Quantifying Catchment Nutrient Modelling Parameters

An analysis using the available New Zealand data

May 2023

Prepared By:

Ton Snelder

Tim Cox

Caroline Fraser

Sandy Elliott

Tim Kerr

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:	2023-03
Report Date:	May 2023

Quality Assurance Statement

Version	Reviewed By	
Final	Simon Harris	

Table of Contents

Figures

Glossary

Term	Definition
Attenuation coefficient	The proportion of the export load that is lost between the sources and the instream observation point
CAEC	Catchment average export coefficient. Area weighted average of export coefficients over a catchment (kg ha^{-1} yr^{-1})
Catchment average export loads	Area weighted average of export coefficients over a catchment multiplied by catchment area (t yr^{-1})
DN2.4	Digital river network, version 2.4
Empirical catchment water quality model	Shortened to 'empirical model' in the report. These models use a purely empirical (statistical) approach to predict instream load or concentration at point in the drainage network based on the proportion of upstream catchment occupied by various Types defined by a typology. There is no attempt to represent either contaminant loss from land or attenuation so the predictions are purely data driven.
ETC	Empirical Type concentration. The expected instream concentration (mg m^{-3}), realised at a point in the drainage network (post attenuation), that is generated by diffuse losses associated with a specific Type that is defined by a typology. These values were estimated by fitting quantile regression models to median concentrations calculated for water quality stations.
ETY	Empirical Type yield. The expected annual load per unit area (kg ha^{-1} yr^{-1}), realised at a point in the drainage network (post attenuation), that is generated by diffuse losses associated with a specific Type that is defined by a typology. These values were estimated by fitting quantile regression models to annual loads calculated for water quality stations.
Export coefficient	Rates of diffuse nutrient loss from land (kg ha^{-1} yr^{-1}) to streams.
Factor	Environmental variables that are used to define Types in typologies including land use/cover, climate (e.g., annual rainfall) and topography (e.g., land slope).
Instream load	In this study, the in-river load calculated for a water quality site from infrequent (e.g., monthly) observations of concentration and daily flow (t y^{-1}).
Instream yield	Instream load standardised by (divided by) catchment area (kg ha yr^{-1}).

Term	Definition
Land use	Categorical description of land use. In this study a total of nine categories were used that include categories that, strictly speaking, are descriptions land cover (i.e., "Natural", "Bare" and "Water").
OLS	Ordinary least squares regression.
Process-based catchment water quality models	Models that represent two separate processes: contaminant loss from land and attenuation to produce a prediction of the instream load or concentration at downstream points in the drainage network. Examples that are discussed in this report are CLUES and SCAMP.
Type	A Type (i.e., a class of land) defined by a typology.
Typology	A system of Types that is used to classify and group land areas that are alike in terms of one or more characteristics of interest (e.g., nutrient loss rates, economic returns, response to management actions/mitigations). In this study, land areas belonging to Types defined by several typologies are considered be alike with respect to their diffuse source nutrient loss rates. These typologies are defined by subdivision of several factors into categories. The factors include land use or cover (sub-divided into categories such as dairy farms, native forest, urban) and environmental factors including soil moisture (sub-divided into categories such as wet, dry), land slope (sub-divided into categories such as flat, steep).

Abstract

Catchment nutrient (Nitrogen (N) and phosphorus (P)) models are integral to implementing the National Policy Statement – Freshwater Management (NPS-FM). Models are used to predict the N and P loads or concentrations at points in the drainage network under both existing catchment conditions and some possible future set of conditions associated with changed land use or management.

Analyses associated with NPS-FM implementation commonly use a class of catchment nutrient models that we call 'process-based' models. Process-based models explicitly represent the processes of contaminant loss in the catchment (source losses), transport to downstream receiving environments, and attenuation (i.e., reduction in loads between the point of discharge and downstream receiving environments by natural processes).

This study investigated how parameters representing diffuse source losses and attenuation in process-based catchment nutrient models contribute to the overall uncertainty of model predictions. A significant challenge in catchment nutrient modelling is robust quantification of uncertainty. Two major sources of uncertainty in process-based catchment nutrient models are quantifying the loss of N and P from land (referred to as diffuse source losses) and attenuation. To assess diffuse source losses, we used three existing typologies describing variation of land use/cover, soil moisture and topography. These typologies are associated with lookup tables of N and P export coefficients (loss rates as mass per unit area and time; $kg\ ha^{-1}\ yr^{-1}$). These export coefficients are generally derived from other models such as OVERSEER. The typologies and lookup tables are commonly used to parameterise diffuse source losses in catchment models. The contributions of the uncertainties of the export coefficients associated with these typologies to the overall uncertainty of catchment nutrient models have not been quantified. Attenuation parameters are commonly calibrated by reconciling estimates of total source losses in catchments with instream TN and TP loads calculated from monitoring data. Loads calculated from monitoring data are uncertain. However, the impact of this uncertainty on the uncertainty of attenuation parameters has not been quantified.

This study also developed and demonstrated a new class of purely empirical catchment models ('empirical models'). This class of model offers some advantages in terms of transparency, ease of implementation, and defensibility as well as more easily estimated model uncertainty.

The study found that there were significant differences in overall diffuse source losses between the existing typologies. For TN, the minimum between-typology difference in diffuse source losses at the catchment scale was $4\ kg\ ha^{-1}\ yr^{-1}$. For a typical catchment, this represents a characteristic between-typology difference of at least 33%. For TP, the between-typology difference in diffuse source losses was $0.4\ kg\ ha^{-1}\ yr^{-1}$, which for a typical catchment represents a characteristic difference of least 66%. These quantifications provide useful information for making first order estimates of the uncertainty of process-based catchment nutrient models. However, it was beyond the scope of this study to identify which set of export coefficients are closest to the truth.

Attenuation coefficients estimated from source loads derived from export coefficients and calculated instream yields had large uncertainties. Across the three typologies and approximately 330 water quality stations with calculated instream yields, 50% of attenuation coefficients had confidence interval widths of 0.18 or larger, 25% had widths of 0.3 or larger and 10% had widths of 0.47 or larger. These represent approximately 20%, 30% and 50% of the physically possible range of an attenuation coefficient (i.e., 0 to 1). These quantifications indicate that attenuation parameter uncertainties make considerable contributions to the

overall uncertainty of process-based catchment nutrient models. In addition, for TN and TP, approximately 20% and 30% of instream yields calculated for water quality stations were greater than the estimated catchment source loss. This means the estimated attenuation coefficients were negative, which is not physically possible. This is evidence that some, or all, of the export coefficients associated with the three typologies are underestimated in some situations.

The study derived satisfactory empirical models that can be used to predict TN concentrations (as median values mg TN m^{-3}) and yields (kg TN ha^{-1} yr^{-1}) as a function of the proportions of the catchment area occupied by 17 and 13 Types defined by specifically developed typologies, respectively. The empirical models for TN concentrations and yields avoid the need to use uncertain export coefficients and to calibrate attenuation parameters. In addition, the empirical models allow the user to estimate the 90% prediction interval as an estimate of the uncertainty of the empirical model predictions for any catchment.

The empirical models for TP were not successful, and we do not recommend their use at this stage.

The empirical models provide a simple and fast approach to developing catchment nitrogen models, at any location in New Zealand. These models are likely to be suitable for at least some assessments of land use and land management impacts.

The empirical models were calibrated to a limited and national dataset, which means there are two general limitations that apply to their use. First, when these models are used at smaller than national scales (e.g., individual catchments), the parameters will be potentially biased (i.e., not represent the conditions associated with the smaller model domain). Second, the number of types that could be represented by the models was limited due to the size of the dataset. This means they have only coarse spatial resolution of landscape factors that are associated with variation in diffuse source N loss.

Executive Summary

Introduction

Catchment nutrient (Nitrogen (N) and phosphorus (P)) models are integral to implementing the National Policy Statement – Freshwater Management (NPS-FM). Analyses associated with NPS-FM implementation commonly use a class of catchment nutrient models that we call process-based models. Process-based models explicitly represent the processes of contaminant loss in the catchment (source losses), transport to downstream receiving environments, and attenuation (i.e., reduction in loads between the point of discharge and downstream receiving environments by natural processes). Process-based models are used to predict N and P loads or concentration at points in the drainage network under both existing catchment conditions and some possible future set of conditions associated with change land use or management.

Setting up process-based catchment nutrient models involves quantifying two types of parameters: (1) representing rates of N and P loss in the catchment and (2) representing rates of attenuation. Typically, the parameters representing N and P loss rates from land are referred to as "diffuse source losses". The parameters representing attenuation are typically calibrated by reconciling the estimated total source losses in catchments with the instream loads observed at water quality monitoring stations. Because many of the impacts of relate to nutrient concentrations, an additional step is often required to translate predicted loads into concentrations.

A significant challenge in catchment nutrient modelling is robust accounting and reporting of uncertainty. Two major sources of uncertainty are the diffuse source losses and the calculated instream loads, which both contribute to uncertainty of the calibrated attenuation parameters. Combining all sources of uncertainty to fully characterise catchment model uncertainty or to estimate the uncertainty of attenuation parameters is difficult and rarely undertaken. However, failing to quantify and report uncertainties can lead to overconfidence in the evidence produced by catchment modelling and limits the ability to make risk management-based decisions.

This study had three aims. The first two aims are tractable steps toward quantifying components of the uncertainty associated with process-based models: (1) assess the level of agreement between three existing datasets that are used to quantify diffuse source losses (i.e., contaminant losses from land) that are currently used to parameterise process-based models, (2) assess the impact of the uncertainty associated with diffuse source losses and instream loads on the uncertainty of attenuation parameter estimates. Quantification of these uncertainties would provide useful information for making first order estimates of the uncertainty of process-based models. The third aim was to investigate the feasibility of fully empirical catchment models as an alternative to process-based models. This class of model offers some advantages in terms of transparency, ease of implementation, and defensibility as well as more easily estimated model uncertainty.

Use of land typologies in catchment modelling

In the context of catchment nutrient modelling in New Zealand, land typologies (hereafter, 'typologies') are often used to represent diffuse source losses in catchment nutrient models. Typologies discriminate landscape-scale variation of the factors that impact diffuse source nutrient loss rates including land use/cover, climate, topography, and soil. Variation in these factors is usually delineated by categorical subdivision (e.g., land uses/covers such as Dairy, Native forest, Urban), climate (e.g., Wet, Dry), and topography (e.g., Steep, Flat). Types are defined by specific combinations of these categories (e.g., Dairy/Wet/Flat). Each Type in a

typology is then associated with an 'export coefficient', which is an estimated average diffuse source loss rate that is recorded in a 'lookup table'. The export coefficients are used as the parameters representing diffuse source losses in process-based models. The values for the export coefficient associated with each Type are generally derived from published information (such as compilations of measured losses from catchments with a homogenous land use) or mechanistic models. Modelled quantification of export coefficients for Types representing agricultural land in catchment modelling studies in New Zealand have typically relied on the farm-scale OVERSEER model (AgResearch 2016).

The use of typologies and export coefficients simplifies setting up catchment nutrient models by providing one of the sets of model parameters (i.e., diffuse source loss rates). The use of an export coefficient for each Type is justifiable on the basis that the variation within each Type is "averaged out" at the scale of catchments. In this study, we examined three existing typologies and associated look-up tables of export coefficients that have been used in catchment modelling in New Zealand, which are herein referred to as: 'Monaghan' (Monaghan *et al.* 2021); 'Srinivasan' (Srinivasan *et al.* 2021); and 'LWP/Bright' (Bright *et al.* 2018).

Data and methods

We calculated median concentrations and instream loads of TN and TP from records of mainly monthly concentration and daily mean flow using a national dataset of water quality monitoring station data. Median concentrations were derived for 783 (TN) and 763 (TP) sites. Mean annual loads and their uncertainties were calculated for 315 sites for TN and TP. For many of the analyses, the instream loads were expressed as yields by dividing by the area of the upstream catchment (kg ha^{-1} yr^{-1}).

We calculated the catchment average export coefficient (CAEC) to represent the total diffuse source loss of TN and TP from the catchments of all segments of the national digital river network (DN2.4), which included the catchments of all water quality monitoring stations. The CAEC was calculated for each typology in three steps. First, the area occupied by each Type in a catchment is multiplied by the corresponding export coefficient. Second, these values are then summed to obtain the catchment diffuse source load. Third, the CAEC is the catchment diffuse source load divided by the catchment area. Because the CAEC is standardised by catchment area, it allows estimates of diffuse source losses to be compared between catchments.

We compared estimates of CAEC values between the three typologies. The agreement of the between-typology CAEC values was quantified by R^2, root mean squared deviation (RMSD, the characteristic difference between two typologies) and bias (the mean difference between two typologies).

We also compared estimates of CAEC values for each typology to instream yields for TN and TP. We expected that CAEC values would be larger than instream yields- due to attenuation. We also calculated attenuation coefficients from CAEC values and instream yields and then estimated the uncertainty of these coefficients based on the uncertainty of the calculated instream yields. We expected attenuation coefficient values to be between zero and one, lower values indicating greater attenuation and vice versa. Negative attenuation coefficients are not physically possible (i.e., CAEC values are less than observed instream yields) and we considered this to be evidence that the diffuse source inputs were underestimated.

We also attempted to derive purely empirical catchment models (hereafter, 'empirical models') directly from observed instream yields or concentrations. These models use specifically developed typologies to discretise landscape-scale variability in diffuse source nutrient losses. Empirical models for either yields or concentrations have only one parameter for each Type

defined by the typology that we refer to as empirical Type yield (ETY) or empirical Type concentration (ETC). An ETY is the expected annual load of nutrient per unit area, realised at a point in the drainage network (post attenuation), generated by diffuse losses associated with a specific Type. ETYs have the same measurement units as export coefficients (kg ha^{-1} yr^{-1}), however, all other things being equal, ETYs are less than export coefficients because the former include attenuation. An ETC is the concentration of nutrient (units mg m^{-3}) realised at a point in the drainage network (post attenuation), generated by diffuse losses associated with a specific Type.

Comparison of export coefficients between typologies and to calculated instream yields

For the catchments of each water quality monitoring station, CAECs for TN estimated using the three typologies were highly correlated. The CAECs for LWP/Bright and Monaghan had an R^2 of 0.84 (correlation between typologies) and minimal bias (2%). However, the characteristic difference in CAEC between these two typologies, as quantified by RMSD, was 4.1 kg ha^{-1} yr^{-1}. Monaghan and Srinivasan had an R^2 of 0.86 and an RMSD of 5.5 kg ha^{-1} yr^{-1}. LWP/Bright and Srinivasan had an R^2 of 0.77 and an RMSD of 5 kg ha^{-1} yr^{-1}. For TN the CAECs produced by Srinivasan were systematically lower than those of LWP/Bright and Monaghan as quantified by biases of 26% and 22%, respectively.

Because the LWP/Bright typology is not associated with export coefficients for TP, CAECs for TP were only compared between Monaghan and Srinivasan. The CAECs for TP for these two typologies had an R^2 of 0.34, an RMSD of 0.4 kg ha^{-1} yr^{-1} and a bias of 15% (i.e., Srinivasan was systematically lower than Monaghan).

For TN, CAECs for 20 – 21% of the water quality station catchments estimated by all three typologies were less than the calculated instream yields. This means the estimated attenuation coefficients were negative. When uncertainties associated with the calculated instream TN loads were considered, attenuation coefficients for between 9% and 21% of water quality monitoring stations were negative with 90% confidence intervals that did not include zero. This suggests that the export coefficients for some or all the Types within all typologies under-estimate diffuse source TN losses for some locations. There was no obvious geographic pattern in water quality stations with negative attenuation coefficients for TN. However, there was evidence that TN export coefficients for Forestry and Natural landcover/use categories are too low.

For TP, CAECs for 30% and 35% of the water quality station catchments derived from Monaghan and Srinivasan, respectively, were less than the calculated instream loads (meaning estimated attenuation coefficients were negative). When uncertainties associated with the calculated instream TP loads were considered, attenuation coefficients for between 12% and 19% of water quality monitoring stations were negative with 90% confidence intervals that did not include zero. This suggests that the TP export coefficients for some or all of the Types within the Monaghan and Srinivasan typologies under-estimate diffuse source TP losses for some locations. There was no obvious geographic pattern in water quality stations with negative attenuation coefficients for TP. There was no evidence that TP export coefficients were systematically too low for any particular landcover/use categories.

The calculated instream yields for the water quality stations had large uncertainties that contribute uncertainty to estimates of catchment attenuation. For TN attenuation, the widths of confidence intervals for 50% of water quality stations were 0.18 or larger, 25% had widths 0.3 or larger, and 10% had widths of 0.47 or larger. These represent approximately 20%, 30% and 50% of the physically possible range of an attenuation coefficient (i.e., 0 to 1). In general,

calculated instream load uncertainties and attenuation coefficients for TP were larger than for TN. Therefore, uncertainties associated with calculated instream loads represent considerable contributions to the uncertainty of process-based catchment nutrient models.

Empirical models

We derived satisfactory empirical models that can be used to predict TN concentrations (as median values mg TN m^{-3}) and yields (kg TN ha^{-1} yr^{-1}) as a function of the proportions of the upstream catchment occupied by 17 and 13 Types (see Table A and B), respectively. These models are expressed mathematically as:

$$Y = ETY_1P_1 + ETY_2P_2 + ETY_3P_3 + \cdots ETY_mP_m + PS_Y$$

$$C = ETC_1P_1 + ETC_2P_2 + ETC_3P_3 + \cdots ETC_mP_m + PS_C$$

where, Y is the yield (kg ha^{-1} yr^{-1}) and C the concentration (mg m^{-3}) at the evaluation point, and PS$_Y$ and PS$_C$ are the yields or concentrations associated with s point sources, $P_1, P_2, P_3, \ldots P_m$ are the proportions of catchment area occupied by each Type in the upstream catchment, and $ETY_1, ETY_2, ETY_3 \ldots ETY_m$ and $ETC_1, ETC_2, ETC_3 \ldots ETC_m$ are the empirically derived parameters for yields and concentrations respectively.

Predictions of TN concentrations and yields, and the lower and upper bounds of the 90% prediction interval, can be made for any catchment in New Zealand with catchment area >10 km^2 using the above equations and parameters (ETY and ETC) shown in Tables A and B. The empirical models can also be used to simulate effects of land use change or management actions on TN concentrations and yields by changing the proportion of catchment occupied by particular Types and by applying appropriate changes to the model parameters.

Table A. ETC parameters for the empirical TN concentration model for each of the 17 Types represented by the model. The values can be interpreted as the contribution of each Type to the best estimate, upper and lower bounds for the 90% prediction intervals of TN concentration (mg m^{-3}).

Type	Best estimate	Prediction interval lower bound	Prediction interval upper bound
Bare	0	0	0
Cropland	4464	759	22771
Dairy_Dry	6130	1856	5636
Dairy_Irrigated	6311	1422	13496
Dairy_Moist	2513	1580	4124
Dairy_Wet	1170	503	1338
Forestry_Dry	1272	402	3290
Forestry_Wet	237	154	771
Natural_Dry	110	-71	334
Natual_Wet	46	-4	111
OrchardVineyard	1102	318	11728
SheepBeef_Dry_Flat	314	31	4352
SheepBeef_Dry_Hill	183	86	232
SheepBeef_Wet_Flat	1193	526	4226
SheepBeef_Wet_Hill	477	143	1126
Urban	1368	584	1388
Water	0	0	0

Table B. ETY parameters for the empirical TN yield model for each of the 13 Types represented by the model. The values can be interpreted as the contribution of each Type to the best estimate, upper and lower bounds for the 90% prediction intervals of TN yield (kg ha^{-1} yr^{-1}).

Type	Best estimate	Prediction interval lower bound	Prediction interval upper bound
Bare	0	0	0
Cropland	4.9	4	84.9
Dairy_Dry	28.5	10.2	41.5
Dairy_Irrigated	29.6	6.6	65.4
Dairy_Moist	17	15.9	47
Dairy_Wet	37.5	31.3	64.1
Forestry	8.5	4.4	27
Natural	2.4	1.4	6.6
OrchardVineyard	18.6	8.7	-45.8
SheepBeef_Flat	8.3	0.4	18.7
SheepBeef_Hill	3.9	0.2	12
Urban	10.7	-7.9	6.8
Water	0	0	0

We failed to define satisfactory empirical models for TP concentrations and yields. The fitted model parameters were inconsistent with expectations (e.g., they sometimes implied ETCs and ETYs that were greater for natural than productive land, or greater for flat than steep land). The failure to define satisfactory empirical TP models is probably because our simple Types do not discriminate variation in important natural processes that determine P concentrations and yields, including: geogenic supply; mobilisation (erodibility, rainfall slope etc); and microbially mediated reduction-oxidation. It is noted that the three typologies examined by this study also do not discriminate variation in natural P concentrations and yields and this will contribute significantly to the uncertainty of their associated export coefficients (i.e., lookup tables).

Case study application of empirical models

We applied the new empirical TN models in a simple case study based on the Manawatū River catchment. For the empirical concentration and yield models, 84% and 97% of observations fell within the 90% prediction intervals, respectively. This indicates that the prediction intervals are reliable estimates of model uncertainties. The empirical TN yield model had satisfactory performance, minimal bias and RMSD of 2.8 kg ha^{-1} yr^{-1}. The empirical concentration model had poorer performance, largely due to a large (-29.3%) bias, indicating consistent over-prediction by the model.

We used the empirical models as well as an existing process-based model that had previously been developed for the Manawatū River catchment to model a range of land use change and mitigation scenarios. Predictions of changes in instream loads under the scenarios made using all models were in close agreement.

Discussion

Given the median CAEC for the monitoring station catchments in this study was 12 kg TN ha^{-1} yr^{-1}, the minimum between-typology difference in CAEC of 4 kg ha^{-1} yr^{-1} represents a characteristic uncertainty in catchment TN loads of at least 33%. Given the median CAEC for the monitoring station catchments was 0.62 kg TP ha^{-1} yr^{-1}, the between-typology difference in CAEC of 0.4 kg TP ha^{-1} yr^{-1} represents a characteristic uncertainty in catchment TP loads of 66%. As well as these random components of uncertainty there were systematic differences between typologies (quantified by bias).

This study cannot indicate which set of export coefficients are closest to the true value, or which Types are most incorrect within each typology. At least a portion of the discrepancy between the typologies will likely be attributable to differences in the details of the OVERSEER applications that were used to quantify the export coefficients for agricultural land. As for all numerical models, differences in the inputs to OVERSEER, even within plausible ranges, can result in appreciable differences in model output. The results of this study highlight that these and other sources of uncertainty associated with export coefficients make appreciable contributions to the uncertainty of catchment nutrient models.

The study indicates that CAECs for TN and TP are often too low compared to the calculated instream loads. It was generally beyond the scope of this study to identify the Types (or other particular conditions within a Type) for which export coefficients are too low. However, there was evidence that TN export coefficients for Forestry and Natural land cover are too low for all three typologies.

If we let the characteristic between-typology difference in CAEC for TN of 4kg TN ha^{-1} yr^{-1} represent a first order estimate of uncertainty of the catchment source loss, we can estimate the contribution of source load uncertainty to the uncertainty of the attenuation coefficient. For

example, for a catchment with a CAEC of 20 kg TN ha^{-1} yr^{-1}, the 95% confidence interval for source load extends from 12 to 28 kg TN ha^{-1} yr^{-1}. If the catchment has an instream load (expressed as a yield) of 10 kg TN ha^{-1} yr^{-1}, the attenuation coefficient is 0.5 and the 95% confidence interval extends from 0.17 to 0.64. The range in the attenuation coefficient is, therefore, almost half of the physically possible range of an attenuation coefficient. This indicates that export coefficient uncertainty represents a considerable contribution to the uncertainty of attenuation parameters of process-based catchment nutrient models. Similarly, the study also showed that uncertainty associated with instream loads makes large contributions to the uncertainty of estimated attenuation coefficients.

The uncertainties associated with export coefficients and instream loads are combined in process-based catchment nutrient models and contribute to their overall uncertainty. It was beyond the scope of this study to assess this combined uncertainty. However, the results of this study provide useful information for making estimates of the uncertainty of process-based models. In general, the study indicates that there are large uncertainties associated with process-based nutrient catchment models that use any of the three typologies examined by this study and/or that are calibrated to instream loads that are calculated from monthly water quality monitoring data.

The alternative 'empirical' approach to parameterising catchment water quality models for nitrogen (TN) concentrations and yields avoids the use of uncertain export coefficients or the requirement to calibrate attenuation parameters. The empirical models allow the user to estimate the 90% prediction interval as an estimate of the uncertainty of the empirical model predictions for any catchment. This provides a simpler and more transparent way to quantify model uncertainty than process-based models.

The empirical models presented in this report provide simple and easily-used tools that can be applied at any location within New Zealand. We have developed a dataset that provides proportions of catchment area occupied by each Type used by the TN concentration and yield models for all segments of national digital river network[1] (catchment area >10 km^2)[2]. These data allow estimates of yield or concentration to be made at any location in New Zealand. In addition, the data could also be used for national- and regional-scale assessments that aim to rapidly assess impacts of land use and land management scenarios at any location in New Zealand.

An important caveat that applies to the empirical models is associated with the national scale of the dataset that was used to derive the parameters. Because the water quality station data were limited, we were only able to derive robust ETCs and ETYs for a limited number of Types. This means the models only coarsely resolve variation in diffuse source loss of N across the landscape. An additional caveat that applies to the empirical models is that as the spatial extent of a modelled domain reduces (e.g., they are used at the scale of individual catchments), there is a reduction in applicability of the model parameters (ETCs and ETYs). This is because the empirical models were fitted to a national dataset and the ETC and ETY values are therefore national in scope. As the model domain decreases, the parameters will be potentially biased (not represent the conditions associated with the smaller model domain).

We note that catchment models are used in scenario analyses. In this type of application, the objective is generally to evaluate relative differences in concentrations and loads between scenarios. It is reasonable to assume that the uncertainty of relative differences will be less

---

[1] The digital network associated with the River Environment Classification (version 2.4) described by Snelder and Biggs (2002).
[2] Available in the Whitiwhiti Ora Data Supermarket: https://landuseopportunities.nz/

than the uncertainty of predictions of absolute quantities. However, this study did not test the validity of this assumption, and this would be a useful direction for future research.

1 Introduction

The National Policy Statement – Freshwater Management (NPS-FM) requires that regional councils identify freshwater objectives and associated limits that define acceptable levels of resource use and support for multiple other values. Because there are potentially many ways that these outcomes can be achieved, finding the most acceptable solution involves exploration of options. Integral to this is the use of catchment models that provide a basis for simulating the impacts of land use and management on contaminant levels in freshwater receiving environments.

Nitrogen (N) and phosphorus (P) are key contaminants associated with land use that regional councils are obliged to consider because they are nutrients that drive the state of several attributes mandated by the NPS-FM, including river periphyton and lake phytoplankton. Catchment models are used to link nutrient losses from multiple sources in a catchment, including diffuse losses from agricultural land, to N and P concentrations and loads in downstream receiving environments (e.g., Elliott *et al.* 2016). Process-based catchment models are based on a mass balance, in which it is assumed that observed instream loads are the sum of the upstream source contributions, less any net loss of mass during transport down the drainage network. The net loss is referred to as "attenuation". Attenuation of N and P occurs due to natural processes such as denitrification (N), sorption (P), and biological uptake (N and P). Setting up mechanistic catchment nutrient models involves defining parameters, which at least quantify rates of N and P loss from multiple sources (including land) and attenuation rates. Typically, the parameters representing N and P loss rates are derived from external sources including literature and other models and attenuation parameters are generally determined by reconciliation of the estimates of losses from land parcels in a catchment with the loads observed at water quality monitoring stations (Semadeni-Davies *et al.* 2020, 2021). Once set up and calibrated, catchment models can be used to explore the potential impact of alternative policy options, or planned mitigation actions, relating to changes in land management or land use.

Because catchment nutrient modelling involves representing contaminant losses at landscape scales, typologies are often used to simplify the parameterisation of diffuse nutrient losses from individual land parcels within a catchment. The use of typologies in environmental modelling is based on the recognition that, while no two locations are the same, it is generally not feasible to treat every location as unique. Therefore, some level of compression of detail (i.e., allocating locations to classes such that all locations in a class can be considered similar) is necessary in most environmental management tasks such as monitoring, reporting and policy development (McMahon *et al.* 2001). Typologies are used to group land areas into *Types*. Land areas belonging to a Type are considered sufficiently alike in terms of one or more characteristics of interest (e.g., nutrient loss rates, economic returns, response to management actions/mitigations) that they can be treated as the same. The grouping process is guided by principles that establish how various environmental factors influence the characteristic(s) of interest and how variation in these factors can be delineated by categories. The process is also guided by the availability of supporting data to characterise the responses expected from each Type and the availability of appropriate spatial data to define factors and their categories.

In the context of catchment nutrient modelling, typologies designed to reflect variation in diffuse nutrient losses are generally defined by combining land use or cover (a factor that may be sub-divided into categories such as dairy farms, native forest, urban) with environmental

factors that influence nutrient loss rates such as climate, land slope, and soil types (e.g., Monaghan et al. 2021; Srinivasan et al. 2021). A challenge in defining a typology is finding the optimal compression of detail. Where there are few Types, variation in the characteristics of interest cannot be represented with a great deal of detail. However, as the number of Types increase, the differences in characteristics between the Types will become less distinct and the increased resolution becomes more difficult to justify. Regardless of what principles are used to develop typologies, they cannot represent all the detail of reality. Therefore, environmental typologies cannot provide optimal discrimination of any individual characteristic and there is no 'correct' solution (Udo de Haes and Klijn 1994).

To date, most catchment nutrient models in New Zealand employ typologies of some kind to spatially represent diffuse losses for the catchments of interest. Typologies are used to first categorise locations that can be considered similar (i.e., a Type). Each Type is then assigned a representative diffuse source contaminant loss rate. Diffuse contaminant loss refers to losses from the surface (runoff), or from the bottom of the root zone (leaching), of a land parcel. We refer to the estimated diffuse loss rate for a Type as an "export coefficient[3]", which has units of kg ha^{-1} yr^{-1}. We refer to "look up tables" that associate each Type defined by a typology with an export coefficient.

Export coefficients are generally derived from published information (such as compilations of measured losses from catchments with a homogenous land use) or mechanistic modelling. Modelled quantification of export coefficients for Types representing agricultural land in catchment modelling studies in New Zealand have typically relied on the farm-scale OVERSEER model (AgResearch 2016). Recently published studies have populated look up tables with OVERSEER-based export coefficients (for TN and TP) for specific typologies that have national coverage (Monaghan et al. 2021; Srinivasan et al. 2021). Thus, typologies provide estimates of variation in diffuse nutrient contributions across the landscape that can be used as parameters in catchment models.

A significant challenge in catchment nutrient modelling is robust accounting and reporting of uncertainty. Two major sources of uncertainty in process-based models are the diffuse source loads and the calculated instream loads, which both then contribute to uncertain estimates of calibrated attenuation parameters. Uncertainties associated with observed instream loads based on monthly monitoring data are often in the order of +/- 20% for TN loads and +/- 40% for TP loads (Snelder *et al.*, 2017). Uncertainties associated with estimates of diffuse source loads derived from typologies and associated export coefficient lookup tables arise from the mechanistic model uncertainty (e.g., 30%, (Shepherd *et al.* 2013; Etheridge *et al.* 2018), the expected within-Type variation associated with the typology (e.g., 20 to 30%, (Srinivasan *et al.* 2021), as well as uncertainties associated with the mapping of Types across the landscape (Etheridge *et al.* 2018). To date, combining all sources of uncertainty to fully characterise catchment model uncertainty or to estimate the uncertainty of attenuation parameters has not been undertaken in New Zealand. Failing to account for and report uncertainties can lead to overconfidence in the evidence produced by catchment modelling and limits the ability to make risk management-based decisions.

There are two tractable first steps toward quantifying components of the uncertainty of process-based models. First, analyses can be conducted to assess the consistency of export coefficients associated with different typologies. The level of consistency between the catchment average values of export coefficients associated with different typologies would

---

[3] Note that some models and authors use 'export coefficient' to mean the yield after attenuation (i.e., the contaminants delivered at some downstream location after attenuation).

provide useful information for making first order estimates of the uncertainty of process-based models. However, this type of analysis has not been undertaken in New Zealand. Second, the uncertainty associated with instream load estimates is rarely taken into account when calibrating attenuation parameters of process-based catchment models. Quantification of uncertainty associated with attenuation parameters would also provide useful information for making first order estimates of the uncertainty of process-based models.

An alternative approach to using process-based catchment models with the associated parameterisation based on export and attenuation coefficients, is to characterise diffuse source contributions purely from observed (empirical) data. There are several hundred water quality stations with observed concentrations and flows throughout New Zealand and comprehensive mapping of land use and environmental factors (to define typologies) across the country that constitutes an extensive set of empirical data. There are advantages in parameterising catchment models directly on these data, rather than relying on export coefficients derived from external sources. We define a new class of purely empirical catchment water quality models (hereafter, 'empirical models') as an alternative to traditional process-based catchment models. These models offer some advantages in terms of transparency, ease of implementation, and defensibility as well as more easily estimated model uncertainty. The main disadvantage with an empirical approach is the number of Types that can be discriminated will be limited by the availability and distribution of water quality data; a smaller range of Types will limit the spatial detail that can be represented by the models.

The development of empirically-based catchment water quality models is predicated on the recognition that observed instream nutrient loads and concentrations represent the combination of upstream export of nutrients from land areas belonging to different Types, the subsequent attenuation of these exports, as well as contributions from point sources. For this study, therefore, we define an empirical Type yield (ETY) as the expected annual load per unit area, realised at a point in the drainage network (post attenuation), generated by diffuse losses associated with a specific Type that is defined by a typology. ETYs have the same measurement units as export coefficients (kg ha^{-1} yr^{-1}), however, all other things being equal, ETYs are less than export coefficients because the former include attenuation. If a look-up table of ETYs (and their uncertainties) for a specific typology can be estimated from empirical data, these values could be used to parameterise catchment nutrient models without the need to rely on uncertain export coefficients or to calibrate parameters representing attenuation. Like mechanistic catchment models, models parameterised using ETYs could be used to make "what if" simulations of changes in catchment nutrient loads under changes in land use and land management.

Empirical models that directly predict nutrient concentrations could also be quantified and used in a similar way to those described above for nutrient loads. In this study, we define the empirical Type concentration (ETC) to be the concentration realised at a point in the drainage network (post attenuation), generated by diffuse losses associated with a specific Type that is defined by a typology. The ETC has the units mg m^{-3}. An advantage of the ETC approach is that there are significantly more sites available across New Zealand for which concentration data is available compared to load estimates. This means that, compared to ETYs, there is more statistical power and the possibility therefore of deriving robust ETCs for a more detailed typology (i.e., more Types and greater environmental specificity).

This study used the available data to explore approaches to parameterising catchment nutrient models in the New Zealand context with two aims. First, we aimed to compare catchment N and P losses estimated using published typologies (and their associated export coefficient

lookup tables) with each other, and with observed instream loads. Comparison of estimated losses between different typologies can provide insight into uncertainties associated with loss rate parameters and comparison to instream loads provides insight into uncertainties associated with parameters representing attenuation. Second, we aimed to derive typology-based ETYs and ETCs for modelling nutrients anywhere in New Zealand. These empirically based model parameters could provide an alternative, or complementary, approaches to process-based catchment nutrient modelling that may be appropriate in some circumstances. We aimed to demonstrate the use of empirical models and compare them to a process-based model to provide an indication of their efficacy and to guide future applications.

2   Overview of data and methods

The analyses undertaken by this study can be divided into two primary streams of work:

1. Evaluation of estimates of catchment losses of TN and TP based on existing typologies and associated export coefficients.

2. Development of empirical catchment water quality models.

The analyses undertaken by these two workstreams were informed by a common set of existing data that are shown schematically as blue parallelograms in Figure 1 and are explained in detail in Section 3. Three existing typologies and their associated export coefficient lookup tables were included in the study. Spatial data describing land use, climate (precipitation) and topography (land slope) were obtained so that all land in New Zealand could be categorised into the Types defined by the existing topologies. The spatial data were also used to define the simple typologies that are part of the empirical water quality models. River water quality data was obtained for long term state of environment monitoring from regional council and NIWA records. Point sources data as annual loads discharged at specific locations were obtained from a national database.

The first analyses that were undertaken by the study involved processing the existing data into the specific inputs that were required by the two workstreams. Step A involved the development of national scale maps of source export coefficients (nutrient losses per unit area). by firstly mapping each typology, and then joining on Type export coefficients from lookup tables. Step A is explained in Section 4.2.1. Steps B and C involved processing the river water quality data and point source data to obtain calculated instream loads and concentrations attributable to diffuse sources (i.e., excluding point sources) for each monitoring station. Steps B and C are explained in Section 4.1.

The first workstream is represented by steps 1, 2 and 3 shown on the left of Figure 1. Step 1 compared the catchment average export coefficients, CAECs (spatially weighted diffuse source export coefficients) associated with each typology and quantified the extent to which the CAECs were consistent between typologies. Step 1 is explained in Section 4.1.2.

Step 2 compared the calculated instream loads for each monitoring station with the CAECs for the upstream catchment that were estimated with each of the typologies. 4.1.3. Our expectation was that the CAECs would be greater than the calculated instream loads because a proportion of the source load is attenuated by catchment processes. Step 3 used the CAECs and calculated instream diffuse source loads to estimate attenuation coefficients and their uncertainties. Step 2 and 3 are explained in Section 4.1.3.

The second workstream is represented by steps 4, 5 and 6 on the right of Figure 1. Steps 4 and 5 combined spatial data and the instream water quality data to derive empirical yield and

empirical concentration water quality models and are explained in Section 4.3. Step 6 applied these empirical models in a case study catchment and evaluated their performance against an existing process-based model of the catchment. Step 6 is explained in Section 4.4.
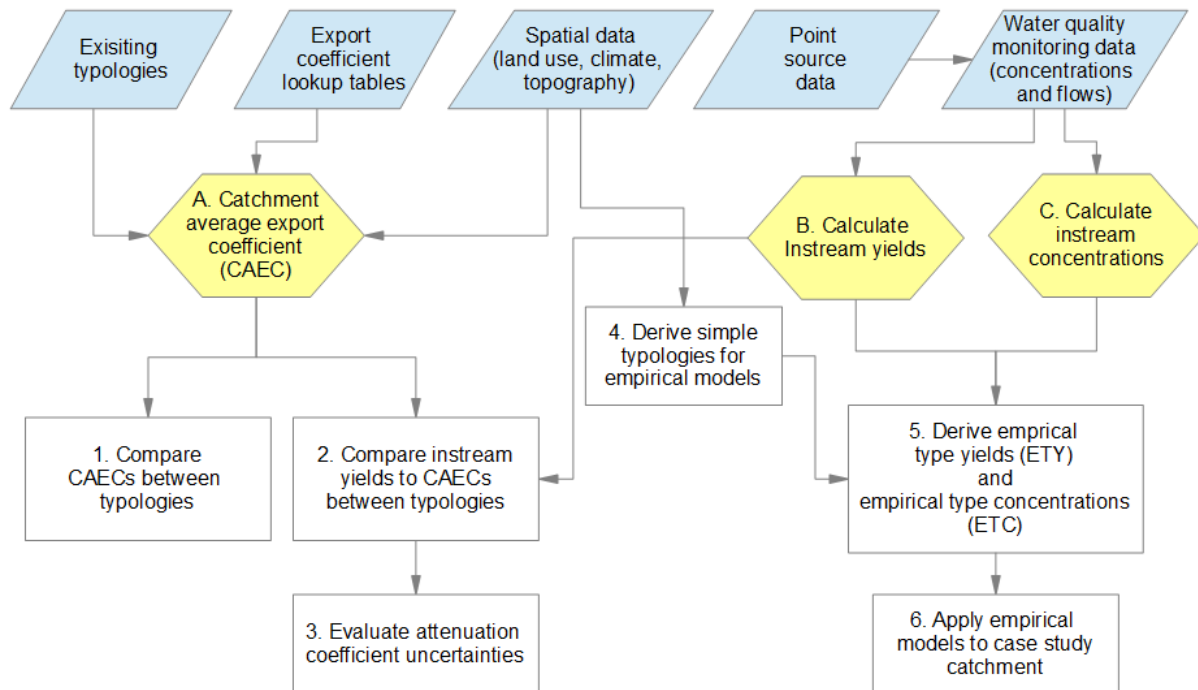
Figure 1. Schematic diagram of the input data and analyses undertaken by this study. The blue parallelograms indicate existing input data. The yellow hexagons indicate preparation of specific input data to subsequent analysis steps. The white rectangles indicate analyses and associated outputs from the study.

3   Data

3.1   River Data

3.1.1   Water quality data

Methods describing the acquisition of river water quality monitoring data and processing are described in Whitehead et al. (2021a). The dataset included 1081 water quality monitoring stations with TN observations and 1074 sites with TP observations.

3.1.2   Flow Data

River water quality data was obtained for long term state of environment monitoring from regional council and NIWA records as part of the most recent national state of the environment assessment (Whitehead, Fraser, Snelder, *et al.* 2021). For the water quality monitoring stations that could be associated with river flow gauging stations, we obtained the entire time series of available mean daily flow data from regional councils and NIWA databases. River flow gauging stations were only reliably identified, and flow records obtained for a subset (~400) of the water quality monitoring stations.

3.1.3 Point source data

Point source contributions of TN and TP in the catchment of each monitoring station were obtained so that the total catchment mass loss of the two nutrients on an annual basis could be estimated (see Methods). Point sources were based on a preliminary catalogue of annual point source loads (t y^{-1}) collated by NIWA (A. Semandi-Davies, *pers comm.*)

3.2 Drainage Network

The hydrological connectivity for the analysis was defined by 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). The digital network was derived from 1:50,000 scale contour maps; in version 2 (herein referred to as DN2.4) 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.

3.3 Land use

Land use is a key characteristic of the typologies used in both streams of this project. Each of the three existing typologies examined in the study was developed and tested with different land use layers. These differences are associated with the source and date of the datasets and how these were combined. As we expect that land use will be a significant contributor to variation in nutrient losses, we chose to define a single land use layer to be used within all typologies in this report, to reduce the contribution of variation in land use mapping on differences in estimated aggregate nutrient losses. Further, our ability to define land use was restricted to the use of publicly available datasets.

Our land use layer is fundamentally defined based on land cover information from the Land Cover Database (LCDB v5.0[4]), with pastoral land cover split between Dairy and Sheep and Beef land uses based on information about the extent of sheep and beef and Dairy farm units obtained from Richard McDowell (AgResearch pers comm). We aggregated LCDB classes into seven simplified land uses, where the aggregation took into consideration the land uses mapped in the three typologies and judgement about expectations of similarity in nutrient loss rates between land covers. The reclassified categories are described in Table 1.

---

[4] https://lris.scinfo.org.nz/layer/104400-lcdb-v50-land-cover-database-version-50-mainland-new-zealand/

Table 1. Re-classification of LCDBV5 Name_2018 classes in land use categories.

Class_2018	LCDBV5 Name_2018	Reclassified category
0	Not land	Bare
1	Built up	Urban
2	Urban Park	Urban
5	Transport Inf	Urban
6	Mines&Dumps	Urban
10	Sand&Gravel	Bare
12	Landslide	Bare
14	Snow&Ice	Bare
15	Alpine Grass	Natural/ Sheep & Beef [2]
16	Gravel&Rock	Bare
20	Lake&Pond	Water
21	River	Water
22	Estuarine	Water
30	Cropland	Cropland
33	Orchard&Vineyard	Orchard&Vineyard
40	High Producing Grass	Dairy/Sheep&Beef [2]
41	Low Producing Grass	Dairy/Sheep&Beef [2]
43	Tussock Grassland	Natural/ Sheep & Beef [1]
44	Depleted Grassland	Natural/ Sheep & Beef [1]
45	Herbaceous Freshwater	Water
46	Herbaceous Saline	Water
47	Flaxland	Natural
50	Fernland	Natural
51	Gorse&Broom	Natural
52	Manuka&Kanuka	Natural
54	Broadleaved Indigenous hardwoods	Natural
55	Sub Alpine Shrubland	Natural
56	Mixed Exotic Shrubland	Natural
58	Grey Scrub	Natural
64	Forest Harvested	Forestry
68	Deciduous Hardwood	Natural
69	Indigenous Forest	Natural
70	Mangrove	Water
71	Exotic Forest	Forestry

Notes:

1. Depleted and Tussock grassland areas that were coincident with Sheep & Beef land use (as defined by Monaghan et al. 2021) were specified as Sheep & Beef; remaining areas were defined as natural.

2. Productive grassland areas that were coincident with Dairy land use (as defined by Monaghan) were specified as Dairy; remaining areas were defined as Sheep & Beef

3.4 Typologies

The three typologies and associated nutrient loss lookup tables explored in this study are herein referred to as "Monaghan" (Monaghan *et al.* 2021), "Srinivasan" (Srinivasan *et al.* 2021) and "LWP/Bright" (Bright *et al.* 2018). The referenced publications provide complete descriptions of the typology development and population of nutrient loss lookup tables. The following section provides a brief description of the key characteristics of each typology and the methods used to extrapolate each Typology to provide national coverage for use in this study.

Each typology comprises a number of Factors (for example land use, soil type, slope etc) and each Factor is then subdivided into a number of Categories (e.g., slope might be categories into "Steep" and "Flat"). Each typology comprises several Types that are defined as unique combinations of different categories across the factors. The lookup tables for all three typologies were predominantly populated with export coefficients derived from OVERSEER model outputs (particularly for pastoral land uses), with some land use or cover categories (e.g., native forest, plantation forest) populated from expert knowledge or other published data.

The Srinivasan typology provides the most complete coverage of New Zealand (all land is assigned a Type and an associated export coefficient, except for urban land use). Spatial layers describing the distribution of the categories of the Srinivasan slope and moisture factors were derived based on the descriptions provided in Srinivasan et al. (2021). These were fundamentally based on slope factors from the Land Resource Inventory (Newsome *et al.* 2008), mean annual rainfall data provided by Ministry for the Environment[5] and mapped irrigation data obtained from the Ministry for the Environment[6]. We used the land use map described in section 3.3 within the typology. Srinivasan does not provide losses for Bare and Water land covers. As these land covers generally have small aerial contributions and no other suitable values were available, TN and TP losses were set to zero. For the urban land cover category, TN and TP export coefficients were derived from Moores et al. (2017).

The Monaghan typology is farm based, where each farm is assigned to a specific Type. These Types are restricted to pastoral (Sheep & Beef and Dairy) land uses. Monaghan subdivides dairy Types by 4 environmental factors: temperature (2 categories), wetness (4 categories), drainage (3 categories) and slope (3 categories). In total there were 72 Types associated with dairy land use. Monagahan defines seventeen Sheep & Beef Types based on variable combinations of region, climate and farming system. Spatial layers assigning land to the Sheep & Beef and Dairy Types defined by Monaghan et al (2021) were obtained from Richard McDowell (AgResearch pers comm). Areas of land not assigned to a Monaghan Type were infilled with Types and associated loss rates from Srinivasan. Further, to ensure that total areas of land use were consistent between typologies, Monaghan Types and loss rates were overwritten with Srinivasan Types and loss rates where the land use map (described in section 3.3) identified land uses other than Sheep & Beef or Dairy.

The LWP/Bright typology is based on three environmental Factors: climate zone (20 categories); irrigability (2 categories); and soil (3 categories), as well as irrigation status (2 categories) and land use (5 categories). The associated nutrient loss lookup tables only provide loss rates for TN, and the land use categories are restricted to productive land uses (Dairy, Sheep & Beef, arable, horticulture and forestry). The original spatial layers describing the distribution of the categories associated with the LWP/Bright typology factors (climate

---

zone, soil class and irrigability) were made available for this study. Mapped irrigation was as per Srinivasan, and we used the land use map in described in section 3.3. Areas of land not assigned to a LWP/Bright Type were infilled with Types and associated loss rates from Srinivasan.

4   Methods

4.1   Preparation of water quality data

4.1.1   Calculating instream loads

We calculated the annual instream loads of TN and TP (i.e., the total mass of TN and TP passing a specific location in a river) at each water quality monitoring station that complied with the following load data requirement criteria:

- Observations in at least 8 years in in the 10 years up to the end of December 2020

- At least 60 total concurrent observations of flow and concentrations

- At least 80% of all quarters (defined as January -March, April-June, July-September, October-December) in the most recent 10 years.

Rating curve methods were used to calculate the instream loads at sites that had concurrent TN and TP concentration observations and river daily mean flow records by (1) identifying the best rating curve method (out of four possible alternatives) for each site (through manual inspection of all possible rating curves for each site), and then(2) calculating loads by combining the best rating curve with the daily flow time series. A full description of the load calculation methodology is provided in Appendix A.

We used all available flow-concentration observations at each site to characterize the rating curves and set temporal trend terms in the underlying rating curve models so that the load calculations represent the expected mean annual load for 2020 (see Appendix A for details). Setting temporal trend terms to a fixed year (for those models that use time variable components), means that trends were accounted for in the calculation of loads. We also estimated 95% confidence intervals for the estimated instream loads, following a bootstrapping procedure (described in Appendix A3).

For the following analysis, instream loads are generally reported as instream yields, which are the instream load divided by the upstream catchments area, with units of kg ha^{-1} yr^{-1}.

4.1.2   Calculating median concentrations

We characterised TP and TN concentrations at each water quality monitoring station, as the median of all monthly observations for the 5-year period ending December 2020. The statistical precision of the median depends on the variability in the water quality observations and the number of observations. We therefore used filtering rules to restrict the sites that were used in our analysis to those for which the median could be calculated with reasonable precision. For a given level of variability, the precision of the median increases with the number of observations. As a general rule, the rate of increase in the precision of compliance statistics reduces for sample sizes greater than 30 (i.e., there are diminishing returns on increasing sample size with respect to precision above this number of observations; McBride 2005). In addition, because water quality observations tend to fluctuate seasonally, the precision of the calculated median is affected by how well each season is represented over the period of

record. Our filtering rules therefore restricted site × variable combinations that were used in the analyses to those with measurements for at least 90% of the sampling intervals in that period (at least 56 of 60 months). Site by variable combinations that did not comply with these rules were excluded from the subsequent analysis. The time period and filtering rules are consistent with those used by Whitehead et al. (2021).

4.1.3 Calculating point source contributions to instream yields and concentrations

Each point source in the dataset described in section 3.1.3 included location information in the form a unique segment identifier (nzsegment). Point sources were assigned to the digital network based on the segment identifier, and load contributions were accumulated in the downstream direction of the network. Point source yield contributions at all segments of the digital network were estimated from the accumulated point source loads divided by upstream catchment area. Point source concentration contributions at all segments of the digital network were estimated by dividing point source loads by estimates of segment site mean flows (sourced from Woods *et al.* 2006), with appropriate units conversion.

4.2 Calculating typology catchment average export coefficients

For each typology, we overlaid the spatial layers associated with the typology factors to generate maps of typology Types. To achieve this, we converted all spatial layers into coincident raster layers with 200m x 200m cells. This was a practical decision made for processing efficiency and took into consideration the requirement for national coverage, differences in the source data precision as well as a spatial scale that was commensurate with the typical smallest productive farm entities. As a test of the imprecision introduced by this choice, we compared estimates of rasterised catchment areas against catchment areas defined for the DN2.4. We found that above a catchment area of approximately 10 km^2 that differences in the estimates were very small (<<1%).

For each typology, maps of nutrient export coefficients were generated by assigning export coefficients from the relevant lookup tables to the typology Type maps. For each Type, areas that were infilled (either from literature or another typology) were marked to allow tracking of the proportion of infilled Types upstream of any location within the digital network.

Each typology export coefficient map was overlaid with a coincident raster layer of the DN2.4 sub-catchments. Sub-catchment diffuse source loads were calculated by summing the products of the raster cell nutrient export coefficient by the raster cell area, over each sub-catchment. Sub-catchment diffuse source loads were then accumulated in the downstream direction of the DN2.4 network to derive an estimated catchment diffuse source load for each nutrient for each network segment. At each network segment, the estimated catchment diffuse source load was expressed as a catchment average export coefficient (CAEC) by dividing the estimated diffuse source load by the upstream catchment area. A similar process was followed to identify the percentage of upstream catchment area that was not based on infilled export coefficients (see section 3.4), for each Typology.

The resultant products for each Typology and nutrient (TN or TP) were:

- Estimates of CAECs at each segment of DN2.4 (kg ha^{-1} yr^{-1})

- Proportion of upstream catchment area where export coefficients were not based on infilled data.

4.3   Examination of typology catchment average export coefficients

4.3.1   Between typology comparison of CAECs

For each typology we produced maps of the estimates of the CAECs as colour coded maps of the DN2.4. These maps provide a visual comparison of the estimates produced from the different typologies and identify regions and environments where these differences are greatest. For each typology, network segments with less than 50% of upstream area accounted for by the typology, or total upstream area less than 10km^2 were excluded from the maps.

We also extracted the estimates of CAEC for each typology and water quality monitoring station. We compared the predictions between each pair of typologies, evaluating the consistency of the estimates qualitatively using scatter plots, and quantitatively based on five statistics: regression R^2, Nash-Sutcliffe efficiency (NSE), bias (BIAS), percent bias (PBIAS), and the root mean square deviation (RMSD).

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

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

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

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

4.3.2   Comparison of typology CAECs with instream yields and estimation of attenuation coefficients

We compared the CAECs from each typology against monitoring station instream yields visually using scatter plots. We expect the sum of all source loads into a catchment (diffuse and point sources) to be greater than the observed instream load. Points that lie above the 1:1 line (i.e., instream yields are greater than CAECs plus point source yields) provide some indication that some or all of the typology export coefficients are underestimated.

The proportion of source load that is lost between the diffuse source and the instream observation point can be attributed to attenuation and used to quantify attenuation coefficients for modelling. This simple attenuation coefficient represents the effects of various processes such as nutrient uptake and transformation along the flow pathways. We expected attenuation coefficients to be in the range of approximately 0.1-0.9. A value of 1 would indicate that all of the load is attenuated and none of the input source load from the catchment of a monitoring

station appears at the monitoring station. A value of zero indicates that there is no attenuation/loss, and all of the source load appears at the monitoring station. An attenuation coefficient of less than zero indicates that instream loads at the monitoring station are greater than the sum of estimated source loads. We evaluated attenuation coefficients at each monitoring station, for each typology and nutrient and used these to explore whether there were any systematic patterns in attenuation coefficient related to certain Types or land uses, which might provide some evidence of issues with some or all the Type export coefficients.

An estimate of attenuation coefficients was made the following equation:

$$Att.coeff = 1 - \frac{Y - PS}{CAEC}$$ (Equation 1)

Where:

Att. Coeff	Attenuation coefficient (-)
Y	Estimated instream yield at the observation site (kg ha^{-1} yr^{-1})
CAEC	Estimated catchment average export coefficient (kg ha^{-1} yr^{-1})
PS	Estimated point source contributions (kg ha^{-1} yr^{-1})

By subtracting the point source contributions from the instream yield, we are making the assumption that point sources are not attenuated, and therefore the calculated attenuation coefficient should represent the diffuse pathway attenuation.

We evaluated upper and lower 95% confidence limits for the attenuation coefficient based on the 95% confidence intervals of the instream load estimate by:

$$Att.coeff\ LCI = 1 - \frac{Y_{UCI} - PS}{CAEC}$$ (Equation 2)

$$Att.coeff\ UCI = 1 - \frac{Y_{LCI} - PS}{CAEC}$$ (Equation 3)

We note that these confidence intervals do not consider the precision of the point source loads or the catchment diffuse source loads.

Calculated attenuation coefficients were mapped to explore whether there were any spatial patterns to explain variability. We also compared the evaluated attenuation coefficients against catchment proportions of land use, to investigate potential explanations for variability in calculated attenuation coefficients, and to evaluate whether there is evidence of underestimation of typology export coefficients for specific land uses.

4.4 Empirical catchment water quality models

We propose a new class of purely empirical catchment water quality models (hereafter 'empirical models') that are similar to traditional process-based water quality catchment models such as CLUES (Semadeni-Davies *et al.* 2020, 2021) and SCAMP (Cox *et al.* 2022) in that water quality is modelled as a function of the sum of catchment land areas weighted (i.e., multiplied) by constants that represent the contribution of contaminants (N and P) from those areas. The primary difference between the empirical models and process-based water quality catchment models (hereafter 'process-based' models) is how the models are parameterised. Process-based models explicitly represent processes of contaminant loss from land (source losses) and attenuation to produce a prediction of the instream load or concentration at downstream points in the drainage network. Source losses may be derived from a typology and associated look up table of export coefficients, directly from a land use and management model such as OVERSEER (Semadeni-Davies *et al.* 2020), or may be derived empirically (Elliott *et al.* 2005). The attenuation coefficients are typically calibrated by matching the sum of all catchment losses (i.e., diffuse losses from all land areas and point sources) to instream loads calculated from water quality monitoring data (Semadeni-Davies et al. 2020). It is important to acknowledge that process-based models represent the processes in very lumped forms (e.g., loss, attenuation) and the parameters are often empirically derived (export coefficients or loss rates).

In contrast, parameters for our new class of empirical models are derived from statistical models that are fitted to water quality data (instream yields and concentrations) using a set of Types (that are similar to the Types defined by typologies) as explanatory variables. In other words, these new models are founded on observed data and do not attempt to represent loss and attenuation as separate processes. As such, they do not require the provision of export coefficients to represent diffuse source losses (as can be the case CLUES and is a requirement for SCAMP), nor the calibration of attenuation coefficients. Instead, empirical models directly relate Types to yields and concentrations that are derived from water quality observations. The empirical models include sets of parameters that are derived for each Type. These parameters represent the outcome of the combination of the diffuse source loss from land and the attenuation, which means that empirical models represent these two processes with a single parameter.

The empirical models represent the yield or concentration of a contaminant attributable to diffuse sources at a location in the drainage network as the weighted sum of the proportion of catchment land area occupied by a series of Types. This is expressed mathematically as follows:

$$Z = \beta_1 P_1 + \beta_2 P_2 + \beta_3 P_3 + \cdots \beta_m P_m \qquad \text{(Equation 4)}$$

where Z represents the concentration or yield at a location in the drainage network, $P_1, P_2, P_{3,} \dots P_m$ are the proportions of catchment area occupied by each Type in the upstream catchment. $\beta_1, \beta_2, \beta_3 \dots \beta_m$ are coefficients derived from statistical regression models. The coefficients can be interpreted as the expected proportional contribution of each Type to concentration or yield. Alternatively, the coefficients can be interpreted as the expected concentration or yield for a catchment comprised of only that Type.

The coefficients $\beta_1, \beta_2, \beta_3 \dots \beta_m$ are derived by fitting linear regression models to the available water quality station data (Equation 5). These regression models have the same form as Equation 4, but Z represents data describing the observed yields or concentration at water quality stations after adjustment for point source contributions in the catchment. The

regression model predictors are the proportion of the catchments of each water quality station that are occupied by each Type. This is expressed mathematically as:

$$\begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix} = \begin{bmatrix} P_{1,1} & \cdots & P_{1,m} \\ \vdots & \ddots & \vdots \\ P_{n,1} & \cdots & P_{n,m} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} \qquad \text{(Equation 5)}$$

where Z is a 1 x n vector of the observed concentrations or yields at the n water quality stations after adjusting for any point source discharges in the catchment upstream, the n x m matrix represents the proportion of the catchment of each of n water quality stations (rows) in each of m land Types (columns), and β is 1 x m vector of the fitted regression coefficients for each of the m land Types. Note that there is one fitted regression coefficient for each Type. Note also that the fitted model has no intercept term, which is consistent with concentration or yield being zero if there is no land.

To be clear that the derived coefficients (i.e., $\beta_1, \beta_2, \beta_3 \dots \beta_m$) are used as parameters in empirical concentration and yield models, we refer to them hereafter as empirical Type yields (ETY) and empirical Type concentrations (ETC). The ETY is the expected annual load per unit area realised at a point in the drainage network (i.e., having been attenuated) that is generated by a specific Type that is defined by a typology. The units of ETYs are the same as export coefficients (i.e., kg ha^{-1} yr^{-1}). The ETC is the expected concentration, realised at a point in the drainage network (i.e., having been attenuated), that is generated by a specific Type that is defined by a typology. The units of ETCs are mg m^{-3}. The Types defined by the new class of models are similar to Types defined by existing typologies and must exhaustively cover all land in the catchment.

The general form of the empirical catchment water quality models for yield and concentration are given by:

$$Y = ETY_1 P_1 + ETY_2 P_2 + ETY_3 P_3 + \cdots ETY_m P_m + PS_Y \qquad \text{(Equation 6)}$$

$$C = ETC_1 P_1 + ETC_2 P_2 + ETC_3 P_3 + \cdots ETC_m P_m + PS_C \qquad \text{(Equation 7)}$$

where, Y is the yield and C the concentration at a point in the drainage network, and PS$_Y$ and PS$_C$ are the yield or concentration forms of the catchment point source contributions (as described in 4.1.3), P_1, P_2, $P_{3,} \dots P_m$ are the proportions of area occupied by each Type in the upstream catchment, and ETY_1, ETY_2, $ETY_3 \dots ETY_m$ and ETC_1, ETC_2, $ETC_3 \dots ETC_m$ are the empirically derived parameters for yields and concentrations respectively.

4.5   Derivation of empirical model parameters

4.5.1   Statistical modelling

We attempted to derive the regression coefficients shown in Equation 5 for each response variable (TN and TP yields and concentrations). Prior to fitting the models, we subtracted an estimate of the point source contribution from each of the water quality station concentrations or yields so that the response (i.e., $Y_{1,n}$) was only representing the attenuated diffuse sources of N and P.

There are two considerations with the process of fitting the statistical model expressed in Equation 5. First, the distribution of site concentrations and yields at the water quality stations will generally not be normally distributed. Normally distributed data (more specifically regression residuals) is a requirement of ordinary linear regression (OLS). It is therefore common to apply transformations, such as a logarithmic transformation, to normalise the

response when fitting OLS models. However, transformation of the concentration or yields would mean that the fitted regression coefficients (i.e., $\beta_1, \beta_2, \beta_3 \ldots \beta_n$) could not be interpreted as ETCs or ETYs for each land Type. We therefore use quantile regression instead of OLS to estimate the regression coefficients. Quantile regression is often used when the conditions of OLS are not met (Cade and Noon 2003). Whereas OLS estimates the conditional mean[7] of the response variable given some predictor variables, quantile regression estimates a specified quantile of the data. We fitted the model represented by Equation 5 to the median (i.e., the 0.5 quantile) value using quantile regression. The prediction from the model should be considered as an estimate of the median, conditional on the predictors (i.e., 50% of cases can be expected to be greater than or less than the prediction).

Because quantile regression is non-parametric, the fitted model does not describe the probability distribution within which prediction will lie. However, quantile regression models can be fitted to any quantile of the data. Therefore, in addition to fitting a model to the median (0.5 quantile), we also fitted models to the 0.05 and 0.95 quantiles (of the site concentrations and yields) to provide the lower and upper bounds of the 90% prediction interval[8]. Quantile regression models were fitted using the quantreg package of the R Statistical Software (R Core Team 2023).

In addition to the quantile regression models, we fitted OLS models to the same sets of predictors. The only information extracted from the OLS model was the R^2 value. The R^2 value of an OLS model indicates the variation in the response that is explained by the model (reported as a percentage of total variation), which is a measure of the fit of an OLS model that is commonly understood. The purpose of reporting the R^2 value was only to provide an appreciation of quality of the association between the water quality station concentrations and yields and the Types. The OLS model was fitted to the log (base 10) transformed response. Because the intercept term was set to zero, the R^2 value was evaluated as the squared correlation between the fitted values and the observed response (after transformation).

The second complication is that the predictors $(P_1, P_2, P_3, \ldots P_m)$ are what is referred to as compositional data. That is, $P_1, P_2, P_3, \ldots P_n$ represent the composition of the catchment land as the proportions occupied by each Type. Because the Types are exhaustive and the predictors represent proportions, they sum to one and, therefore the set of all proportions includes redundant information (e.g., $P_n = 1 - \sum_{i=1}^{i=n-1} P_i$). This means that another condition of multivariable regression, that the predictors are independent, is violated. Non-independence of predictors is referred to as multicollinearity because the implication is that there is correlation between the predictors.

When there is multicollinearity in the predictors of a regression model, the estimated coefficients (i.e., values of $\beta_1, \beta_2, \beta_3 \ldots \beta_m$) can become sensitive to small changes in the model. For example, small changes in the predictors or cases that are included in the model can dramatically change the coefficient values or even their signs. This means that multicollinearity reduces the precision of the estimated coefficients and increases their *p*-values (i.e., decreasing their statistical significance and reducing confidence in the estimated values). Multicollinearity can therefore make it difficult to justify the model, and this increases as the severity of the multicollinearity increases. It is noted that multicollinearity is a problem for

---

[7] The conditional mean of a random variable is its expected value – the value it would take "on average" over an arbitrarily large number of occurrences – given a certain set of "conditions". In a multiple linear regression model, these conditions are defined by the values of the independent (i.e., predictor) variables.

[8] The prediction interval indicates the range a future individual observation will fall.

interpretation of the estimated coefficients but does not affect the predictions or the goodness-of-fit (performance) statistics of the model (Neter *et al.* 2004).

An option to avoid the problem of collinearity is to remove some of the strongly correlated predictors. In this analysis, we were wanting to evaluate the coefficients for all predictors, to provide parameter values for all land Types and, therefore, this approach was not an option. However, the problems caused by multicollinearity reduce with increasing dataset size because sampling error reduces and precision increases as sample size increases (Mason and Perreault 1991). Because the datasets in this project were reasonably large, we adopted the approach of retaining all predictors and carefully inspecting the fitted coefficients and their standard errors to ensure that they were generally reasonable (i.e., were not so large as to render the coefficient unreliable). We also used cross validation to generate multiple instances of the fitted coefficients and used these to evaluate the sensitivity of the coefficients to the fitting data.

4.5.2 Definition of land Types

We defined Types to be used as predictors (i.e., $P_1, P_2, P_3, \dots P_m$ in Equation 5) by combining the land use categories described in Section 2.2 with the slope and moisture categories defined by the typology of Srinivasan et al. (2021). This resulted in nine possible land use categories (Urban, Forestry, Dairy, Orchard & Vineyard, Natural, Sheep & Beef, Cropland, Water, Bare), four slope categories (Flat, Rolling, Easy Hill, Steep), and four climate categories (Dry, Wet, Moist, Irrigated). The combination of all possible categories produces a total of 9 x 4 x 4 = 144 potential land Types.

Regression coefficients for all potential (144) Types cannot be reliably estimated because the number of classes are large compared to the fitting dataset size (i.e., ~300 for yields and ~900 for concentrations). However, all other things being equal, the utility and credibility of a catchment model that includes many Types is higher than the converse because it accounts for spatial variation in nutrient diffuse sources and allows for simulation of more nuanced management actions. We therefore derived sets of Types that comprised differing numbers of Types (i.e., each set had a differing value of *m* in Equation 4). One set of Types was defined based on only the land use categories. For the TN concentration and yield models, we excluded the Water and Bare land use categories based on our judgement that these categories make a negligible contribution to catchment nitrogen loss (i.e., we expected ETY and ETC for these Types to be zero). We also defined Types by successively subdividing some of the land use categories based on aggregated Srinivasan slope and climate categories or by simply using aggregated Srinivasan slope and climate categories. We did not know in advance how many regression coefficients could be reliably estimated (by the statistical modelling process) for each response variable (i.e., concentrations and yields of TN and TP). We therefore defined between six and nine sets of potential Types (depending on the response variable) that comprised a variable number of Types, fitted models to each of these sets (referred to as model 1, 2, 3 etc) and inspected the fitted coefficients to determine the "best" model (see section 4.5.3 below).

The definition of the sets of potential Types was subjective and was guided by expert opinion. We considered the information available in the fitting dataset as well as which land use Types are likely to exhibit appreciable differences in unit contributions under differing climate or slope categories. Regression coefficients are most likely to be reliably estimated for land Types that are consistently occurring and have wide variation in occupancy across the fitting datasets (i.e., for which the predictor P_m covers a wide range of non-zero values). The land use categories Dairy, Sheep & Beef, Forestry and Natural were the most consistently occurring

non-zero and variable land use categories. We therefore included differing coarse subdivisions of each of these land use categories by slope and climate in the nine sets of potential land Types. For example, we included subdivision of Dairy into coarse climate categories defined by Dry, Irrigated, and the combination of Moist and Wet. We included subdivision of Sheep & Beef into climate categories and coarse slope categories defined by Flat and the combination of Easy Hill, Rolling and Steep. The land use categories Orchard Vineyard, Urban, Cropland, Bare and Water consistently had low occupancy (e.g., the predictor P_m was generally a low or zero value). We therefore included these land use categories in the nine sets of potential land Types but did not further subdivide them by climate or slope categories.

The sets of potential land Types (referred to as Set 1 to 9) differed between the TN and TP models and comprised differing numbers of land Types (from 4 to 35). A complete description of the sets of potential land Types that were used for TN and TP is contained in Appendix B.

4.5.3 Determination of the best model

For each response variable (TN and TP yields and concentrations), we fitted separate models to each of the sets of Types (i.e., $P_1, P_2, P_3, \dots P_m$ in Equation 5). We inspected the fitted models and noted several considerations. First, we used the adjusted R^2 value of the equivalent OLS model as a measure of explanatory power of the models. Second, we examined the values of the fitted coefficients for each model. We considered coefficient values of credible models would be positive (i.e., we expected all land Types to contribute N and P). We also considered that coefficient values of credible models would be consistent with prior knowledge of nutrient loss by different types of land use and physiographic categories. In other words, we used prior knowledge to assess whether derived coefficients made sense. For example, we expected that N losses would be higher for Dairy than Sheep & Beef, and both land uses would have higher losses than Natural land cover. We also expected that, everything else being equal, regression coefficients for N and P would be higher for wet land Types than dry land Types and higher for steep land Types than flat land Types. We also expected that these conceptual relationships would often be confounded by cross-correlations and variables not explicitly considered (e.g., those involving stocking rates and farming intensity, soil type, etc.) and the fact that conditions that resulted in high loss rates might also result in high attenuation, and vice versa. Third, we considered the significance of the fitted coefficients and interpreted these as measures of confidence in their representation of the true value of the ETY or ETC. We considered that, all other things being equal, significant fitted coefficients were preferable to non-significant coefficients and models with greater numbers of significant coefficients were preferable to the converse. We also expected that the number of significant coefficients would decrease (i.e., p-values would increase) with increasing numbers of predictors (I.e., land Types) due to decreasing statistical power.

For each response variable, we considered that the "best" model represents a trade-off between the number of land Types (i.e., the discrimination of variation in land use, climate and slope), the R^2 value, the consistency of the coefficient values with our prior understanding, and the proportion of the parameters that were significant ($p < 0.05$). We also made the judgement *a priori* that all coefficient values for 50% quantile of the best model must be positive. We note that the best model is a judgement that does not mean that other sets of Types and their coefficients are not useful or better in some circumstances. In addition, we note that given different or updated datasets, different models would be derived. Therefore, we declare a "best" model in this study to demonstrate the approach and propose that this be regarded as an example and not the only possible model.

4.5.4 Objective evaluation of the models

For each response variable (TN and TP concentrations and yields), we evaluated four aspects of the "best" model: (1) the predictive performance compared to criteria for fit metrics, (2) the predictive performance compared to alternative frequently used models (3) the ability to estimate the 90% confidence interval, and (4) the stability of the fitted coefficients. These evaluations were carried out based on independent predictions of the response variables made for each water quality station by cross validation. Cross validation was carried out by first subdividing the dataset (representing the concentrations and yields at each water quality station) randomly into 10 equally sized subsets that are hereafter referred to as "folds". We fitted 10 "realisations" of each model (i.e., of the 0.05, 0.5 and 0.95 quantiles) by excluding one fold each time (the held-out fold). We used each of the 10 fitted models to predict the response for the associated held-out fold to obtain objective predictions (i.e., predictions for water quality stations that were not used in fitting the model) for each quantile and each water quality station.

We evaluated the predictive performance of the "best" model using two statistics: Nash-Sutcliffe efficiency (NSE), and percent bias (PBIAS). The normalisation associated with NSE and PBIAS allows the performance of the models to be compared to criteria proposed by Moriasi et al. (2015), outlined in Table 2.

Table 2. Performance ratings for the measures of model performance used in this study. The performance ratings are from Moriasi et al. *(2015).*

Performance Rating	NSE	PBIAS
Very good	NSE > 0.65	|PBIAS| <15
Good	0.50 < NSE ≤ 0.65	15 ≤ |PBIAS| < 20
Satisfactory	0.35 < NSE ≤ 0.50	20 ≤ |PBIAS| < 30
Unsatisfactory	NSE ≤ 0.35	|PBIAS| ≥ 30

The second evaluation was a comparison of the NSE and PBIAS for the 0.5 quantile models with the same performance statistics achieved for equivalent random forest (RF) models. RF is a machine-learning method based on and ensemble of regression trees (Breiman 2001; Cutler *et al.* 2007). Because RF models can include many predictor variables and automatically fit non-linear relationships and high-order interactions, they achieve high accuracy. This means that RF models are an accepted method of making model-based predictions of current river concentrations and yields based on data obtained for water quality stations (e.g., Snelder et al. 2020; Whitehead et al. 2021b). RF based models were fitted to the same response variable data as used in this study using a large set of predictors representing describing various aspects of the climate, topography, geology, land cover, and land use of the catchments of the water quality stations (see Whitehead et al. 2021b for details). We expected that the RF models would perform better than our 0.5 quantile models but note that RF models do not produce interpretable coefficients that can be used as parameters in catchment nutrient water quality models. The purpose of the RF models, therefore, is to provide a fair benchmark against which to compare model performance.

The third evaluation was of the estimation of the 90% prediction interval by the 0.05 and 0.95 quantile models. From the cross-validation outputs we evaluated the proportion of the predictions of the median that were less than or greater than the predicted 0.05 and 0.95 quantiles, respectively (i.e., the proportion of the predictions of the median that fell outside the

90% prediction interval). We expected that on average (over the 10 cross validation realisations) 10% of the estimates of the median would lie outside of the 90% prediction interval.

The fourth evaluation was of the stability of the fitted coefficients. From the cross-validation outputs, we retained the fitted coefficients (i.e., values of $\beta_1, \beta_2, \beta_3 \dots \beta_m$) for each realisation. We compared the mean values and the standard deviation of the fitted coefficients over the 10 realisations to the coefficients and their standard errors estimated for the full models (i.e., the models fitted to the entire dataset). We interpreted agreement of the mean and standard deviation of the coefficients estimated from the cross validation with their counterparts estimated from the full dataset to indicate that collinearity in the predictors was not causing sensitivity in the estimated coefficients (i.e., they were stable and reliable).

We undertook a final evaluation of the plausibility of the fitted model coefficients by comparing them to the response variable (TN and TP yields and concentrations) at the water quality stations. For each Type we plotted the response variable against the proportion of catchment occupied. For each Type, the fitted model coefficient was plotted at the position indicating a proportion occupancy of 1. We expected that, for each Type, the fitted model coefficient would tend to be consistent with the observed response variables at sites having high occupancy (i.e., the fitted coefficients for each Type would be similar to the observed response variables in catchments that are dominated by that Type). This expectation is consistent with the physical meaning of the fitted model coefficients as the expected proportional contribution of each Type to concentration or yield.

4.6 Application of empirical models

The empirical models, and their parameters (ETCs and ETYs) may have value as stand-alone tools or as supporting adjuncts to traditional process-based models. We undertook a small case study to demonstrate how the new empirical models might provide practical and sound support for analyses of catchment nutrient management in New Zealand. We also explored using the ETY within a semi-distributed framework with parameter adjustments to better match observed yields. We refer to this application as SDEM. The objectives of the work described in this section were to demonstrate potential "real-world" applications of the empirical models and to identify strengths, limitations, and points of difference across the empirical models and SDEM as compared to process-based modelling approaches.

The Manawatū River basin (Figure 1) was selected as the case study area for our application. The Manawatū is characterised by relatively high nutrient levels and a landscape dominated by pastoral agriculture. In addition, the Manawatū catchment has many point source discharges associated with municipal wastewater treatment plants and factories. There are extensive monitoring data available for locations throughout the basin. The basin is well studied, including a previous mass balance catchment nutrient modelling study performed by the authors of this report (Cox *et al.* 2022).
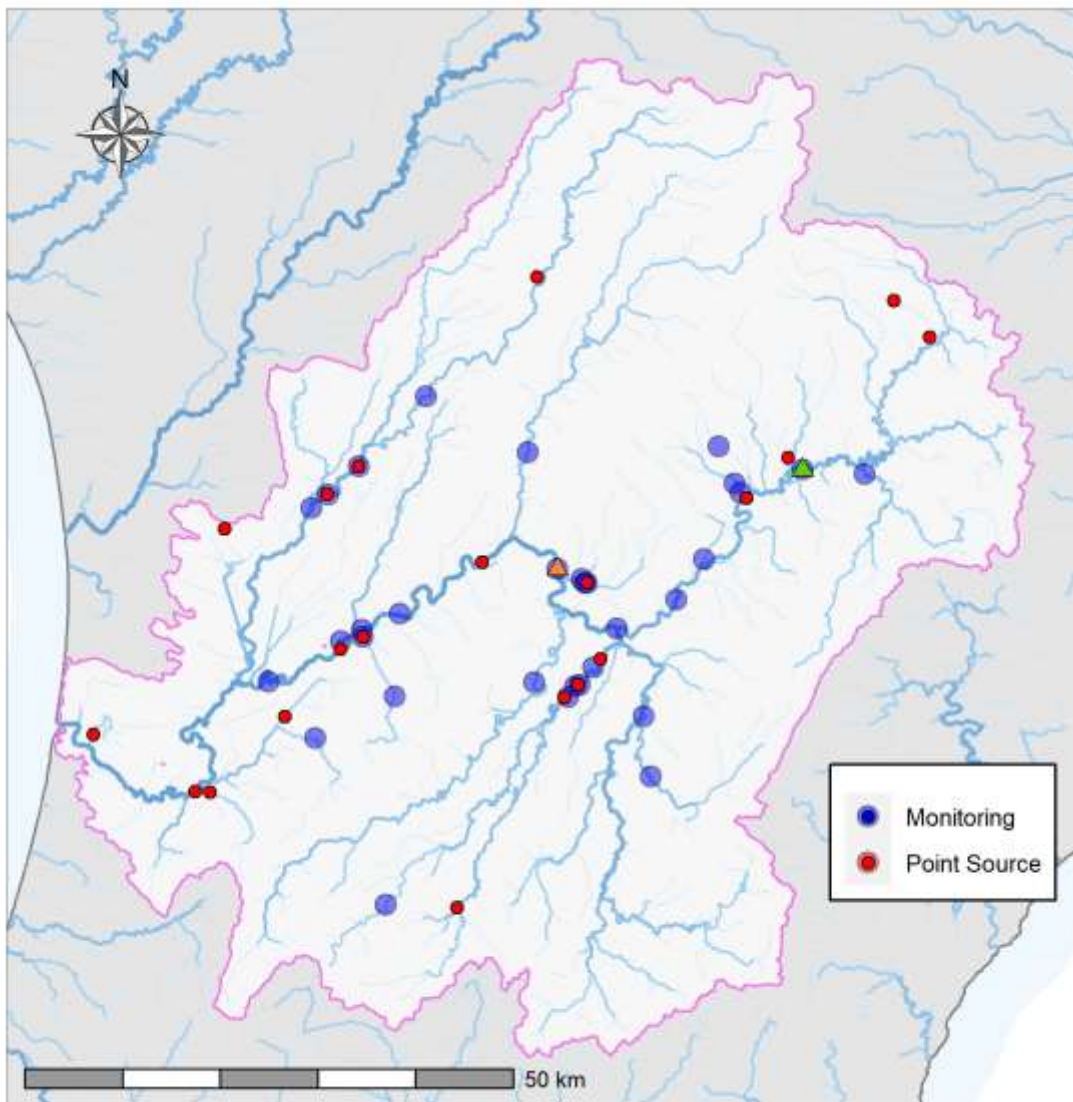
Figure 2: Manawatū River basin model domain showing the location of water quality monitoring stations and point source discharges. Monitoring locations at Manawatū at Upper Gorge and Manawatū at Weber Road are highlighted with orange and green triangles, respectively.

For this case study, we applied both yield and concentration empirical models to estimate nitrogen yields and concentrations at monitoring stations in the basin. To allow a comparison with the modelling work of Cox et al. (2022), we used the land use layer used within that study to quantify the proportion of catchment area in each Type for the empirical models. Other typology factors and the methods used to derive the Types were otherwise the same as described in section 4.5. Point source contributions were derived following section 4.1.3. The Type and point source information for each monitoring station catchment was summarised in a spreadsheet and Equation 6 was used to estimate yields and Equation 7 to estimate median concentrations at each station. Results were compared to observations of concentration and yield from Cox et al. (2022), which were developed following the same methods described in section 4.1, but relate to an end year of 2018 (rather than 2020). The locations of the monitoring stations in the Manawatū River basin are shown in Figure 2. Performance was

evaluated using the performance measures described in section 4.3.1 and by the percentage of sites for which the observations lay within the 90% prediction intervals of the models.

In addition to direct application of the empirical models, we used the ETYs (i.e., regression model coefficients) as parameters within a semi-distributed catchment model. A semi-distributed representation of the Manawatu River basin was defined by sub-dividing the catchment into several nested sub-catchments associated with a subset of the monitoring stations. Loads from each sub-catchment, estimated from the ETYs were then routed down the drainage network. This type of semi-distributed application provides the same results as the empirical yield model. However, the structure allows adjustments to the parameter values (i.e., ETYs) at the sub-catchment scale to better align with observed instream yields. In our application, adjustments to the ETYs were made uniformly across the Types using a single adjustment factor for each model sub-catchment to achieve agreement with calculated annual loads within ±10% at the 32 monitoring stations. This threshold was subjectively deemed as adequate agreement for this exercise and aligns with the approach taken in a previous modelling study (Cox et al. 2022). The adjustment process proceeded in an upstream to downstream direction. In this way, upstream adjustments were accounted for in the ETY adjustments for downstream catchments. We herein refer to this model as the semi-distributed empirical model (SDEM).

We compared the SDEM and empirical yield model predictions to predictions made with a previously developed process-based model of the Manawatu River basin (herein referred to as HRC-SCAMP). This model was developed for Horizons Regional Council by Cox et al. (2022) and uses export coefficients derived from several sources including the Monaghan values (Monaghan *et al.* 2021) for pastoral landuse, Moores et al. (2017) for urban land uses and Drewry (2018) and (Bloomer *et al.* 2020) for horticultural and arable land uses.

The HRC-SCAMP model was calibrated at 32 monitoring stations (one station which is not included in our updated load estimates), based on loads pertaining to an end date of 2018 (two years earlier than the load estimates described in this report). The HRC-SCAMP model attenuation coefficients were quantified as part of a separate calibration exercise by Cox et al. (2022) based on the monitoring station data. As the SDEM model used the nested sub-catchment structure of the SCAMP model, it was also calibrated to the 2018 load estimates. This also allowed a direct comparison of the predicted yields between the two models.

We compared the relative predicted contributions to yield from different sources (land uses and point sources) at the monitoring stations between the SDEM model and HRC-SCAMP. This was achieved by generating stacked bar charts that display the proportion of yield contribution from land uses and point sources for each model.

The suite of models was used to simulate yields and concentrations under current conditions and a set of scenarios. The objective of these simulations was to demonstrate the utility of the models for simulating various catchment mitigation actions and to highlight and explain the range and variability of predictions produced by the various models.

The first two scenarios (Scenarios 1 and 1a) incorporate a set of mitigation actions applied to pastoral (dairy and sheep/beef) farms, as described by McDowell et al. (2021). The scenarios represent full implementation of both "established" and "developing" farm mitigations. We applied projected national area-weighted mean reduction values, provided McDowell et al. (2021), of 59% and 13%, for dairy and sheep/beef, respectively. All reductions were applied as a percent reduction from current values (either export coefficients, ETCs or ETYs), for the appropriate land use categories. No other parameters, or point source loads, were modified.

Scenario 1 and 1a differed with respect to the assumed spatial extent of mitigation. In Scenario 1, the mitigation reductions are applied to all pastoral farmland in the Manawatū River basin. In Scenario 1a, the mitigation is assumed to occur only in an isolated upper basin sub-catchment (upstream of the Manawatū River at Weber Road monitoring station) with a single downstream assessment point, with respect to mitigation impacts, located approximately in the middle of the basin (Manawatū River at Upper Gorge).

The implementation of scenario 1a for empirical models requires an adjustment to Equation 4 to accommodate mitigation to a sub-catchment rather than the entire basin. The adjustment to Equation 4 is expressed mathematically as:

$$Z = \beta_1 P_1 \times \left(1 - m_1 \frac{P_{1a}A_a}{P_1 A}\right) + \beta_2 P_2 \times \left(1 - m_2 \frac{P_{2a}A_a}{P_2 A}\right) + \ldots \beta_m P_m \times \left(1 - m_m \frac{P_{ma}A_a}{P_m A}\right) \text{ (Equation 8)}$$

Where Z is the yield or concentration at a specified evaluation point, $P_1, P_2, P_3, \ldots P_m$ are the proportions of catchment area occupied by each Type, $\beta_1, \beta_2, \beta_3 \ldots \beta_m$ ETY or ETC values, $m_1, m_2, \ldots m_m$ are the proportional reductions associated with mitigation for each of the Types, $P_{1a}, P_{2a}, \ldots P_{ma}$ are the area proportions of each Type in the mitigated sub-catchment, A is the total catchment area upstream of the evaluation point and A_a is the total area of the mitigated sub-catchment.

The second set of scenarios (Scenarios 2 and 2a) simulated a land use change for 50% of all dairy farms to native bush. No other changes were simulated. As above, the difference between Scenario 2 and 2a is only the spatial extent of the assumed land use change (full basin vs. Weber Road sub-catchment).

For all scenarios, the predictions are presented as percent changes relative to the model predictions for current conditions. For Scenarios 1 and 2, results are presented for all 32 assessment points distributed throughout the basin. For Scenarios 1a and 2a, as noted above, predicted outcomes are presented at a single selected assessment point, roughly located in the middle of the basin (Manawatu River at Upper Gorge).

5   Results

5.1   Water quality monitoring station concentrations and instream yields

Table *3* provides a summary of the total number of water quality monitoring stations used in this study, i.e., they met the minimum data requirements for concentration and yield evaluation and had a catchment area >10km^2, for each of TN and TP.

Table 3*: Summary of water quality monitoring station numbers that met the minimum data requirements for calculating concentrations and yields and were used in this study.*

Variable	Concentration	Yield
TN	783	315
TP	763	315

Maps of the locations of water quality monitoring stations for which we evaluated concentrations and instream yields are shown in Figure 3 and Figure 4. Cumulative distributions of TN and TP concentrations across the water quality monitoring stations are shown in Figure 5, and cumulative distributions of instream yields (including uncertainties) are shown in Figure 6. The distribution of concentrations and instream yields for both TN and TP were approximately log-normal. The uncertainty bounds are generally larger (relatively) for the TP instream yield estimates compared to those for TN.
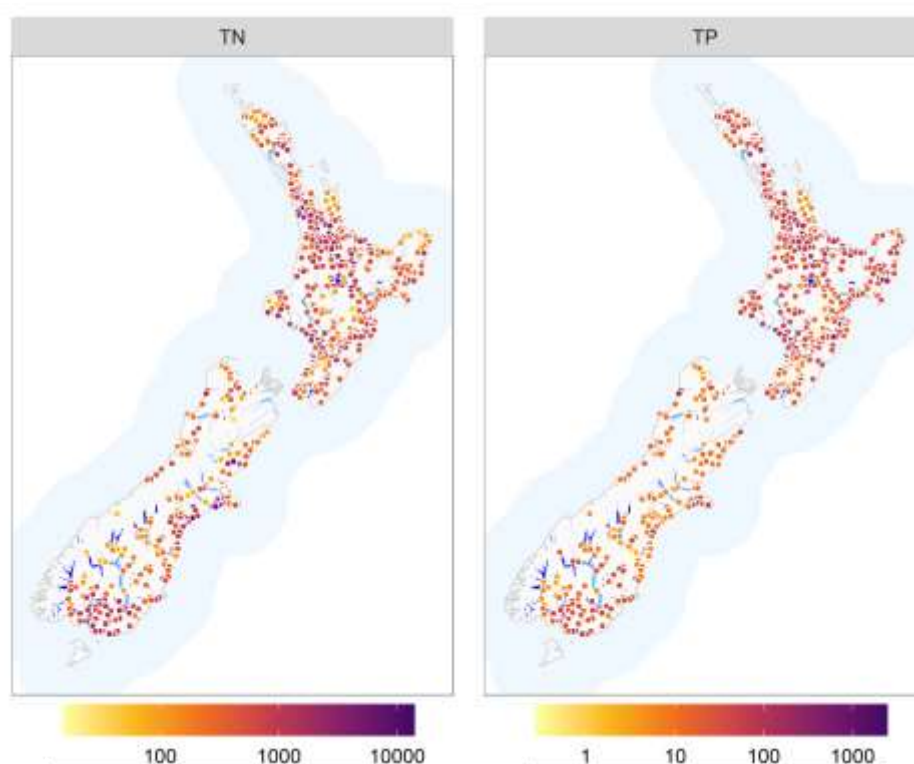
Figure 3: Locations of water quality monitoring stations for median TN and TP concentrations. The sites are coloured to indicate the evaluated site median concentrations (mg m^{-3}).
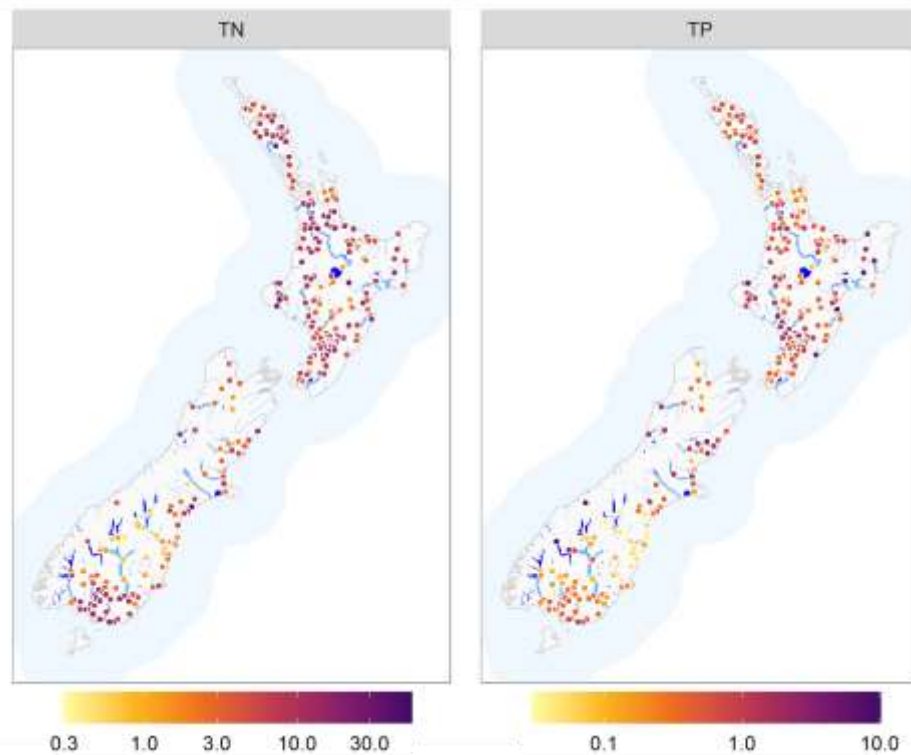
Figure 4: Locations of water quality monitoring stations for TN and TP yields. The sites are coloured to indicate the evaluated site instream yields, (kg ha^{-1}yr^{-1}).
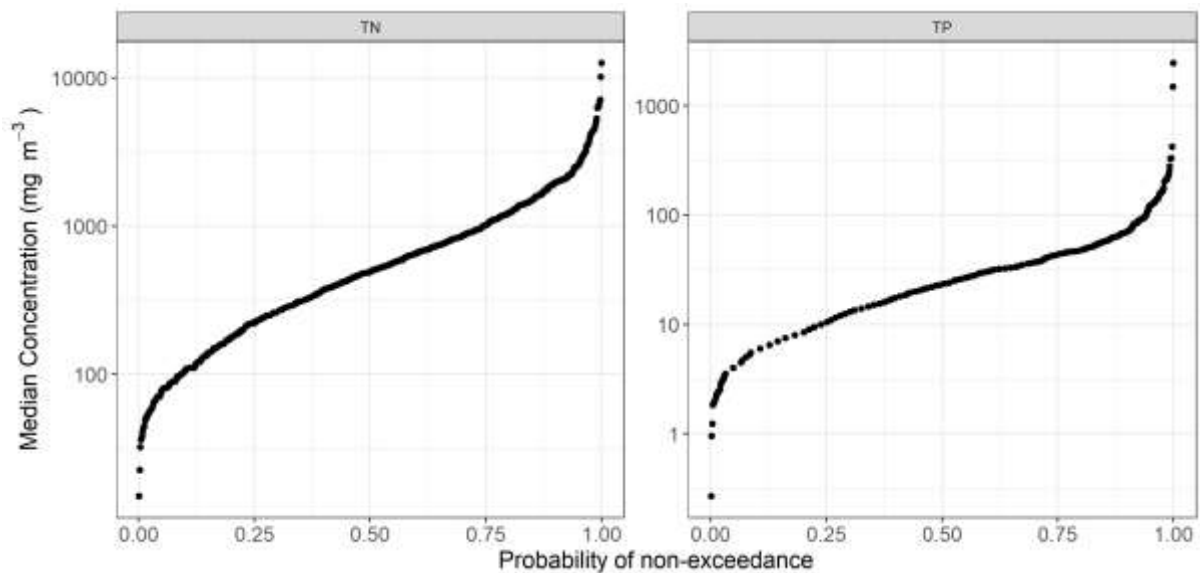
Figure 5: Cumulative distribution of TN and TP concentrations. See Figure 3 for site locations. Note that the y-axis has a log-scale.
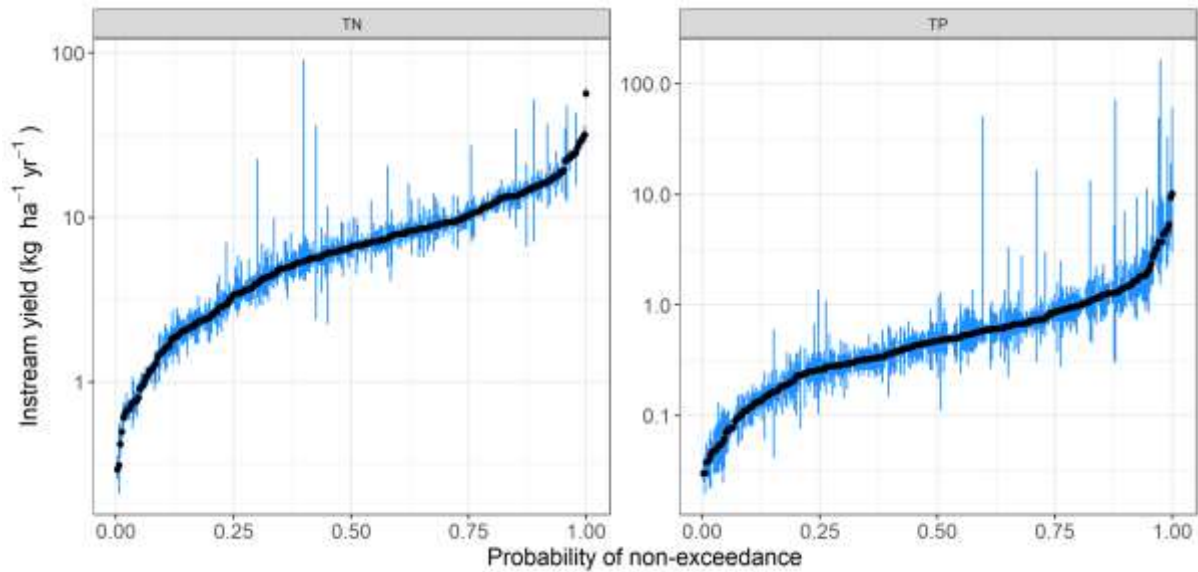
Figure 6: Cumulative distribution of estimated TN and TP instream yields See Figure 4 for site locations. The error bars indicate the 95% confidence interval for the instream yields. Note that the y-axis has a log-scale.

5.2   Between typology comparison of catchment average export coefficients

5.2.1   Mapped catchment average TN export coefficients

There are similarities in the overall distribution of high and low TN CAECs estimates across all three typologies (Figure 7). The largest values occurred in intensively farmed areas such as western Taranaki, northern Waikato, Canterbury plains and southern Southland. The smallest values occurred in South Island hill country and moderate values in the North Island east coast.

The largest absolute differences in TN CAECs between Monaghan and LWP/Bright were on the Canterbury plains (Figure 8). There were also some systematic differences, such as Monaghan consistently having higher values in Taranaki and the North Island east coast, and LWP/Bright having higher estimates in Manawatu. Srinivasan CAECs were consistently lower than the CAECs associated with the other two typologies across the North Island. Srinivasan CAECs were higher for the South Island hill country relative to the CAECs associated with both other typologies.

Figure 7: Maps of river network indicating estimates of upstream catchment average TN export coefficients (CAECs) for each typology. Network segments are not coloured when less than 50% of the upstream area is covered by the typology, or when the upstream area is less than 10 km².

Figure 8: Maps of river network indicating differences in estimates of catchment average TN export coefficients (CAECs) between pairs of typologies. Streams are not coloured when less than 50% of the upstream area is covered by one or more of the pairs of typologies, or when the upstream area is less than 10 km².

5.2.2 Catchment average TN export coefficients at monitoring stations

Of the 315 unique monitoring stations with observed TN instream yields, TN CAECs could be estimated based on more than 50% of the upstream area covered by the typology (see section 4.2) for 149, 220 and 315 sites for the Monaghan, LWP/Bright, Srinivasan typologies.

Scatter plots of the CAECs for each pair of typologies are shown in Figure 9. Overall, the consistency between the TN CAECs was highest between LWP/Bright and Monaghan, with a NSE of 0.84 and minimal bias. However, the characteristic difference in the CAECs between these two typologies, as quantified by RMSD, was 4.11 kg ha^{-1} (Table 4). As the median CAEC for the monitoring station catchments was 12kg TN ha^{-1} yr^{-1}, this between typology difference in estimates represents a characteristic uncertainty of 33%. Monaghan and Srinivasan were the most strongly correlated (R^2 = 0.86); however, there was a systematic difference between these two typologies, as quantified by a bias of 3.72 kg ha^{-1} yr^{-1}, with Monaghan estimating larger CAECs than those estimated by Srinivasan (Table 4).
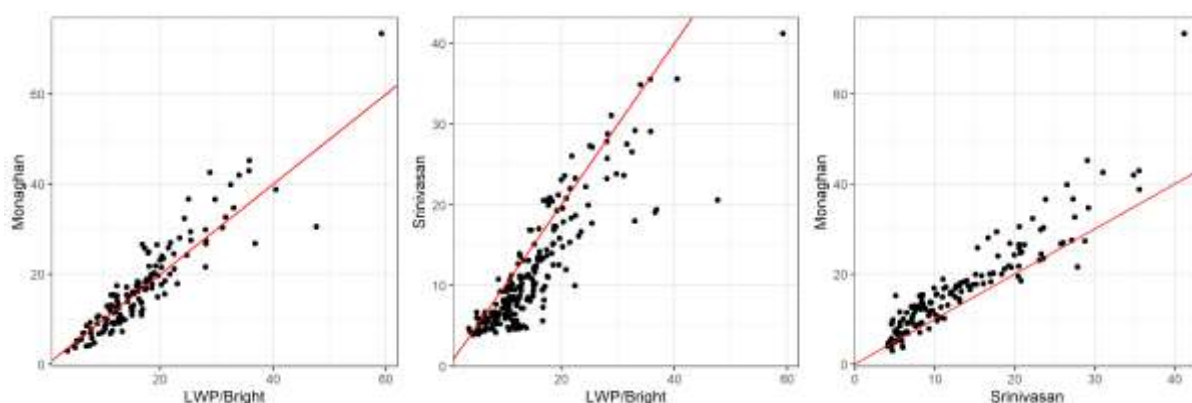
Figure 9: Scatter plot showing monitoring station TN CAECs calculated using alternative typologies. Only shown for sites with >50% upstream area covered by both typoloies. Red line is 1:1

Table 4: Performance statistics describing the consistency between estimates of TN CAECs for each pair of typologies.

Typology 1	Typology 2	N	R^2	NSE	Bias	PBIAS	RMSD
LWP/Bright	Monaghan	149	0.84	0.84	0.30	1.76	4.11
LWP/Bright	Srinivasan	220	0.77	0.53	-3.12	-26.11	4.95
Srinivasan	Monaghan	149	0.86	0.70	3.72	21.82	5.51

5.2.3 Mapped catchment average TP export coefficients

The difference between the TP CAECs (Figure 10, panel 3) demonstrates that the Monaghan estimates are generally higher than those from Srinivasan for almost all the north island (apart from Taranaki). South Island estimates are more similar, with the exception of Srinivasan estimating much higher TP CAECs for northern high-country areas. Note, TP export coefficients are not provided with the LWP/Bright, so results are only presented for Monaghan and Srinivasan.
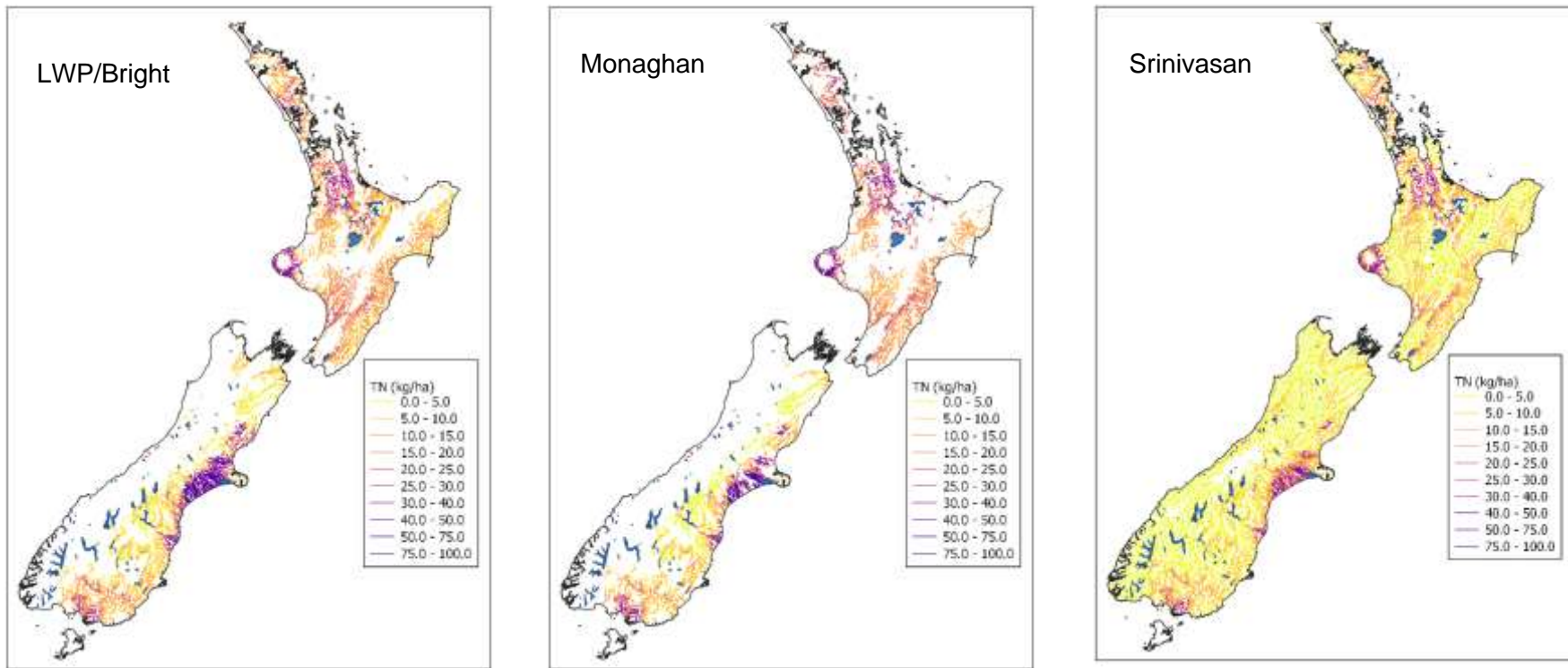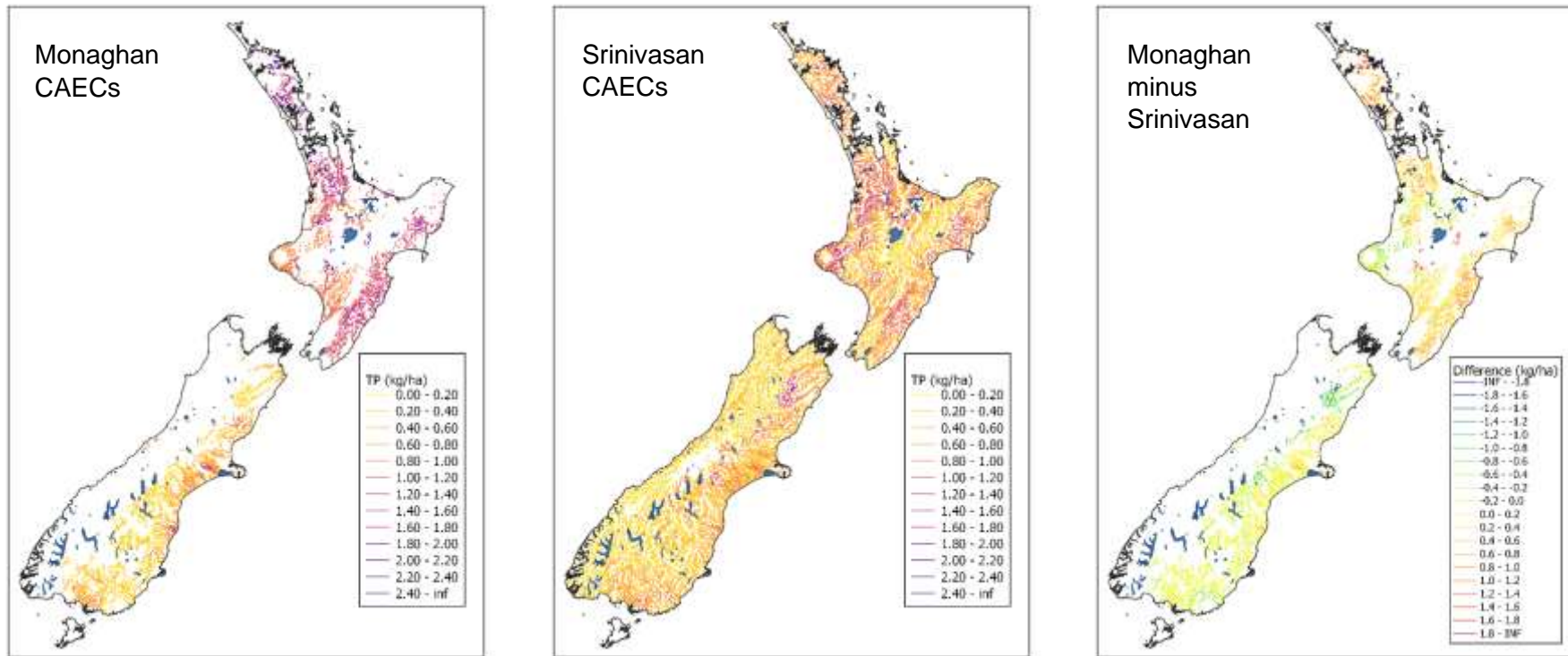
Figure 10: Maps of river network indicating estimates of catchment average TP export coefficients (CAECs) derived from each typology (first two plots). The furthest right plot shows the difference between first two plots. Streams are not coloured when less than 50% of the upstream area is covered by the typology, or when the upstream area is less than 10km².
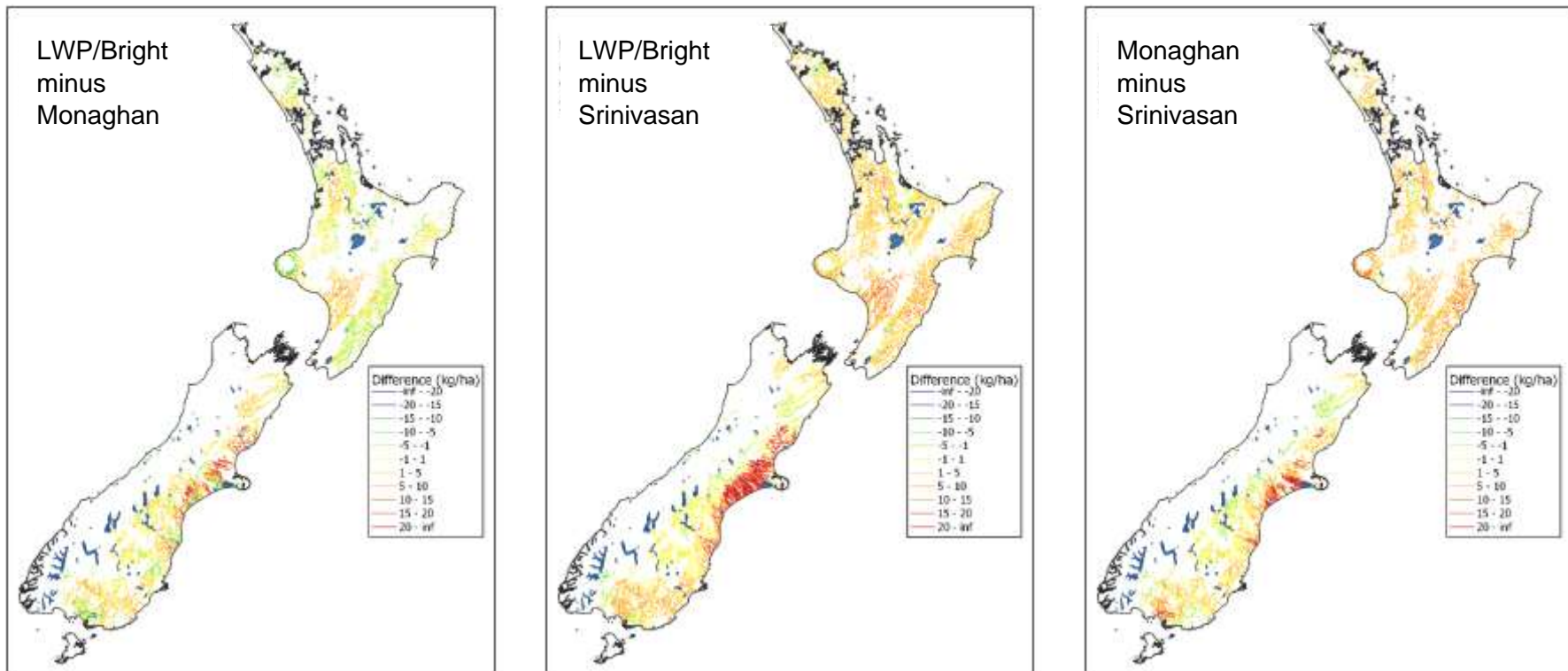
Of the 315 unique monitoring locations, TP CAECs could be estimated based on more than 50% of the upstream area defined by the typology for 149 and 315 sites for Monaghan, and Srinivasan, respectively. Note that TP export coefficients are not provided with the LWP/Bright typology.

Figure 11 shows a scatter plot of the estimated TP CAECs for Srinivasan and Monaghan. Performance statistics describing the consistency between Srinivasan and Monaghan are provided in Table 5. Overall the agreement between the typology CAECs was low, with R^2 of 0.34, NSE of 0.25 and RMSD of 0.4 kg TP ha^{-1} yr^{-1}. As the median CAEC for the monitoring station catchments was 0.62 kg TP ha^{-1} yr^{-1}, this between typology difference in estimates represents a characteristic uncertainty of 66%. There was a systematic difference in the TP CAECs between these typologies as quantified by a bias of 0.13 kg ha^{-1} yr^{-1}, with Monaghan greater than Srinivasan.
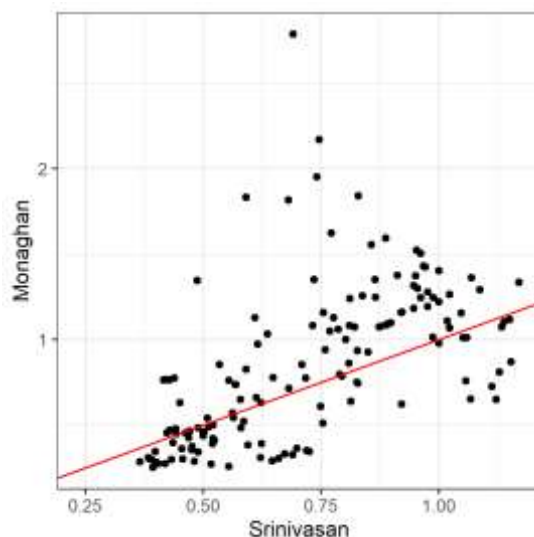
Figure 11: Scatter plot of estimated monitoring station TP CAECs between Srinvasan and Monaghan . Only for shown for sites with >50% upstream area covered by both typologies. Red line is 1:1

Table 5: Performance statistics describing the consistency in estimated TP CAECs between Srinivasan and Monaghan.

Dataset 1	Dataset 2	N	R^2	NSE	Bias	PBIAS	RMSD
Srinivasan	Monaghan	149	0.34	0.25	0.13	15.1	0.39

5.3 Comparison of typology CAECs with instream yields and estimation of attenuation coefficients

5.3.1 TN CAECs and instream yields

Between 20-21% of all source yields (CAECs plus point sources) were evaluated to be less than the observed instream yield (above the black line in Figure 12). Even accounting for the instream yield uncertainty and excluding sites with less than 50% upstream area defined by

the typology, 9%, 7% and 21% of sites had source yields that were less than the lower confidence interval of the instream yields, for LWP/Bright, Monaghan and Srinivasan, respectively.

Figure 12: Scatter plot of instream TN yield versus typology TN catchment average export coefficients plus point source contributions. Black line is 1:1. The error bars indicate the instream yield uncertainty (95% confidence interval).

5.3.2 TN attenuation coefficients

As well as 20-21% of all attenuation coefficients being evaluated to be less than zero, uncertainties in the evaluated TN attenuation coefficients were variable and moderately wide in relation to the expected feasible range (Figure 13). 50% of TN attenuation coefficients had uncertainty ranges that were larger than 0.17,0.18 and 0.22 for LWP/Bright, Monaghan and Srinivasan, respectively. 11%, 12% and 18% of TN attenuation coefficients had uncertainty ranges that were larger than 0.5 for LWP/Bright, Monaghan and Srinivasan, respectively.
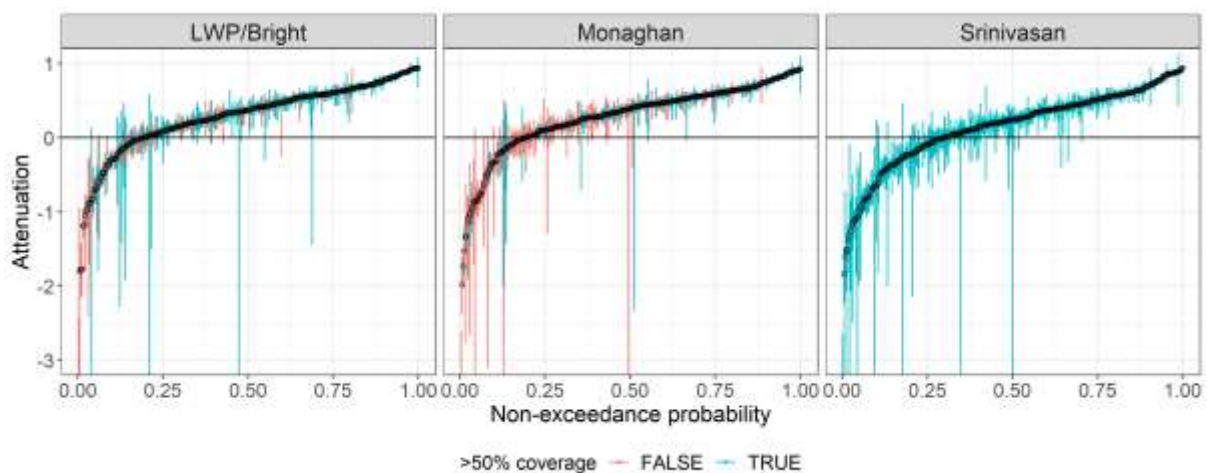
Figure 13: Cumulative distributions of evaluated TN attenuation coefficients (and their uncertainties) for each of the three typologies. The error bars indicate the uncertainty (95% confidence interval) for the attenuation coefficient associated with calculated instream yield uncertainty.

The spatial distribution of TN attenuation coefficients and the distribution of those that were evaluated to be negative, did not appear to have immediately obvious spatial patterns (Figure

14). This result may have come about due to underestimation of typology export coefficients, or issues with the instream yield estimate.

Figure 14: Maps of estimated TN attenuation coefficients. Sites shown with green dots had negative attenuation coefficients.

Spatial patterns in the calculated TN attenuation coefficients were explored by plotting these values against the proportion of catchment area occupied by the four most well occupied land land use categories (Figure 15). The plots indicate that there is a relationship between the probability that attenuation coefficients are negative, and the proportion of upstream catchment occupied by the native or forestry land use category. This suggests that the export coefficients for these land uses are systematically underestimated by all typologies (Natural with an export coefficient of 2 kg TN ha^{-1} yr^{-1} and Forestry 4 kg TN ha^{-1} yr^{-1}). We note that the Natural export coefficient is only derived from Srinivasan (and was used to infill losses for this land use for both other typologies). Similarly, the Forestry land use export coefficient used in the Monaghan CAEC estimates is derived from Srinivasan and is the same as the value used in LWP/Bright.
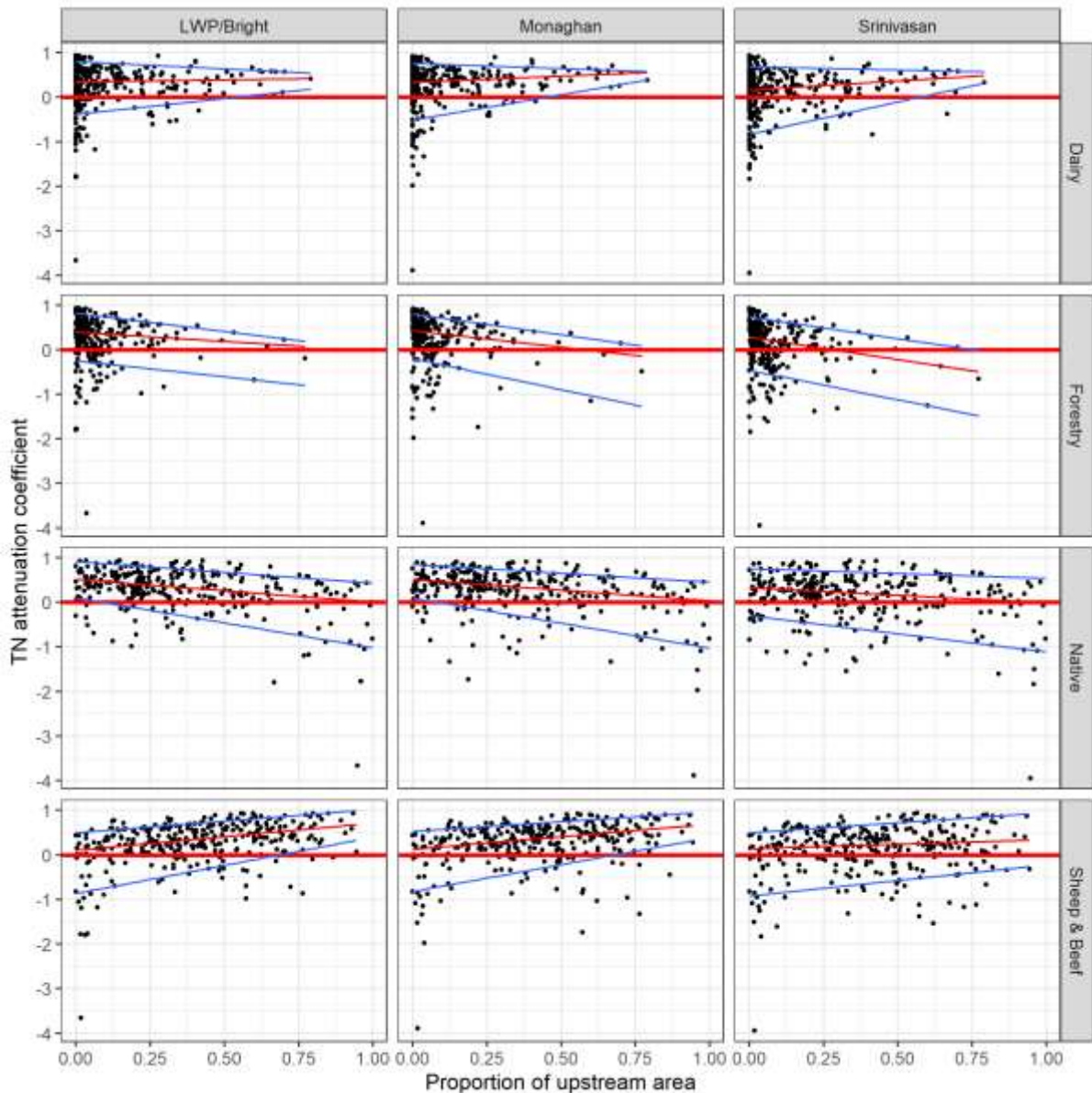
Figure 15: Scatter plots of evaluated TN attenuation coefficients compared to the proportion of upstream catchment occupied by differing land use categories. Coloured lines are linear quantile regressions of the median (red) and 10% and 90% quantiles (blue).

5.3.3 TP CAECs and instream yields

In the analysis for TP, we found that the estimated point sources were greater than the observed instream loads at 3 sites. We excluded these three sites from the following analysis. For Monaghan and Srinivasan, 30% and 35% of all source yields (CAECs plus point sources) were evaluated to be less than the instream yields, respectively (above the black line in Figure 12). Even accounting for the instream yield uncertainty and excluding sites with less than 50% upstream area defined by the typology, 12% and 19% of sites had source yields that were less than the lower confidence interval of the instream yields for Monaghan and Srinivasan, respectively.
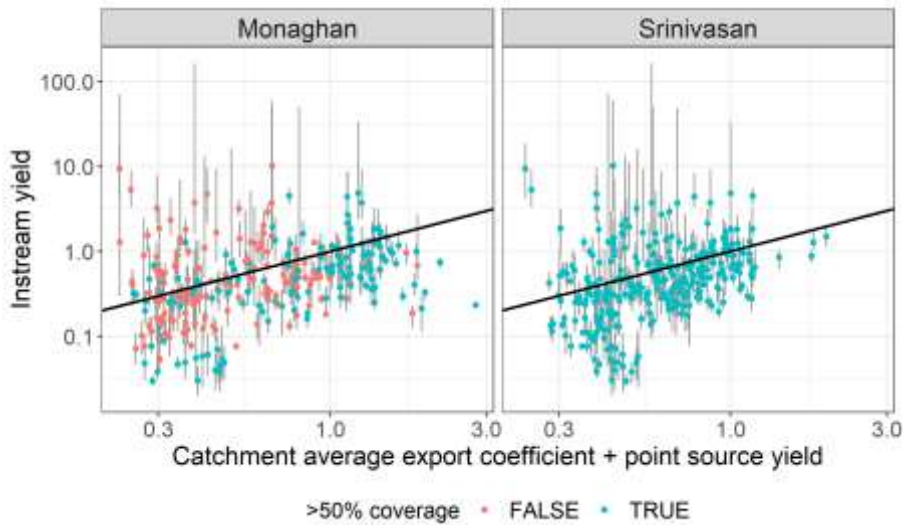
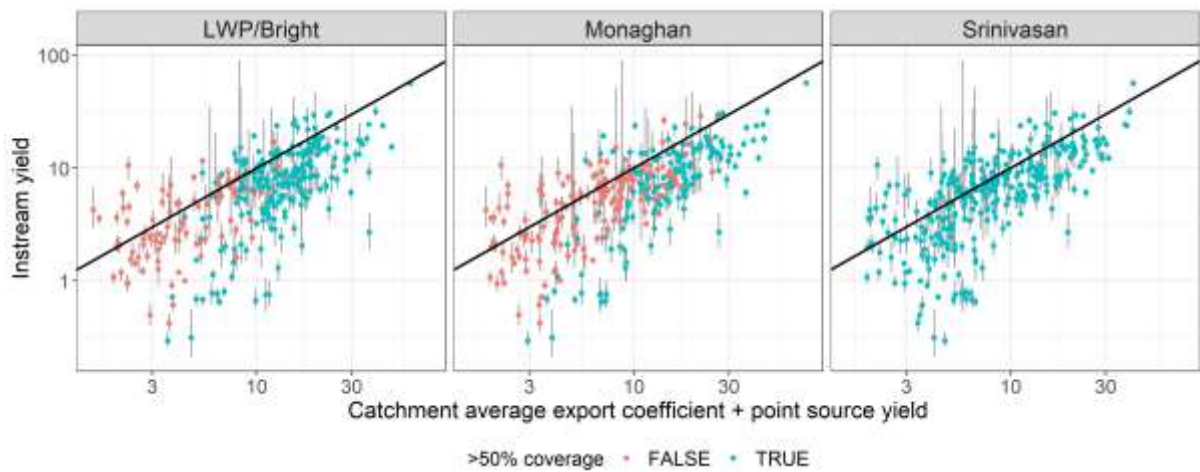*Figure 16: Scatter plot of instream TP yield versus typology TP catchment average export coefficients plus point source yields. Black line is 1:1. The error bars indicate the instream yield uncertainty (95% confidence interval).*

5.3.4 TP attenuation coefficients

As well as 30% (Monaghan) and 35% (Srinivasan) of all attenuation coefficients evaluated to be less than zero, uncertainties in the evaluated TP attenuation coefficients were variable and wide in relation to the expected feasible range (Figure 17). 50% of TP attenuation coefficients had uncertainty ranges that were larger than 0.42, and 0.43 for Monaghan and Srinivasan, respectively. 43% of TP attenuation coefficients had uncertainty ranges that were larger than 0.5 for both Monaghan and Srinivasan. We also note that the uncertainty estimates of the attenuation coefficients are generally very wide, and typically larger than for the TN attenuation coefficients.
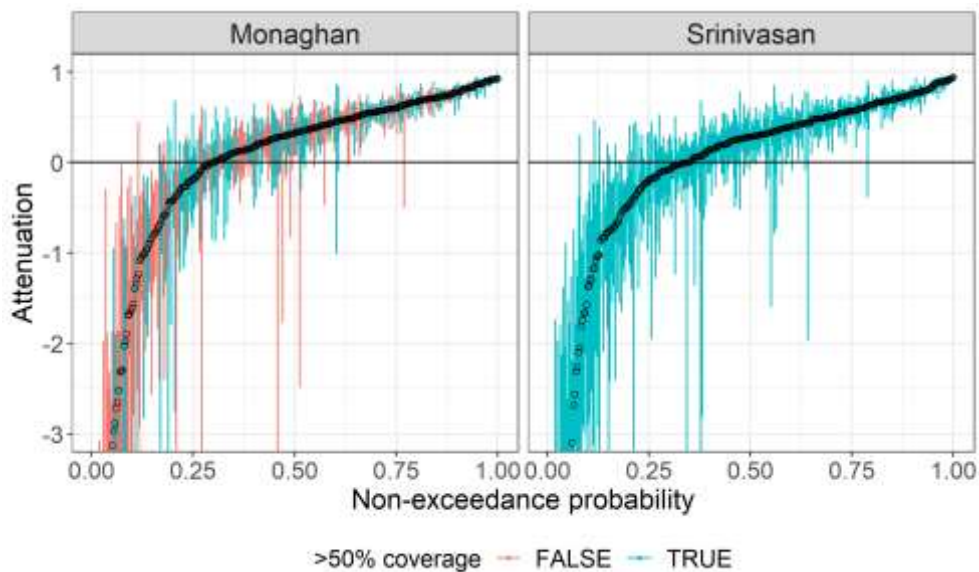
Figure 17: Cumulative distributions of evaluated TP attenuation coefficients (and their uncertainties) for the Monaghan and Srinivasan typologies.

The spatial distribution of TP attenuation coefficients and the distribution of those that were evaluated to be negative, did not appear to have immediately obvious patterns, although Manawatu and Taranaki did have very large proportions of negative coefficients (Figure 18). This result may have come about due to underestimation of typology export coefficients, or issues (particularly underestimation) with the instream yield estimate.
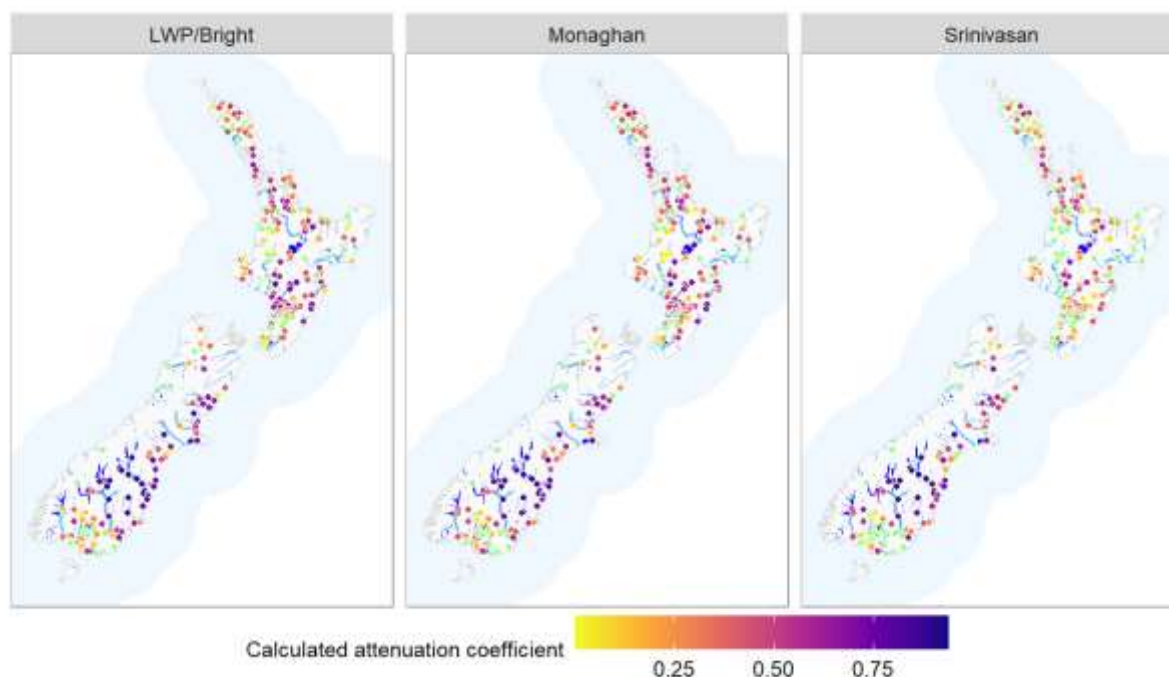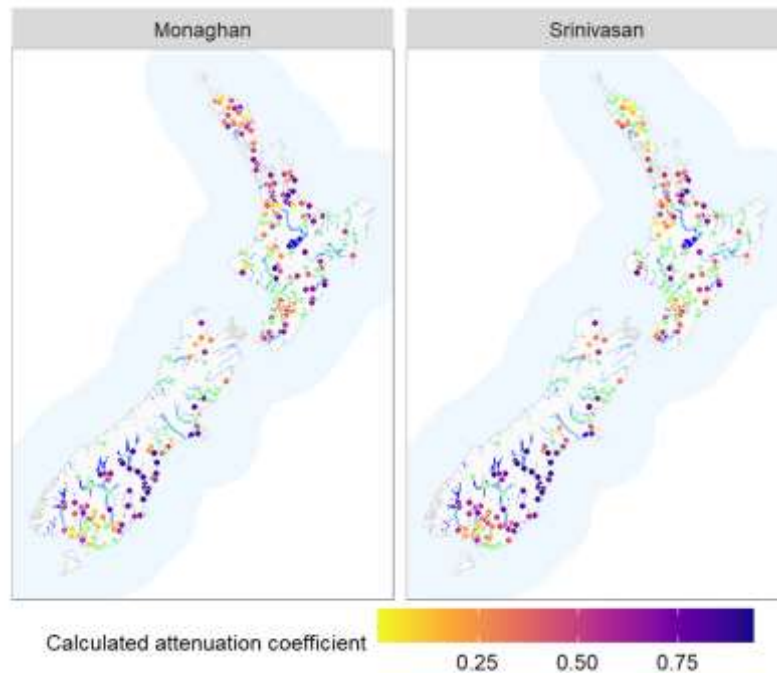

Figure 18: Maps of estimated TP attenuation coefficients. Sites shown with green dots had negative attenuation coefficients.

Spatial patterns in the calculated TP attenuation coefficients were explored by plotting these values against the proportion of catchment area occupied by the four most well occupied land land use categories (Figure 19). The relationships in these plots are less clear than the same plots for TN, although there are weak relationships that suggest TP export coefficients are underestimated for areas occupied by native land cover. Underestimation of diffuse TP loss rates may be because P export coefficients, particularly when derived from OVERSEER, may be missing a significant portion of the total P export associated with infrequent, but largescale, erosion events (Parfitt *et al.* 2007, 2013; Gray *et al.* 2016).
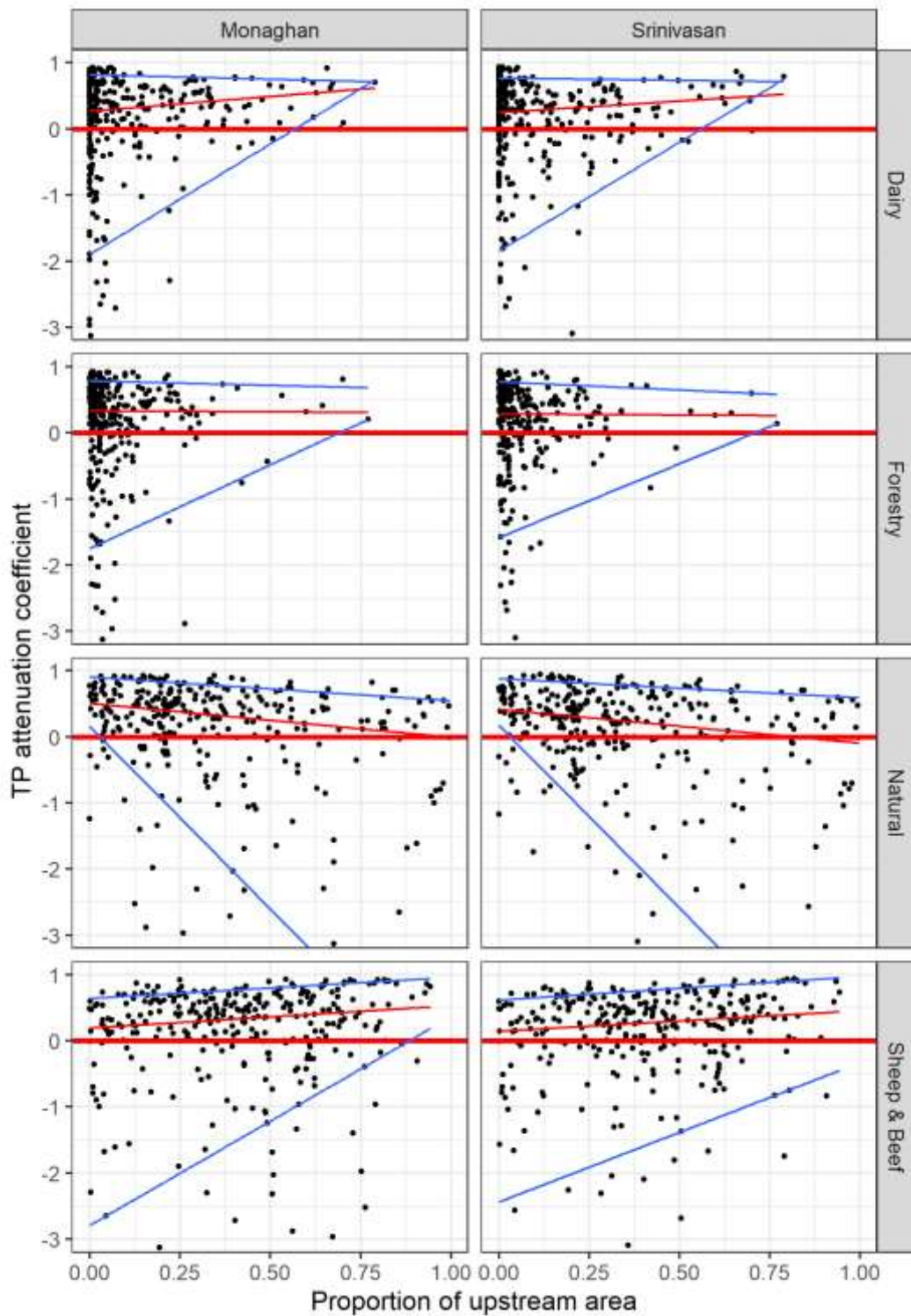
Figure 19: Scatter plots of evaluated TP attenuation coefficients against proportion of upstream land uses. Coloured lines are linear quantile regressions of the median (red) and 10% and 90% quantiles (blue).

5.4 Empirical Type concentration for TN

The median TN concentrations (model response) for the 778 sites were not normally distributed (right skewed), justifying the use of quantile regression (Figure 5). The proportion of significant fitted coefficient values decreased with increasing numbers of predictors for models pertaining to all three quantiles (i.e., median (0.5 quantile), 0.05 and 0.95 quantiles; Figure 20). There was also a trend in negative fitted coefficient values as the number of predictors included in the models increased (Figure 20).
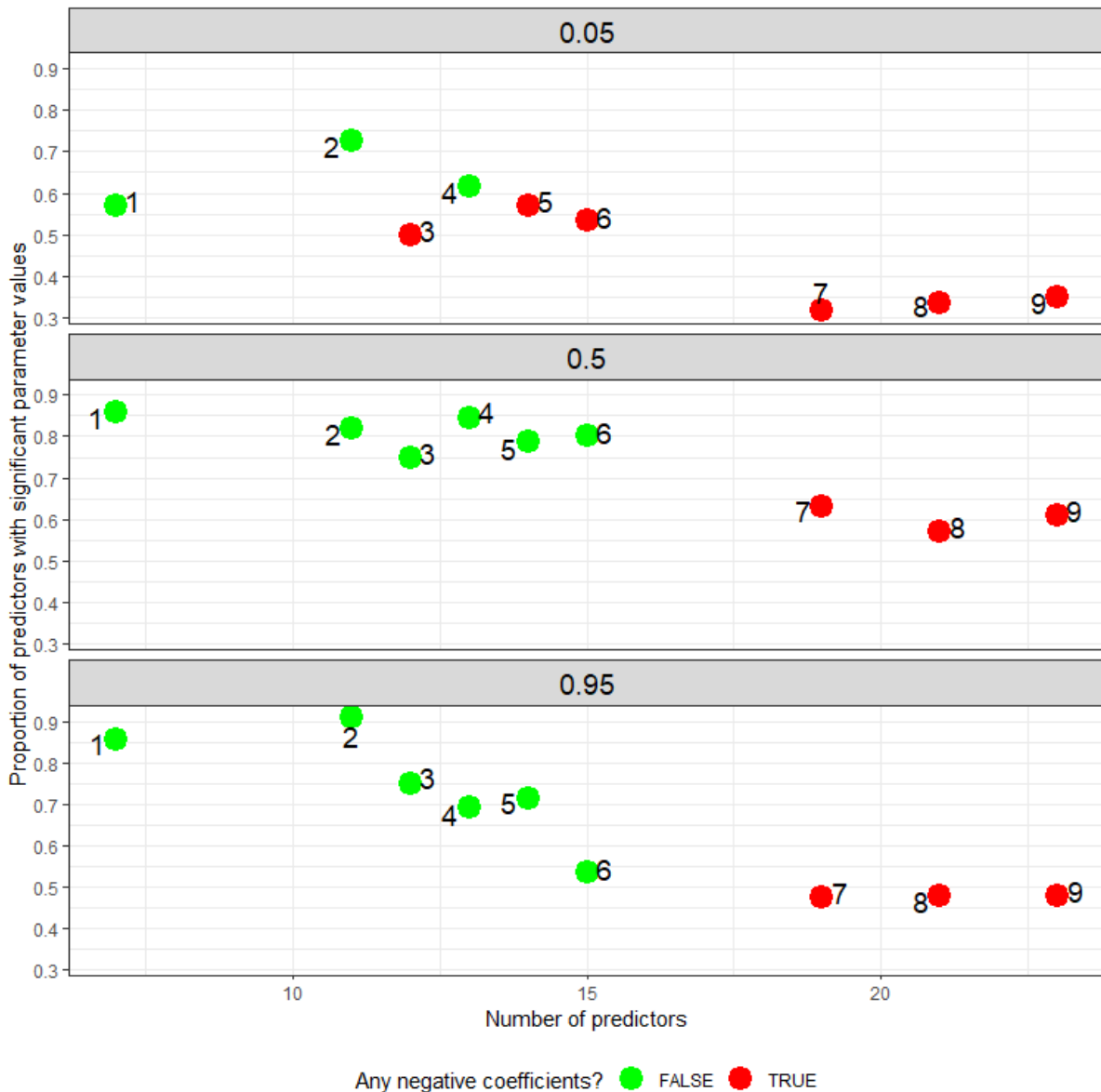
Figure 20. Proportion of significant coefficients versus number of predictors for quantile models for TN concentration fitted to nine sets of predictors (Types) for models pertaining to the 0.05, 0.5, and 0.95 quantiles. The numbers beside each point indicate the model number (1 to 9).

From nine potential quantile regression models, we judged Model 6 to be the best (see Table 15 in Appendix B). Quantile regression model 6 included 15 Types (excluding the Bare and

Water Types) for which, 53%, 73% and 60% of the fitted parameters were significant for the 0.05, 0.5, and 0.95 quantile models. There were no negative coefficients for the 0.5 and 0.95 quantile models and two negative coefficients for the 0.05 quantile model (Figure 21). The OLS models that included the same predictors as the quantile regression models had R^2 values between 62% (Model 1) to 67% (Model 9). The OLS model that had the same predictors as the Model 6 (the best model) had an R^2 value of 66%.

The fitted coefficients for Model 6 were generally consistent with expectations (see Section 4.5.3, Figure 21 and Table 6). For example, the highest values were associated with Cropland, Dairy and Sheep & Beef land uses, and the lowest values were associated with Natural land cover. The values for Forestry, compared with, for example, Sheep & Beef were higher than indicated by existing lookup tables and typologies. The standard errors for the fitted coefficients were largest for the Orchard & Vineyard land Type, which is consistent with the low occupancy of this land Type in our dataset.
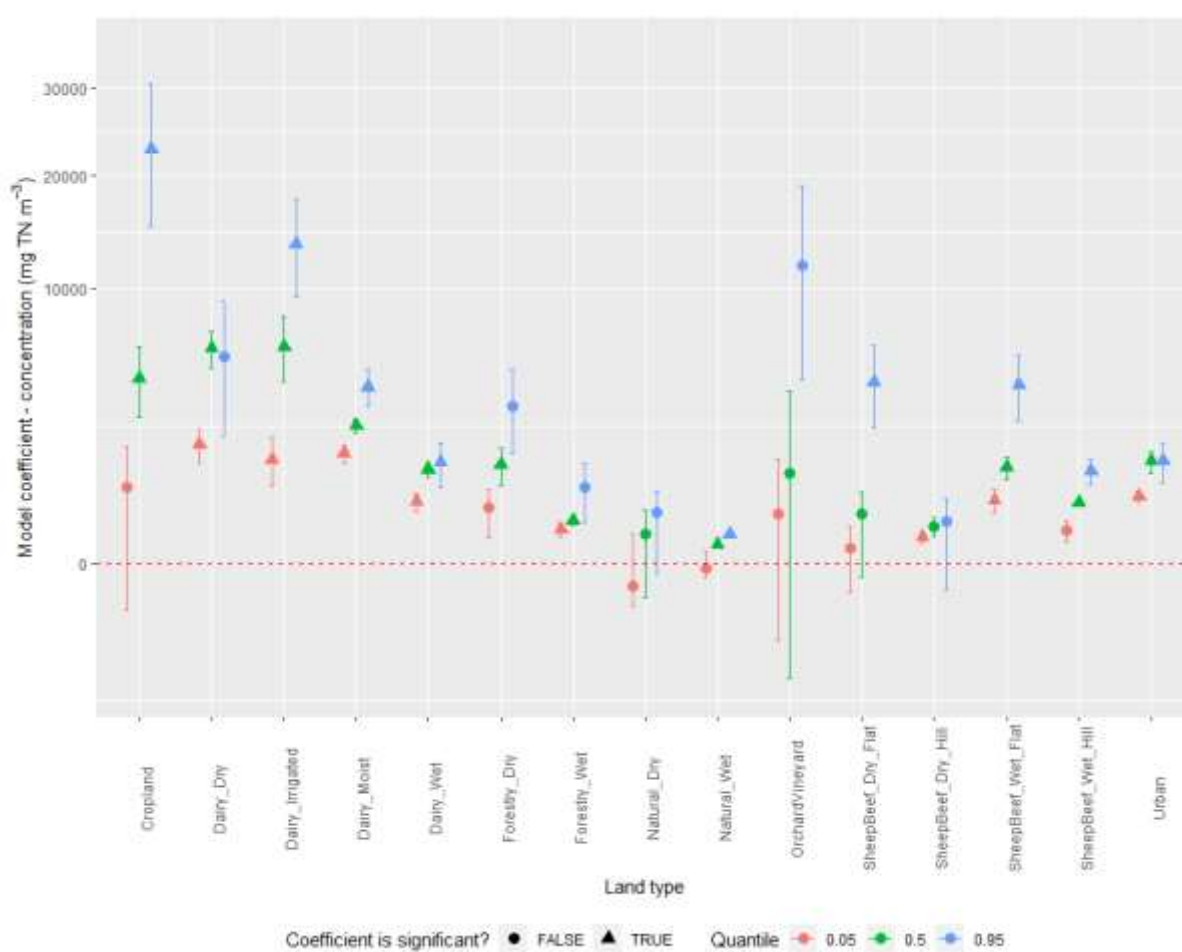
Figure 21. Fitted coefficients for the best TN concentration model (Model 6). Note that the y-axis is transformed to provide greater resolution of values with low magnitudes compared to higher magnitudes. The error bars indicate the standard errors for the fitted coefficients.

Table 6. ETC parameters for each of the 17 Types derived from the best empirical TN concentration model. The values can be interpreted as the contribution of each Type to catchment TN concentration (mg m⁻³). Note that the Types Bare and Water were excluded from the regression model and are assumed to have ETC values of zero.

Type	Best estimate	Prediction interval lower bound	Prediction interval Upper bound
Bare	0	0	0
Cropland	4464	759	22771
Dairy_Dry	6130	1856	5636
Dairy_Irrigated	6311	1422	13496
Dairy_Moist	2513	1580	4124
Dairy_Wet	1170	503	1338
Forestry_Dry	1272	402	3290
Forestry_Wet	237	154	771
Natural_Dry	110	-71	334
Natual_Wet	46	-4	111
OrchardVineyard	1102	318	11728
SheepBeef_Dry_Flat	314	31	4352
SheepBeef_Dry_Hill	183	86	232
SheepBeef_Wet_Flat	1193	526	4226
SheepBeef_Wet_Hill	477	143	1126
Urban	1368	584	1388
Water	0	0	0

The best empirical TN concentration model had good performance (0.50 < NSE ≤ 0.65; Table 2) based on the cross-validation analysis and the criteria of Moriasi et al. (2015; Table 8, Figure 22). The performance of the RF model was very good (NSE > 0.65; Table 2) based on the same criteria, which indicates the predictions of the empirical model are good compared to this benchmark. The mean of the proportion of predictions that were within the 90% prediction interval over the 10 cross validation folds was 88% (range 83% to 95%). This indicates that the 90% prediction interval is a reliable measure of the uncertainty of the empirical TN concentration model predictions.

Table 7. Performance of the best empirical model and RF models of TN concentration.

Model	NSE	PBIAS
Empirical (best estimate)	0.63	-0.6
RF	0.78	1.44
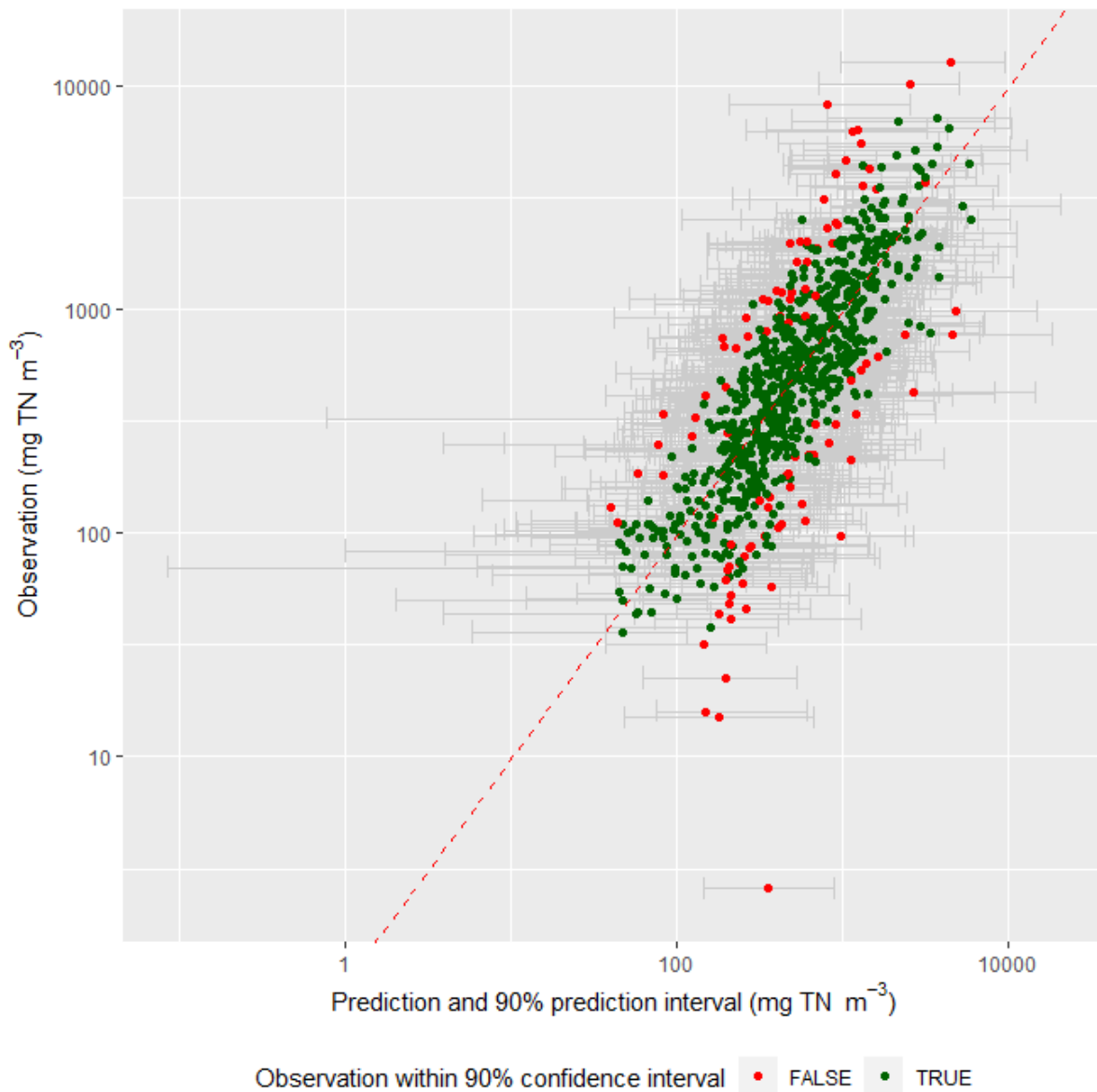
Figure 22. Observed versus predicted site median TN concentrations (points) and 90% prediction interval (grey error bars). The predictions and the estimated 90% prediction interval are independently derived for each water quality station by the cross validation. The green points indicate the observations that are within the 90% prediction interval and the red points indicate the observations that are outside of the 90% prediction interval. Note that the observed values are plotted on the Y-axis and predicted values on the X-axis, following Piñeiro et al. *(2008). Red line: one-to-one line.*

The mean of each of the coefficient values for each quantile (i.e., 5th, 50th and 95th quantiles) over the 10 versions of the best model fitted by cross validation were generally consistent with the fitted coefficients for the full models as indicated by points lying close to the one-to-one line in Figure 23. In addition, the standard deviations of the coefficient values over the 10 realisations of the models fitted by cross validation were approximately equal to the corresponding standard errors (see also Figure 23) for the fitted coefficients for the full model. This is an objective indication of the stability and reliability of the fitted parameters.
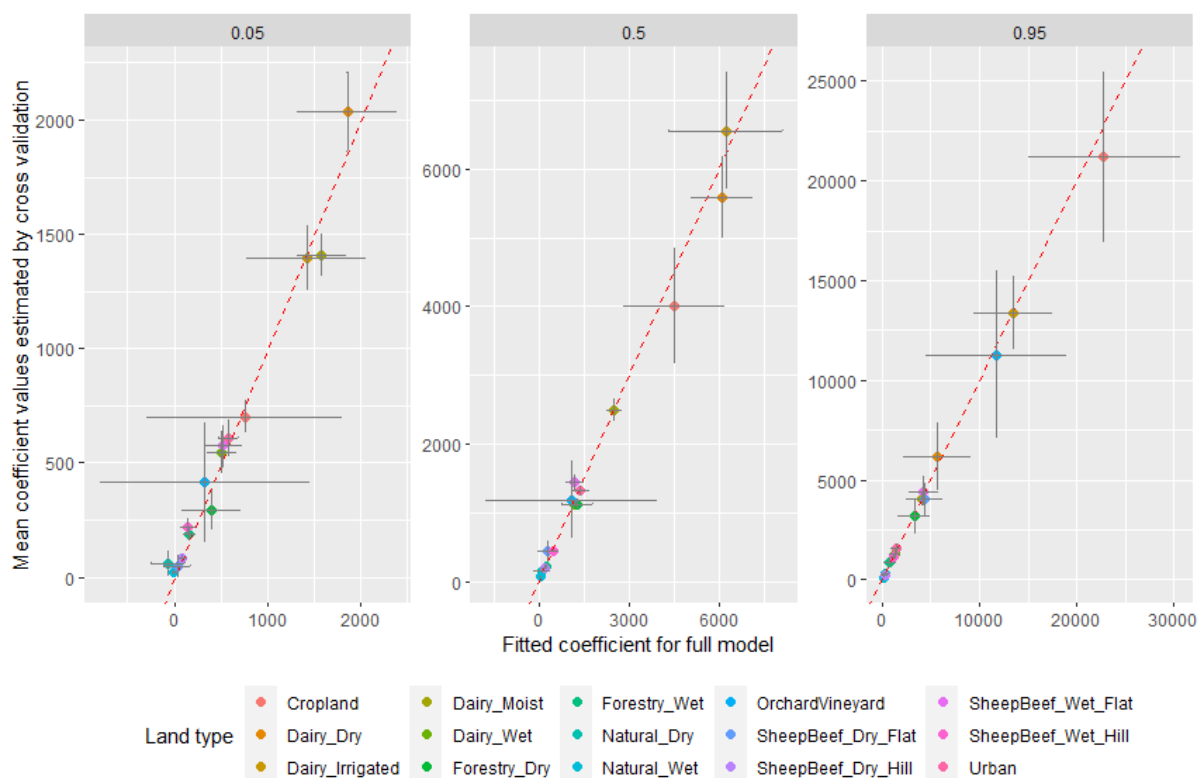
Figure 23. Comparison of coefficients fitted to each Type in the full TN concentration model with the mean of 10 realisations of the same coefficients fitted by cross validation. The vertical error bars indicate the standard deviation of the coefficients over the 10 cross validation folds. The horizontal error bars are the standard errors of the coefficients fitted in the full models. The red dashed line is one to one and indicates perfect agreement.

Finally, for each Type, the fitted model coefficients were reasonably consistent with the observations of TN concentrations at the water quality stations having high occupancy by that Type (Figure 24). For some Types, the data did not include many or any water quality stations with high (e.g., >0.7) occupancy. However, for some Types there was good representation by sites with high occupancy (e.g., Natural_Wet, Urban, SheepBeef_Wet_Hill, SheepBeef_Dry_Hill). In these cases, as the proportion occupancy increased, the central tendency of the observed TN concentrations converged on the fitted model coefficients (Figure 24). In addition, where there was not good representation by sites with high occupancy, the fitted coefficients were generally consistent with the extrapolated value of trend lines (red lines, Figure 24) fitted to the data.
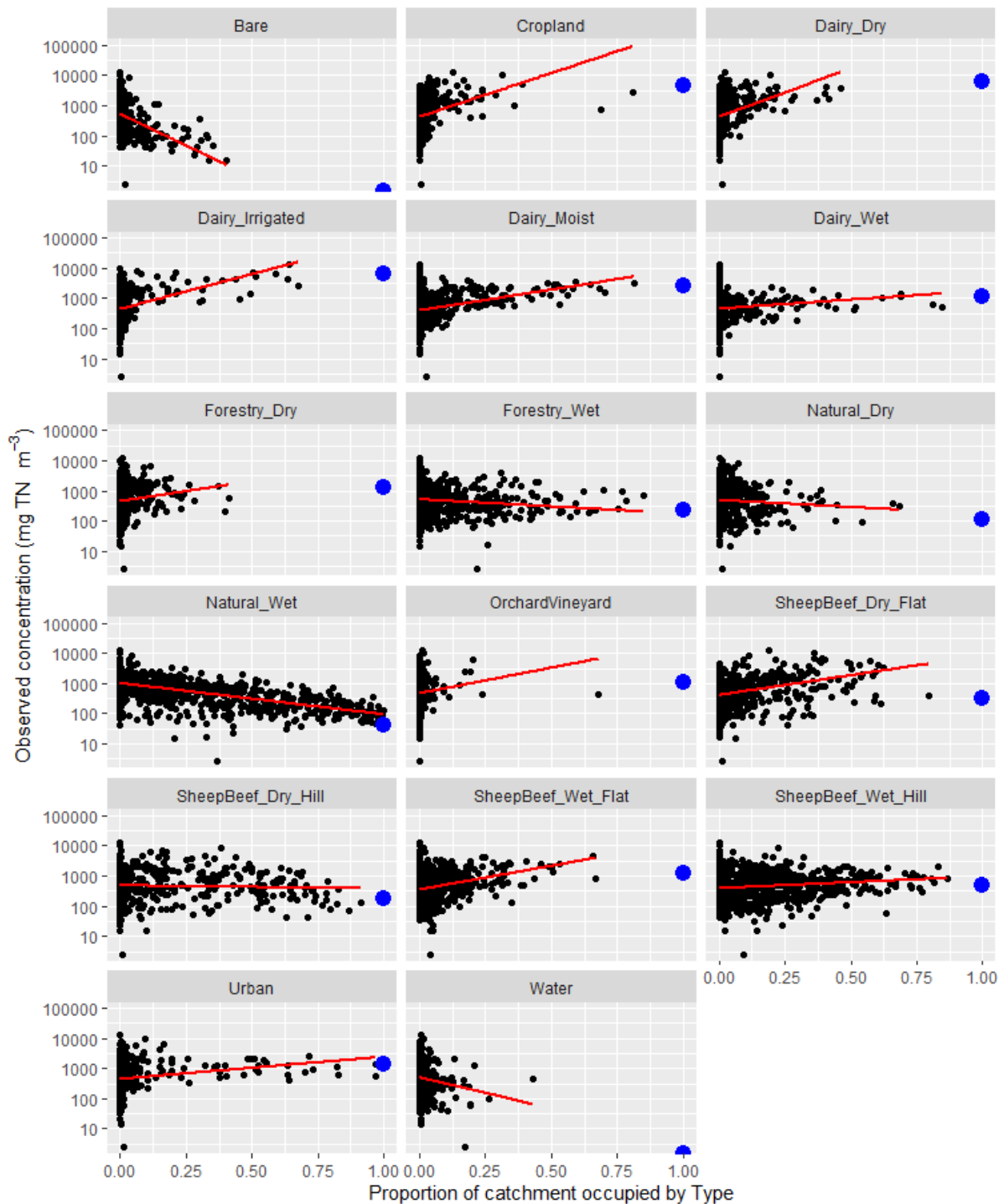
Figure 24. Observed TN concentrations versus the proportion of catchment occupied by each Type (panels). For each Type, the ETC values for the best model are indicated as a blue dot at the position on the x-axis indicating a proportion occupancy of 1. The red line is a linear regression indicating the expected value of TN concentration conditional on the proportion of catchment occupied by the Type. Note that the Types Bare and Water were excluded from the regression model and are assumed to have ETC values of zero.

5.5 Empirical Type yield for TN

The median TN yields (model response) for the 314 sites were not normally distributed (right skewed), justifying the use of quantile regression (Figure 6). The proportion of significant fitted coefficient values decreased with increasing numbers of predictors for models pertaining to all three quantiles (i.e., 0.5, 0.05 and 0.95 quantiles; Figure 25). For the 0.5 quantile (i.e., median) models, only Models 1 and 2 had no negative fitted coefficient values (Figure 25).
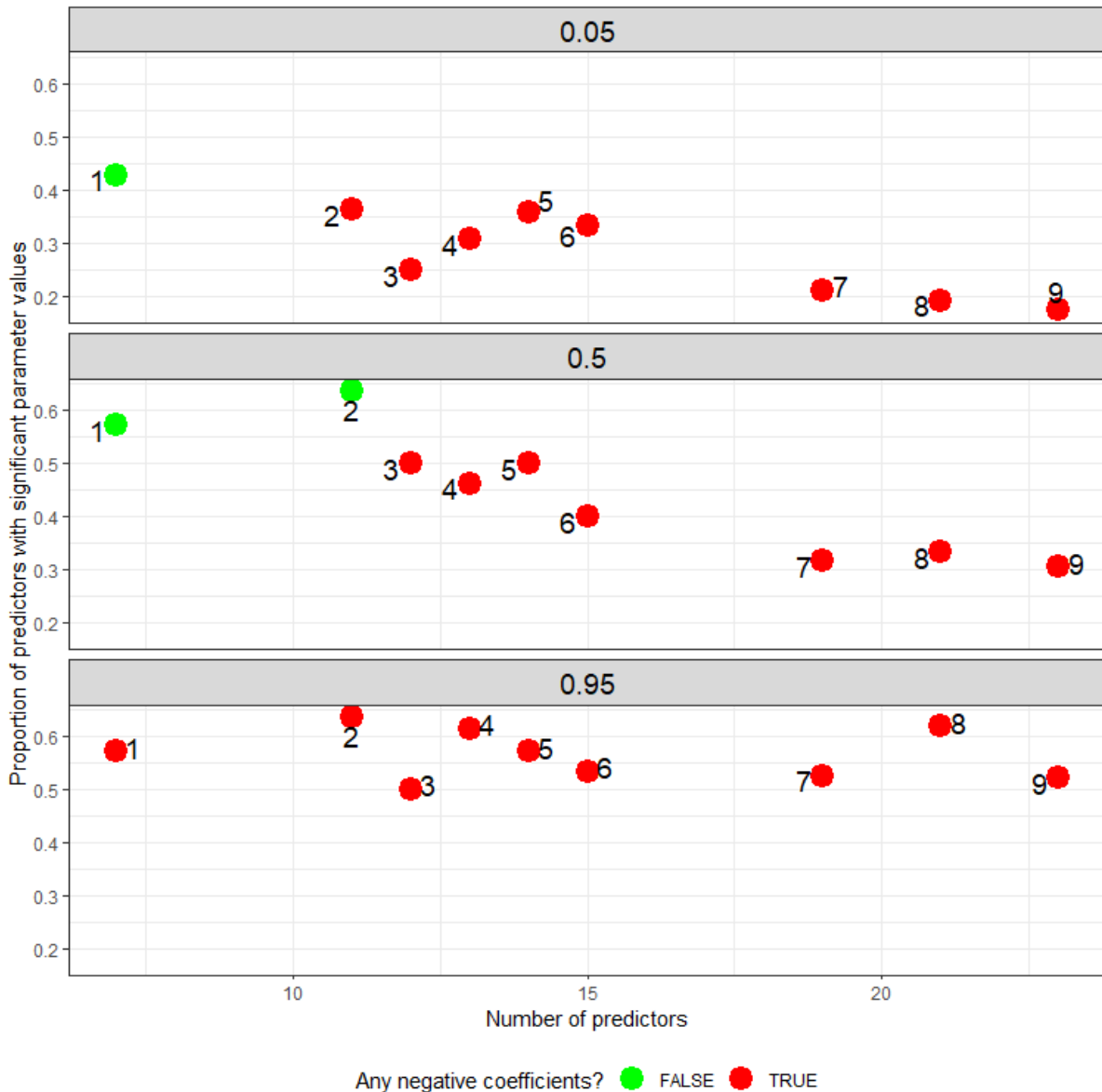
Figure 25. Proportion of significant coefficients versus number of predictors for quantile models for TN yield models fitted to nine sets of predictors (Types) for models pertaining to the 0.05, 0.5, and 0.95 quantiles. The numbers beside each point indicate the model number (1 to 9).

From the nine sets of possible models, we judged the Model 2 to be the best. For Model 2 there were no negative coefficients for the 0.5 quantile model and only one negative coefficient for the 0.05 and 0.95 quantile models. Model 2 comprised 11 Types (excluding the Bare and Water Types) for which, 36%, 73% and 64% of the fitted parameters were significant for the

0.05, 0.5 and 0.95 quantile models. The OLS models that included the same predictors as the quantile regression models had R^2 values between 39% (Model 1) to 62% (Model 9). The OLS model that had the same predictors as Model 2 (the best model) had an R^2 value of 40%.

The fitted coefficients for the Model 2 were generally consistent with expectations (see Section 4.5.3, Figure 26 and Table 8). For example, the highest values were associated with Dairy, followed by Orchard & Vineyard and Sheep & Beef land uses and the lowest values were associated with Natural. The coefficient for Dairy were highest in the wet and irrigated moisture Types, which is consistent with expectations. We considered that lower losses for hill Sheep & Beef compared to flat is consistent with the expectation that the former Type is generally occupied by lower intensity farm systems. The values for Forestry, compared with, for example, Sheep & Beef were higher than indicated by existing published lookup tables of typology-based export coefficients. We consider that this may be a realistic representation of losses from forestry considering that the derived coefficient will represent the contribution of various stages of the forestry growth and harvesting cycle, not just the contribution from mature trees. The standard errors for the fitted coefficients were largest for the Orchard & Vineyard followed by Cropland Types, which is consistent with the low occupancy of these Types in our dataset and may also reflect variation of land management within these Types (e.g., intensive market gardens vs arable).
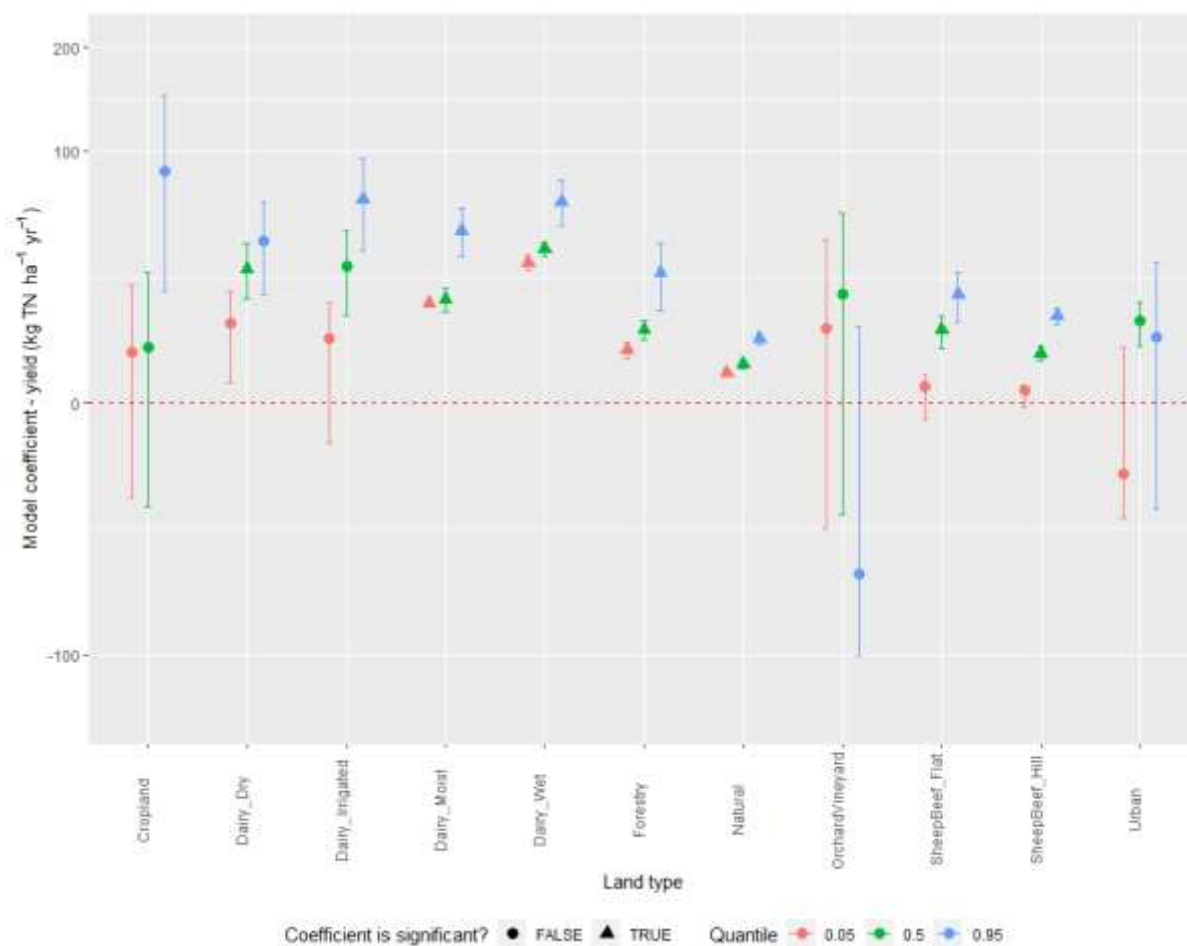
Figure 26. Fitted coefficients for the best TN yield model (Model 2). Note that the y-axis is transformed to provide greater resolution of values with low magnitudes compared to higher magnitudes. The error bars indicate the standard errors for the fitted coefficients.

Table 8. ETY parameters for each of the 13 Types derived from the best empirical TN yield model. The values can be interpreted as the contribution of each Type to catchment TN yield (kg ha^{-1} yr^{-1}). Note that the Types Bare and Water were excluded from the regression model and are assumed to have ETY values of zero.

Type	Best estimate	Prediction interval lower bound	Prediction interval upper bound
Bare	0	0	0
Cropland	4.9	4	84.9
Dairy_Dry	28.5	10.2	41.5
Dairy_Irrigated	29.6	6.6	65.4
Dairy_Moist	17	15.9	47
Dairy_Wet	37.5	31.3	64.1
Forestry	8.5	4.4	27
Natural	2.4	1.4	6.6
OrchardVineyard	18.6	8.7	-45.8
SheepBeef_Flat	8.3	0.4	18.7
SheepBeef_Hill	3.9	0.2	12
Urban	10.7	-7.9	6.8
Water	0	0	0

The best TN yield model (Model 2) had satisfactory performance (0.35 < NSE ≤ 0.50; Table 2) based on the cross-validation analysis and the criteria of Moriasi et al. (2015; Table 8, Figure 22). The performance of the RF model was very good (NSE > 0.65; Table 2) based on the same criteria, which indicates the predictions of the 0.5 quantile model are fair compared to this benchmark. The mean of the proportion of predictions that were within the 90% prediction interval over the 10 cross validation folds was 86% (range 74% to 93%). This indicates that the 90% prediction interval is a reliable measure of the uncertainty of the empirical TN yield model predictions.

Table 9. Performance of the 0.5 quantile and RF models of TN yield.

Model	NSE	PBIAS
Empirical (best estimate)	0.36	-4.7
RF	0.72	-0.8
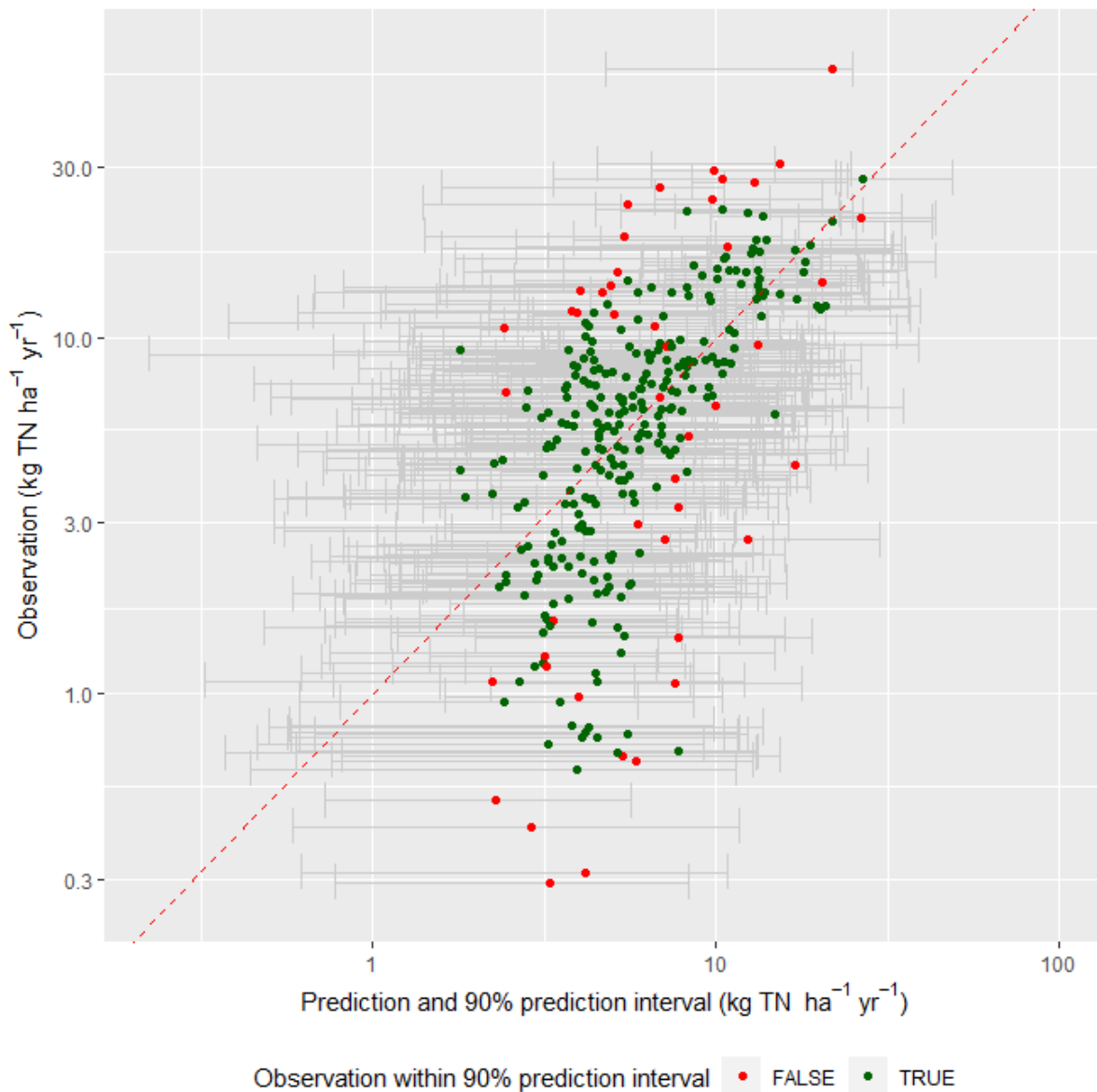
Figure 27. Observed versus predicted site median TN yield (points) and 90% prediction interval (grey error bars). The predictions and the estimated 90% confidence interval are independently derived for each water quality station by the cross validation. The green points indicate the observations that are within the 90% prediction interval and the red points indicate the observations that are outside of the 90% prediction interval. Note that the observed values are plotted on the Y-axis and predicted values on the X-axis, following Piñeiro et al. (2008). Red line: one-to-one line.

The mean of each of the coefficient values for each quantile (i.e., 0.05, 0.5 and 0.95 quantiles) over the 10 versions of the best model fitted by cross validation were generally consistent with the fitted coefficients for the full models as indicated by points lying close to the one-to-one line in Figure 28. In addition, the standard deviations of the coefficient values over the 10 realisations of the models fitted by cross validation were approximately equal to the corresponding standard errors (see also Figure 28) for the fitted coefficients for the full model. This is an objective indication of the stability and reliability of the fitted parameters.

Figure 28. Comparison of coefficients fitted to each land Type in the full TN yield model with the mean of 10 realisations of the same coefficients fitted by cross validation. The vertical error bars indicate the standard deviation of the coefficients over the 10 cross validation folds. The horizontal error bars are the standard errors of the coefficients fitted in the full models. The red dashed line is one to one and indicates perfect agreement.

Finally, for each Type, the fitted model coefficients were reasonably consistent with the observations of TN yields at the water quality stations having high occupancy by that Type (Figure 30). For some Types, the data did not include many or any water quality stations with high (e.g., >0.7) occupancy. However, for some Types there was good representation by sites with high occupancy (e.g., Natural, Urban, SheepBeef_Hill). In these cases, as the proportion occupancy increased, the central tendency of the observed TN yields converged on the fitted model coefficients (Figure 30). In addition, where there was not good representation by sites with high occupancy, the fitted coefficients were generally consistent with the extrapolated value of trend lines (red lines, Figure 30) fitted to the data.
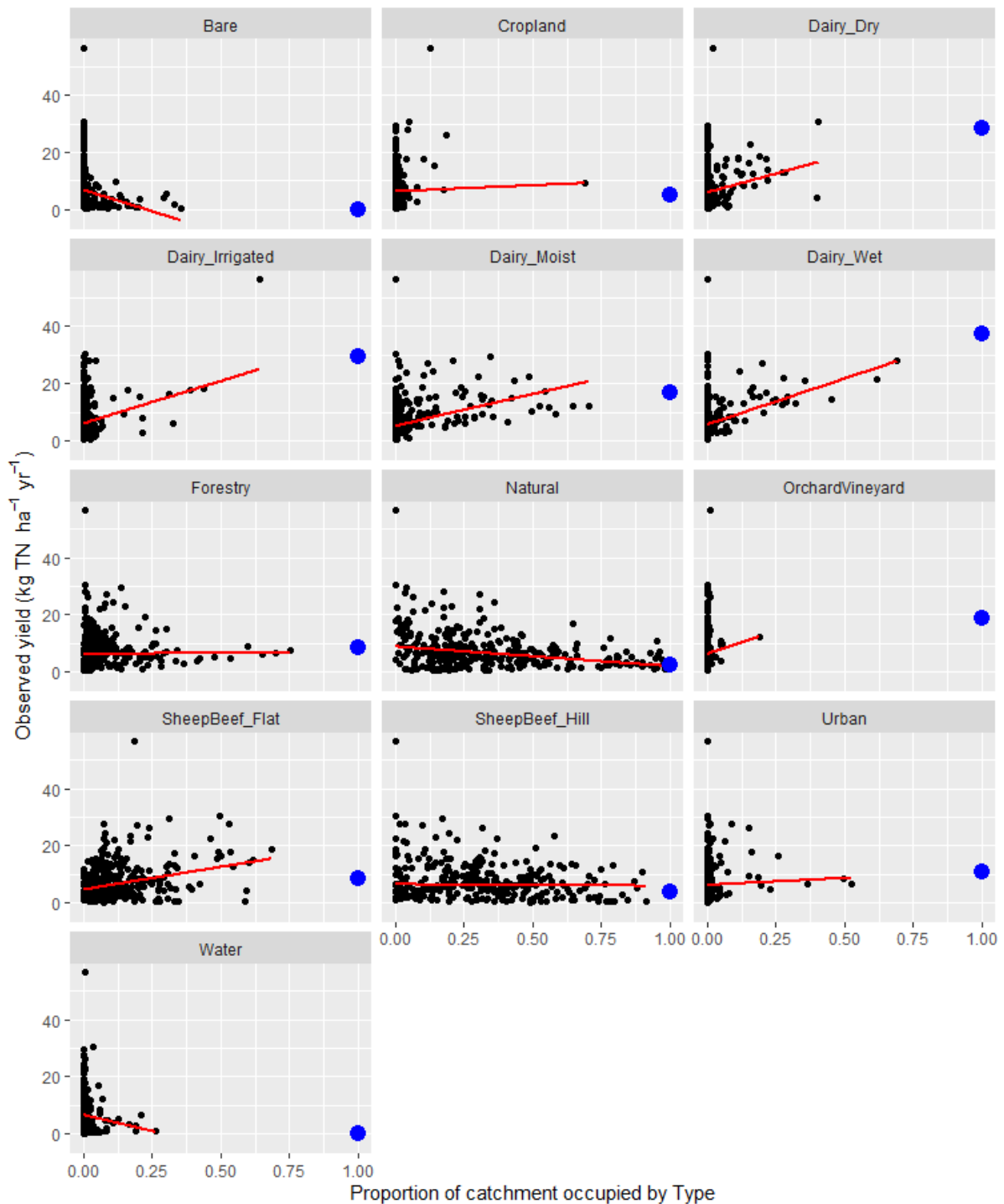
Figure 29. Observed TN yields versus the proportion of catchment occupied by each Type (panels). For each Type, the ETY values for the best model are indicated as a blue dot at the position on the x-axis indicating a proportion occupancy of 1. The red line is a linear regression indicating the expected value of TN yield conditional on the proportion of catchment occupied by the Type. Note that the Types Bare and Water were excluded from the regression model and are assumed to have ETY values of zero.

5.6   Empirical Type concentration for TP

The median TP concentrations (model response) for the 749 sites were not normally distributed (right skewed), justifying the use of quantile regression (Figure 5). The proportion of fitted coefficient values that were significant decreased with increasing numbers of predictors for models pertaining to all three quantiles (i.e., 0.05, 0.5 and 0.95 quantiles; Figure 30). There were negative fitted coefficient values in all of the models (Figure 20). All models contained the land Types Bare and Water and these were always associated with negative coefficients.
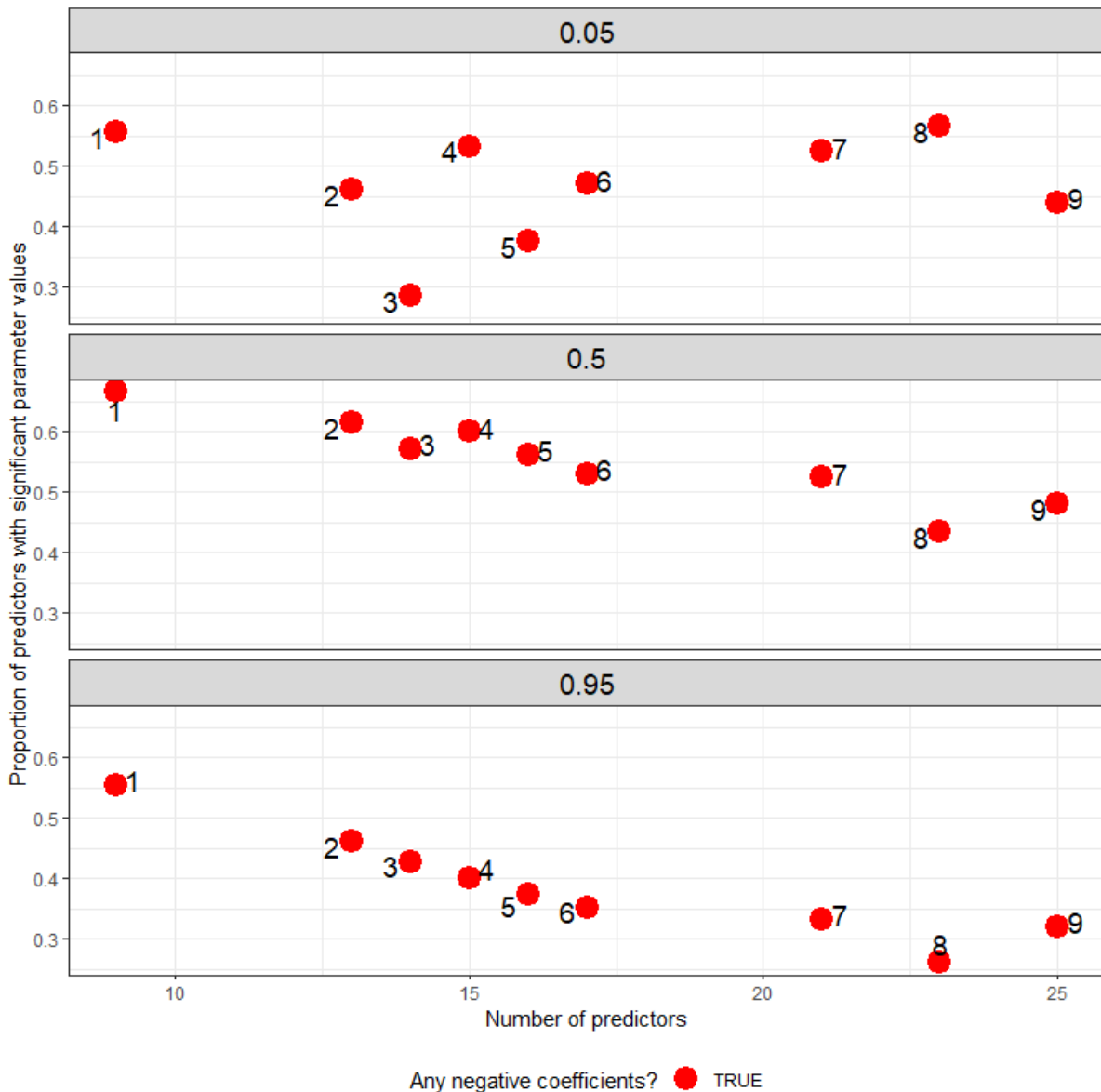
Figure 30. *Proportion of significant coefficients versus number of predictors for quantile models for TP concentration fitted to nine sets of predictors (Types) for models pertaining to the 0.05, 0.5, and 0.95 quantiles. The numbers beside each point indicate the model number (1 to 9).*

From the nine models, we judged Model 2 to be the best. This implies an acceptance of the negative coefficients fitted to the Bare and Water land Types. The negative contribution of the

Water may have some physical justification in terms of trapping of TP in impoundments. However, the negative coefficient for Bare has no obvious physical justification. The model pertaining to Set 2 (i.e., the best model) comprised 13 land Types for which, 46%, 62% and 46% of the fitted coefficients were significant for the 0.05, 0.5, and 0.95 quantile models (Figure 31). The OLS models that included the same predictors as the quantile regression models had R^2 values between 32% (Model 1) to 50% (Model 9). The OLS model that had the same predictors as Model 2 (the best model) had an R^2 value of 39%.

The fitted coefficients representing the unit contributions of the Types to TP concentrations (Figure 31 and Table 10) were inconsistent with expectations. For example, for the 0.5 quantile model, the coefficients associated with Dairy ranged between 174 (Dairy_Dry; Table 10) and 10 (Dairy_Irrigated; Table 10). This is a large variation within the Dairy land use category for which we have no physical justification. In addition, the coefficient associated with Natural was 7 however, the coefficients for Dairy_Irrigated and SheepBeef_Wet_Flat were also very low 10, and 1, respectively (Table 10). We have no physical justification for lower TP contributions by productive land uses compared to natural state (i.e., Natural land cover). We consider this is because our simple Types do not discriminate variation in important natural processes that determine P concentrations including geogenic geogenic supply, mobilisation (erodibility, rainfall slope etc) and microbially mediated reduction-oxidation, which we discuss in our conclusions.

Table 10. ETC parameters for each of the 13 Types derived from the best empirical TP concentration model. The values can be interpreted as the contribution of each Type to catchment TP concentration (mg m^{-3}).

Type	Best estimate	Prediction interval lower bound	Prediction interval upper bound
Bare	-22	-14	-19
Cropland	19	7	1117
Dairy_Dry	174	85	564
Dairy_Irrigated	10	9	138
Dairy_Moist	130	61	287
Dairy_Wet	17	-17	54
Forestry	31	10	92
Natural	7	3	21
OrchardVineyard	45	17	44
SheepBeef_Flat	1	2	-11
SheepBeef_Hill	33	5	82
Urban	55	45	70
Water	-18	-7	162
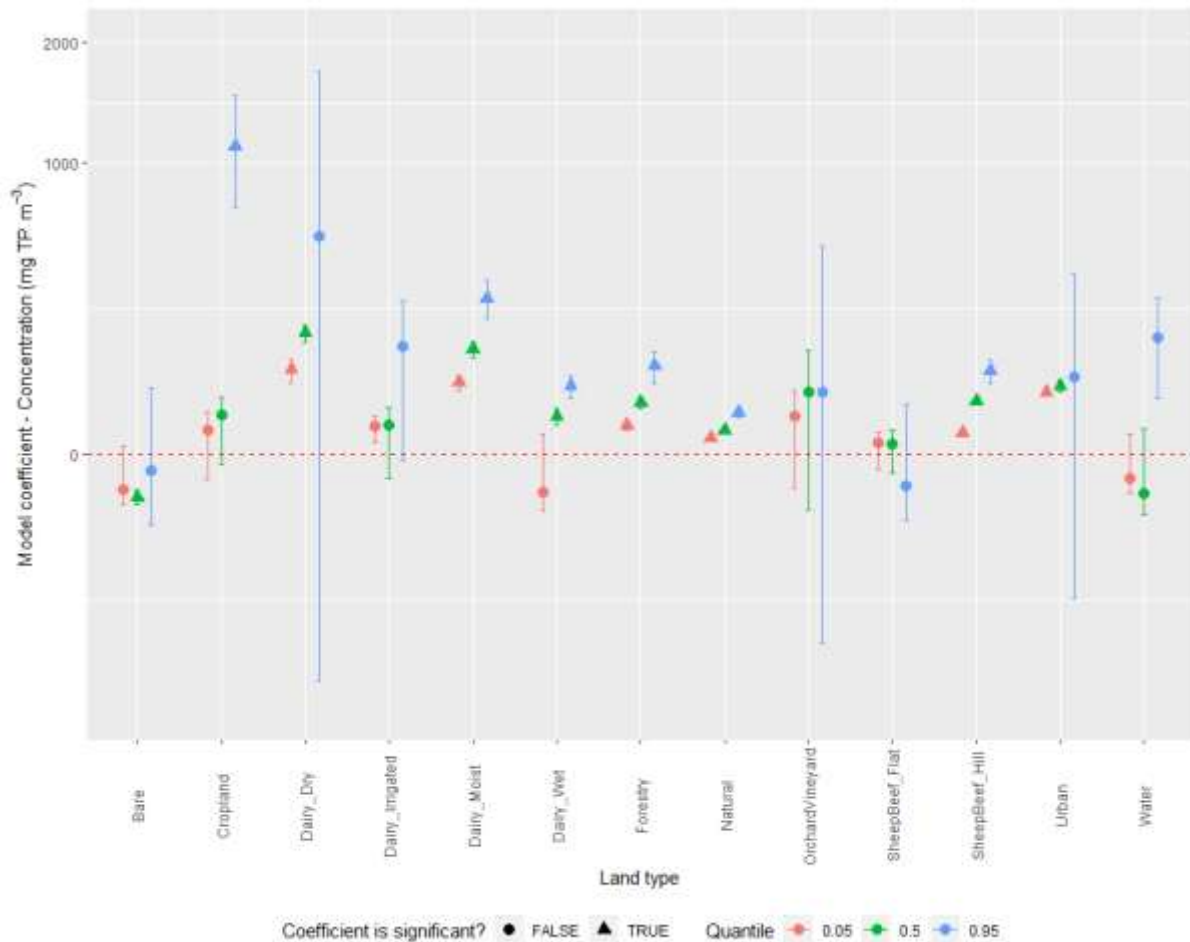
Figure 31. Fitted coefficients for the best TP concentration model (Model 5). Note that the y-axis is transformed to provide greater resolution of values with low magnitudes compared to higher magnitudes. The error bars indicate the standard errors for the fitted coefficients.

The best TP concentration model (Model 2) had unsatisfactory performance (NSE < 0.35; Table 2) based on the cross-validation analysis and the criteria of Moriasi et al. (2015; Table 11, Figure 32). The performance of the RF model was very good based on the same criteria, which indicates the predictions of the 0.5 quantile model are poor compared to this benchmark. The mean of the proportion of predictions that were within the 90% prediction interval over the 10 cross validation folds was 12.7% and the 95% prediction interval extended between 8% and 17% (i.e., contained the expected proportion of 10%). This indicates that the 90% prediction interval is a reliable measure of the imprecision of the quantile regression model predictions.

Table 11. Performance of the 0.5 quantile and RF models of TP concentration.

Model	NSE	PBIAS
0.5 quantile	0.30	0.3
RF	0.68	0.13
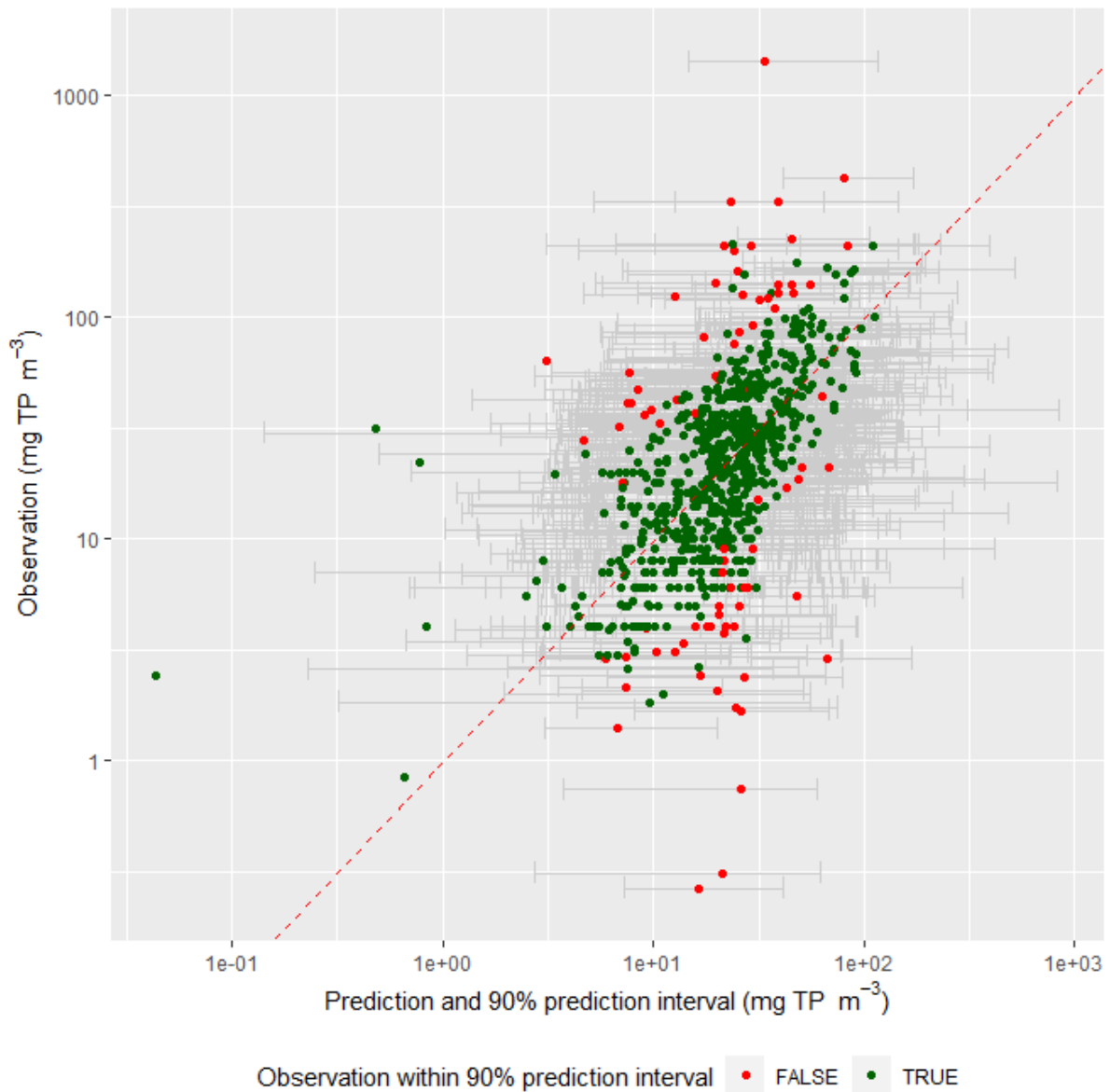
Figure 32. Observed versus predicted site median TP concentrations (points) and 90% prediction interval (grey error bars). The predictions and the estimated 90% prediction interval are independently derived for each water quality station by the cross validation. The green points indicate the observations that are within the 90% prediction interval and the red points indicate the observations that are outside of the 90% prediction interval. Note that the observed values are plotted on the Y-axis and predicted values on the X-axis, following Piñeiro et al. (2008). Red line: one-to-one line.

The mean of each of the coefficient values for each quantile (i.e., 0.05, 0.5 and 0.95 quantiles) over the 10 realisations of the best model fitted by cross validation were consistent with the fitted coefficients for the full models as indicated by points lying close to the one-to-one line in Figure 33. In addition, the standard deviations of the coefficient values over the 10 realisations of the models fitted by cross validation were approximately equal to the corresponding standard errors (Figure 33) for the fitted coefficients for the full model.
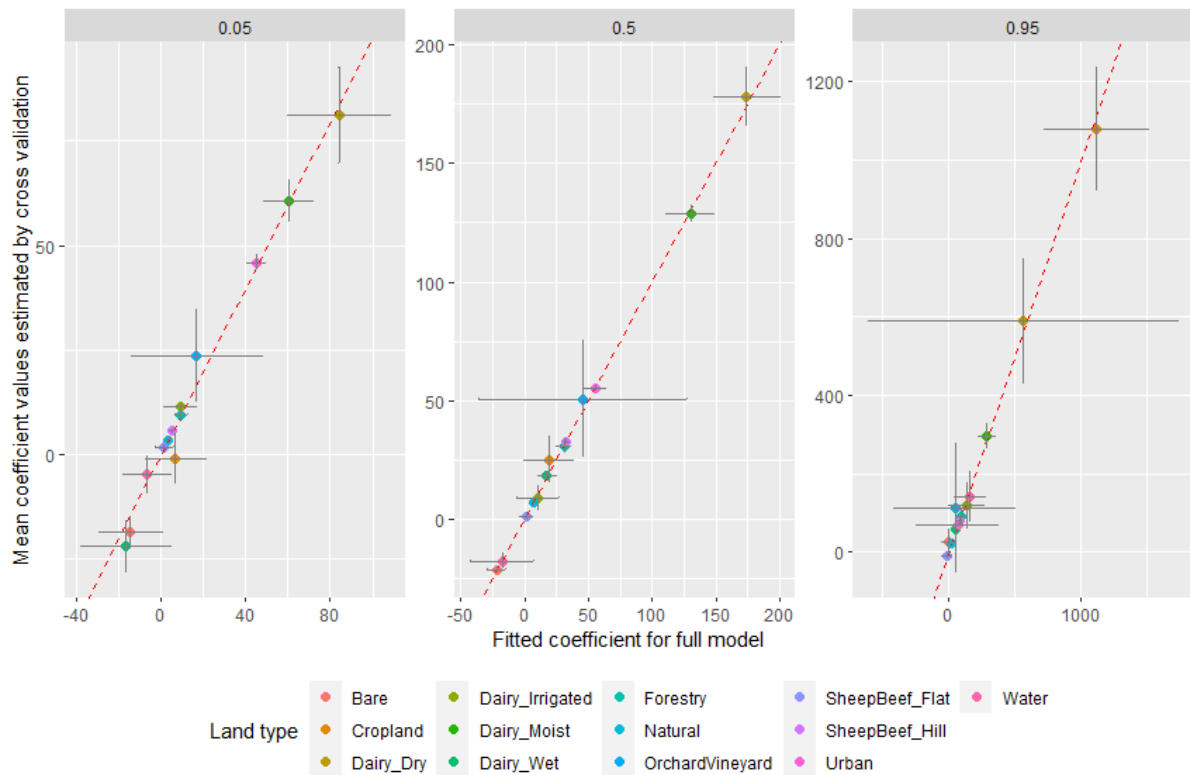
Figure 33. Comparison of coefficients fitted to each Type in the full TP concentration model with the mean of 10 realisations of the same coefficients fitted by cross validation. The vertical error bars indicate the standard deviation of the coefficients over the 10 cross validation folds. The horizontal error bars are the standard errors of the coefficients fitted in the full models. The red dashed line is one to one and indicates perfect agreement.

5.7 Empirical Type yield for TP

The median TP yields (model response) for the 312 sites were not normally distributed (they were right skewed), justifying the use of quantile regression (Figure 6). For the TP yield models, Sets 1, 2 and 3 comprised between four and six land Types that were combinations of climate and slope categories (see Appendix B). Model 4 comprised only the nine land use Types. Set 5 and 6 comprised 18 and 19 Types, respectively that were combinations of land use, climate and slope categories. The proportion of significant fitted coefficient values generally decreased with increasing numbers of predictors for models pertaining to all three quantiles (i.e., 0.5, 0.05 and 0.95 quantiles; Figure 34).

For the 0.5 quantile (i.e., median) models, Models 1, 2, and 3 had no negative fitted coefficient values (Figure 34). The OLS model that included the same predictors as Model 4 (which comprised the nine land use categories), had an adjusted R^2 value of 4%. This indicates that the signal of land use Type barely rises above the overall between-site variability in TP yields. The maximum R^2 values for the OLS model was only 28% indicating that there was appreciable unexplained variation in TP yields.
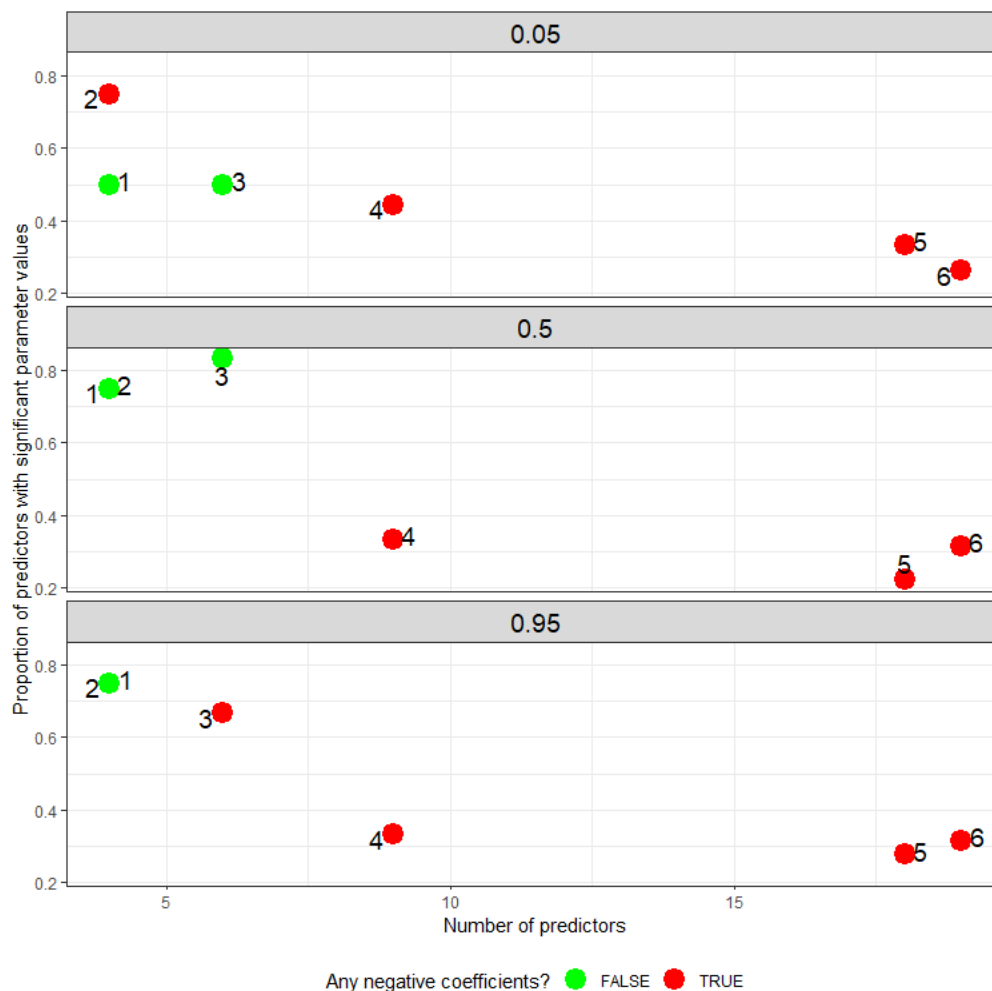
Figure 34. Proportion of significant coefficients versus number of predictors for quantile models for TP yield fitted to six sets of predictors (Types) for models pertaining to the 0.05, 0.5, and 0.95 quantiles. The numbers beside each point indicate the model number (1 to 9).

From the six models, we judged the Models 3 to be the best. For Model 3, 50%, 83% and 66% of the fitted coefficients were significant for the 0.05, 0.5 and 0.95 quantile models respectively. The were no negative coefficients for the 0.05 and 0.5 quantile models and only one negative coefficient for the 0.95 quantile model. The OLS model that had the same predictors as Model 3 had an R^2 value of 20%.

The fitted coefficients for the best model were generally inconsistent with expectations for the unit contributions of the land Types to TP yields (Figure 35 and Table 12). For example, the coefficients for land Types that comprised the Flat slope category were higher than for the Hill slope category, which is inconsistent with the general understanding of P loss being strongly influenced by sediment loss which is greater for hilly areas (Sharpley *et al.* 2001). We consider this is because our simple Types do not discriminate variation in important natural processes that determine P yields including geogenic geogenic supply, mobilisation (erodibility, rainfall slope etc) and microbially mediated reduction-oxidation, which we discuss in our conclusions.
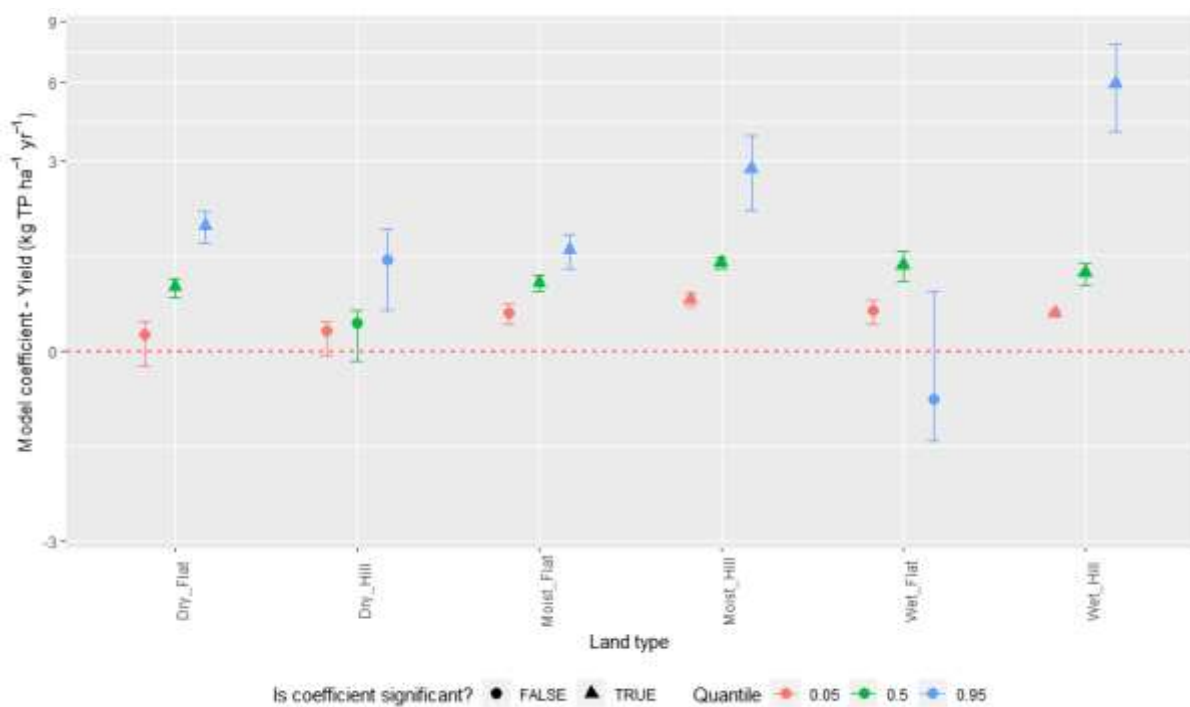
Figure 35. Fitted coefficients for the best TP yield model (Model 3). Note that the y-axis is transformed to provide greater resolution of values with low magnitudes compared to higher magnitudes. The error bars indicate the standard errors for the fitted coefficients.

Table 12. ETY parameters for each of the 6 Types derived from the best empirical TP yield model. The values can be interpreted as the contribution of each Type to catchment TP yield (kg ha^{-1} yr^{-1}).

Type	Best estimate	Prediction interval lower bound	Prediction interval upper bound
Dry_Flat	0.33	0.02	1.28
Dry_Hill	0.06	0.03	0.68
Moist_Flat	0.39	0.12	0.85
Moist_Hill	0.64	0.22	2.73
Wet_Flat	0.61	0.14	-0.2
Wet_Hill	0.5	0.12	5.9

The best TP concentration model (Model 3) had unsatisfactory performance (NSE < 0.35; Table 2) based on the cross-validation analysis and the criteria of Moriasi et al. (2015; Table 11, Figure 32). The performance of the RF model was good (0.50 < NSE ≤ 0.65; Table 2) based on the same criteria, which indicates the predictions of the 0.5 quantile model are poor compared to this benchmark. The mean of the proportion of predictions that were within the 90% confidence interval over the 10 cross validation folds was 11% and the 95% confidence interval extended between 2% and 20% (i.e., contained the expected proportion of 10%). This indicates that the 90% confidence interval is a reliable measure of the imprecision of the quantile regression model predictions.

Table 13. Performance of the 0.5 quantile and RF models of TP concentration.

Model	NSE	PBIAS
0.5 quantile	0.19	-3.8
RF	0.53	2.1
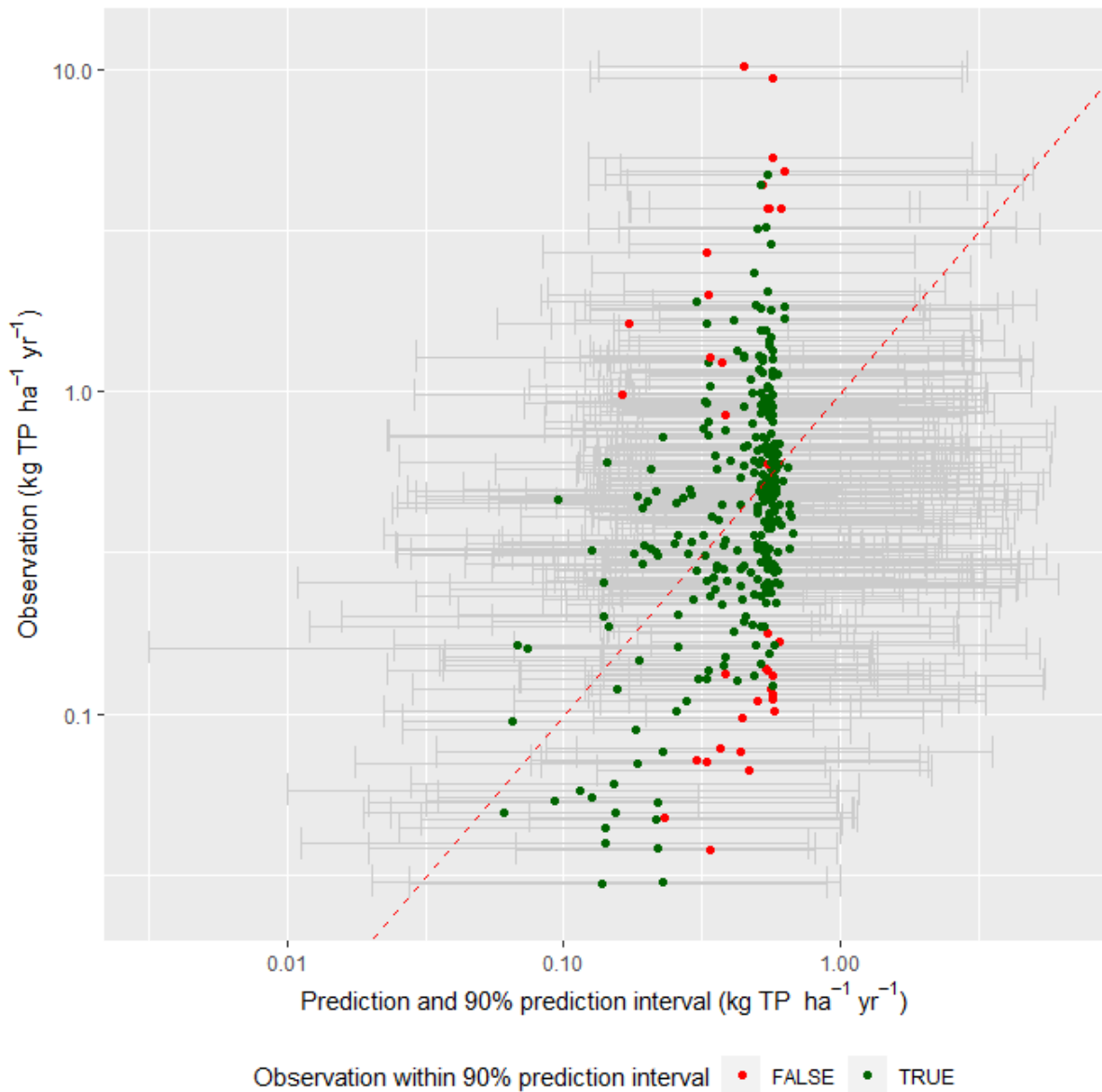
Figure 36. Observed versus predicted site median TP yield (points) and 90% prediction interval (grey error bars). The predictions and the estimated 90% prediction interval are independently derived for each water quality station by the cross validation. The green points indicate the observations that are within the 90% prediction interval and the red points indicate the observations that are outside of the 90% prediction interval. Note that the observed values are plotted on the Y-axis and predicted values on the X-axis, following Piñeiro et al. (2008). Red line: one-to-one line.

The mean of each of the coefficient values for each quantile (i.e., 0.05, 0.5 and 0.95 quantiles) over the 10 versions of the best model fitted by cross validation were generally consistent with the fitted coefficients for the full models as indicated by points lying close to the one-to-one line in Figure 28. In addition, the standard deviations of the coefficient values over the 10 realisations of the models fitted by cross validation were approximately equal to the corresponding standard errors (see also Figure 28) for the fitted coefficients for the full model. This is an objective indication of the stability and reliability of the fitted parameters.

Figure 37. Comparison of coefficients fitted to each Type in the full TP yield model with the mean of 10 realisations of the same coefficients fitted by cross validation. The vertical error bars indicate the standard deviation of the coefficients over the 10 cross validation folds. The horizontal error bars are the standard errors of the coefficients fitted in the full models. The red dashed line is one to one and indicates perfect agreement.

5.8   Empirical model application

Direct application of the empirical concentration model to the Manawatū River basin resulted in agreement, within ±10%, between predicted and measured instream concentrations at six of thirty-two (19%) monitoring stations. For the empirical yield model, eleven of thirty-two (34%) predictions achieved ±10% agreement with measured instream yields. Scatter plots of observations versus predictions for the empirical models are shown in Figure 38 and Figure 39,. Figure 38 indicates a consistent over prediction by the empirical concentration model (i.e., negative bias). This is confirmed by the calculated PBIAS value for the empirical concentration model (-29%) (Table 13). Both the NSE (reproducing observations) and PBIAS (bias) metrics indicate an "unsatisfactory" performance rating (Table 3) for this application of the empirical concentration model. The empirical yield model achieved a "satisfactory" rating for NSE and a "very good" rating for PBIAS.

Table 14. TN empirical model performance summary for Manawatū River basin. Bias and RMSD are in units of mg m⁻³ and kg ha⁻¹yr⁻¹ for concentration and yield, respectively.

Empirical Model	Proportion of predictions within the 90% PI	R^2	NSE	Bias	PBIAS	RMSD
Concentration	84%	0.13	-1.7	-210	-29%	486
Yield	97%	0.55	0.47	0.25	2.6%	2.8
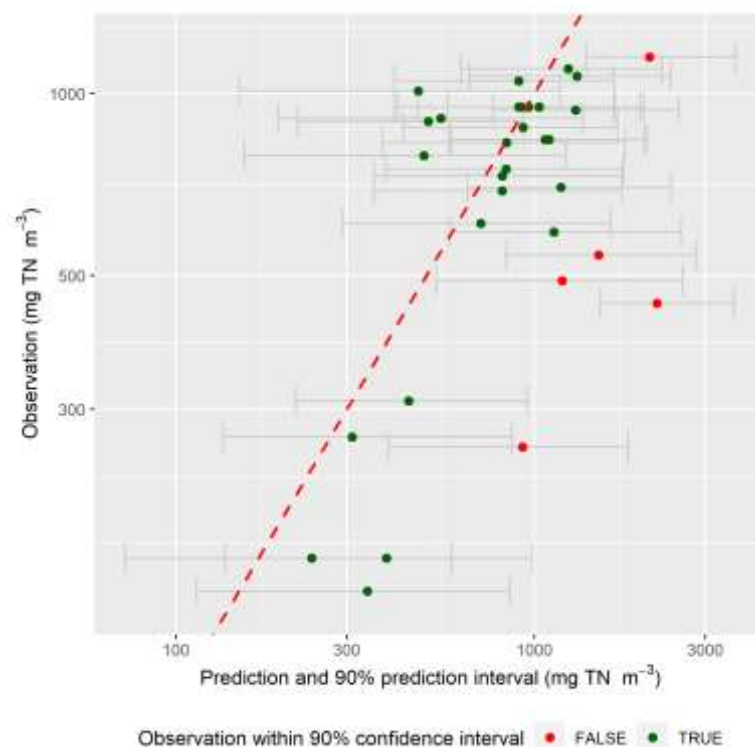
Figure 38. Observations versus empirical concentration model predictions of median TN concentration for sites within the Manawatu basin. The green points indicate the observations that are within the 90% prediction interval (grey error bars) and the red points indicate the observations that are outside of the 90% prediction interval. Note that the observed values are plotted on the Y-axis and predicted values on the X-axis, following Piñeiro et al. (2008). Red line: one-to-one line.
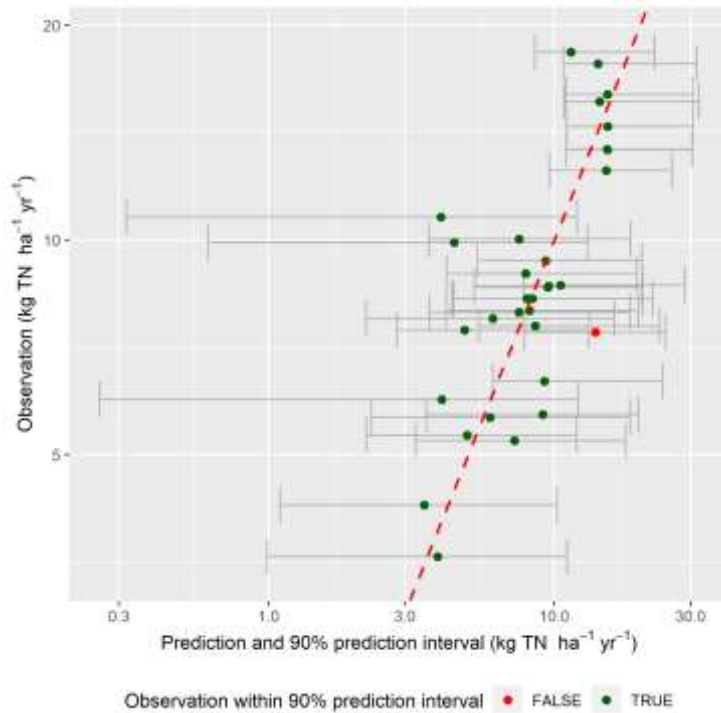
Figure 39. Observations versus empirical yield model predictions of TN yield for sites within the Manawatu basin. The green points indicate the observations that are within the 90% prediction interval (grey error bars) and the red points indicate the observations that are outside of the 90% prediction interval. Note that the observed values are plotted on the Y-axis and predicted values on the X-axis, following Piñeiro et al. *(2008). Red line: one-to-one line.*

The adjustment factors applied to the SDEM of the Manawatū River basin are shown in Figure 40. The SDEM appeared to perform best for mainstem Manawatū River sub-catchments, compared to lower order streams. Minimal (if any) adjustments of the national scale parameter set, applied to sub-catchments, were required to achieve acceptable agreement with measured loads at mainstem monitoring stations. Lower order tributaries of the mainstem typically required larger adjustments, with consistent over-predictions in the north and under-predictions in the south and east.
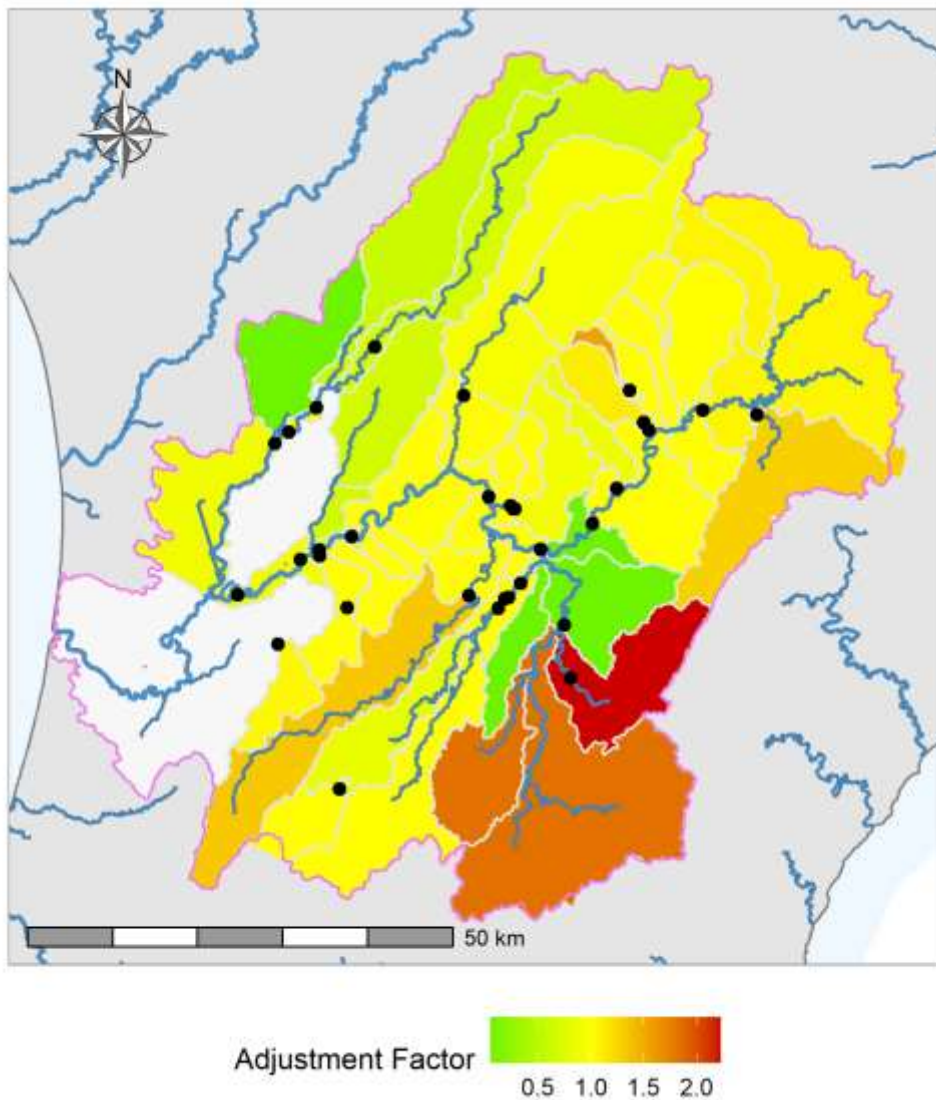
Figure 40: Adjustment factors applied to the SDEM of the Manawatu River basin. Grey boundaries indicate sub-catchments used in the model. Black dots are monitoring stations.

Results using the SDEM indicate significantly higher contributions from dairy, compared to HRC-SCAMP, and significantly lower contributions from sheep and beef (Figure 41). The SDEM also indicates larger contributions from forestry and native bush compared to the SCAMP model.

The scenario results are summarised in Table 14, for a single mid-basin assessment point, and Figure 42, for the entire basin (Scenarios 1 and 2 only). Results show generally close agreement in relative change predictions for the three new models presented here. The HRC-SCAMP model predicts smaller changes for all scenarios.
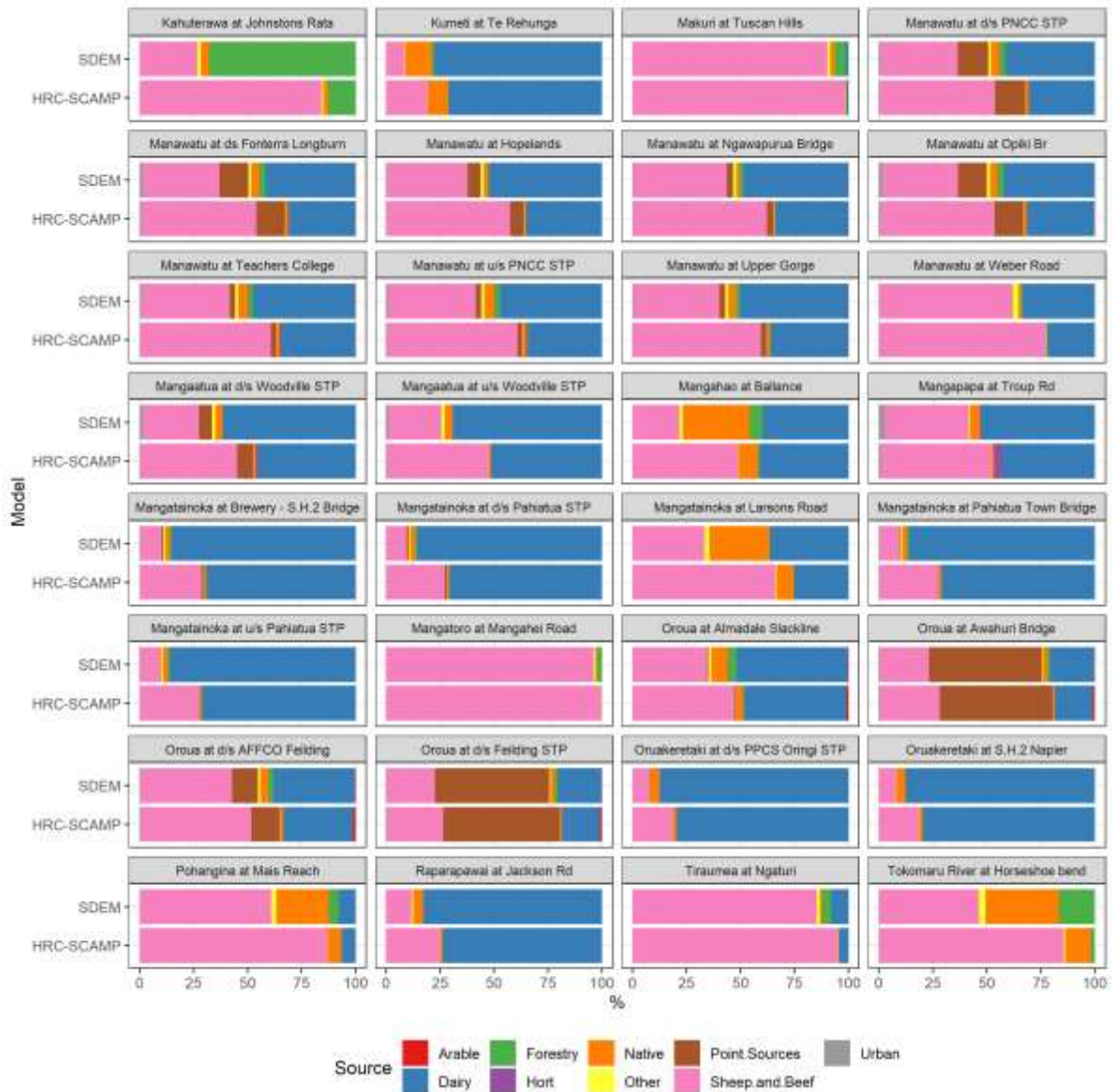
Figure 41. Comparison of modelled TN source yield percentage contributions for the Manawatū River basin between the SDEM and HRC-SCAMP models.

Table 14. Scenario simulation results for the Manawatū River at Upper Gorge monitoring station. All values represent percent changes from baseline in TN yield or median concentration.

Scenario	Empirical model		SDEM	HRC-SCAMP
	Concentration	Yield		
Scenario 1.	-33%	-35%	-33%	-29%
Scenario 1a.	-5.30%	-5.00%	-4.90%	-4.70%
Scenario 2	-23%	-23%	-23%	-17%
Scenario 2a.	-3.10%	-2.80%	-2.70%	-2.20%
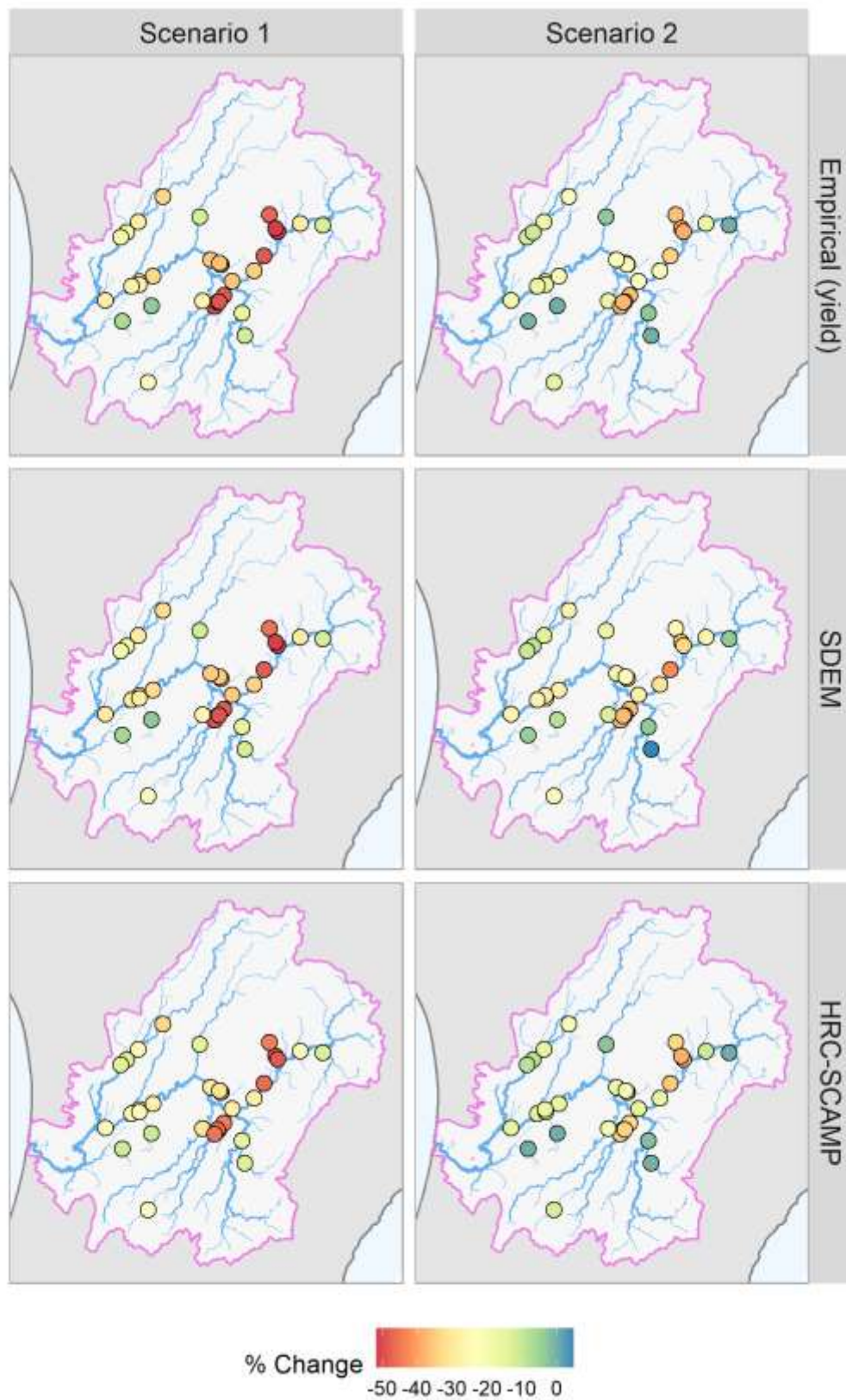
Figure 42: Estimated changes in yield at monitoring locations in the Manawatū River basin for Scenario 1 and Scenario 2, for the empirical yield model, SDEM and HRC-SCAMP.

6   Discussion

6.1   Large differences in catchment TN loads estimated using different typologies

We evaluated the consistency of estimates of catchment average export coefficients (CAECs) for both TN and TP that were derived using three alternative typologies. We discuss the inconsistencies observed for TP later in this discussion; the following section focuses on TN.

The three typologies examined in this report (Srinivasan, LWP/Bright and Monaghan) showed reasonable agreement with each other in terms of variation in CAECs of TN across the country when viewed as national scale maps (Figure 8, Figure 9). Correspondingly, we found reasonable correlation in CAECs estimated using all three typologies for TN ($R^2 \sim 80\%$, Table 4). However, we found that estimates based on Srinivasan were systematically lower by 3 – 4kg TN ha^{-1} yr^{-1} than those derived from the other typologies (Table 4). This means that 50% of CAECs estimated using Srinivasan are more than 3 – 4kg TN ha^{-1} yr^{-1} less than estimates made using LWP/Bright and Monaghan. In addition, even for estimates derived from the two typologies that were in closest agreement (LWP/Bright and Monaghan), the characteristic differences (RMSD) were approximately 4kg TN ha^{-1} yr^{-1}. This means that for one third of catchments, differences in CAECs estimated by any two typologies can be expected to differ from each other by at least 4kg TN ha^{-1} yr^{-1}. Given that 95% of catchments have CAECs <34kg TN ha^{-1} yr^{-1} and the median is 12kg TN ha^{-1} yr^{-1}, the differences in estimates associated with different typologies are large. Succinctly, this means that there are large uncertainties in accounting for sources of N in catchments using any of the three typologies examined by this study.

This study cannot indicate which set of export coefficients (i.e., which typology) is closest to the true value or which Types are most incorrect within each typology. Our analysis indicates that the TN export coefficients for Forestry and Natural may be too low for all three typologies, which is discussed below. At least a portion of the discrepancy between typologies will be attributable to differences in how OVERSEER was applied to quantify the export coefficients associated with pastoral land. Presumably, there were different input data and assumptions made when the OVERSEER model was used to estimate farm loss rates. These differences likely include climate and topographic input differences (within a given land Type), as well as differences in assumed pasture type, animal type and numbers, and farm management practices associated with the representative farms. As for all numerical models, input differences, even within plausible ranges, can result in appreciable differences in model output. This is typically referred to as model and parameter uncertainty. The results of this study suggest that uncertainty in estimated export coefficients may contribute to significant uncertainty in estimates of total catchment diffuse source nitrogen loads.

6.2   Source loads and instream loads are uncertain and therefore so are attenuation parameters

Over 75% of monitoring stations had observed instream yields of less than 10 kg TN ha^{-1} yr^{-1} (Figure 5). This means that the discrepancies in CAECs estimated using different Typologies represent substantial uncertainty in calibrated attenuation parameters. For example, consider an estimated CAEC of 20 kg TN ha^{-1} yr^{-1} with an uncertainty (RMSD) of 4kg TN ha^{-1} yr^{-1} (95% confidence interval from 12 to 28 kg TN ha^{-1} yr^{-1}) and an instream yield of 10 kg TN ha^{-1} yr^{-1}. This equates to an attenuation coefficient of 0.5 with a 95% confidence interval of 0.17 to 0.64. This range is almost half of the physically possible range of an attenuation coefficient and therefore represents considerable uncertainty. We note that this simple estimate of CAEC

uncertainty does not consider additional sources of uncertainty that are encountered when setting up catchment models such as uncertainties associated with land use mapping.

Across the monitoring stations, the confidence intervals for TN attenuation coefficients that were derived by consideration of only the instream yield uncertainty were also wide (Figure 13). The width of the TN attenuation coefficient confidence intervals increased as attenuation coefficients decreased, due to the nature of Equation 1. Across attenuation coefficients calculated for all typologies, 50% had confidence interval widths of 0.18 or larger, 25% had widths of 0.3 or larger and 10% had widths of 0.47 or larger. These represent approximately 20%, 30% and 50% of the physically possible range of an attenuation coefficient (i.e., 0 to 1). Therefore, uncertainty associated with instream load estimates represent a considerable contribution to the uncertainty of process-based catchment nutrient models.

Because of both model and parameter uncertainty described above and the uncertainty in calculated instream yields, accurately and precisely defining attenuation parameters is challenging. The results presented here imply that there is limited ability to detect significant spatial differences in attenuation across catchments. This has implications for how water quality models are used for decision making. The justification for defining spatially variable attenuation coefficients within such catchment models may need to rely largely on expert judgement. At the very least, model predictions should be presented with acknowledgement of this uncertainty and appropriate cautions regarding the use of the predictions.

6.3   Underestimation of TN export coefficients for Forestry and Natural land

Comparisons of the proportion of upstream area in different land use categories with estimated TN attenuation coefficients for each typology indicated that there was a pattern of decreasing attenuation coefficient with increasing proportions of upstream areas in Forestry or Natural. In addition, there was a greater proportion of negative attenuation coefficients for water quality stations whose catchments were occupied by more than 50% of either land use, compared to those with less than 50%. Negative attenuation coefficients indicate that source loads are less than observed instream loads and indicate the aggregate of the export coefficients underestimates the true TN diffuse source loss rates.

In our study, only Srinivasan included Natural as a land use category. We used Srinivasan's export coefficient for Natural by to represent the Natural category in the other two typologies. Both Srinivasan and LWP/Bright included an export coefficient for Forestry. Srinivasan has a median TN export coefficient for Natural of 2 kg ha^{-1} yr^{-1}, and a range of 1 - 7.1 kg ha^{-1} yr^{-1}. Srinivasan and LWP/Bright have TN export coefficients of 4 kg ha^{-1} yr^{-1} for Forestry, with Srinivasan indicating a range of 1 - 28 kg ha^{-1} yr^{-1}. However, the empirical yield models for TN suggest a median TN yield for Forestry of 8.5 kg ha^{-1} yr^{-1} (Table 8). It is important to keep in mind that our estimates of ETY for Forestry represent the national median value at the scale of whole catchments. In addition, areas categorised as Forestry in our study will have comprised areas that are at various stages of the growing cycle including areas undergoing new planting, thinning and harvesting. Although N concentrations and losses in streams whose catchments are dominated by mature planted forests are generally low, losses can be considerably higher where forests are established on pasture land (Davis 2014). It is also clear that N loss from areas of plantation forest are not constant and that there are times during the growing and harvesting cycle when nitrogen leaching is much higher than mature forest including fertilisation, planting, thinning and harvesting (Davis 2014; Hughes and Quinn 2019). The temporal variation in N loss from Forestry is particularly pronounced in forests planted on coastal sands or land that was previously pasture (Davis 2014). The ETY estimates for TN yield for Forestry associated with the empirical models are therefore probably representative

of losses at the scale of catchments with patches of plantation forestry that are at various stages of the growing cycle.

It is likely that the most effort and best modelling available to characterise export coefficients has been focused on the agricultural and horticultural land uses, and in particular pastoral land uses. However, catchment models require robust export coefficients associated with other land uses including Natural and Forestry because those other land uses make up significant proportions of many catchments in New Zealand. For Natural areas, load contributions represent unmanageable loads (i.e., land use or land management change are unlikely to occur or to cause significant reductions in N loss). Whether the unmanageable load is 10 or 20% of the total load, for example, has important implications for the load reductions required from other land uses to achieve a required catchment load reduction. Underestimation of export coefficients from Natural and Forestry will lead to underestimation of calibrated attenuation coefficients. We therefore recommend further investigation of losses from Forestry and Natural land to help improve accuracy of water quality catchment models.

6.4 Empirical catchment water quality models are an alternative to process-based models

In this study, we developed an alternative class of empirical catchment model for nitrogen (TN) concentrations and yields. The empirical model approach provides a direct relationship between Types and concentrations and yields and avoids the reliance on uncertain estimates of export coefficients that are derived from other sources (e.g., OVERSEER). The empirical approach also removes the requirement to calibrate parameters that represent attenuation. In other words, empirical models are a more parsimonious alternative to setting up catchment N models that may be appropriate for at least some applications. In addition, the empirical models allow the user to estimate the 90% prediction interval as an estimate of the imprecision of the empirical model predictions for any catchment. This provides a simpler and more transparent way to quantify model uncertainty than process-based models.

The empirical models presented in this report provide simple and easily used tools that can be applied at any location within New Zealand. We have developed a dataset that provides proportions of catchment area occupied by each Type used by the TN concentration and yield models for all segments of the DN2.4 (>10 km^2). These data allow estimates of yield or concentration to be made at any location in New Zealand very easily. In addition, the data could also be used for national- and regional-scale assessments that aim to rapidly assess impacts of land use and land management scenarios at any location in New Zealand.

An important caveat that applies to the empirical models is associated with the national scale of the quantile regression models that were used to derive the ETCs and ETYs. Because the water quality station data were limited, we were only able to derive robust ETCs and ETYs for a limited number of Types. This limits the spatial resolution of the empirical models. In addition, as the spatial extent of a modelled domain reduces, the applicability of our ETCs and ETYs. will diminish. We note that site-specific adjustments of the nationally-derived parameters may be appropriate on a case-by-case basis (i.e., "local calibration"). We also note that the most likely applications of this type of model would be in scenario analysis, where the objective may not be to evaluate absolute concentrations or yields, but rather to evaluate relative differences in these outcomes between scenarios. The accuracy of these relative differences relies on the assumption that the relativities between ETC or ETY estimates between Types are relevant and applicable for the catchment of interest. We note that testing the validity of these assumptions was beyond the scope of this study but would be a useful direction for future research.

Defining the Types that were included in the empirical models involved expert judgement and was strongly influenced by the structure of the Srinivasan *et al.* (2021) typology. The definition of the 'best' empirical models was also based on expert judgement, and involved trade-offs between the model performance, the proportion of significant parameters, and optimising for the total number of Types to allow for greater flexibility for scenarios. Nevertheless, the applications presented in this report demonstrate the potential for use of empirical catchment water quality (nitrogen) models. Future research or applications could refine the approach, potentially with more exhaustive exploration of Type sets (typologies, their factors and their categories), incorporating the observation uncertainties into the fitting process, using updated observed water quality datasets, and using criteria for defining the "best" model based on specific model purpose.

6.5 The derived empirical models can be used for several types of simulation

This study demonstrates the application of the empirical TN models to several types of simulations that are relevant to supporting regional council decision making. Relevant simulations include those associated with assessing impacts on water quality associated with land use mitigations, setting target attribute states, and assessing allocation regimes for nutrients.

The empirical TN yield model performed satisfactorily in the case study application, with performance measures that were similar to that of the national scale model. The empirical concentration model did not achieve satisfactory results as measured by the NSE and consistently underpredicted stream concentrations, resulting in an unsatisfactory PBIAS rating. These results suggest scalability challenges associated with the ETC model as compared to the ETY model. We surmise that these complications may be related to hydrologic variability and possibly the point source representation within the concentration model. Catchment hydrology is complex and plays an important role in determining stream contaminant concentrations. Hydrology is only partially represented in the empirical concentration model typology (e.g., wet vs. dry). Caution should therefore be exercised when applying the empirical concentration model at less than the national scale. This model may be best suited for coarse approximations of relative changes in concentration due to land use change.

Compared to the HRC-SCAMP model, the semi-distributed implementation of the empirical model (SDEM) estimated that dairy, forestry, and native bush makes a larger contribution to catchment TN load and sheep and beef makes a lower contribution to catchment TN load (Figure 41). While HRC-SCAMP estimates the largest TN source contribution at the bottom of the Manawatū River basin to be sheep and beef farmland, the SDEM suggests that the largest load contribution is from dairy farmland. Similarly, estimates of the TN load relative contributions from forestry and native bush are higher using the SDEM (5.5%), compared to the HRC-SCAMP model (1.5%). The differences have implications for catchment management, mitigation strategies and nitrogen management. For example, the results for a set of simulations (Scenarios 1 and 2) using the new empirical models show greater water quality impacts from changing land use from Dairy to Natural compared to the previously developed HRC-SCAMP model.

The simulated water quality impacts for the scenarios produced with three implementations of the empirical models (empirical concentration model, empirical yield model, and SDEM) were in close agreement with each other and with HRC-SCAMP (Table 14). These results suggest that the three empirical model implementations can be applied with equal confidence for these

types of simulations, when the focus is on relative differences (i.e., a percentage change between the simulated baseline and a simulated scenario).

The empirical yield model is amenable to being used in semi-distributed mass balance catchment models. Such models allow for mass balance tracking and accounting, and the separation of mass and flow rates within a modelling framework. These types of models can be used to simulate prescribed changes in mass (e.g., treatment removal options or point source modifications) or changes in flow (e.g., climate change forecasting or water use projections). Site specific calibration, guided by monitoring data, may be desirable for such models as a way of refining spatial variability and reducing bias in the yield parameters. For the example presented in this study, the relativity of yields was maintained across Types by applying a uniform adjustment rate to all Types within sub-catchments. Conceptually, these uniform adjustments represent an assumption that sub-catchment attenuation rates are higher, or lower, than those implied by the national parameter set. Such an assumption could be supported by knowledge of localised groundwater conditions that support denitrification.

The empirical models are more parsimonious and transparent than a process model and is, in theory, easier to construct and calibrate. An advantage of the empirical models is that they are not dependent on externally derived and uncertain export coefficients such as those derived using OVERSEER or expert opinion. The empirical concentration model has the advantage of directly estimating stream TN concentrations anywhere in the country. Instream concentrations, rather than loads, are typically the response variable of concern with respect to defined water quality objectives.

Although the prediction intervals derived using our empirical models are wide, their direct derivation from observational data provides an easily estimated quantification of uncertainty that is less easily achieved with process-based models. Uncertainties can be derived for process-based models but their quantification is more complicated due to the greater number of model inputs and components (Semadeni-Davies *et al.* 2020, 2021). In addition, process-based models often rely on inputs derived from external models such as OVERSEER for which uncertainties are not well quantified (Elliott *et al.* 2016; Etheridge *et al.* 2018).

6.6   Catchment TP models have high uncertainty

The uncertainty of instream load estimates was generally higher for TP than for TN. This occurs because, compared to TN, high TP concentrations are generally associated with high flows. When samples are taken punctually and monthly (as is generally the case with SOE monitoring) high flow samples are generally sparse and have high variance (Snelder *et al.*, 2017). Because load is calculated by multiplying concentration and flow, a large component of the TP load is associated with high flows. However, the high flow component of the TP load is very uncertain due to the sparse data and high variance of TP concentrations associated with high flow. Poorly characterised TP loads have flow-on effects with respect to the precision of variables calculated based on loads, such as attenuation coefficients and empirical yield coefficients.

Our analysis showed large differences in catchment average TP export coefficients between the Srinivasan and Monaghan typologies (Figure 11). In comparison to TP, catchment average TN export coefficients estimated using Srinivasan and Monaghan were more consistent. The larger differences for TP might be due to differences in typology definition (i.e., factors and categories) and/or more variability in the methods used to derive the export coefficients for each Type.

The TP export coefficients for Srinivasan and Monaghan typologies were predominantly derived using the OVERSEER model. OVERSEER does not account for a significant portion of the total phosphorus export associated with infrequent, but large-scale, erosion events (Parfitt et al. 2007, 2013; Gray et al. 2016). Underestimation of TP associated with large-scale erosion events is likely to be contribute to the uncertainty of all TP export coefficients that are derived from OVERSEER. Evidence for this was provided by our study's large proportion (~30% of sites) with negative attenuation coefficients. This indicates that export coefficients for at least some Types in both typologies are underestimating the true diffuse source TP loss rates.

The empirical TP yield and concentration models had poor performance (Figure 30, Figure 34) compared to the performance of the same Types of TN models. For example, the empirical TP yield model that used only land use/cover categories as predictors explained very little of the between-site variability in TP yields (adjusted R^2 value of 4%). Our conclusion is that the typologies used to derive the empirical yield and concentration models, as well as the OVERSEER modelling used to generate the TP export coefficients associated with the Srinivasan and Monaghan typologies, are lacking important explanatory variables. The missing variables may be associated with natural processes that determine P concentrations. Generally, there are three important natural processes that influence P concentrations: geogenic supply, mobilisation and transport, and microbially mediated reduction-oxidation (redox) which influences mobility, and speciation of N and P (Boomer and Bedford 2008; Maynard *et al.* 2011; Parsons *et al.* 2017). Porder and Ramachandran (2013) found a 30-fold difference in median P concentration among rock types, ranging from 120 ppm (several ultramafic rocks) to >3,000 ppm (several alkali basalts). Mage and Porder (2013) showed that parent material explained the most variation in P availability in soils (56% of variation explained) and topographic position (ridges, slopes or valleys) explained an additional (10-15% of variation). Many of these rock types are found in NZ and there can be significant variation in these types even within large catchments.

We conclude from our findings, and the above literature, that catchment modelling that does not account for the natural processes that influence P loss, and in particular parent material (rock type and topographic position), are going to poorly represent P loads and concentrations in downstream aquatic environments. We propose that a way forward may be to derive national scale maps that account for some of the natural processes that govern TP generation, transport and attenuation and to investigate their incorporation into typologies used for catchment water quality modelling. An interesting outcome of the empirical TP modelling is the apparent limited influence of land use on observed instream TP. This has consequences for the way that decision makers need to define policies and actions to reduce catchment TP yields. For example, these results may suggest that it is difficult currently to justify limits for resource use based on P loss and that it would be better to focus on targeted mitigations such as reducing erosion and better management of critical-source areas (McDowell and Srinivasan 2009). We recommend that further research on quantifying catchment TP losses and loads is needed.

6.7 Catchment water quality models and simulations are uncertain

The uncertainties associated with water quality models, and their use to make simulations of the impact of land management actions on water quality, cannot be reduced appreciably in the short to medium term. However, catchment water quality models will generally need to be used to inform decision makers about appropriate responses to water quality issues including actions such as limiting resource use, requiring mitigations and land use changes. This report

shows that these decisions will ultimately need to be made in the face of considerable uncertainty.

To some extent the large uncertainties associated with estimates of absolute loads or concentrations are less important than the relative difference between two simulations (e.g., between a baseline and a mitigation scenario). In other words, users of model outputs are likely to be more interested in the predicted relative change in load or concentrations between scenarios than the absolute values of the predictions. This is advantageous because there is likely some commonality in the sources of uncertainty between scenarios (e.g., because some uncertainty is due to within-Type variability of land or management characteristics that are used in the estimation of loss rates) and this means that the uncertainty in the relative change will be less than the absolute values. However, methods for understanding and quantifying the uncertainty of these relative differences for both our empirically based approach and for process-based catchment models have not been developed. Defining and quantifying uncertainties in relative differences between scenarios presents a considerable technical challenge that needs further research.

Acknowledgements

References

Bloomer D, O'Brien G, Posthuma L (2020) Modelled Loss of Nutrients From Vegetable Growing Scenarios In Horowhenua. Page Bloomer Associates, New Zealand.

Boomer KMB, Bedford BL (2008) 'Groundwater-induced redox-gradients control soil properties and phosphorus availability across four headwater wetlands, New York, USA' *Biogeochemistry* **90**, 259–274.

Breiman L (2001) 'Random Forests' *Machine Learning* **45**, 5–32.

Bright J, Ford S, Irving C (2018) Water Allocation Economics Analysis: Land/Water Use Modelling. Aqualinc client report C17020/1. Aqualinc Research Limited, Christchurch, New Zealand.

Cade BS, Noon BR (2003) 'A gentle introduction to quantile regression for ecologists' *Frontiers in Ecology and the Environment* **1**, 412–420.

Cohn TA (2005a) 'Estimating contaminant loads in rivers: An application of adjusted maximum likelihood to type 1 censored data' *Water Resources Research* **41**,. Available at http://onlinelibrary.wiley.com/doi/10.1029/2004WR003833/full [Verified 21 January 2016]

Cohn TA (2005b) 'Estimating contaminant loads in rivers: An application of adjusted maximum likelihood to type 1 censored data' *Water Resources Research* **41**,.

Cohn TA, Caulder DL, Gilroy EJ, Zynjuk LD, Summers RM (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 TA, Delong LL, Gilroy EJ, Hirsch RM, Wells DK (1989) 'Estimating constituent loads' *Water resources research* **25**, 937–942.

Cox T, Kerr T, Snelder T, Fraser C (2022) Manawatū-Whanganui Region Catchment Nutrient Models: Model Updates. Supporting Regional Land and Water Management. LWP Client Report 2022–02. LWP Ltd, Christchurch, New Zealand.

Cutler DR, Edwards JTC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) 'Random forests for classification in ecology' *Ecology* **88**, 2783–2792.

Davis M (2014) 'Nitrogen leaching losses from forests in New Zealand' *New Zealand Journal of Forestry Science* **44**, 1–14.

Drewry JJ (2018) Nitrogen and phosphorus loss values for selected land uses. Landcare Research Contract Report LC3367. Landcare Research, New Zealand.

Duan N (1983) 'Smearing estimate: a nonparametric retransformation method' *Journal of the American Statistical Association* **78**, 605–610.

Efron B (1981) 'Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods.' *Biometrika* **68**, 589–599.

Elliott AH, Alexander RB, Schwarz GE, Shankar U, Sukias JPS, McBride GB (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 AH, Semadeni-Davies AF, Shankar U, Zeldis JR, Wheeler DM, Plew DR, Rys GJ, Harris SR (2016) 'A national-scale GIS-based system for modelling impacts of land use on water quality' *Environmental Modelling & Software* **86**, 131–144.

Etheridge Z, Fietje L, Metherell A, Lilburne L, Mojsilovich O, Robson M, Steel K, Hanson M (2018) 'Collaborative expert judgement analysis of uncertainty associated with catchment-scale nitrogen load modelling with OVERSEER®' *Farm Environmental Planning–Science, Policy and Practice* 14.

Fraser CE, Snelder T (2019) Test of methods for calculating contaminant loads in the Manawatū-Whanganui region: Supplementary report. 2019–08. LWP Ltd, Christchurch.

Gray CW, Wheeler DM, McDowell R, Watkins NL (2016) 'Overseer and phosphorus: Strengths and weaknesses' *Integrated nutrient and water management for sustainable farming Fert Lime Res Ctr, Massey Univ, Palmerston North, New Zealand* 1–32.

Hirsch RM, Archfield SA, De Cicco LA (2015) 'A bootstrap method for estimating uncertainty of water quality trends' *Environmental Modelling & Software* **73**, 148–166. doi:10.1016/j.envsoft.2015.07.017

Hirsch RM, De Cicco LA (2015) User guide to Exploration and Graphics for RivEr Trends (EGRET) and dataRetrieval: R packages for hydrologic data. US Geological Survey,

Hirsch RM, Moyer DL, Archfield SA (2010) 'Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay river inputs 1' *JAWRA Journal of the American Water Resources Association* **46**, 857–880.

Hughes AO, Quinn JM (2019) 'The effect of forestry management activities on stream water quality within a headwater plantation Pinus radiata forest' *Forest Ecology and Management* **439**, 41–54.

Johnes PJ (2007) 'Uncertainties in annual riverine phosphorus load estimation: impact of load estimation methodology, sampling frequency, baseflow index and catchment population density' *Journal of Hydrology* **332**, 241–258.

Mage SM, Porder S (2013) 'Parent material and topography determine soil phosphorus status in the Luquillo Mountains of Puerto Rico' *Ecosystems* **16**, 284–294.

Mason CH, Perreault Jr WD (1991) 'Collinearity, power, and interpretation of multiple regression analysis' *Journal of marketing research* **28**, 268–280.

Maynard JJ, O'Geen AT, Dahlgren RA (2011) 'Sulfide induced mobilization of wetland phosphorus depends strongly on redox and iron geochemistry' *Soil Science Society of America Journal* **75**, 1986–1999.

McBride GB (2005) 'Using Statistical Methods for Water Quality Management: Issues, Problems and Solutions (Vol. 19).' (John Wiley & Sons)

McDowell RW, Monaghan RM, Smith C, Manderson A, Basher L, Burger DF, Laurenson S, Pletnyakov P, Spiekermann R, Depree C (2021) 'Quantifying contaminant losses to water from pastoral land uses in New Zealand III. What could be achieved by 2035?' *New Zealand Journal of Agricultural Research* **64**, 390–410.

McDowell RW, Srinivasan MS (2009) 'Identifying critical source areas for water quality: 2. Validating the approach for phosphorus and sediment losses in grazed headwater catchments' *Journal of Hydrology* **379**, 68–80.

McMahon G, Gregonis SM, Waltman SW, Omernik JM, Thorson TD, Freeouf JA, Rorick AH, E. KJ (2001) 'Developing a spatial framework of common ecological regions for the conterminous United States' *Environmental Management* **28**, 293–316.

Monaghan R, Manderson A, Basher L, Smith C, Burger D, Meenken E, McDowell R (2021) 'Quantifying contaminant losses to water from pastoral landuses in New Zealand I. Development of a spatial framework for assessing losses at a farm scale' *New Zealand Journal of Agricultural Research* **64**, 344–364.

Moores J, Easton S, Gadd J, Sands M (2017) Te Awarua-o-Porirua Collaborative Modelling Project. Customisation of urban contaminant load model and estimation of contaminant loads from sources excluded from the core models. NIWA Client Report No. 2017050AK. NIWA and Jacobs for Greater Wellington Regional Council, Auckland.

Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL (2007) 'Model evaluation guidelines for systematic quantification of accuracy in watershed simulations' *Transactions of the ASABE* **50**, 885–900.

Moriasi DN, Gitau MW, Pai N, Daggupati P (2015) 'Hydrologic and water quality models: Performance measures and evaluation criteria' *Transactions of the ASABE* **58**, 1763–1785.

Nash JE, Sutcliffe JV (1970) 'River flow forecasting through conceptual models part I—A discussion of principles' *Journal of hydrology* **10**, 282–290.

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (2004) 'Applied linear statistical models.' (McGraw-Hill: Chicago, IL)

Newsome PFJ, Wilde RH, Willoughby EJ (2008) land resource information system spatial data layers. Data Dictionary. Landcare Research New Zealand Ltd, Palmerston North. Available at https://lris.scinfo.org.nz/document/9162-lris-data-dictionary-v3/ [Verified 22 March 2022]

Parfitt R, Dymond J, Ausseil A, Clothier B, Deurer M, Gillingham A, Gray R, Houlbrooke D, Mackay A, McDowell R (2007) Best practice phosphorus losses from agricultural land. Landcare Research Client report LC0708/012. Landcare Research, Palmerston North, NZ.

Parfitt RL, Frelat M, Dymond JR, Clark M, Roygard J (2013) 'Sources of phosphorus in two subcatchments of the Manawatu River, and discussion of mitigation measures to reduce the phosphorus load' *New Zealand journal of agricultural research* **56**, 187–202.

Parsons CT, Rezanezhad F, O'Connell DW, Van Cappellen P (2017) 'Sediment phosphorus speciation and mobility under dynamic redox conditions' *Biogeosciences* **14**, 3585–3602.

Piñeiro G, Perelman S, Guerschman J, Paruelo J (2008) 'How to evaluate models: Observed vs. predicted or predicted vs. observed?' *Ecological Modelling* **216**, 316–322.

Porder S, Ramachandran S (2013) 'The phosphorus concentration of common rocks—a potential driver of ecosystem P status' *Plant and soil* **367**, 41–55.

Preston SD, Bierman VJ, Silliman SE (1989) 'An evaluation of methods for the estimation of tributary mass loads' *Water Resources Research* **25**, 1379–1389.

Quilbé R, Rousseau AN, Duchemin M, Poulin A, Gangbazo G, Villeneuve J-P (2006) 'Selecting a calculation method to estimate sediment and nutrient loads in streams: application to the Beaurivage River (Québec, Canada)' *Journal of Hydrology* **326**, 295–310.

R Core Team (2023) 'R: A language and environment for statistical computing.' Available at https://www.R-project.org/.

Roygard JKF, McArthur KJ, Clark ME (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.

Semadeni-Davies A, Jones-Todd C, Srinivasan MS, Muirhead R, Elliott A, Shankar U, Tanner C (2020) 'CLUES model calibration and its implications for estimating contaminant attenuation' *Agricultural Water Management* **228**, 105853.

Semadeni-Davies AF, Jones-Todd CM, Srinivasan MS, Muirhead RW, Elliott AH, Shankar U, Tanner CC (2021) 'CLUES model calibration: residual analysis to investigate potential sources of model error' *New Zealand Journal of Agricultural Research* **64**, 320–343.

Sharpley AN, McDowell RW, Kleinman PJ (2001) 'Phosphorus loss from land to water: integrating agricultural and environmental management' *Plant and soil* **237**, 287–307.

Shepherd M, Wheeler D, Selbie D, Buckthought L, Freeman M (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.

Snelder TH, Biggs BJF (2002) 'Multi-scale river environment classification for water resources management' *Journal of the American Water Resources Association* **38**, 1225–1240.

Snelder TH, McDowell RW, Fraser CE (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 TH, Whitehead AL, Fraser C, Larned ST, Schallenberg M (2020) 'Nitrogen loads to New Zealand aquatic receiving environments: comparison with regulatory criteria' *New Zealand Journal of Marine and Freshwater Research* **54**, 527–550.

Srinivasan MS, Muirhead RW, Singh SK, Monaghan RM, Stenger R, Close ME, Manderson A, Drewry JJ, Smith LC, Selbie D (2021) 'Development of a national-scale framework to characterise transfers of N, P and Escherichia coli from land to water' *New Zealand Journal of Agricultural Research* **64**, 286–313.

Udo de Haes HA, Klijn F (1994) Environmental policy and ecosystem classification. In 'Ecosyst. Classif. Environ. Manag.' (Ed F Klijn) pp. 1–21. (Kluwer Academic Publishers.: Dordrecht, Netherlands)

Whitehead A, Fraser CE, Snelder TH (2021) Spatial modelling of river water-quality state. Incorporating monitoring data from 2016 to 2020. 2021303CH. NIWA, Christchurch.

Whitehead A, Fraser CE, Snelder TH, Walter K, Woodward S, Zammit C (2021) Water quality state and trends in New Zealand Rivers. Analyses of national data ending in 2020. 2021296CH. NIWA, Christchurch.

Woods RA, Hendrikx J, Henderson R, Tait A (2006) 'Estimating mean flow of New Zealand rivers' *Journal of Hydrology (New Zealand)* **45**, 95–110.

Appendix A   Water quality site load calculations

A1   General approach

Mean annual TN and TP loads at all water quality sites were derived from monthly TN and TP concentrations and observed daily flows. Load calculation methods generally comprise 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 2 but in the explanation that follows, we describe step 2 first.

If flow and concentration observations were available for each day, the instream yield, (the mean annual load, standardised by the upstream catchment area) would be the summation of the daily flows multiplied by their corresponding concentrations:

$$L = \frac{K}{A_c N} \sum_{j=1}^{N} C_j Q_j \qquad \text{(Equation A11)}$$

where L: mean annual instream yield (kg yr^{-1} ha^{-1}), A_c: catchment area, 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.

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 only be calculated directly from measurements for these days. However, flow is generally observed continuously 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 A1) for each day of the entire sampling period (Cohn *et al.* 1989). The estimated flow and concentration pairs are then multiplied to estimate unit loads, and these are then summed and transformed by K, N and A_c to estimate mean annual instream yield (i.e., step 2 of the calculation method; Equation A1).

There are a variety of approaches to defining rating curves. Identifying the most appropriate approach to defining the rating curve requires careful inspection of the available data for each site and contaminant. The details of the approaches and the examination of the data are summarised below. Further details are provided by Fraser and Snelder (2019).

A2   Load calculation methods

A2.1   L7 model

Two regression model approaches to defining rating curves of (Cohn *et al.* 1989, 1992) and (Cohn 2005a) are commonly used to calculate loads. The regression models relate the natural 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:

$$ln(\widehat{C_i}) = \beta_1 + \beta_2 \left[ln(q_i) - \overline{(ln(q))} \right] + \beta_3 \left[ln(q_i) - \overline{(ln(q))} \right]^2 + \beta_4 (t_i - \bar{T}) \qquad \text{(Equation A2)}$$
$$+ \beta_5 (t_i - \bar{T})^2 + \beta_6 sin(2\pi t_i) + \beta_7 cos(2\pi t_i)$$

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 $\hat{C}_i$: 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, $ln(\hat{C}_j)$. The resulting estimates of $ln(\hat{C}_j)$ are back-transformed (by exponentiation) to concentration units. Because the models are fitted to the log transformed concentrations the back-transformed predictions are corrected for retransformation bias. We used the smearing estimate (Duan 1983) as a correction factor (S):

$$S = \frac{1}{n}\sum_{i=1}^{n} e^{\hat{\varepsilon}_i} \qquad \text{(Equation A3)}$$

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:

$$L = \frac{KS}{N}\sum_{j=1}^{N}\hat{C}_j Q_j \qquad \text{(Equation A4)}$$

If the fitted model is not significant, $\hat{C}_j$ is replaced by the mean concentration and S is one.

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.

$$ln\left(\widehat{C_j^Y}\right) = \beta_1 + \beta_2\left[ln(q_j) - \overline{(ln(q))}\right] + \beta_3\left[ln(q_j) - \overline{(ln(q))}\right]^2 + \beta_4(Y_j - \bar{T}) \qquad \text{(Equation A5)}$$
$$+ \beta_5(Y_j - \bar{T})^2 + \beta_6 sin(2\pi Y_j) + \beta_7 cos(2\pi Y_j)$$

where $\widehat{C_j^Y}$ is the estimated j^{th} concentration for the estimation year, and Y_j is the transformed date of the ith observation, and all other variables are as per equation A2. 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 A5 to the observation dataset. It follows that the estimated load for the year of interest can be calculated by:

$$L^Y = \frac{KS}{N}\sum_{j=1}^{N}\hat{C}_j^Y Q_j \qquad \text{(Equation A6)}$$

A2.2   L5 model

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

$$ln(\hat{C}_i) = \beta_1 + \beta_2(ln(q_i)) + \beta_3(t_i) + \beta_4 sin(2\pi t_i) + \beta_5 cos(2\pi t_i) \qquad \text{(Equation A7)}$$

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:

$$ln\left(\widehat{C_j^Y}\right) = \beta_1 + \beta_2\left[ln(q_j)\right] + +\beta_4\left(Y_j - \bar{T}\right) + \beta_6 sin(2\pi Y_j) + \beta_7 cos(2\pi Y_j) \qquad \text{(Equation A8)}$$

A2.3 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 quantile-based categories (flow decile bins)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 A9, providing an estimate of average annual load over the observation time period.

$$L = \frac{K}{N} \sum_{j=1}^{N} \hat{C}_j Q_j \qquad \text{(Equation A9)}$$

where $\widehat{C_j}$ is calculated mean concentration associated with the flow quantile bin of the flow $Q_{j\cdot}$, and all other variables are as per equation A5.

A2.4 Flow stratification with trend

We included a modified version of the flow stratification method to account for trends in water quality. This is useful when loads are required to be estimated for a particular point in time, rather than as an average over the complete observation period, particularly when there is a strong trend evident. We detrended the observation data by fitting Equation A10 to the concentration time series.

$$ln(\widehat{C_\iota}) = \beta_1 + \beta_2(t_i) \qquad \text{(Equation A10)}$$

We then used the concentration residuals to develop a non-parametric rating curve. $\widehat{C_j}$ is calculated as the mean residual concentration associated with the flow quantile bin of the flow $Q_{j\cdot}$, plus the predicted value of concentration at time T_j, which is multiplied by the smearing coefficient to account for the log transformation of Equation A10).

A3 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.

A4 Selection of best load estimation methodology

TN and TP loads were calculated for all sites using each of the four load estimation methods. We evaluated the performance of each rating curve method for predicting observed concentrations, using a range of model performance measures (see Fraser and Snelder (2019) for details). We identified site loads and method combinations that had any of:

1. unrealistically large instream yield values (i.e., site load divided by catchment area);

2. large differences in the loads calculated using different methods.

For these site and method 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 method for each site that were outside of the two 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. For the remainder of the site and nutrient variable combinations that were not flagged by the above criteria (and for which the diagnostic plots were not inspected), the most appropriate load estimation method was selected aas the rating curve method that yielded the lowest RMSD.

A5 Verification of loads

Load estimation involves subjective decisions, such as the choice of method. We sought to verify our load estimates (i.e., demonstrate they were reasonable) by calculating them using an alternative method. We undertook the validation of our N and P load estimates by applying a new sophisticated load estimation method called Weighted Regressions on Time, Discharge, and Season (WRTDS; Hirsch et al., 2015, 2010).

The WRTDS method provides for considerable flexibility in representing the long-term trend, seasonal components, and discharge-related components of the behaviour of the water-quality variable of interest. However, this flexibility comes at the expense of requiring more data. Fitting a WRTDS model requires that the number of samples collected at the sampling site is more than 100 and the period of sample collection is at least 10 years. In addition, model fitting requires a complete record of daily flow values for the site over the entire period being modelled.

The WRTDS method expresses concentration as a function of time, discharge, and season with the following form:

$$ln(\widehat{C}) = \beta_0 + \beta_1 ln(Q) + \beta_2(t_i) + \beta_3 sin(2\pi t) + \beta_4 cos(2\pi t) + \varepsilon \qquad \text{(Equation A11)}$$

where, $\widehat{C}$ is the predicted concentration of the water quality variable, Q is the flow rate, the β values are fitted parameters, t is the time in years and ε is the unexplained variation. The functional form is linear in t, linear in $ln(Q)$, and sinusoidal on an annual period (i.e., season). However, the method of fitting the model means that the parameter values are not constant throughout the entire domain of the data but vary over the explanatory variable space defined by Q and t. This is achieved by weighting the observations based on their relevance to the point in the explanatory variable space being considered (referred to by Hirsch et al. 2010 as

an estimation point Q_0, T_0). Thus, observations that are close to Q_0, T_0 (in any of flow, year, or season) have a strong influence on the parameter values at that point in the explanatory variable space and the influence decreases the further the observation is from the estimation point. This approach has the following advantages over the methods described above:

1. The concentration – flow relationship is allowed to change smoothly over time.

2. The trend component is not constrained to be any particular functional form and is allowed to change smoothly over time.

3. There is no assumption that the seasonal pattern repeats but rather the shape of the seasonal pattern is allowed to change smoothly over time.

These advantages mean that a WRTDS model can detect and fit both long term (persistent) trend, short term fluctuations, as well as cyclic seasonal variability that evolves over time. Collectively this allows for more realistic representation of how water quality changes and increases the potential to understand the drivers of change.

A WRTDS model includes a "flow-normalisation" procedure that controls for the association between concentration and flow regime variation that happen to have occurred during the monitoring period and thereby describe the concentrations that would have occurred under "average" flow regime. The weighted regression approach to fitting a WRTDS model means "flow-normalised" predictions are not simply adjustments for instantaneous flows but account for flow regime variability at longer timescales.

The performance of a fitted WRTDS model is assessed using "leave-one-out cross validation" (Hirsch and De Cicco 2015). This procedure leaves one observation out of the fitting dataset, fits a model to the remaining observations and uses that model to estimate the concentration for the left-out observation. This step is repeated for all observations in the dataset producing a set of independent predictions for each observation. These independent predictions can be used to quantify various measures of model performance (Hirsch and De Cicco 2015). In this study, we fitted a linear regression of the observations against the predictions and used the R^2 value of this regression to describe the performance of the model. We note that WRTDS can also be used to assess trends by assessing the magnitude and significance of differences in predicted concentrations between dates of interest (Hirsch *et al.* 2015). However, we did not make use of this capability of WRTDS in this study.

We calculated TN and TP loads at each site using WRTDS with the same input data described above. Calculations were performed with the Exploration and Graphics for RivEr Trends (EGRET) and data Retrieval R package (Hirsch *et al.* 2015). The outputs of the EGRET package are numerous and complicated. We obtained from the output the flow normalised flux (TN and TP kg day^{-1}) for 2020. Flow normalised flux is a representation of flux that integrates over the probability distribution of discharge in order to remove the effect of year-to-year variation in discharge. It is therefore consistent with the approach we used to calculate loads which integrated unit loads over the entire flow time series. Using the flux estimate for 2020 was consistent with our load estimates which pertain to 2020. We converted the flux to the units of TN and TP yields used by this study (kg ha^{-1} yr^{-1}) by dividing by catchment area of the water quality stations and multiplying by the appropriate unit conversion factor.

It should be noted that there are many methods of load calculation that differ in how the underlying relationships are represented in subtle ways. Many studies have documented differences in loads calculated from the same data but using different methods and it is often shown these differences can be large (e.g., Cohn, 2005; Johnes, 2007; Preston et al., 1989;

Quilbé et al., 2006; Snelder et al., 2017). Therefore, differences between our loads and those estimated by WRTDS are expected. Notwithstanding this, the plot shown in Figure 43 and Figure 44 indicates strong correspondence between the two sets of loads. The majority of the 95% confidence intervals estimated by this study intersect the one-to-one line indicating that the two sets of estimates are consistent. Our conclusion is that the loads calculated and used in this study are reasonable and are the best estimates that we could produce, given the data.

Figure 43. Comparison of TN loads (expressed as yields) calculated by this study (x-axis) with loads calculated using WRTDS. The error bars show the 95% confidence intervals for both sets of estimates. The red line indicates perfect correspondence (one-to-one).
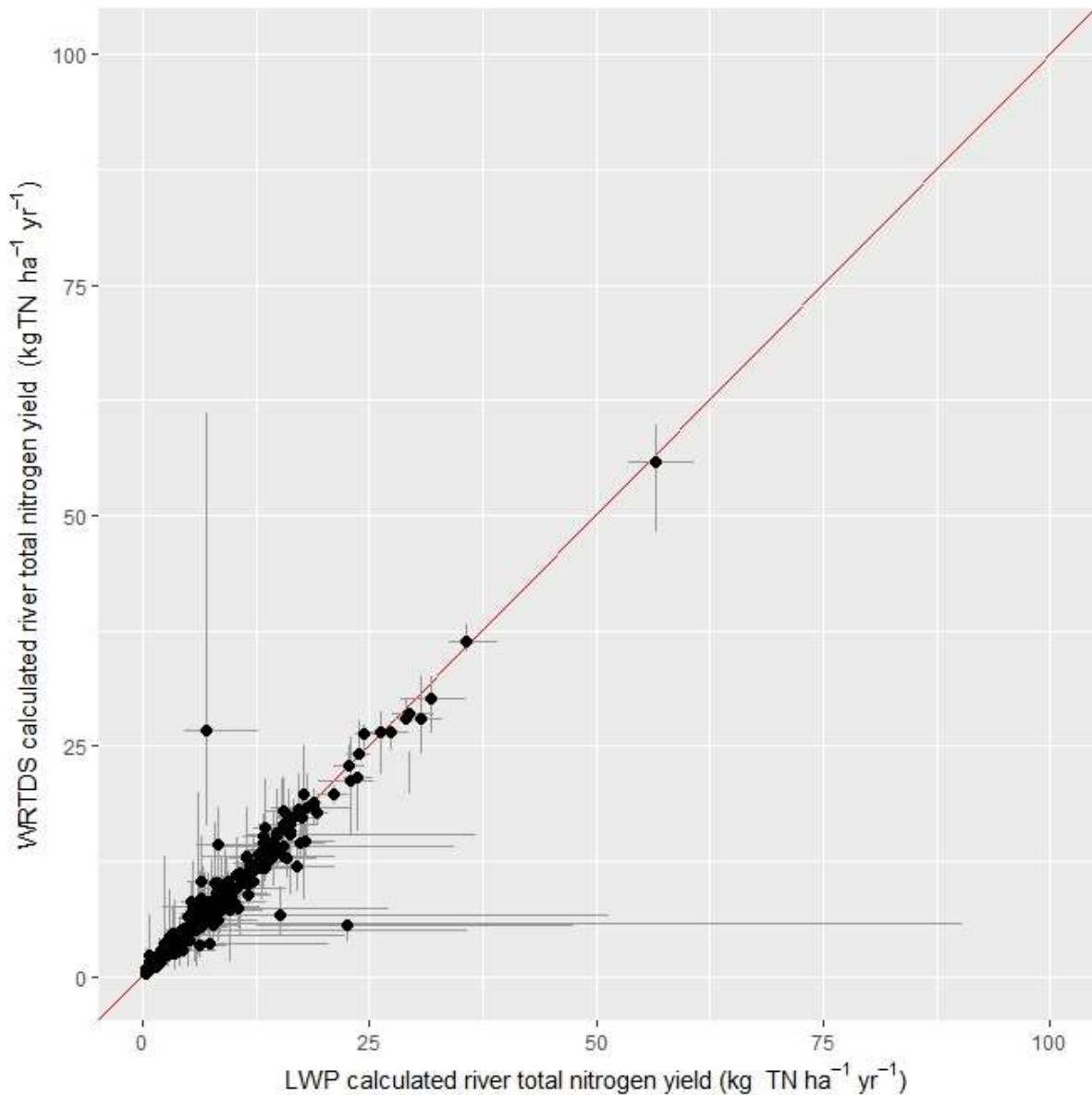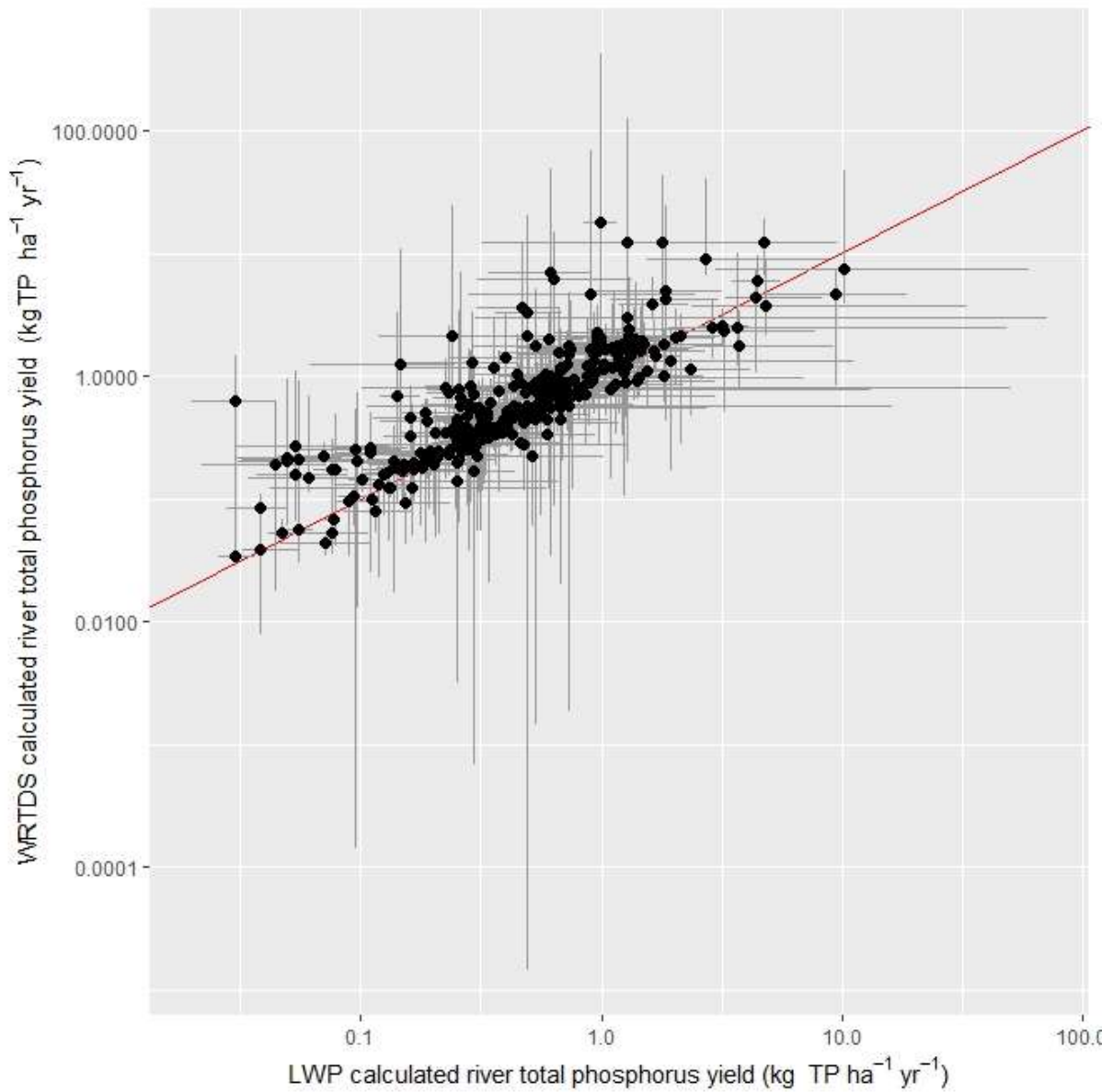
Figure 44. Comparison of TP loads (expressed as yields) calculated by this study (x-axis) with loads calculated using WRTDS. The error bars show the 95% confidence intervals for both sets of estimates. The red line indicates perfect correspondence (one-to-one).

Appendix B   Definition of Types for the empirical water quality models

Table 15. Sets of potential Types for the TN concentration and yield models and the TP concentration model. Category names that have a subscript 'S' (e.g., Steep$_S$) indicate a category defined as per Srinivasan et al. (2021). Category names linked by a plus sign (e.g. EasyHill$_S$ + Steep$_S$) indicate the aggregation of categories defined by Srinivasan et al. (2021). Land use definitions are defined in Table 1. Note that Bare and Water were excluded for the TN concentration and yield models and therefore the number of Types represented in the models was decreased by two.

Set (no. types)	Types used by this study	Details of how the Types are defined
1 (9)	Bare, Cropland, Dairy, Forestry, Natural, OrchardVineyard, SheepBeef, Urban, Water	Land use/cover categories only. The categories Water, Bare were removed for the TN models
2 (13)	Bare, Cropland, Dairy_Dry, Dairy_Irrigated, Dairy_Moist, Dairy_Wet, Forestry, Natural, OrchardVineyard, SheepBeef_Flat, SheepBeef_Hill, Urban, Water	Dairy subdivided into Moisture categories defined by Dry$_S$, Moist$_S$, Wet$_S$ and Irrigated$_S$. Sheep & Beef subdivided into Hill and Flat defined as EasyHill$_S$ + Rolling$_S$ + Steep$_S$ and Flat$_S$, respectively.
3 (14)	Bare, Cropland, Dairy_Dry, Dairy_Wet, Forestry, Natural_Dry, Natural_Wet, OrchardVineyard, SheepBeef_Dry_Flat, SheepBeef_Dry_Hill, SheepBeef_Wet_Flat, SheepBeef_Wet_Hill, Urban, Water	Dairy and Natural subdivided into Dry and Wet defined as Dry$_S$ and Moist$_S$ + Wet$_S$. Sheep & Beef subdivided into Dry_Hill, Dry_Flat, Wet_Hill and Wet_Flat where Hill and Flat are EasyHill$_S$ + Steep$_S$ and Flat$_S$ + Rolling$_S$, respectively, and where Dry and Wet are Dry$_S$ and Moist$_S$ + Wet$_S$ respectively.
4 (15)	Bare, Cropland, Dairy_Dry, Dairy_Irrigated, Dairy_Moist, Dairy_Wet, Forestry, Natural, OrchardVineyard, SheepBeef_Dry_Flat, SheepBeef_Dry_Hill, SheepBeef_Wet_Flat, SheepBeef_Wet_Hill, Urban, Water	Dairy subdivided into Dry, Wet and Irrigated defined as Dry$_S$, Moist$_S$ + Wet$_S$ and Irrigated$_S$. Sheep & Beef subdivided into Dry_Hill, Dry_Flat, Wet_Hill and Wet_Flat where Hill and Flat are EasyHill$_S$ + Steep$_S$ and Flat$_S$ + Rolling$_S$, respectively, and where Dry and Wet are Dry$_S$ and Moist$_S$ + Wet$_S$ respectively.
5 (16)	Bare, Cropland, Dairy_Dry, Dairy_Irrigated, Dairy_Moist, Dairy_Wet, Forestry, Natural_Dry, Natural_Wet, OrchardVineyard, SheepBeef_Dry_Flat, SheepBeef_Dry_Hill, SheepBeef_Wet_Flat, SheepBeef_Wet_Hill, Urban, Water	As for Set 4 with Natural subdivided into Wet and Dry defined as Dry$_S$ and Moist$_S$ + Wet$_S$.
6 (17)	Bare, Cropland, Dairy_Dry, Dairy_Irrigated, Dairy_Moist, Dairy_Wet, Forestry_Dry, Forestry_Wet, Natural_Dry, Natural_Wet, OrchardVineyard, SheepBeef_Dry_Flat, SheepBeef_Dry_Hill, SheepBeef_Wet_Flat, SheepBeef_Wet_Hill, Urban, Water	As for Set 5 with Forestry subdivided into Wet and Dry defined as Srinivasan Dry and Moist + Wet.
7 (21)	Bare, Cropland, Dairy_Dry, Dairy_Irrigated, Dairy_Moist, Dairy_Wet, Forestry_Dry, Forestry_Wet, Natural_Dry, Natural_Wet, OrchardVineyard, SheepBeef_Dry_EasyHill, SheepBeef_Dry_Flat, SheepBeef_Dry_Rolling, SheepBeef_Dry_Steep, SheepBeef_Wet_EasyHill, SheepBeef_Wet_Flat, SheepBeef_Wet_Rolling, SheepBeef_Wet_Steep, Urban, Water	As for Set 6 with Sheep & Beef subdivided into Dry and Wet defined as Dry$_S$, Moist$_S$ + Wet$_S$ and Flat$_S$, Rolling$_S$, EasyHill$_S$ and Steep$_S$.
8 (23)	Bare, Cropland, Dairy_Dry, Dairy_Irrigated, Dairy_Moist, Dairy_Wet, Forestry_Dry, Forestry_Wet, Natural_Dry_Flat, Natural_Dry_Hill, Natural_Wet_Flat, Natural_Wet_Hill, OrchardVineyard, SheepBeef_Dry_EasyHill, SheepBeef_Dry_Flat, SheepBeef_Dry_Rolling, SheepBeef_Dry_Steep, SheepBeef_Wet_EasyHill, SheepBeef_Wet_Flat, SheepBeef_Wet_Rolling, SheepBeef_Wet_Steep, Urban, Water	As for Set 7 with Natural subdivided into Dry_Flat, Dry_Hill, Wet_Flat and Wet_Hill defined as Dry$_S$, Moist$_S$ + Wet$_S$ and Flat$_S$, Rolling$_S$ + EasyHill$_S$ + Steep$_S$.
9 (25)	Bare, Cropland, Dairy_Dry, Dairy_Irrigated, Dairy_Moist, Dairy_Wet, Forestry_Dry_Flat, Forestry_Dry_Hill, Forestry_Wet_Flat, Forestry_Wet_Hill, Natural_Dry_Flat, Natural_Dry_Hill, Natural_Wet_Flat, Natural_Wet_Hill, OrchardVineyard, SheepBeef_Dry_EasyHill, SheepBeef_Dry_Flat, SheepBeef_Dry_Rolling, SheepBeef_Dry_Steep, SheepBeef_Wet_EasyHill, SheepBeef_Wet_Flat, SheepBeef_Wet_Rolling, SheepBeef_Wet_Steep, Urban, Water	As for Set 8 with Forestry subdivided into Dry_Flat, Dry_Hill, Wet_Flat and Wet_Hill defined as Dry$_S$, Moist$_S$ + Wet$_S$ and Flat$_S$, Rolling$_S$ + EasyHill$_S$ + Steep$_S$.

Table 16. Sets of potential land Types for the TP yield model. Category names that have a subscript 'S' (e.g., Steep$_S$) indicate a category defined as per Srinivasan et al. (2021). Category names linked by a plus sign (e.g. EasyHill$_S$ + Steep$_S$) indicate the aggregation of categories defined by Srinivasan et al. (2021).

Set (no. Types)	Types used by this study	Details of how the Types are defined
1 (4)	Dry_Flat, Dry_Hill, Wet_Flat, Wet_Hill	Dry and Wet defined as Dry$_S$ and Moist$_S$ + Wet$_S$, respectively. Flat and Hill defined as Flat$_S$ and EasyHill$_S$ + Rolling$_S$ + Steep$_S$.
2 (4)	Dry_Steep, EasyHillRolling, Flat, Wet_Steep	Dry and Wet defined as Dry$_S$ + Moist$_S$ and Wet$_S$, respectively.
3 (6)	Dry_Flat, Dry_Hill, Moist_Flat, Moist_Hill, Wet_Flat, Wet_Hill	Dry, Moist and Wet defined as Dry$_S$, Moist$_S$ and Wet$_S$, respectively. Flat and Hill defined as Flat$_S$ + Rolling$_S$ and EasyHill$_S$ + Steep$_S$, respectively.
4 (9)	Bare, Cropland, Dairy, Forestry, Natural, OrchardVineyard, SheepBeef, Urban, Water	Land use/cover categories only.
5 (18)	Bare_Dry, Bare_Wet, Cropland_Dry, Cropland_Wet, Dairy_Dry, Dairy_Wet, Forestry_Dry, Forestry_Wet, Natural_Dry, Natural_Wet, OrchardVineyard_Dry, OrchardVineyard_Wet, SheepBeef_Dry, SheepBeef_Wet, Urban_Dry, Urban_Wet, Water_Dry, Water_Wet	Land use/cover categories subdivided into Dry and Wet defined as Dry$_S$ and Moist$_S$ + Wet$_S$. .
6 (19)	Bare, Cropland, Dairy_Dry_Flat, Dairy_Dry_Hill, Dairy_Moist_Flat, Dairy_Moist_Hill, Dairy_Wet_Flat, Dairy_Wet_Hill, Forestry, Natural, OrchardVineyard, SheepBeef_Dry_Flat, SheepBeef_Dry_Hill, SheepBeef_Moist_Flat, SheepBeef_Moist_Hill, SheepBeef_Wet_Flat, SheepBeef_Wet_Hill, Urban, Water	As for 5 but include Moist defined by Moist$_S$ for Dairy and Sheep & Beef.

Appendix C Fitted coefficients for empirical models

Table 17. Best TN concentration model (Model 6) fitted coefficients (mg m^{-3}). Coefficients for each quantile, their standard errors (St Error) and p-values.

Type	Quantile	Coefficient	St Error	P value
Cropland	0.5	4464	1578	0.005
OrchardVineyard	0.5	1102	3058	0.719
Dairy_Dry	0.5	6130	963	0.000
Forestry_Wet	0.5	237	48	0.000
Natural_Dry	0.5	110	260	0.674
Urban	0.5	1368	287	0.000
SheepBeef_Dry_Hill	0.5	183	98	0.064
Forestry_Dry	0.5	1272	496	0.011
SheepBeef_Wet_Hill	0.5	477	57	0.000
Natural_Wet	0.5	46	11	0.000
Dairy_Irrigated	0.5	6311	2025	0.002
Dairy_Moist	0.5	2513	272	0.000
Dairy_Wet	0.5	1170	195	0.000
SheepBeef_Wet_Flat	0.5	1193	280	0.000
SheepBeef_Dry_Flat	0.5	314	357	0.379
Cropland	0.05	759	1086	0.485
OrchardVineyard	0.05	318	1063	0.765
Dairy_Dry	0.05	1856	549	0.001
Forestry_Wet	0.05	154	79	0.051
Natural_Dry	0.05	-71	163	0.663
Urban	0.05	584	128	0.000
SheepBeef_Dry_Hill	0.05	86	32	0.007
Forestry_Dry	0.05	402	334	0.229
SheepBeef_Wet_Hill	0.05	143	77	0.064
Natural_Wet	0.05	-4	24	0.875
Dairy_Irrigated	0.05	1422	677	0.036
Dairy_Moist	0.05	1580	276	0.000
Dairy_Wet	0.05	503	146	0.001
SheepBeef_Wet_Flat	0.05	526	199	0.008
SheepBeef_Dry_Flat	0.05	31	135	0.819
Cropland	0.95	22771	7364	0.002
OrchardVineyard	0.95	11728	6983	0.093
Dairy_Dry	0.95	5636	3324	0.090
Forestry_Wet	0.95	771	530	0.146
Natural_Dry	0.95	334	389	0.391
Urban	0.95	1388	536	0.010
SheepBeef_Dry_Hill	0.95	232	323	0.473
Forestry_Dry	0.95	3290	2108	0.119
SheepBeef_Wet_Hill	0.95	1126	326	0.001
Natural_Wet	0.95	111	19	0.000
Dairy_Irrigated	0.95	13496	3327	0.000
Dairy_Moist	0.95	4124	879	0.000
Dairy_Wet	0.95	1338	702	0.057
SheepBeef_Wet_Flat	0.95	4226	1503	0.005
SheepBeef_Dry_Flat	0.95	4352	1767	0.014

Table 18. Best TN yield model fitted coefficients (kg ha^{-1} yr^{-1}). Coefficients for each quantile, their standard errors (St Error) and p-values.

Type	Quantile	Coefficient	Std..Error	P value
SheepBeef_Hill	0.5	3.9	1.2	0.002
Cropland	0.5	4.9	24.7	0.843
OrchardVineyard	0.5	18.6	38.6	0.631
Dairy_Dry	0.5	28.5	11.3	0.012
Dairy_Irrigated	0.5	29.6	15.1	0.050
Dairy_Moist	0.5	17.0	4.1	0.000
Forestry	0.5	8.5	2.1	0.000
Natural	0.5	2.4	0.6	0.000
Dairy_Wet	0.5	37.5	3.4	0.000
Urban	0.5	10.7	4.7	0.024
SheepBeef_Flat	0.5	8.3	3.7	0.026
SheepBeef_Hill	0.05	0.2	0.3	0.441
Cropland	0.05	4.0	18.1	0.825
OrchardVineyard	0.05	8.7	32.4	0.787
Dairy_Dry	0.05	10.2	10.2	0.315
Dairy_Irrigated	0.05	6.6	10.8	0.541
Dairy_Moist	0.05	15.9	1.2	0.000
Forestry	0.05	4.4	1.4	0.002
Natural	0.05	1.4	0.3	0.000
Dairy_Wet	0.05	31.3	3.3	0.000
Urban	0.05	-7.9	12.4	0.524
SheepBeef_Flat	0.05	0.4	1.0	0.692
SheepBeef_Hill	0.95	12.0	2.0	0.000
Cropland	0.95	84.9	50.1	0.091
OrchardVineyard	0.95	-45.8	43.8	0.297
Dairy_Dry	0.95	41.5	23.2	0.075
Dairy_Irrigated	0.95	65.4	29.5	0.027
Dairy_Moist	0.95	47.0	13.3	0.000
Forestry	0.95	27.0	12.9	0.038
Natural	0.95	6.6	1.3	0.000
Dairy_Wet	0.95	64.1	13.8	0.000
Urban	0.95	6.8	19.3	0.724
SheepBeef_Flat	0.95	18.7	9.0	0.038