statistic}$. Blank cells indicate that the predictor was not included in the 'reduced' random forest model.

5.5.3 Model predictions

We used the catchment characteristics models to make predictions of R for the whole country (Figure 15, Figure 16, Figure 17). Patterns in the value of R are evident in these maps; notably high R values in the central North Island and on the Canterbury plains, and low R values associated with lower altitude areas for NO3N. In addition, R varies strongly between rivers with catchments that are dominated by mountain, hill and lowland topography, particularly for TN and TP.

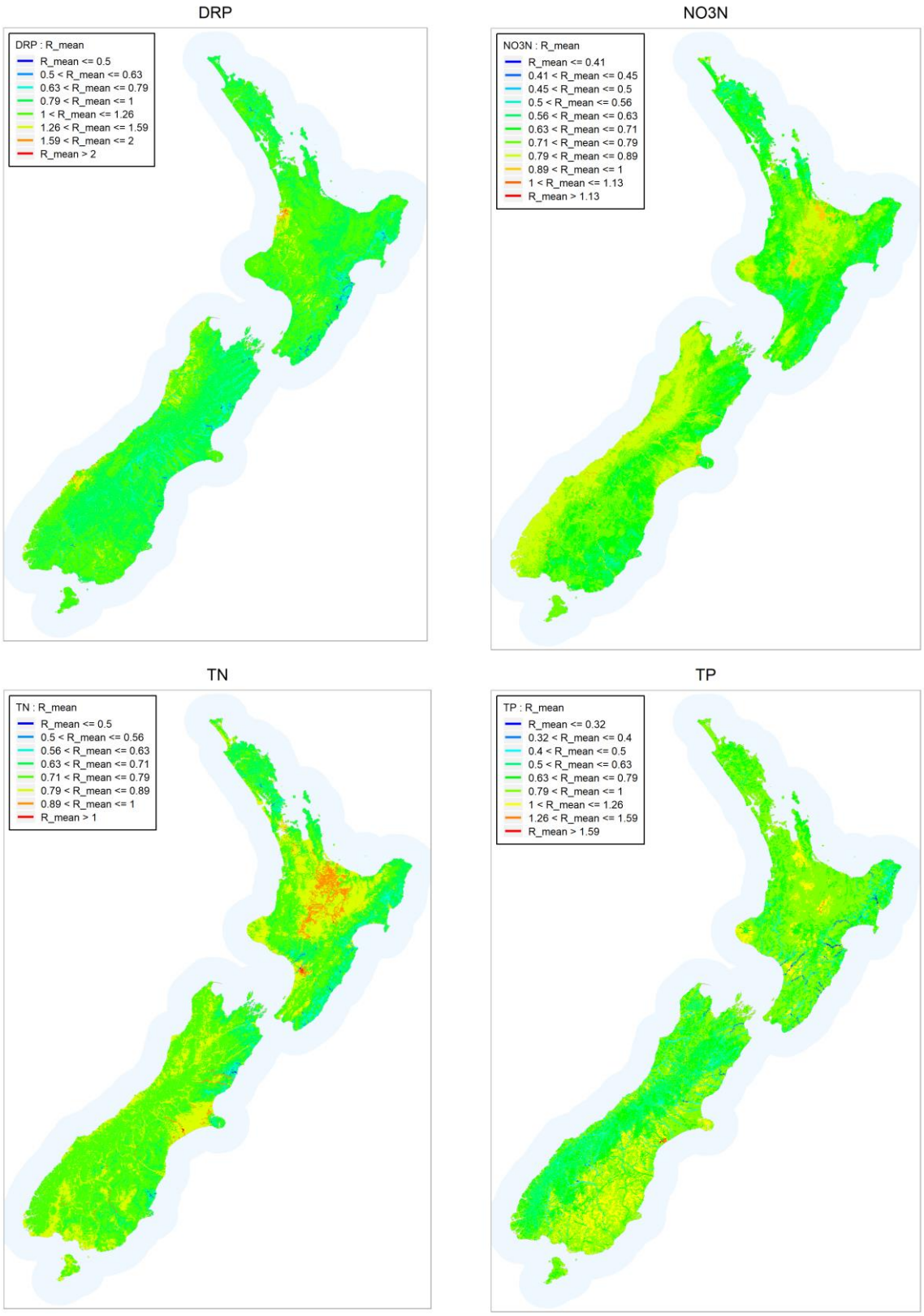

Figure 15: Maps of predicted R_{mean} values from catchment characteristics model. Maps show all rivers from order 1 and above.

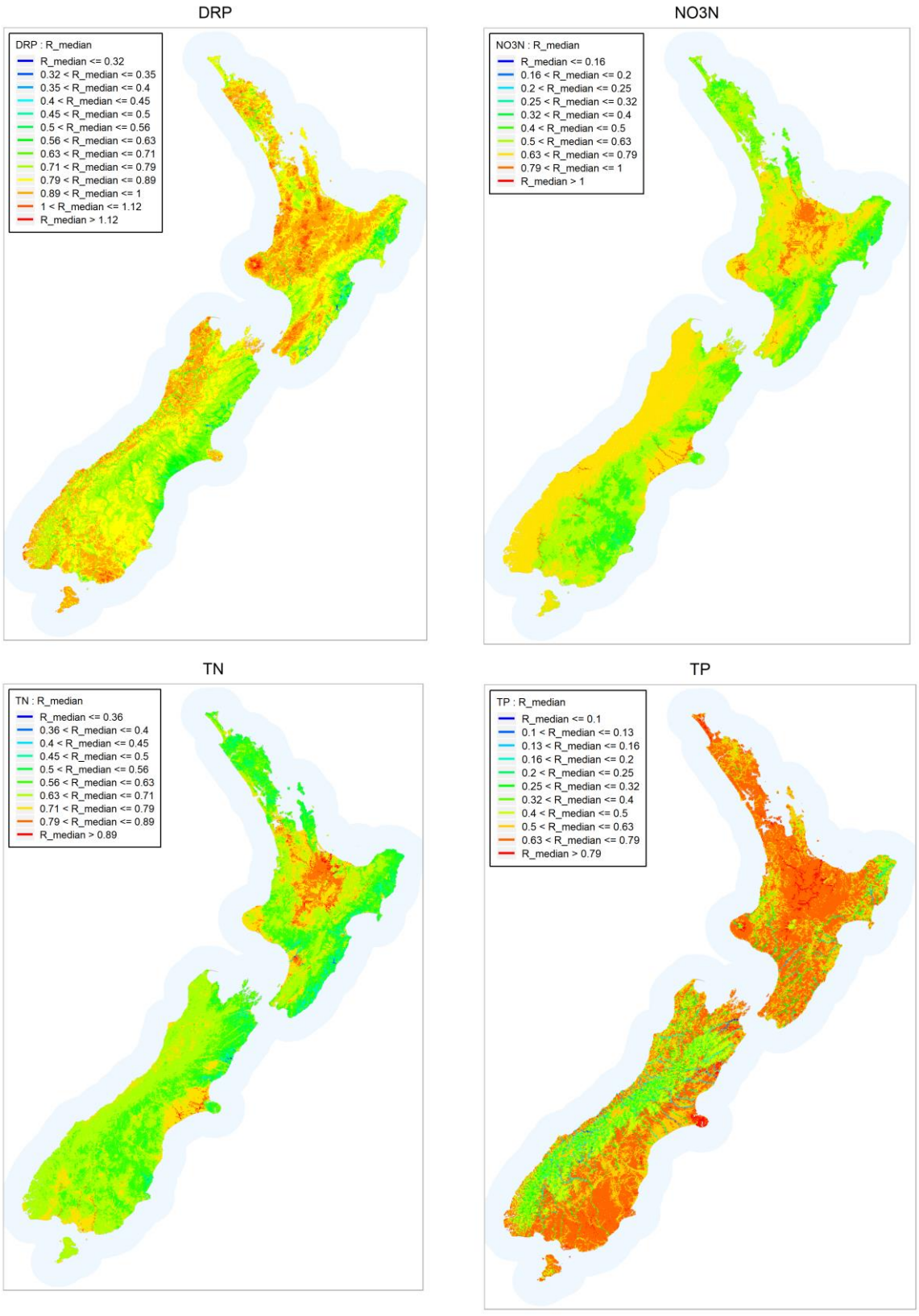

Figure 16: Maps of predicted R_{median} values from catchment characteristics model. Maps show all rivers from order 1 and above.

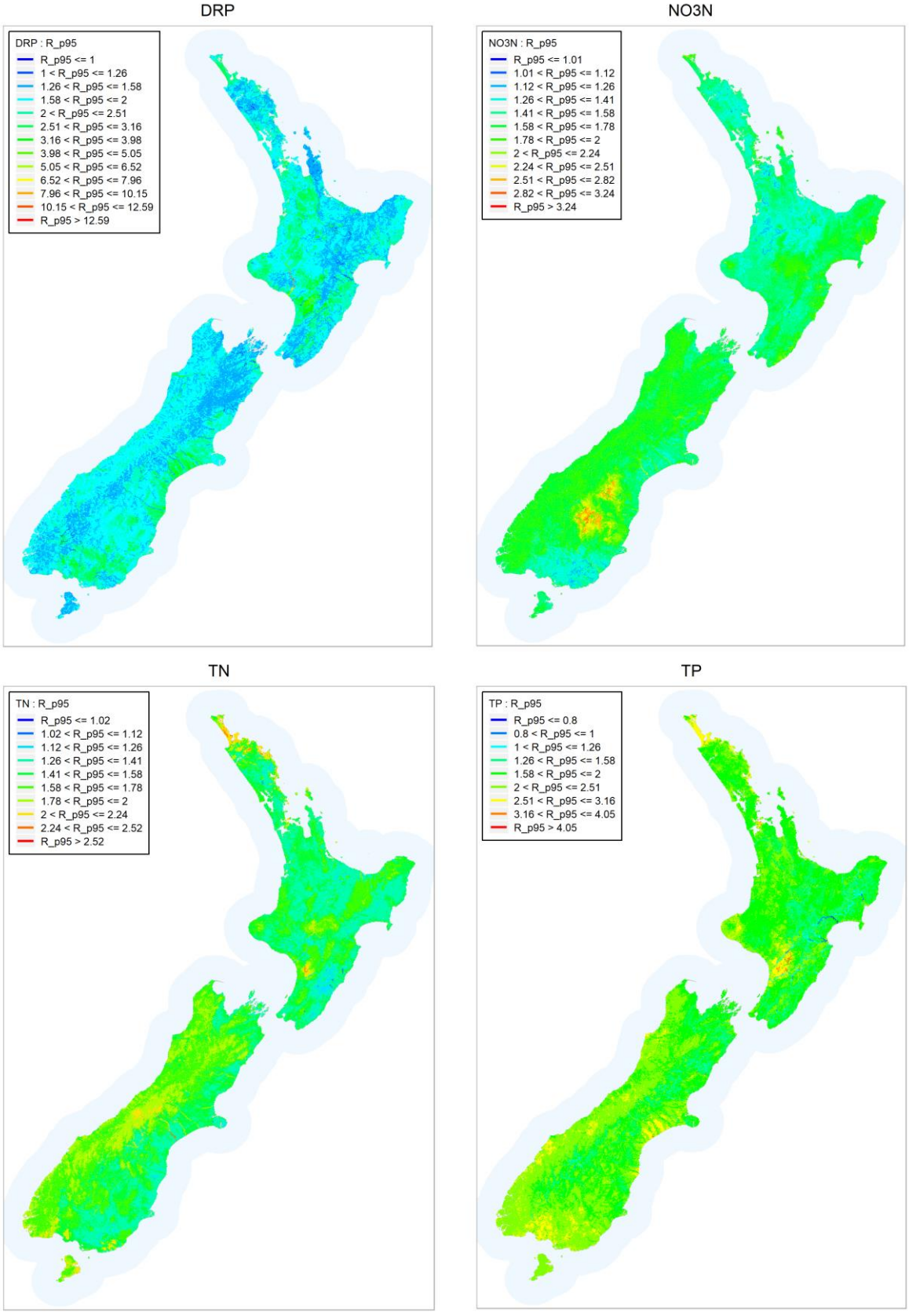

Figure 17: Maps of predicted R_{p95} values from catchment characteristics model. Maps show all rivers from order 1 and above.

5.6 Distributional characteristics model

5.6.1 Model predictors

The values of each of the seven predictor variables that were used as predictors in the distributional characteristics model are mapped in Figure 18. These predictors describe characteristics of flow, concentration and their individual and joint distributions.

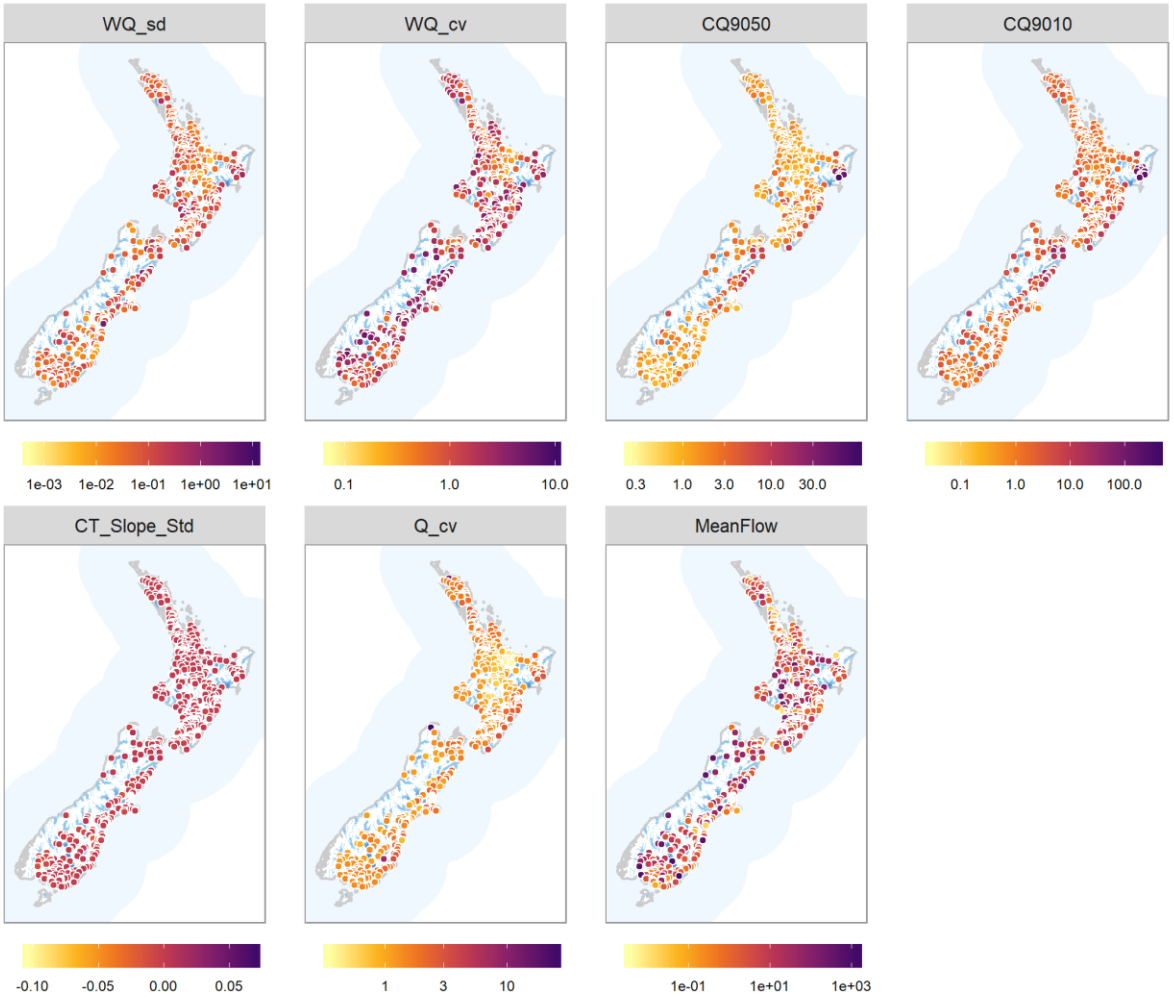

Figure 18: Maps of the predictors used in the distributional characteristics model.

5.6.2 Model performance

The performance of the distributional characteristics model is summarised in Figure 19 and Table 10. There were four models that was determined to be “very good” (TN, NO3N and TP R_{median} and TN R_{mean}) and one to be “good” (NO3N and TN R_{mean}). Three models were satisfactory (DRP R_{median} and TP and DRP R_{mean}) . All models for R_{p95} were classified as “unsatisfactory”. All models were an improvement on both the global and catchment characteristics models in terms of characteristic errors (i.e., RMSD and MAE).

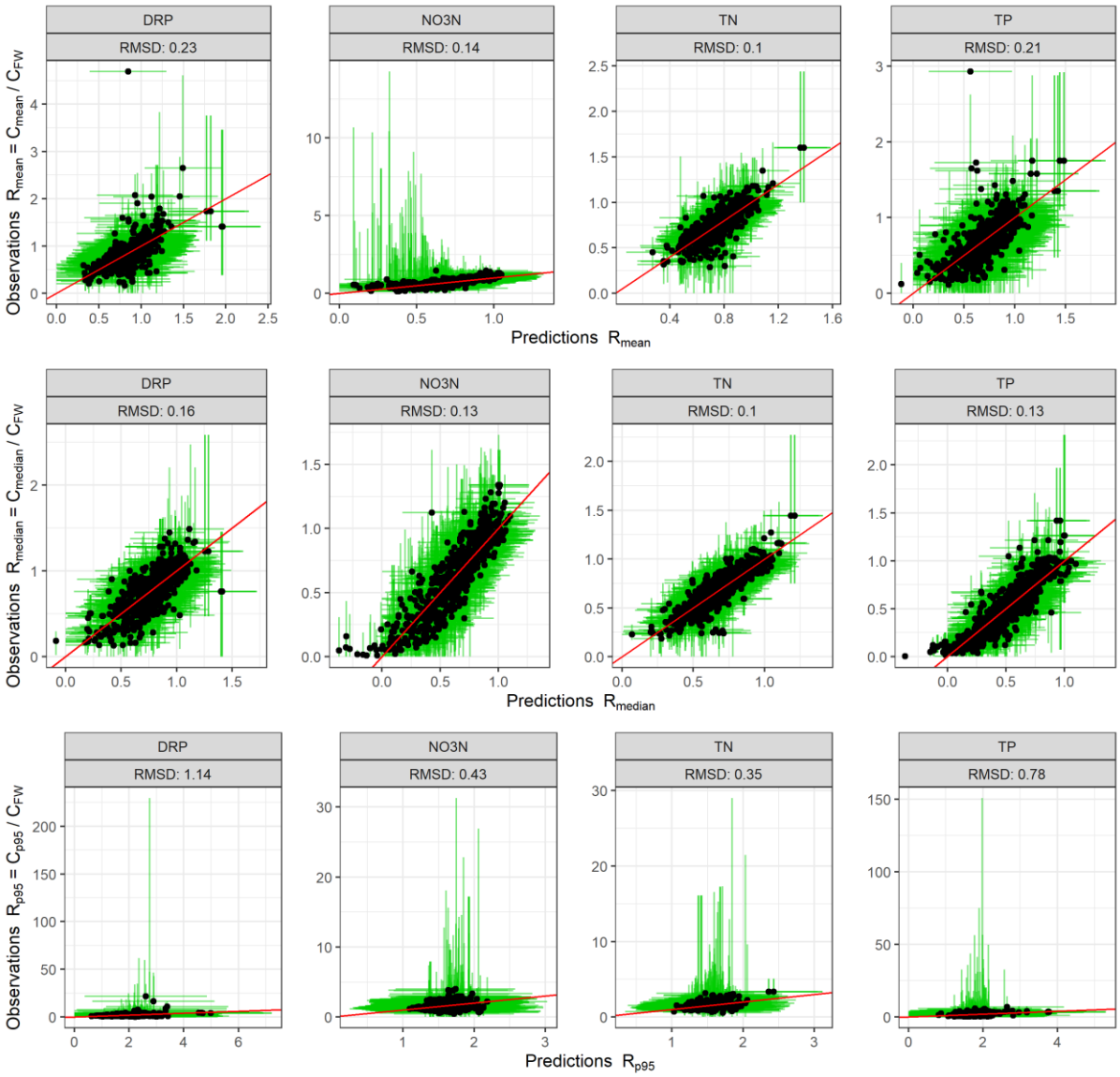

Figure 19: Scatter plots of the predicted R (mean, median or p95) from the distributional characteristics model versus observed R (C/C_{FW}). Red line shows 1:1 line as a comparison point. Green lines indicate the 95% confidence intervals of the observations and predictions (based on prediction +/- 1.96xRMSD).

Table 10. Performance of the catchment characteristics model of R (mean, median or p95). Performance was determined using independent predictions (i.e., sites that were not used in fitting the models) generated from the cross-validation procedure. 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, MAE = median absolute error.

	Nutrient variable	R^2	NSE	PBIAS	RSR	RMSD	RMAE	MAE
R_{mean}	DRP	0.41	0.41	-0.43	0.77	0.232	0.30	0.09
	NO3N	0.63	0.63	0.30	0.61	0.142	0.38	0.09
	TN	0.71	0.71	0.02	0.54	0.101	0.30	0.05
	TP	0.53	0.53	0.04	0.69	0.205	0.37	0.11
R_{median}	DRP	0.57	0.57	0.29	0.66	0.160	0.40	0.10
	NO3N	0.79	0.79	0.73	0.46	0.134	0.29	0.08
	TN	0.79	0.79	0.07	0.46	0.096	0.28	0.06
	TP	0.80	0.80	0.53	0.45	0.127	0.27	0.08
R_{p95}	DRP	0.16	0.15	-0.98	0.92	1.143	0.25	0.32
	NO3N	0.11	0.11	0.66	0.94	0.432	0.51	0.23
	TN	0.22	0.22	-0.26	0.88	0.351	0.50	0.19
	TP	0.11	0.10	-0.62	0.95	0.782	0.55	0.46

5.6.3 Modelled relationships

Figure 20 presents a summary of the direaction and magnitude of influence of the predictors on the response of $R_{statistic}$ by nutrient variable. The bar chart shows the magnitude and direction of the regression coefficients for $R_{staistic}$ when the predictors have been standardised to have mean of zero and variance of one. By doing this, the regression coefficients can be interpreted as the relative sensitivity of the response (i.e. $R_{statistic}$) to each of the predictors. When the regression coefficients are negative, this indicates that $R_{statistic}$ decreases as the predictor variable increases, and when they are positive, $R_{statistic}$ increases as the predictor variable increases.

For R_{mean} the most sensitive predictor variable is CQ9010. This predictor describes the overall steepness of the C-Q rating curve. For all nutrient variables, R_{mean} decreases as CQ9010 increases. Large values of CQ9010 are associated with sites where most of the load is associated with high flows. R_{median} also has a strong dependency on CQ9010, although for the nitrogen variables, particularly NO3N, CQ9050 also plays an important role in defining R. CQ9050 help to distinguish those sites where the concentration data indicates dilution occurs at high flows. When CQ9050 is smaller, this can indicate that dilution is occurring at higher flows, and this is associated with larger values of R_{median}. The coefficient of variation of the concentrations (WQ_cv) is also important for both R_{median} and R_{p95}, although the directions of the influence differ: increasing WQ_cv is associated with decreasing R_{median}, and increasing R_{p95}. This indicates that as the WQ_cv increases, the component of load associated with high flows generally increases (in most cases concentration increases with flow) and the 95th percentile value also increases, whereas the median concentration is not strongly correlated

with WQ_cv. Hence for R_{median}, increasing WQ_cv leads to an increase in the denominator of equation 1a, and a corresponding reduction in R_{median}. Whereas, for R_{p95}, WQ_cv influences both the numerator and denominator, and for highly skewed distributions, this tends to lead to increases in R_{p95} with increases in WQ_cv.

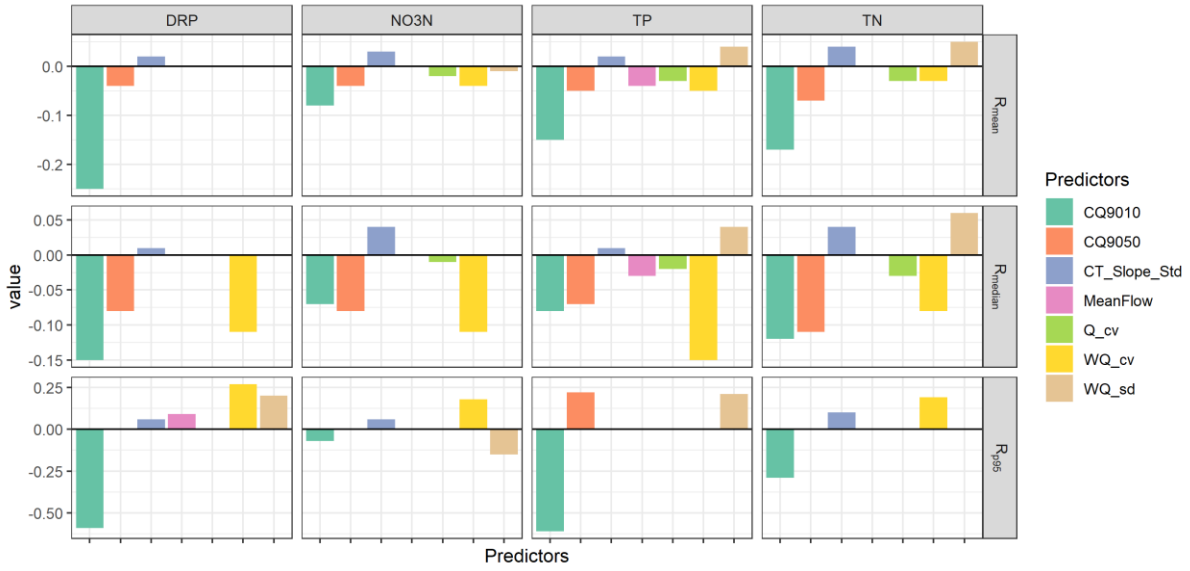

Figure 20: Summary of sensitivity of R to distributional characteristics model predictors. Note sensitivities are based on regression coefficients using variables that have been transformed and standardised as described in the text.

5.7 Summary of model coefficients

Table 11 and Table 12 summarise the model parameters for the global model (i.e. the median of R values, as used in equation 1a) and the distributional characteristics model (to use in equation 12 and subsequently equation 1a). Values of R (R_{mean}, R_{median} and R_{p95}) for each nutrient variable are available as a table of predictions from the RF model associated with every segment of the digital river network. The 95% confidence intervals for the estimates of R (from any one of the three methods) can obtained as R ±1.96 x the model characteristic error (as defined by the RMSD in Table 13). This approach is relatively coarse, yet simple to apply. For the random forest models prediction errors could potentially be derived that varied spatially, but this is beyond the scope of this piece of work.

Table 11: Summary of the Global model prediction of $R_{statistic}$

Nutrient variable	R_{mean}	R_{median}	R_{p95}
DRP	0.882	0.785	1.529
NO3N	0.721	0.633	1.449
TN	0.764	0.683	1.371
TP	0.732	0.517	1.596

Table 12: Summary of regression coefficient ($\beta_{1,2,..8}$ for equation 12) for the distributional characteristics model

		β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
		(Intercept)	log10_WQ_sd	log10_WQ_cv	log10_CQ9050	log10_CQ9010	log10_Q_cv	log10_MeanFlow	CT_Slope_Std_T
R_{mean}	DRP	0.955	NA	NA	-0.192	-0.551	NA	NA	0.468
	NO3N	0.845	-0.019	-0.141	-0.163	-0.166	-0.096	NA	0.629
	TN	0.945	0.061	-0.093	-0.284	-0.367	-0.159	NA	0.744
	TP	0.957	0.056	-0.162	-0.238	-0.339	-0.171	-0.040	0.319
R_{median}	DRP	0.768	NA	-0.342	-0.344	-0.324	NA	NA	0.216
	NO3N	0.745	NA	-0.346	-0.334	-0.144	-0.055	NA	0.816
	TN	0.820	0.072	-0.274	-0.491	-0.264	-0.142	NA	0.737
	TP	0.745	0.051	-0.481	-0.308	-0.179	-0.129	-0.030	0.219
R_{p95}	DRP	2.490	0.258	0.860	NA	-1.292	NA	0.096	1.133
	NO3N	1.617	-0.189	0.595	NA	-0.149	NA	NA	1.063
	TN	1.789	NA	0.606	NA	-0.625	NA	NA	1.858
	TP	2.217	0.267	NA	0.949	-1.339	NA	NA	NA

5.8 Comparison of performance of the different models

The characteristic errors of all models, represented by the RMSD and RMAE, are show in Table 13. For all statistic and nutrient variable combinations, the global model had the largest RMSD and the distributional characteristics model had the lowest RMSD. To determine whether these differences were statistically significant, we compared the residuals of each pair of models using the F-test, which compares the variances of the two distributions. For all possible pairs, we found *p*-values <0.01, indicating that there were statistically significant differences between the variances. The pattern of highest to lowest characteristic error associated with the global, catchment characteristic and distributional characteristics models was observed also for the RMAE statistic. We note that the RMAE is lower for some of the R_{p95} models. This is largely due to the more skewed distributions of R_{p95} (Figure 9), which does not strongly influence the magnitude of the MAE, but does lead to larger RMSD values and standard deviations (which were used to standardise the MAE and RMSD values to define their relative versions).

Table 13:Summary table of RMSD and RMAE values for alternative models of R. Colour scales are unique to each performance measure and indicate worst performance (red) to best performance (green).

Nutrient variable		RMSD				RMAE		
		Global	Catchment Characteristics	Distributional Characteristics		Global	Catchment Characteristics	Distributional Characteristics
R_{mean}	DRP	0.3	0.28	0.23		0.49	0.39	0.3
	NO3N	0.23	0.19	0.14		0.76	0.48	0.38
	TN	0.18	0.14	0.10		0.7	0.44	0.3
	TP	0.3	0.23	0.21		0.63	0.39	0.37
R_{median}	DRP	0.24	0.20	0.16		0.68	0.49	0.4
	NO3N	0.29	0.22	0.13		0.76	0.51	0.29
	TN	0.2	0.16	0.10		0.73	0.47	0.28
	TP	0.28	0.18	0.13		0.77	0.39	0.27
R_{p95}	DRP	1.31	1.32	1.14		0.23	0.26	0.25
	NO3N	0.47	0.42	0.43		0.52	0.47	0.51
	TN	0.39	0.33	0.35		0.46	0.47	0.5
	TP	0.86	0.81	0.78		0.47	0.55	0.55

5.9 Example information

In this section we present a worked example for a case study site. We consider nutrient concentration criteria for TN and DRP to achieve a periphyton biomass objective and derive target loads to achieve these criteria. The nutrient concentration criteria used are defined by Snelder et al. (2019). These criteria differ depending on the REC Source-of-flow class of the target sites. This example is based on site EW-00060 which belongs to the "Cool-Wet Lowland (CWL)" Source-of-flow class. We chose nutrient concentration criteria to achieve the periphyton biomass threshold of 200 mg chlorophyll m^{-2} with a spatial exceedance criteria of 20% (see Snelder et al. (2019) for details). Table 14 summarises the concentration criteria and current concentrations (both expressed as median concentrations).

Target loads to achieve these criteria will be derived as flow weighted concentrations (C_{FW}), following equation 2. The current values of C_{FW} for the site, calculated as part of this study, are included in Table 14 for reference, but it is noted that this information may not necessarily be available in a typical application.

Table 14: Concentration criteria and observed median concentrations and loads (as CFW) for both TN and DRP at the case study site.

Nutrient variable	Concentration criteria $C_{criteria}$ (mg L^{-1})	Current Median Concentration, C_{median} (mg L^{-1})	Current C_{FW} (mg L^{-1}) (95% C.I. in brackets)
DRP	0.0691	0.09	0.089 (0.088-0.092)
TN	0.351	2.115	2.39 (2.30-2.49)

For the same site, we also present the method to estimate median concentrations (and their uncertainties) given an estimated site load (expressed as a flow weighted concentration), using equation 3. For the example we convert the current C_{FW} to a median concentration using the estimates of R and compare against the observed current median concentration.

5.10 Estimates of R using models

For each model we evaluated R for the site and estimated the uncertainty as the R±1.96xRMSD. For cases where the lower confidence interval of R<0, we recommend using the smaller of half the estimated R value, or the minimum observed R value from the testing dataset (Table 15).

Table 15: Minimum observed R values by statistic type and nutrient variables

Nutrient Variable	R_{mean}	R_{median}	R_{p95}
DRP	0.16	0.13	0.26
NO3N	0.13	0.01	0.45
TN	0.29	0.18	0.55
TP	0.11	0.01	0.25

5.10.1 Global model

Predictions from the global model of R_{median} for the example site are taken from Table 11, and uncertainties for these predictions are estimated based on the RMSD values summarised in Table 13.

Table 16: Estimates of R_{median} from the global model for the case study site, including uncertainties.

Nutrient variable	R_{median}	95% Confidence intervals
DRP	0.79	0.31 - 1.26
TN	0.68	0.29 - 1.07

5.10.2 Catchment characteristics model

Predictions of R_{median} for the case study site are taken from lookup tables for all segments (outputs from the random forest models), and uncertainties for these predictions are estimated based on the RMSD values summarised in Table 13.

Table 17: Estimates of R_{median} from the catchment characteristics model for the case study site, including uncertainties

Nutrient variable	R_{median}	95% Confidence intervals
DRP	0.96	0.59 - 1.33
TN	0.89	0.59 - 1.20

5.10.3 Distributional characteristics model

In order to use the distributional characteristics models, paired C-Q observations and statistics derived from the long-term distribution of flows at the site. This information is summarised for the case study site in Figure 21, Figure 22 and Figure 23. The inputs required for the model regression equations are shown in Table 18, which, when combined with the regression coefficients in Table 12, can be used to derive the predictions of R_{median}, presented in Table 19. Uncertainties for these predictions are estimated based on the RMSD values summarised in Table 13.

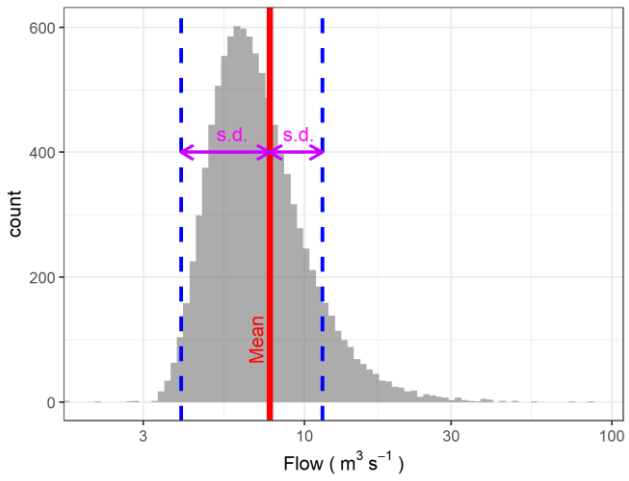

Figure 21: Histogram of flow at the case study site. The red line indicates the mean flow and the blue dashed lines are plus and minus one standard deviation from the mean.

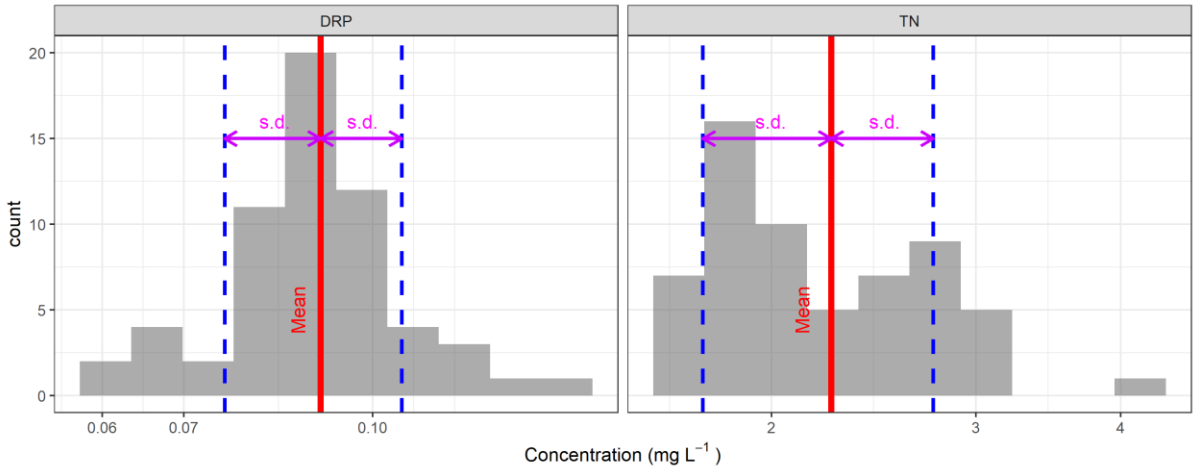

Figure 22: Histogram of concentrations at the case study site. The red line indicates the mean concentration and the blue dashed lines are plus and minus one standard deviation from the mean.

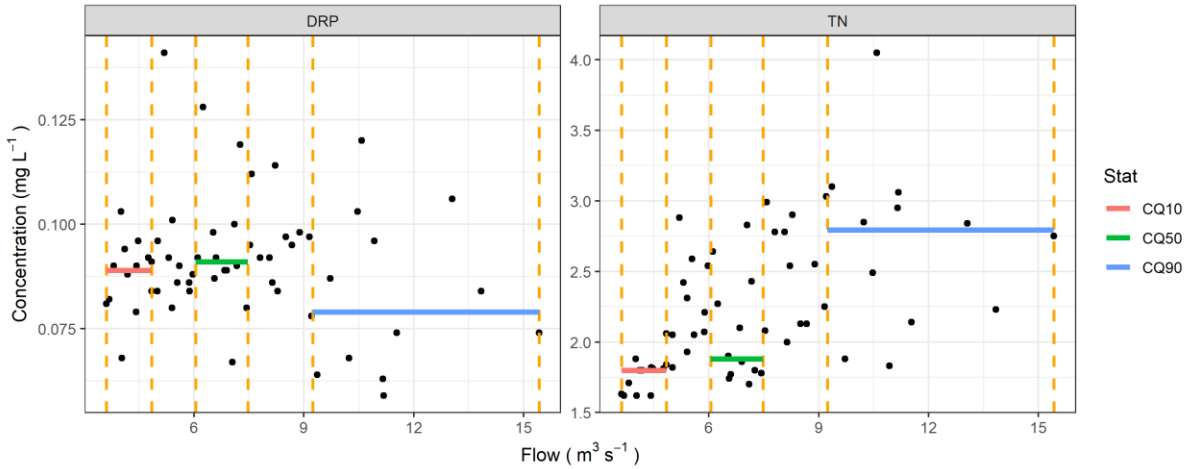

Figure 23: Joint flow-concentration relationship at the case study site. The orange dashed lines indicate the 0,20,40,60,80, and 100% quantiles of flow. The solid horizontal lines show the median concentrations associated with the three flow quantile ranges used to develop the ratios CQ9010, and CQ9050.

Table 18: Summary of input variables for the distributional characteristics models of R for the case study site. Values of the input variables are shown in their original units, as well as following transformation for use in equation 12.

	DRP		TN	
Predictor Name	**Value**	**Transformed value**	**Value**	**Transformed value**
WQ_sd	0.015	-1.823	0.506	-.295
WQ_cv	0.165	-0.780	0.225	-.648
CQ9050	0.868	-0.061	1.487	0.172
CQ9010	0.888	-0.052	1.553	0.191
CT_Slope_Std	-3 e-4	-0.017	1.2e-04	0.011
Q_cv	0.476	-0.322	0.476	-0.322
MeanFlow	7.682	0.886	7.682	0.886

Table 19:Estimates of R_{median} from the distributional characteristics model for the case study site, including uncertainties.

Nutrient variable	R_{median}	95% Confidence intervals
DRP	0.99	0.68 – 1.30
TN	0.97	0.78 - 1.15

5.11 Estimates of C_{FW} and required reductions in loads

The estimates of R_{median} and C_{FW} (i.e., the target load associated with the nutrient concentration criteria) are presented in Figure 24 and Figure 25, respectively. For comparative purposes, estimates of R and C_{FW} derived directly from the load and concentration data at the site are also presented.

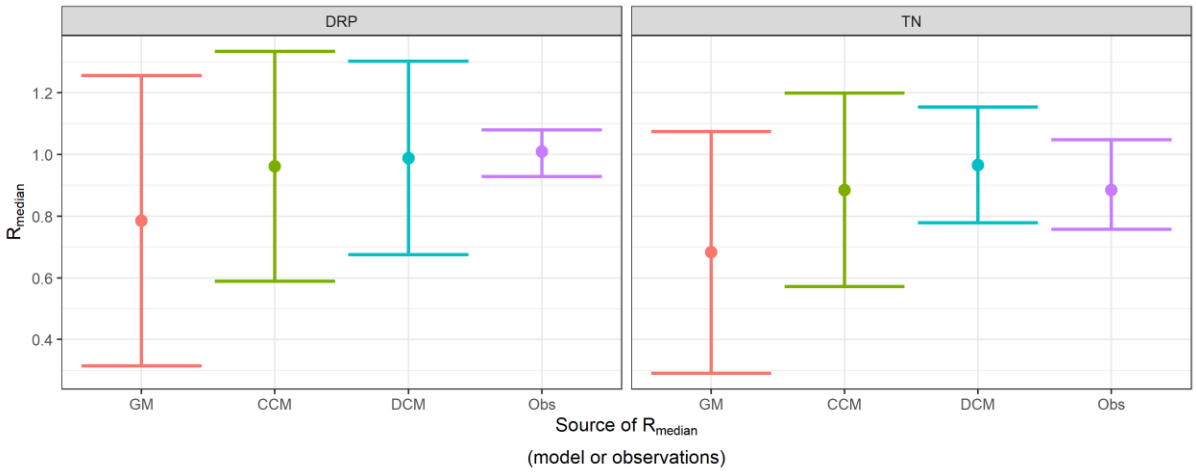

Figure 24: Estimates of R_{median} for the case study site based on the three alternative models (GM: global model; CCM: catchment characteristics model; DCM: distributional characteristics model) and on the observed load and median concentrations (Obs). The error bars indicate the 95% confidence intervals.

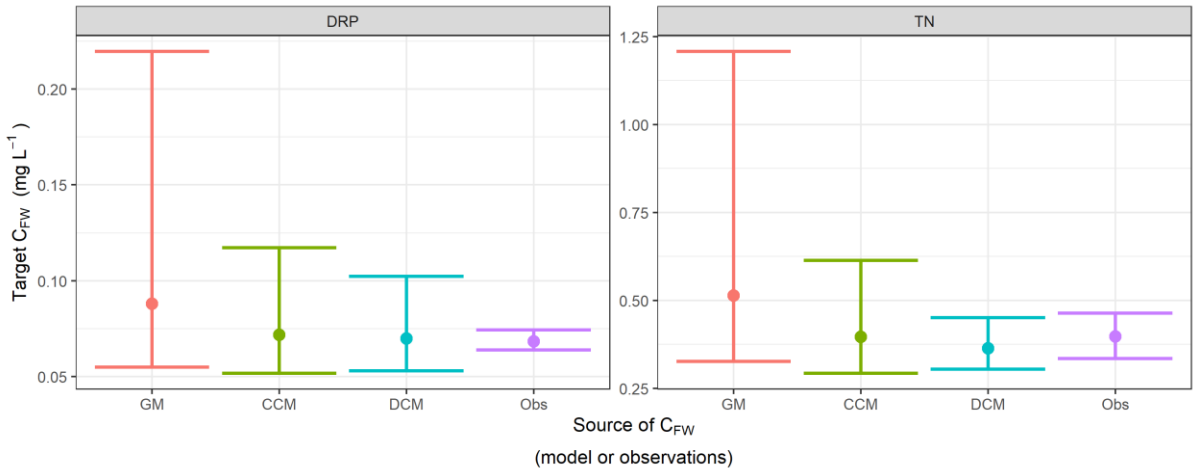

Figure 25: Estimates of target C_{FW} for the case study site to meet the concentration criteria in Table 14. C_{FW} is calculated following equation 1a and using R_{median} derived from each of the three models and the observed R_{median} value. The error bars indicate the 95% confidence intervals.

We note that there may be situations, such as scenario modelling, where the target load does not need to be known in absolute terms, it may be sufficient to know only the percentage reduction in load that is required to achieve the concentration criteria. In these situations, it is not necessary to estimate R and the percentage reduction can be estimated as follows:

$$LoadReduction\% = \frac{C_{median} - C_{criteria}}{C_{median}} \times 100 \qquad \text{(Equation 13)}$$

Where $LoadReduction\%$ is the required percentage reduction, C_{median} is the observed median load and $C_{criteria}$ is the concentration criteria. The $LoadReduction\%$ to achieve the concentration criteria shown in Table 14 for the case study site and calculated using equation

13 is summarised in Table 20. The uncertainty of $LoadReduction\%$ is calculated from the uncertainty of C_{median} (as described in section 4.2.3).

Table 20: Summary of percentage reductions in loads required to achieve the concentration criteria for the case study site.

Nutrient variable	$C_{criteria}$ (mg L^{-1})	C_{median}(mg L^{-1})	LoadReduction%(%)
DRP	0.0691	0.09 (0.086-0.092)	23% (20-25%)
TN	0.351	2.115 (1.90-2.42)	83% (82-85%)

5.12 Estimates of C_{median} given a load estimate

Assuming an example where we did not know C_{median}, but did have an estimate of C_{FW} it is possible to estimate the median concentration C_{median} following equation 1. Estimates of C_{median} based on R estimated from the three alternative models, along with the 95% confidence intervals are demonstrated in Figure 26

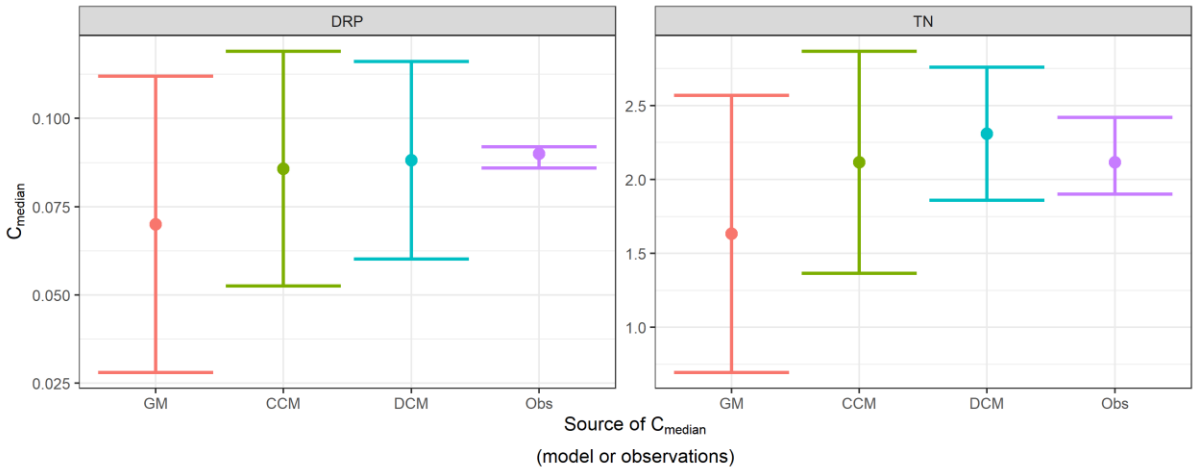

Figure 26: Estimates of C_{median} for the case study site C_{median} is calculated following equation 1 and using R_{median} derived from each of the three models. The "Obs" C_{median} value is the observed median value at the study site. The error bars indicate the 95% confidence intervals.

6 Discussion

This study confirms results of an earlier study (described in Oehler and Elliott, 2011) and shows that R does not generally have a value of unity and that it is spatially variable. This indicates that the target load for a catchment is not well estimated by taking the concentration criteria (expressed as a median concentration) and multiplying by a mean average flow. The models we have derived provide alternative, more accurate, methods for deriving target loads from concentration criteria, or alternatively, deriving a concentration given a load.

We have most confidence in the models associated with the median statistic (predictions of R_{median}) and the least confidence in the models associated the 95th percentile statistic (prediction of R_{p95}). This arises from the uncertainty of estimates of the statistics themselves, combined with the uncertainties in the load estimates. The 95th percentile value is imprecisely estimated from five years of monthly data and the mean is also subject to considerable uncertainty due to the leverage exerted by the most extreme observations. Although the 95th percentile value is used in regulations (e.g., it defines the attribute state for nitrate toxicity in the 2017 NPS-FM; NZ Government, 2017), its value is very uncertain when estimated from five years of monthly observations. In addition, the estimated confidence intervals for the 95th percentile values are themselves uncertain due to the small sample size. We also note that the methods that combine the uncertainties for C_{FW} and $C_{statistic}$ to evaluate the uncertainty of $R_{statistic}$ assume that the component errors are uncorrelated. This assumption potentially leads to an overestimate of the true uncertainty for R.

We defined three types of models for predicting R that provide differing compromises between ease of application and uncertainty. The global models simply assign a constant value of R for each nutrient variable and R statistic (i.e., R_{mean}, R_{median} and R_{p95}). These models are simple to use and require no additional detail about the catchment or distribution of flow and concentration, however, the global models have the greatest uncertainty.

The catchment characteristics models are based on catchment information that is associated with the REC digital drainage network. These models have national coverage; however, the estimates of R must be taken from a lookup table, rather that calculated by the user. The catchment characteristics models have benefits compared to the global model in terms of both accuracy and precision.

The distributional characteristics models require paired observations of flow and concentration at a site as well as information about the long-term flow distribution. The model is based on a linear regression and is expressed as a formula that enables R to be calculated by the user. This approach to the estimation of R has the least uncertainty but the data requirements mean this method can only be applied at sites with water quality and continuous flow data. It is noted that R can be estimated directly at sites with water quality and continuous flow data. The primary aim of the distributional characteristics model was to identify the characteristics of the flow, concentration and joint flow concentration distributions that most significantly influence R, and hence to provide some insights in the mechanisms that cause variation in R.

From the global model, we observe that the majority of site and nutrient variable combinations have values of R_{median} that are less than one. This implies that studies where load targets have been set based on a median concentration limit multiplied by the mean flow are likely to under-estimate the load target that will achieve the concentration criteria (i.e. they are overly restrictive, or conservative). This is particularly the case for TP, which has a median R_{median} of 0.55. The relationships defined by both the catchment characteristics and distributional characteristics models indicate that large R_{median} values tended to be associated with locations

with baseflow-dominated or stable flow regimes (e.g., spring fed rivers of the Canterbury plains and Bay of Plenty), where dilution at high flows occurs. In these types of rivers, setting load targets based on a median concentration limit multiplied by the mean flow will likely overestimate the load target (i.e., the load target will not be sufficiently low to achieve the concentration criteria).

A significant assumption associated with the use of R is that it is constant at a site over a range of loads and associated concentrations (i.e., it is assumed the ratio of loads to concentration stays the same at a site when the load changes). This study did not aim to test the robustness of this assumption, but the models do indicate that between-site variation in R is primarily associated with hydrological variation. For example, the most important predictors in the catchment characteristics models were hydrological characteristics (Figure 14). If actions that change nutrient loads are unlikely to significantly change hydrological regimes, it seems reasonable to assume that R is a constant at a site over a range of contaminant loads. For TN and NO3N, the proportion of stock units attributable to dairy cows was also an important predictor; as this proportion increases so does R. This might indicate that the source of the nutrient or aspects of nutrient supply from the landscape to the river are important determinants of R. Investigating this further was outside the scope of this study but this seems to be a relevant research question given that changes in nutrient loads are often associated with changes in types of stock and farm systems. The sensitivity analysis of the distributional characteristics model (Figure 20) indicates that R is most influenced by the shape of the C-Q rating curve. The C-Q rating curve is the outcome of the flow regime interacting with the source of contaminants. This suggests that the assumption that R is constant at a site requires considering whether the flow-concentration relationships might change in response to management actions that change contaminant loads. No change to the flow-concentration relationship may be a reasonable assumption if actions are not changing the source of the nutrient or aspects of nutrient supply from the landscape to the river. However, changes such as the removal of a point source contribution, or extensive edge of field mitigation measures, might have significant effects on C-Q rating curves, and consequently may change R at a site. We note that all our models have large uncertainty bounds for R estimates. This means that if variation in R at a site is associated, at least partly, with the supply of contaminants and not just hydrological regimes, at-site changes in R may be within the range of uncertainties of our estimates of R.

Our methods for estimating R will have application in studies that use catchment models based on annual loads (e.g., the CLUES model; Elliott et al., 2016). Our models can be used to translate the predictions of loads by these models to concentrations. This transformation of loads to concentration will generally be a requirement of the process of developing options for objectives and limits; and this was the reason for the original development of R by Oehler and Elliott (2011). However, in objective and limit setting processes there will often be a known current concentration and an estimated current load. In addition, there will often be one or more prescribed concentration criteria that are associated with possible objectives. In these circumstances, the question will be: “what is the load reduction that will achieve the concentration criteria?”. The analysis may often be facilitated by assuming that the required percentage change in the load is equivalent to the percentage change in the concentration statistic (to go from the current concentration to the criteria). This analysis does not require knowing R but does assume that R is constant at a site over a range of loads and associated concentrations.

Acknowledgements

We thank the people who assisted in the preparation of the national water quality dataset including Abi Loughnan at LAWA and many regional council staff who provided additional data and information about monitoring sites. We thank Graham Bryers (NIWA) for assistance with NRWQN data. We thank NIWA for the TopNet hydrological model predictions of daily flows at water quality monitoring sites. Thank you also to Sandy Elliott for reviewing this report.

References

Akaike, H., 1973. Information Theory and an Extension of the Maximum Likelihood Principle. B. N. Petrov and F. Csaki (Editors). Springer Verlag, ed. Akademiai Kiado: Budapest., pp. 267–281.

Anastasiadas, S., S. Kerr, C. Arbuckle, S. Elliot, J. Hadfield, B. Keenan, R. McDowell, T. Webb, and R. Williams, 2013. Understanding the Practice of Water Quality Modelling. Motu Economic and Public Policy Research.

Baker-Galloway, M., 2013. The Inconsistent Regional Management of Farming Effects on Waterways.

Beven, K., 1993. Prophecy, Reality and Uncertainty in Distributed Hydrological Modelling. Advances in Water Resources 16:41–51.

Blyth, J., 2018. Water Quality Modelling of the Ruamahanga Catchment: Ruamāhanga Whaitua Committee - Scenario Modelling. Jacobs.

Blyth, J., L. Cetin, and S. Estaon, 2018. Water Quality Modelling of the Ruamahanga Catchment: Baseline Model Build and Calibration Report. Jacobs.

Booker, D.J. and T.H. Snelder, 2012a. Comparing Methods for Estimating Flow Duration Curves at Ungauged Sites. Journal of Hydrology 434:78–94.

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

Booker, D.J. and R.A. Woods, 2014. Comparing and Combining Physically-Based and Empirically-Based Approaches for Estimating the Hydrology of Ungauged Catchments. Journal of Hydrology 508:227–239.

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.

Cohn, T.A., 2005. Estimating Contaminant Loads in Rivers: An Application of Adjusted Maximum Likelihood to Type 1 Censored Data. Water Resources Research 41. http://onlinelibrary.wiley.com/doi/10.1029/2004WR003833/full. Accessed 21 Jan 2016.

Cohn, T.A., D.L. Caulder, E.J. Gilroy, L.D. Zynjuk, and R.M. Summers, 1992. The Validity of a Simple Statistical Model for Estimating Fluvial Constituent Loads: An Empirical Study Involving Nutrient Loads Entering Chesapeake Bay. Water Resources Research 28:2353–2363.

Cohn, T.A., L.L. Delong, E.J. Gilroy, R.M. Hirsch, and D.K. Wells, 1989. Estimating Constituent Loads. Water Resources Research 25:937–942.

Crawley, M.J., 2002. Statistical Computing: An Introduction to Data Analysis Using S-Plus. John Wiley & Sons Inc, Chichester, United Kingdom.

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.

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

Durney, P., J. Dodson, and N. Calder-Steele, 2016. An Integrated Hydrological Model for the Orari Plains, Canterbury. Environment Canterbury.

Efron, B., 1981. Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods. Biometrika 68:589–599.

Elliott, A.H., R.B. Alexander, G.E. Schwarz, U. Shankar, J.P.S. Sukias, and G.B. McBride, 2005. Estimation of Nutrient Sources and Transport for New Zealand Using the Hybrid Mechanistic-Statistical Model SPARROW. Journal of Hydrology (New Zealand) 44:1.

Elliott, A.H., A.F. Semadeni-Davies, U. Shankar, J.R. Zeldis, D.M. Wheeler, D.R. Plew, G.J. Rys, and S.R. Harris, 2016. A National-Scale GIS-Based System for Modelling Impacts of Land Use on Water Quality. Environmental Modelling & Software 86:131–144.

Fenemor, A., 2013. Marine Production Modelling and Economic Analysis: Prepared for ICM Programme; 2000-2011. Landcare Research.

Gassman, P.W., M.R. Reyes, C.H. Green, and J.G. Arnold, 2007. The Soil and Water Assessment Tool: Historical Development, Applications, and Future Research Directions. Transactions of the ASABE 50:1211–1250.

Graham, D.N. and M.B. Butts, 2005. Flexible, Integrated Watershed Modelling with MIKE SHE. Watershed Models 849336090:245–272.

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

Holdgate, M.W., 1980. A Perspective of Environmental Pollution. CUP Archive.

Julian, J.P., K.M. de Beurs, B. Owsley, R.J. Davies-Colley, and A.-G.E. Ausseil, 2017. River Water Quality Changes in New Zealand over 26 Years: Response to Land Use Intensity. Hydrology and Earth System Sciences 21:1149–1171.

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

Larned, S., T. 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.

McMillan, H.K., E.Ö. Hreinsson, M.P. Clark, S.K. Singh, C. Zammit, and M.J. Uddstrom, 2013. Operational Hydrological Data Assimilation with the Recursive Ensemble Kalman Filter. Hydrology and Earth System Sciences 17:21–38.

MFE, 2012. Land-Use and Carbon Analysis System: Satellite Imagery Interpretation Guide for Land-Use Classes (2nd Edition). Ministry for the Environment, Wellington, New Zealand.

MFE and StatsNZ, 2019. New Zealand's Environmental Reporting Series: Environment Aotearoa 2019. Available from Www.Mfe.Govt.Nz and Www.Stats.Govt.Nz. Ministry for the Environment & Statistics NZ, Wellington, New Zealand.

Ministry for Environment, 2017. National Policy Statement for Freshwater Management 2014 (Amended 2017). http://www.mfe.govt.nz/publications/fresh-water/national-policy-statement-freshwater-management-2014-amended-2017.

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.

Moriasi, D.N., M.W. Gitau, N. Pai, and P. Daggupati, 2015. Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. Transactions of the ASABE 58:1763–1785.

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.

Newsome, P., J. Shepard, D. Pairman, S. Belliss, and A. Manderson, 2018. Establishing New Zealand's LUCAS 2016 Land Use Map. Contract Report, Manaaki Whenua – Landcare Research, New Zealand.

NIWA, 2009. National River Water Quality Network (NRWQN). NIWA. https://www.niwa.co.nz/freshwater/water-quality-monitoring-and-advice/national-river-water-quality-network-nrwqn. Accessed 28 May 2019.

Norton, N. and D. Kelly, Current Nutrient Loads and Options for Nutrient Load Limits for a Case Study Catchment: Hurunui Catchment. NIWA Client Report.

NZ Government, 2017. National Policy Statement for Freshwater Management 2014 (Amended 2017).

Oehler, F. and A.H. Elliott, 2011. Predicting Stream N and P Concentrations from Loads and Catchment Characteristics at Regional Scale: A Concentration Ratio Method. Science of the Total Environment 409:5392–5402.

Olsson, U., 2005. Confidence Intervals for the Mean of a Log-Normal Distribution. Journal of Statistics Education 13:null-null.

Palliser, C., S. Elliot, S. Yalden, and U. Shankar, 2015. Waitaki Water Quality Catchment Modelling. NIWA.

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.

Roberts, A.H.C. and N. Watkins, 2014. One Nutrient Budget to Rule Them All–the OVERSEER® Best Practice Data Input Standards. Nutrient Management for the Farm, Catchment and Community, Occasional Report No 27.

Roygard, J. and K. McArthur, 2008. A Framework for Managing Non-Point Source and Point Source Nutrient Contributions to Water Quality: Technical Report to Support Policy Development. Horizons Regional Council.

Roygard, J.K.F., K.J. McArthur, and M.E. Clark, 2012. Diffuse Contributions Dominate over Point Sources of Soluble Nutrients in Two Sub-Catchments of the Manawatu River, New Zealand. New Zealand Journal of Marine and Freshwater Research 46:219–241.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, 2008. Global Sensitivity Analysis: The Primer. John Wiley & Sons.

Scott, L., 2013. Hinds Water Quality Modelling for the Limit Setting Process. Environment Canterbury.

Semadeni-Davies, A.F., S. Elliot, and S. Yalden, 2015. Modelling Nutrient Loads in the Waikato and Waipa River Catchments. Development of Catchment-Scale Models. NIWA.

Semadeni-Davies, A.F. and A. Sunil Kachhara, 2017. Te Awarua-o-Porirua (TAoP) Collaborative Modelling Project: CLUES Modelling of Rural Contaminants. NIWA.

Shepherd, M. and D. Wheeler, 2013. How Nitrogen Is Accounted for in OVERSEER® Nutrient Budgets. Accurate and Efficient Use of Nutrients on Farms. Occasional Report No 26.

Shepherd, M., D. Wheeler, D. Selbie, L. Buckthought, and M. Freeman, 2013. Overseer®: Accuracy, Precision, Error and Uncertainty. Currie, LD, and Christensen, CL, Accurate and Efficient Use of Nutrients on Farms, Massey University, Palmerston North:1–8.

Singh, R., T. Wagener, K. van Werkhoven, M.E. Mann, and R. Crane, 2011. A Trading-Space-for-Time Approach to Probabilistic Continuous Streamflow Predictions in a Changing Climate–Accounting for Changing Watershed Behavior. Hydrology and Earth System Sciences 15:3591–3603.

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

Snelder, T.H., S.T. Larned, and R.W. McDowell, 2018. Anthropogenic Increases of Catchment Nitrogen and Phosphorus Loads in New Zealand. New Zealand Journal of Marine and Freshwater Research 52:336–361.

Snelder, T.H., R.W. McDowell, and C.E. Fraser, 2017. Estimation of Catchment Nutrient Loads in New Zealand Using Monthly Water Quality Monitoring Data. JAWRA Journal of the American Water Resources Association 53:158–178.

Snelder, T.H., C. Moore, and C. Kilroy, 2019. Nutrient Concentration Targets to Achieve Periphyton Biomass Objectives Incorporating Uncertainties. JAWRA Journal of the American Water Resources Association.

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.

Trafford, G. and S. Trafford, 2011. Farm Technical Manual. Caxton Press, Christchurch, New Zealand.

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

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

Upton, S., 2018. Overseer and Regulatory Oversight: Models, Uncertainty and Cleaning up Our Waterways. Parliamentary Commissioner for the Environment.

Wade, A.J., C. Soulsby, S.J. Langan, P.G. Whitehead, A.C. Edwards, D. Butterfield, R.P. Smart, Y. Cook, and R.P. Owen, 2001. Modelling Instream Nitrogen Variability in the Dee Catchment, NE Scotland. Science of the Total Environment 265:229–252.

Wheeler, D., M. Shepherd, M. Freeman, and D. Selbie, 2014. OVERSEER® Nutrient Budgets: Selecting Appropriate Timescales for Inputting Farm Management and Climate Information. Nutrient Management for the Farm, Catchment and Community, Occasional Report No 27.

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 Details about stocking density calculations

The area occupied by different land uses is mapped nationally by the Land Use and Carbon Analysis System: LUCAS (Newsome *et al.*, 2018). LUCAS was developed for reporting and accounting of carbon fluxes and greenhouse gas emissions, as required by the United Nations Framework on Climate Change and the Kyoto Protocol (MFE, 2012). LUCAS maps include 12 land use classes, including plantation forests and grazed grassland, and pertain to four time periods, including 2016 which we used in this study. Land use classes were prepared using high resolution imagery from the SPOT-5 satellite (MFE, 2012). Auxiliary datasets were also used to differentiate and map the land use classes, in particular grazed grasslands, as described by (Newsome et al. 2018). Land use classes are mapped at a minimum mapping area of 1 ha and include natural and plantation forests and grazed grassland, which is differentiated into high and low producing subclasses.

The numbers of animals in four stock type categories (dairy cows, beef cows, sheep and deer) are collected on all enterprises involved in livestock farming by Statistics New Zealand as part of an annual agricultural production census (APC). We used these data to produce measures of the density of animals on land used for pastoral agriculture in the monitoring site catchments and further modified these to indicate the land use intensity. To produce these measures, we used the highest resolution versions of APC data that are publicly available, which are associated with a spatial layer comprising 960 hexagonal grid cells (35,000 ha) that cover all New Zealand (https://statisticsnz.shinyapps.io/livestock_numbers/). For each cell, we used numbers of animals in each stock type categories for the most recent census year (2017). We converted the stock numbers associated with the hexagonal grid and each census year into the density of animals of each stock type in every sub-catchment defined by a digital representation of New Zealand's river network in five steps. First, all areas categorized by LUCAS as grazed grassland (combining high and low producing subclasses) were intersected with the hexagonal grid. Second, the animals of each stock type in each hexagonal grid cell were evenly distributed over the grazed grassland within the cell and the animal numbers were converted to densities (number of animals per hectare of grazed grassland). The animal numbers in 2017 were distributed to the grazed grassland associated with the 2016 version of LUCAS. Third, the combined hexagonal grid cells and grazed grassland were intersected with sub-catchment boundaries defined by the digital drainage network. Fourth, the numbers of animals of each stock type in each sub-catchment were calculated as the sum of each area of grazed grassland within each sub-catchment multiplied by the corresponding animal densities. Fifth, the numbers of animals of each stock type upstream of each segment of the network were calculated by summing the animals (density multiplied by area) in that segment's sub-catchment and all upstream sub-catchments.

We then converted the animal numbers to 'stock units', which are a measure of metabolic demand by livestock that is commonly used in New Zealand (Parker, 1998) and used these values as indicators of land use intensity. The numbers of animals upstream of each segment of the river network were converted to SU density (SU ha^{-1}) by multiplying the numbers of animals of each stock type its SU equivalent (Table 21). The SU density was used to represent catchment land use intensity.

Table 21. Stock unit equivalents per animal for 2017. Derived from Parker (1998) and Trafford and Trafford (2011).

StockType	2017
Sheep	1.35
Beef	6.9
Dairy	8
Deer	2.3

Each version of LUCAS defines areas that are occupied by plantation forest. We used these data to calculate the proportion of catchment area occupied by plantation forest for each monitoring site in three steps. First we intersected the areas categorized as plantation forest with the sub-catchment boundaries of the digital drainage network. Second, the area of plantation forest in each sub-catchment was calculated for each sub-catchment. Third, the proportion area occupied by plantation forest upstream of each segment of the network was calculated by summing the area of forest in that segment's sub-catchment and all upstream sub-catchments and dividing by the total catchment area (%).

Appendix B Maps of concentration statistics

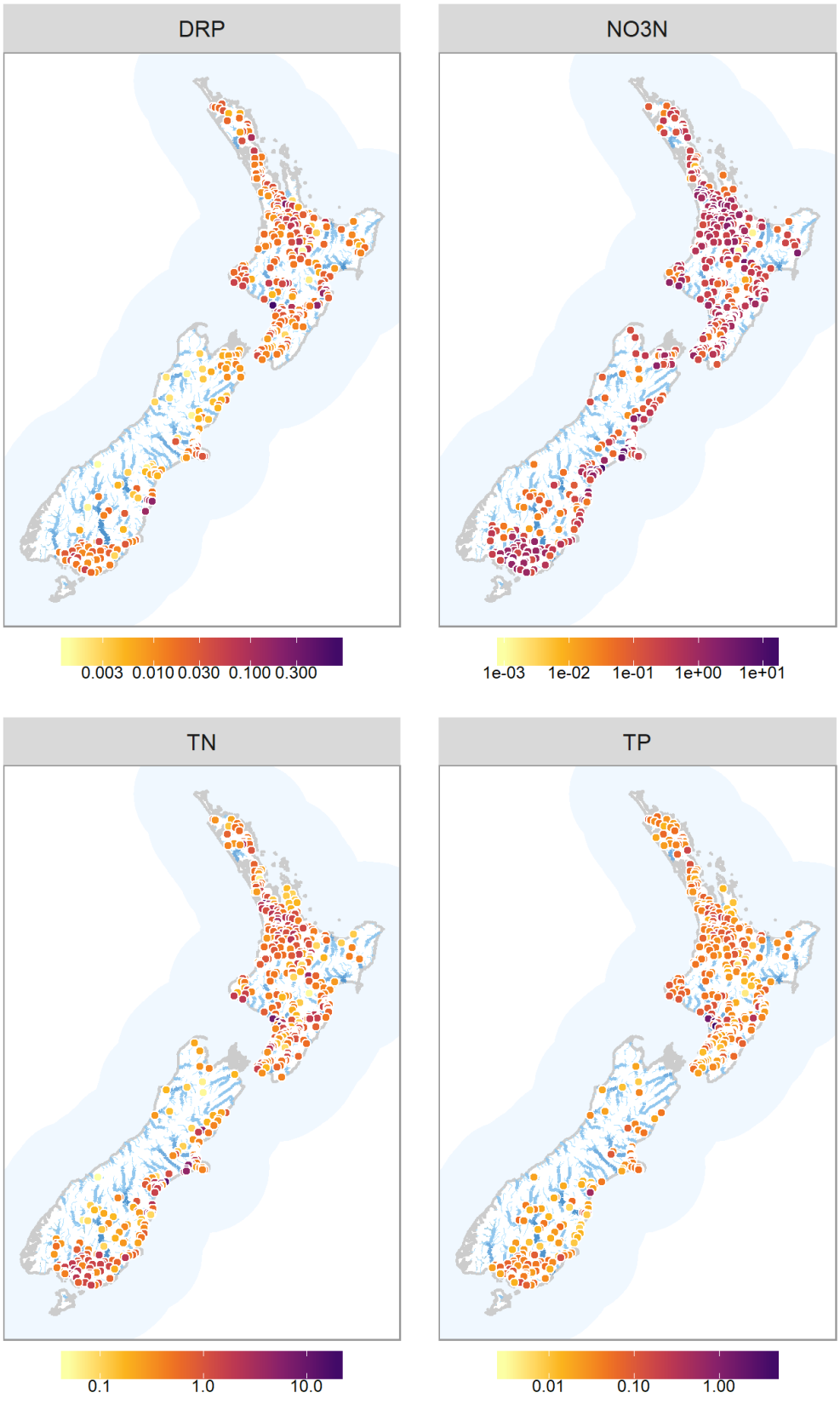

Figure 27: Maps of observed mean concentrations across observation sites

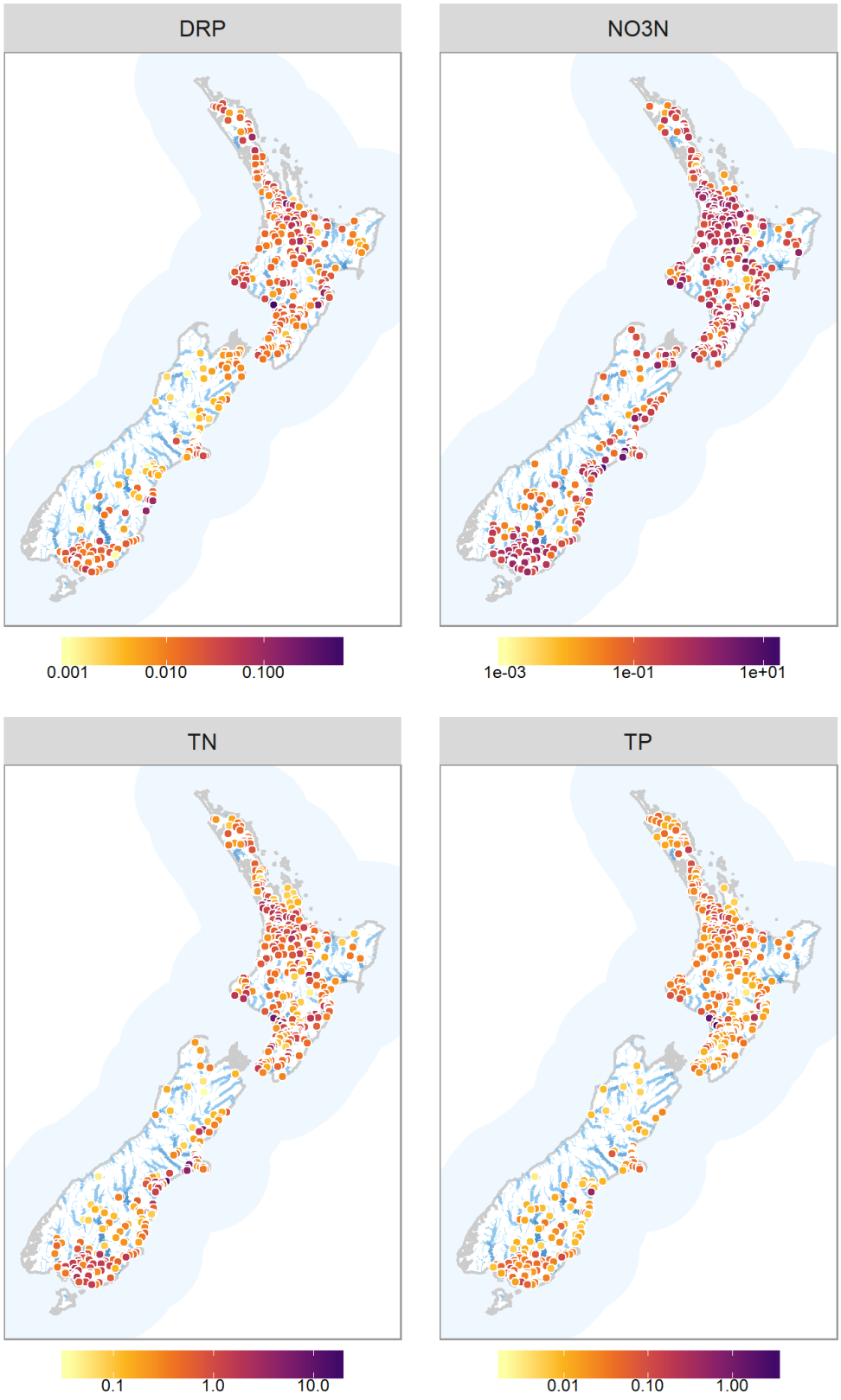

Figure 28: Maps of observed median concentrations across observation sites

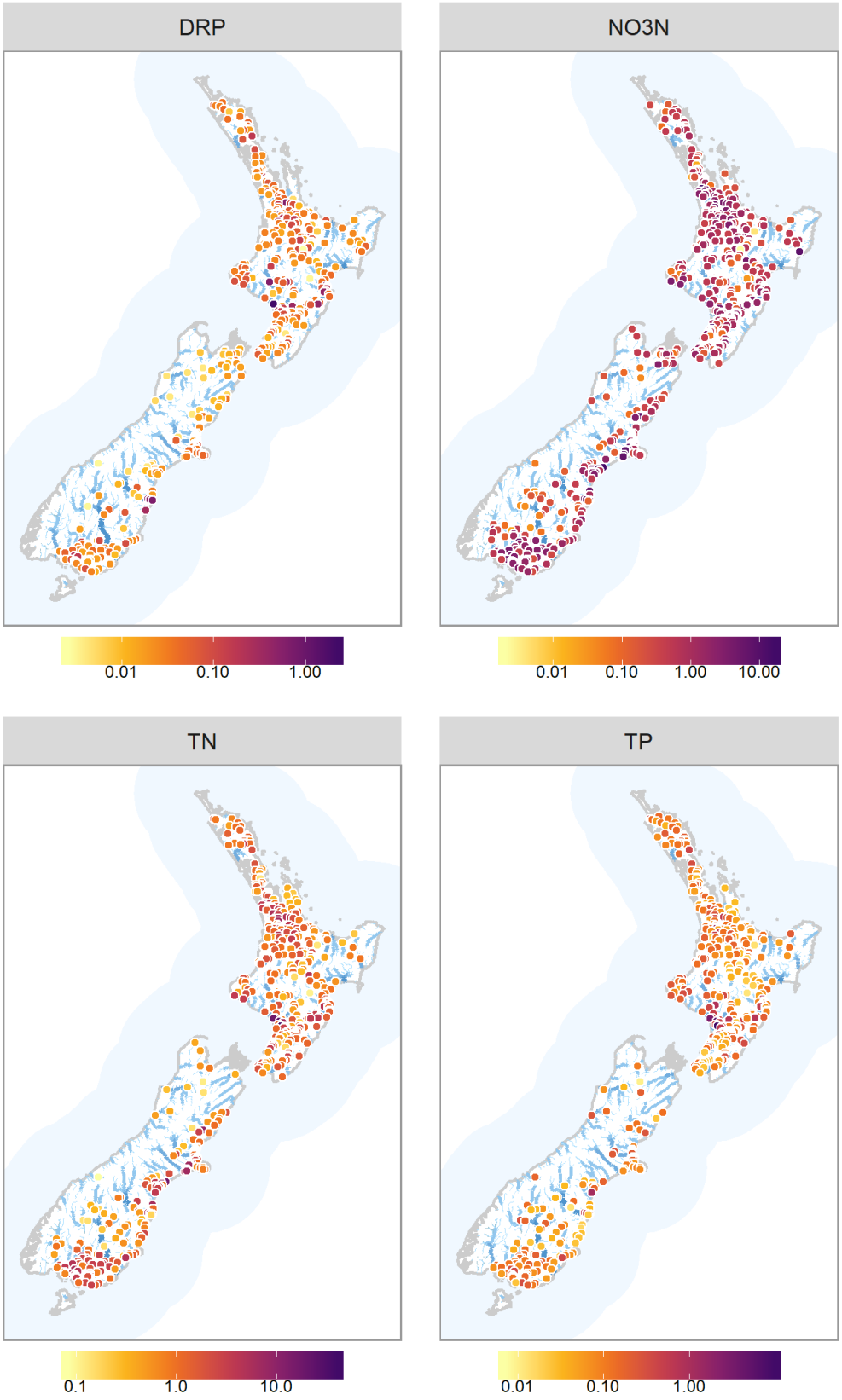

Figure 29: Maps of observed 95th percentile across observation sites

Appendix C Maps of calculated Loads

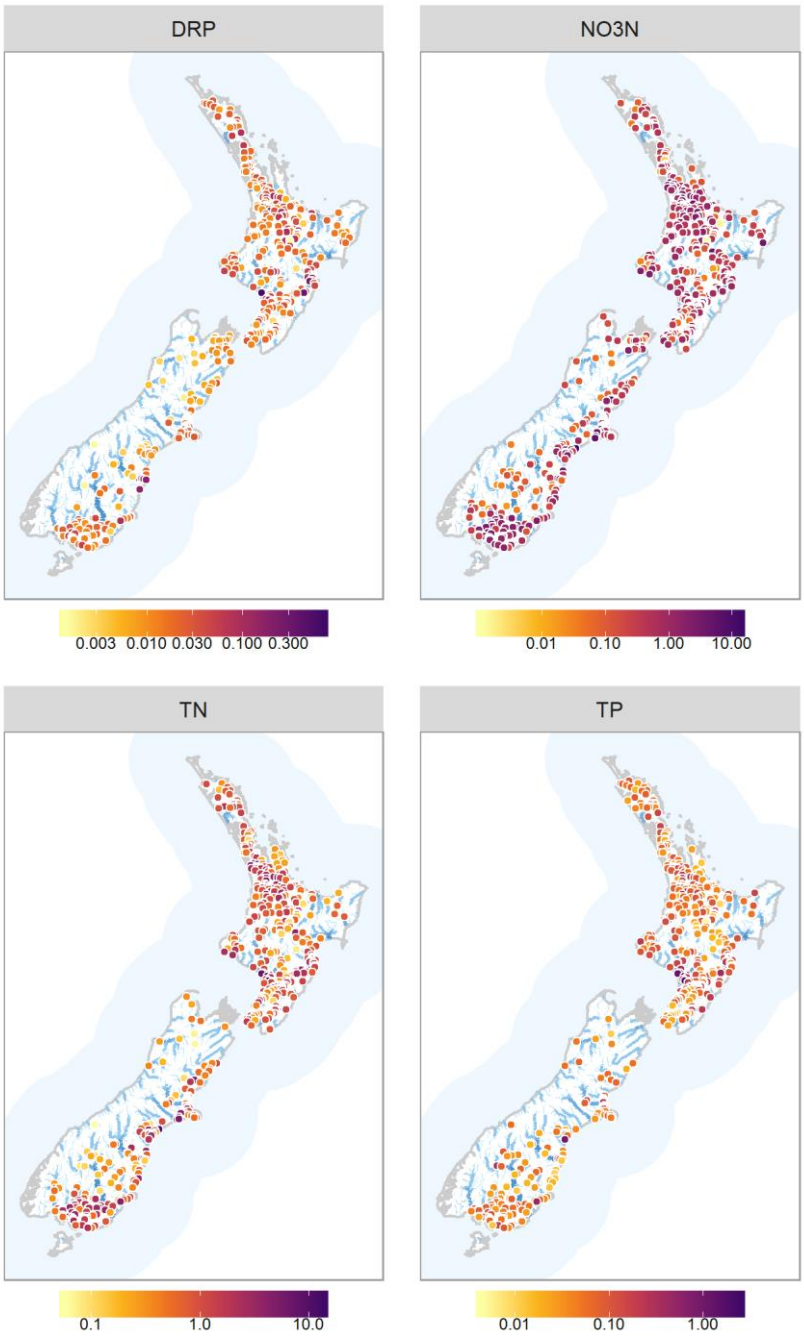

Figure 30: Maps of estimated loads normalised by mean flows, (flow weighted concentrations), across observation sites.

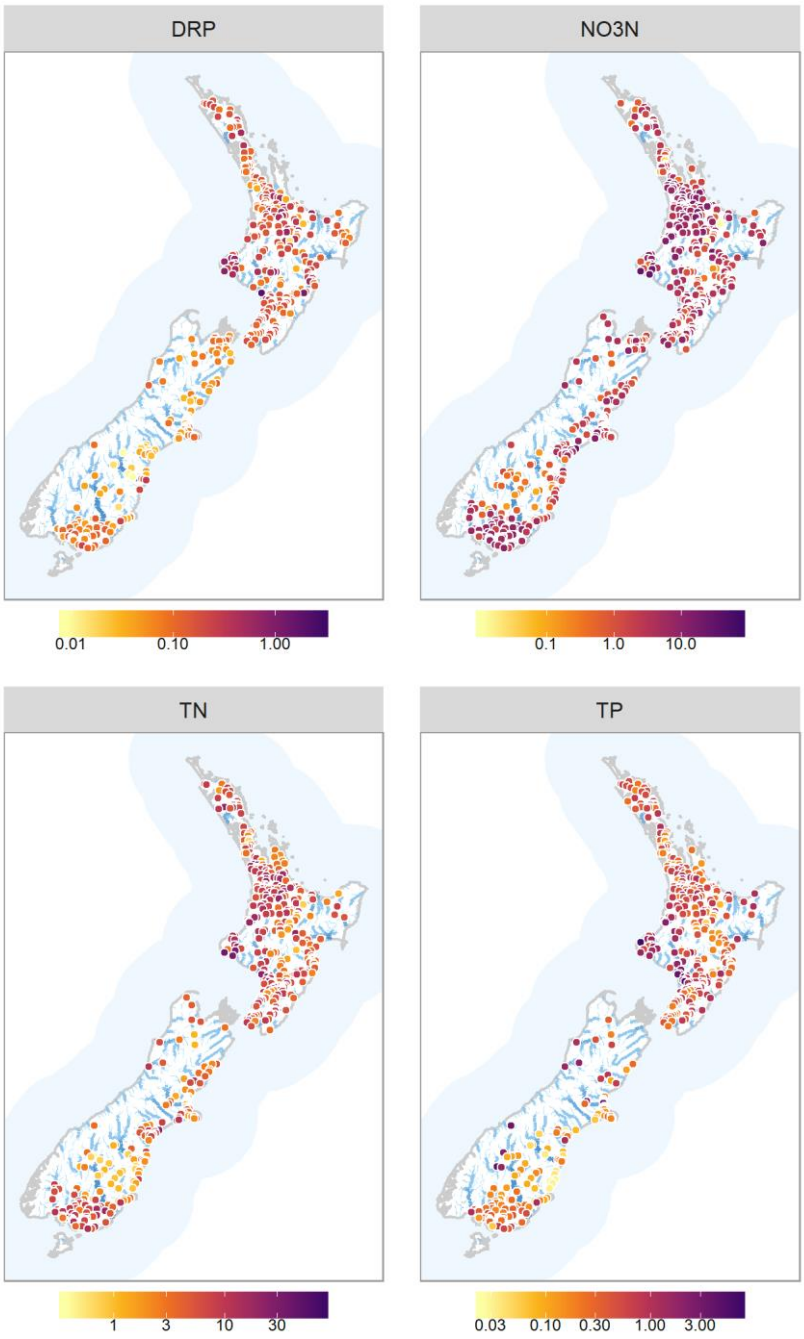

Figure 31: Maps of estimated loads normalised by catchment area, (export coefficients) across observation sites.