Sediment load reductions to meet suspended and deposited sediment thresholds

Prepared for Ministry for the Environment

June 2019

Prepared by:
D M Hicks
A Haddadchi
A Whitehead
U Shankar

For any information regarding this report please contact:

Murray Hicks

+64-3-343 7872
murray.hicks@niwa.co.nz

National Institute of Water & Atmospheric Research Ltd
PO Box 8602
Riccarton
Christchurch 8011

Phone +64 3 348 8987

NIWA CLIENT REPORT No: 2019100CH
Report date: June 2019
NIWA Project: MFE19502

Quality Assurance Statement		
	Reviewed by:	Rick Stoffels
	Formatting checked by:	Rachel Wright
	Approved for release by:	Scott Larned

Contents

Tables

Figures

Executive summary

Background and aims

Excess fine sediment negatively impacts freshwater ecosystems, either when suspended in the water column or by depositing on stream beds. Towards including fine sediment in the National Objectives Framework (NOF), the Ministry for the Environment (MfE) has pursued a series of complimentary workstreams to understand links between catchment sediment loads, fine sediment related environmental state variables (ESVs, including visual clarity, turbidity, and deposited sediment) and ecological responses, and to develop threshold values for these ESVs based on their effects on freshwater fish and macroinvertebrates. The investigation reported here extends these previous workstreams by quantifying the reductions in mean annual sediment load required to meet the turbidity, visual clarity, and deposited sediment thresholds where predictive modelling datasets indicate they have been breached. This is to support MfE's impact analysis for potential sediment regulations.

Work components

The work comprised two components: one focussing on the ESVs relating to suspended sediment (turbidity, visual clarity), the other addressing deposited fine sediment (which is sediment less than 2 mm in grain size).

The deposited fine sediment component aimed to establish and model relationships between sediment load and the in-stream areal coverage of deposited fine sediment, and, if model development was successful, estimate load reduction requirements to meet deposited sediment bottom lines. This staged approach was to accommodate the high risk, based on past experience, that suitable relationships could not be established between sediment load and the preferred deposited sediment indicator. As things turned out, the poor performance of the model developed in the first stage resulted in the decision to not progress with later stages.

ESV thresholds

Threshold values of median clarity, turbidity, and deposited fine sediment cover were supplied to this study from a parallel study being undertaken for MfE. These thresholds correspond to the proposed values of these three variables at the boundary between the C and D environmental quality bands. Thus, they may be regarded as bottom lines.

Predicting sediment load reduction required to meet visual clarity and turbidity thresholds

We used simplified versions of existing relations between sediment loading and turbidity and visual clarity to map where their proposed thresholds are expected to be exceeded, and then to estimate the reduction in mean annual sediment load required to meet the threshold values. These relations link visual clarity (V) and turbidity (T) to sediment loads via sediment rating curves (i.e., relation between discharge and suspended sediment concentration, C) and flow duration curves and link C with T and V.

The proportional reduction in catchment sediment load (R) required to increase visual clarity or decrease turbidity to a target value was expressed as a simple function of the existing and target median clarity or turbidity (V_{50} and V_{t50}, or T_{50} and T_{t50}, respectively) and the exponents (*d* or *f*, respectively) in the relations between C and V or C and T.

The *d* and *f* exponents applied nationally were estimated as the average of the values of these exponents observed at 77 sites around the country. These national averages of *d* and *f* performed as well as random forest models that predicted at-a-site values of *d* and *f* off catchment characteristics.

Random forest models were used to predict V_{50} and T_{50} across the national drainage network from catchment physical and hydrological characteristics. With these models, the strongest predictor variable was the discharge-weighted average suspended sediment concentration (equal to mean sediment load divided by mean discharge), confirming the positive relationship between catchment sediment load and V_{50} and T_{50} that underpins the overall approach. The uncertainty levels on predicted V_{50} and T_{50} at any site were significant but not unworkable.

Absolute values of sediment load reduction (t/yr) may be derived by multiplying the proportional load reduction by the contemporary catchment load estimated from models such as the one recently updated for MfE (Hicks et al. 2019).

Which river segments exceed the thresholds for turbidity and clarity?

Over the national stream network, 18.4% of segments exceeded the clarity thresholds, 16.4% exceeded the turbidity thresholds, 10.3% exceeded both thresholds (indicating moderate concordance between the two thresholds), and 24.5% exceeded one or other or both thresholds. Thus, the majority of stream segments across New Zealand require no sediment load reduction to meet the turbidity and visual clarity bottom lines.

What load reduction is required to achieve the thresholds for turbidity and clarity?

After discussion with MfE, the analysis of sediment load reduction was focussed on that required to meet the turbidity thresholds. Of the 16.4% of segments nationally exceeding the turbidity thresholds, 70% required a load reduction factor (i.e., [current load – load to meet threshold]/current load) of less than 0.4.

The required load reduction factor estimates carry significant uncertainty, largely due to the uncertainty on the modelled T_{50} values. This indicates a significant risk that many of the segments calculated as requiring a relatively small load reduction may not need it (while some that do have been missed). When averaging or totalling results over multiple segments (e.g., regionally or nationally), such errors are likely to be systematic at the catchment scale but vary randomly between catchments. This is because the median turbidity values at stream segments are estimated off upstream catchment characteristics, and so, within a catchment, linked segments share common upstream characteristics. Thus, the turbidity estimate (and its error) at one segment will not be independent of the estimates at segments elsewhere in the same catchment that are linked to it.

Areas where erosion mitigation would likely be focussed can be identified on stream network maps with the load reduction factor in each segment colour-coded, or by mapping "pour-point" catchments. Pour-points were set at the first segment in a catchment where an above-zero load reduction factor was encountered in a trace upstream from the coast. Both options were provided as ArcGIS shape files. The DOC estate was excluded, to preserve the natural state of rivers, as were mountain catchments with glaciers, because it is not possible to mitigate turbid glacial melt.

Deposited fine sediment

Previous attempts to model deposited fine sediment from New Zealand datasets have been found wanting, both through limited performance and a lack of rational dependence on catchment sediment load. In this study we attempted an alternative, more physically-based modelling approach where we considered the time-averaged fine sediment cover (FSC) should relate to four factors:

- The sediment supply factor recognises that deposition can only occur during flow recessions or baseflows, thus it is the supply and concentration of suspended sediment under these waning/base flow conditions that is important. This will depend on the relative phasing of the sediment and water delivered to the stream network upstream from the segment of interest, which should relate to factors such as the erosion terrain, land cover and land use, and source of flow. It should also relate inversely to runoff "flashiness", since this will control the duration of elevated concentrations during recessions.
- The sediment trap-efficiency factor relates to the local hydraulic conditions at the reach and the sediment grainsize.
- The probability-of-occurrence factor sets the likelihood that deposited sediment will be observed during a monitoring program.
- The space factor depends inversely on the size of the streambed framework material.

Using predictor variables to represent these factors, and with a dataset compiled from 467 sites where FSC had been measured by the Instream Visual method, Genetic Algorithm Optimisation was used to calibrate FSC models from the full dataset and the dataset partitioned by source-of-flow classes from the River Environment Classification. While these models confirmed the importance of most of the factors described above, their performance remained relatively poor. Probable reasons for the disappointing performance include measurement error, sampling error, uncertainty in the predictor variables (several of which were predicted themselves), missing key variables, and process complexity and variability.

Towards managing FSC by regulating catchment sediment loads

Two problems arise for the impacts assessment workstream and for future policy application: the FSC model's predictive capability is generally too inaccurate to adequately resolve which stream segments are over- or under-threshold, and there is minimal dependence on catchment average annual sediment load.

A way forward may be to abandon the focus on managing catchment average sediment loads but, instead, manage the supply of sediment to the late stages of flood recessions and during baseflows – which is when fine sediment deposition in baseflow channels occurs. Further research on this concept is recommended.

1 Introduction

1.1 Background

Excess fine sediment is widely acknowledged to impact negatively on freshwater ecosystems, either when suspended in the water column or by depositing on stream beds (e.g., Davies-Colley et al. 2015). Relevant fine-sediment related environmental state variables (ESVs) include visual water clarity and turbidity (both determined by suspended sediment concentration and physical characteristics) and streambed cover of deposited fine sediment (DFS).

Towards including fine sediment in the National Objectives Framework (NOF), the Ministry for the Environment (MfE) has pursued a series of complimentary workstreams. This has included work to understand links between catchment sediment loads and fine sediment ESVs (e.g., Hicks et al. 2016), to characterise relationships between fine sediment ESVs and ecological responses (Depree et al. 2018), and to develop ESV threshold values, including bottom lines, by analysing the relationships between fine sediment ESVs and ecological response variables (Franklin et al. 2019).

The investigation reported here extends these previous workstreams by quantifying the reductions in mean annual sediment load required to meet the in-stream suspended and deposited sediment indicator thresholds where predictive modelling datasets indicate they have been breached. This is to support MfE's regulatory impact analysis for potential sediment regulations.

The primary risk foreseen for the project was that model relationships would not be established between sediment load and the preferred deposited sediment indicator. This risk was to be mitigated by providing alternative research approaches and a staged approach.

1.2 Scope

In overview, the investigation covers both suspended and deposited sediment indicators nationwide, with the outputs to show sediment load reduction requirements per river reach and at other appropriate scales to meet bottom line thresholds provided by MfE. It is to include quantitative sensitivity analyses and information (quantitative and/or qualitative) on the sources of variability.

In detail, the project has multiple technical work components.

Component 1 covers water clarity and turbidity as indicators of suspended sediment and has two parts:

1. Model refinement and technical analysis of national relationships between suspended sediment load and visual clarity and turbidity. This involves using or enhancing predictive models and datasets for visual clarity and turbidity developed in Larned et al. (2018) and Whitehead (2019) to identify where there are breaches of bottom line thresholds for visual clarity and turbidity.

2. Estimating the suspended sediment load reduction required to meet turbidity and/or visual clarity bottom lines. This will use an appropriate variation of the method outlined in Hicks et al. (2016) and Dymond et al. (2017) to calculate the required suspended sediment load reductions to meet bottom lines, and will also use the suspended sediment yield data and model developed in Hicks et al. (2019).

Component 2 addresses deposited fine sediment (which is sediment of sand grade and finer, with diameter less than 2 mm). It aims to establish and model relationships between sediment load and the in-stream areal coverage of deposited fine sediment, and, if model development is successful, estimate load reduction requirements to meet deposited sediment bottom lines. Since, based on past attempts (e.g., Hicks et al. 2016), there is low confidence in finding such a model, a staged analysis was planned, with decision points on how best to continue, if at all, to be made at the end of each stage. The intended stages were:

1. Develop a model to produce relationships between sediment load and deposited fine sediment using the Visual Instream definition of areal coverage (i.e., the "SAM2" protocol of Clapcott et al. 2011) and produce summary statistics on model performance.
2. If the model's performance is not satisfactory, then develop a model to produce relationships between sediment load and deposited fine sediment using the Suspendable Inorganic Sediment (SIS) indicator (i.e., the "SAM4" protocol of Clapcott et al. 2011) and produce summary statistics on model performance. If this step provides significant results, use appropriate methods to convert from the SIS values to Visual Instream values.
3. Given a useful dataset and model from Stages 1 or 2, identify where there are breaches of deposited sediment bottom line thresholds, then use the general method outlined in Hicks et al. (2016) and Dymond et al. (2017) to calculate the required suspended sediment load reductions needed to meet deposited sediment attribute bottom lines.

As things turned out, the poor performance of the model developed in Component 2 / Stage 1 and discussion with MfE resulted in the decision to not progress with Stages 2 and 3.

Lastly, the applied aspects of the Component 1 work were repeated using an alternative set of turbidity and clarity bottom line values. Since these were provided after the work with the initial set of bottom line values was reported, the results of the re-analysis are presented in Appendix C.

1.3 Structure of this report

This report documents these investigations generally in the order they are listed above.

2 Thresholds / bottom-lines

Threshold values of median clarity, median turbidity, and median deposited fine sediment cover were supplied to this study from a parallel study being undertaken for MfE (Franklin et al. 2019). These thresholds correspond to the values of these three variables at the boundary between the C and D environmental quality bands, thus they may be regarded as “bottom lines”.

The procedure for developing these thresholds is detailed by Franklin et al. (2019). In brief, the river segments within the REC2[1] digital river network were first classified into 12 climate-topography-geology (CTG) classes using a clustering analysis, then different thresholds were developed for each class. This spatially-varying threshold approach was required to deal with natural variability in these variables around the country. Independent CTG classifications were developed for DFS and the two suspended-sediment concentration related variables (i.e., visual clarity and turbidity). The spatial distributions of the CTG classes are mapped in Figure 2-1. The C/D thresholds by CTG class are listed in Table 2-1.

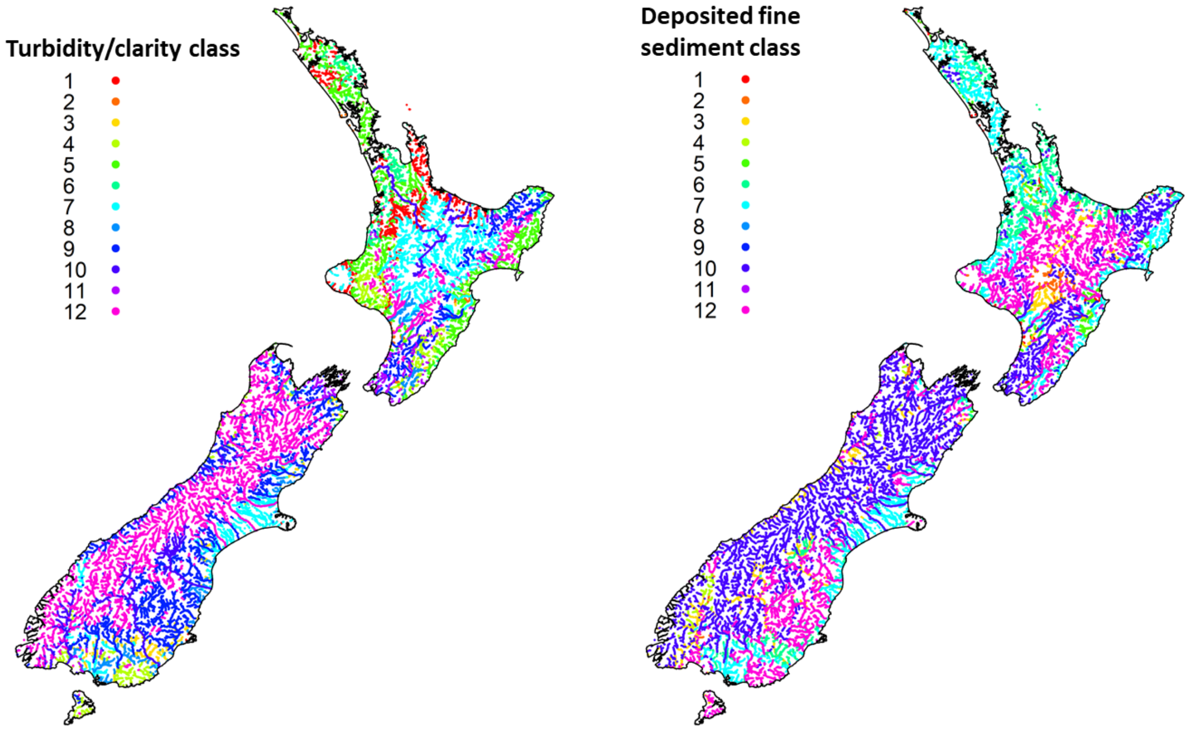

Figure 2-1: Maps of climate-topography-geology classes used for defining thresholds for turbidity and visual clarity and deposited fine sediment. Source: Franklin et al. (2019).

Turbidity and visual clarity both depend on the concentration and size characteristics of suspended sediment (with turbidity increasing and clarity decreasing as sediment size gets finer) and are inversely related to each other (Hicks et al. 2016). We checked that the supplied turbidity and clarity thresholds conformed with this expected relationship by over-plotting (Figure 2-2) the matching turbidity and clarity thresholds by CTG band and the 95% confidence intervals around the relationship between observed median turbidity and clarity from 580 sites (from the dataset used in Section 3.3 to develop national predictive models of turbidity and clarity). This showed that the

[1] The Franklin et al. (2019) work was undertaken for the REC1 river network; the classifications and results were mapped onto the REC2 network for this study by Dr D Booker, NIWA.

threshold pairs for CTG classes 2, 3, and 8 lie outside the 95% confidence interval, with the class 2 point a notable outlier. This indicates potential issues with the threshold values set for these three classes, but it is unclear whether the turbidity or clarity thresholds (or both) make these points outliers.

Table 2-1: C/D band thresholds for median turbidity, visual clarity, and deposited fine sediment cover for climate-topography-geology (CTG) classes. Note that the same 12-class CTG classification applies to the turbidity and visual clarity thresholds but a separate 12-class CTG classification was developed for the deposited fine sediment thresholds. Source: Franklin et al. (2019).

CTG Class	Turbidity threshold (NTU)	Visual clarity threshold (m)	Deposited fine sediment cover threshold (proportion of streambed)
1	3.21	1.55	0.97
2	10.45	1.65	0.21
3	2.02	1	0.6
4	4.83	1.02	0.23
5	13.11	0.42	0.92
6	8.29	0.7	0.46
7	3.32	1.3	0.56
8	6.42	0.44	0.45
9	1.6	2.35	0.61
10	1.49	2.51	0.29
11	1.56	2.06	0.89
12	3.14	2.23	0.45

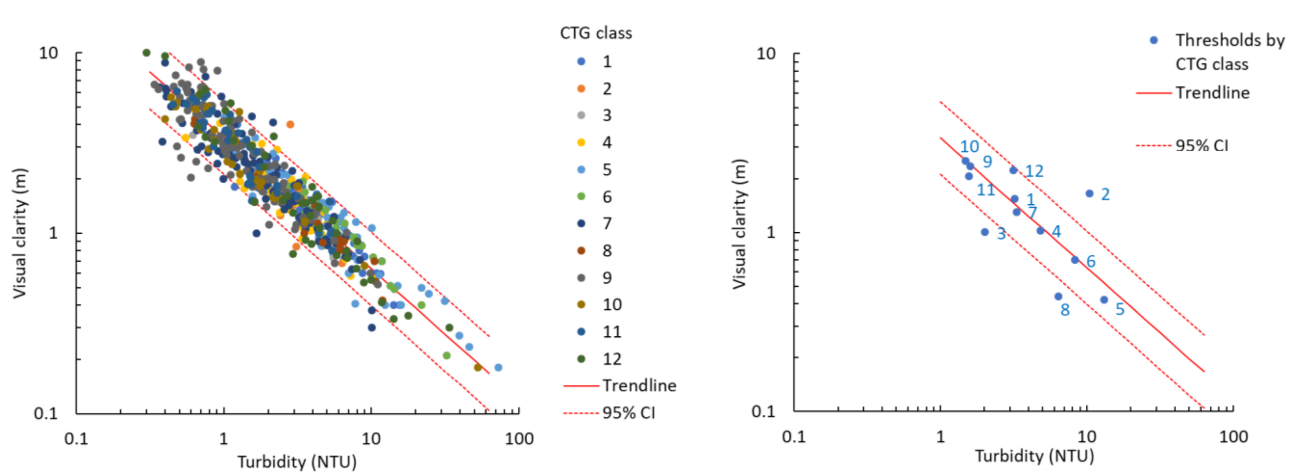

Figure 2-2: Relationships between visual clarity and turbidity for C/D band thresholds (right) and for training dataset used to developed national predictor models of clarity and turbidity (left). Points are distinguished by CTG class. Solid red line shows regression fit to training dataset; dotted red lines show 95% confidence interval around regression line. These red lines are over-plotted on the threshold plot (right).

3 Suspended sediment attributes: visual clarity and turbidity

3.1 General approach

To map nationwide where the proposed bottom line thresholds for visual clarity (V) and turbidity (T) are expected to be exceeded, and then to estimate the reduction in sediment load required to improve stream V and T, we employed simplified versions of the relations between sediment loading and environmental indicators that were developed by Hicks et al. (2016) and published by Dymond et al. (2017). This simplified approach is enabled because the thresholds for V and T are proposed in terms of median values of V and T (i.e., V_{50} and T_{50}), whereas in Hicks et al. (2016) the exceedance percentiles were undeclared. For this study, we only need to focus on medians.

The essential elements of this simplified approach are as follows:

- We expect (based on the findings of Hicks et al. 2016) that V (units: m) and T (NTU) will generally fit power-law functions of suspended sediment concentration (C: mg/l):

  $$V = gC^{d} \quad (1)$$
  $$T = eC^{f}, \quad (2)$$

  where *d*, *e*, *f* and *g* are site-specific empirically-derived coefficients. Thus $V_{50} = gC_{50}^{d}$ and $T_{50} = eC_{50}^{f}$.

- Catchment sediment load is derived using a sediment rating curve and the flow frequency distribution. The sediment rating curve is usually expressed in the form $C = aQ^{b}$, where *a* and *b* are site-specific coefficients and Q is water discharge (l/s), thus the existing catchment sediment load (L: mg/s) is calculated as:

  $$L = \sum p_i\, aQ_i^{b+1} = aQ^{*} \quad (3)$$

  where p_i are the proportions of time that discharges are within each discharge band (Q_i) and $Q^{*} = \sum p_i\, Q_i^{b+1}$.

  Q* may be regarded as a function of the catchment hydrology, which, most simply, can be assumed not to change if the sediment load is reduced (as discussed in Hicks et al. 2016, in reality the Q* function may change, either due to a change in the rating curve slope *b* or a land use-driven change in the flow frequency distribution, but this will be ignored because we cannot currently quantify this at a national scale).

- From the above, for V, we get, for each catchment:

  coefficient $a = (C_{50}/Q_{50}^{b})$
  $C_{50} = (V_{50}/g)^{1/d}$ and $C_{50} = (T_{50}/e)^{1/f}$

  thus

  $$L = aQ^{*} = Q^{*}(V_{50}/g)^{1/d} / Q_{50}^{b} = Q^{*}(T_{50}/e)^{1/f} / Q_{50}^{b} \quad (4)$$

- If we let L_{tv} be the target sediment load for visual clarity and V_{t50} is the defined threshold for V_{50}, then:

  $$L_{tv} = Q^{*}(V_{t50}/g)^{1/d} / Q_{50}^{b} \quad (5)$$

Similarly, for turbidity:

$$L_{tt} = Q^*(T_{t50}/e)^{1/f} / Q_{50}^{b} \quad (6)$$

- Finally, the load reduction factor for visual clarity may be expressed as:

$$R_v = (L- L_{tv})/L = 1- L_{tv}/L = 1 - (V_{t50}/V_{50})^{1/d} \quad (7)$$

and for turbidity:

$$R_t = (L - L_{tt})/L = 1- L_{tt}/L = 1 - (T_{t50}/T_{50})^{1/f} \quad (8)$$

In other words, the proportional reductions in catchment sediment load required to increase visual clarity or decrease turbidity of river water to a target value are simple functions of the existing and target median clarity and turbidity and the exponents *d* and *f* – which can be derived locally or can be assumed to take national average values. This simplification develops because the terms Q*, *g, e,* and *b* at any location are assumed constant, so they cancel out for the ratio of the existing and target cases.

The following sub-sections develop this approach by:

- deriving the *d* and *f* exponents in the relationships between suspended sediment concentration and visual clarity and turbidity from available datasets
- developing predictive models for V_{50} and T_{50} as functions of catchment characteristics in the River Environment Classification (REC) that can be assumed to apply to every reach in the rest of the country, and
- applying these models and relationships in Equations (7) and (8) to predict R_v and R_t across the river network.

3.2 Relationships between suspended sediment concentration and visual clarity and turbidity

3.2.1 Method

The exponents *d* and *f* in the relations between C and V and T (Equations 1 and 2, respectively) were obtained from analysis undertaken by Hicks et al. (2016). This fitted at-a-site V vs C and T vs C relations to data from sites in the National River Water Quality Network (NRWQN). The relations were fitted using Standard Major Axis (SMA) regression of the log-transformed data. Two approaches for developing a national model of these relations were pursued by Hicks et al. (2016). The first simply used the averages of the at-a-site exponents as national estimates of *d* and *f* and used their standard deviations as measures of the uncertainty. The second involved developing random forest models that predicted the exponents as a function of catchment characteristics.

3.2.2 Data

As detailed in Hicks et al. (2016), the NRWQN dataset included concurrently-collected measures of visual clarity (measured by black-disc), nephelometric turbidity (bench-measured on water samples using the same turbidimeter calibrated to formazin standards, with units of NTU[2]), and suspended sediment concentration (measured using the Total Suspended Solids, or TSS, laboratory procedure).

[2] NTU = Nephelometric Turbidity Units

After scrutiny of data quality in other potential datasets from regional councils, the dataset used was limited to 77 sites in NIWA's NRWQN. It was also noted that 13 of those 77 sites had upstream lakes and reservoirs which could potentially influence the relations because of their effect on filtering-out all but very fine sediment from their outflows. For this reason, both the "full" dataset (including segments downstream of lakes) and the "reduced" dataset (excluding segments downstream of lakes) were analysed to see if any significant differences appeared between the two.

Across the full dataset, TSS ranged from 0.1 to 10,500 mg/l, visual clarity from 0.005 to 18 m, and turbidity from 0.15 to 3,500 NTU.

3.2.3 At-a-site exponent statistics

Table 3-1 shows the at-a-site statistics for the clarity and turbidity exponents for the reduced and full datasets. This indicates that on a national average basis d = -0.76 ± 0.12-0.13 while f = 0.98 ±0.17-0.19, irrespective of whether the full or reduced dataset is used. Both d and f are normally distributed at the 5% significance level (K-S test).

Table 3-1: Statistics of exponents *d* and *f* of relations between suspended sediment concentration and clarity and turbidity for full and reduced datasets. The standard deviations provide similar measures of model performance to the root-mean-square error of the random forest predictive models for *d* and *f*.

Statistic	Full dataset – *d* (clarity)	Full dataset – *f* (turbidity)	Reduced dataset – *d* (clarity)	Reduced dataset – *f* (turbidity)
Count	77	77	64	64
Minimum	-1.07	0.5	-1.07	0.5
Maximum	-0.38	1.75	-0.38	1.75
Mean	-0.76	0.98	-0.76	0.98
Standard deviation	0.13	0.19	0.12	0.17
Random forest model RMSE	0.13	0.20	0.12	0.18

3.2.4 At-a-site exponent random forest models

For the random forest models, the exponents d and f were related to a range of explanatory variables, including various measures of catchment hydrology, size, climate, land-cover, lithology, and topology (as detailed in Hicks et al. 2016). However, as shown by the performance statistics reproduced here in Table 3-2, none of the models performed well: explaining little of the variance in the dataset (out-of-bag R^2 typically 0.05-0.07), performing only marginally better than the simple statistics based model using only the mean of the observed exponents (Nash Sutcliffe Efficiency 0.02-0.07), and with root-mean-square errors (RMSE) no better (or slightly worse) than the standard deviation of the observed exponents (Table 3-1). On that basis we chose to use the simple statistics-based estimators of d and f as given in Section 3.2.3.

Table 3-2: Performance results for random forest models predicting the exponents for the relationships of suspended sediment concentration (analysed with TSS procedure) with visual clarity and turbidity. Results provided for two datasets: full (n=77) and reduced (n= 64, with lake outflow impacted sites excluded). OOB R^2 = out-of-bag R^2; NSE = Nash Sutcliffe Efficiency; RMSE = Root Mean Square Error.

	Reduced dataset (n = 64)			Full dataset (n = 77)		
Regression Coefficients	OOB R^2	NSE	RMSE	OOB R^2	NSE	RMSE
TSS- Visual clarity slope	0.05	0.02	0.12	0.04	0.07	0.13
TSS – turbidity slope	0.05	0.07	0.18	0.06	0.05	0.20

3.3 National models predicting median visual clarity and turbidity

3.3.1 Data

River water quality data

The monitoring sites and data used for this study were compiled as part of a national-scale study of river water quality state (Larned et al. 2018). The water quality data consisted of measurements of visual clarity and turbidity from river monitoring sites in council SOE networks and the NRWQN sites (Table 3-3). Detailed methods for processing the water quality data are given in Larned et al. (2018). The monitoring sites had the following properties: i) less than 50% of the values for a variable were censored; ii) values for at least 90% of monthly or quarterly sampling dates were available, including imputed values; iii) at least 30 values were distributed over four of the five years from 2013 to 2017. All monitoring sites were projected onto the REC2 digital river network, then manually checked. In the final dataset used for random forest (RF) modelling, 587 visual clarity and 878 turbidity sites met the inclusion criteria (Table 3-3). Median values of turbidity and visual clarity were extracted for these sites.

The geographic distribution of river monitoring sites used for modelling is shown in Figure 3-1. They are reasonably well-distributed, although there are gaps in the central North Island and west coast of the South Island.

Table 3-3: River water quality variables, measurement units and site numbers used to develop random forest models.

Variable	Abbreviation in model	Units	Number of monitoring sites
Median visual clarity	CLAR	m	587
Median turbidity	TURB	NTU	878

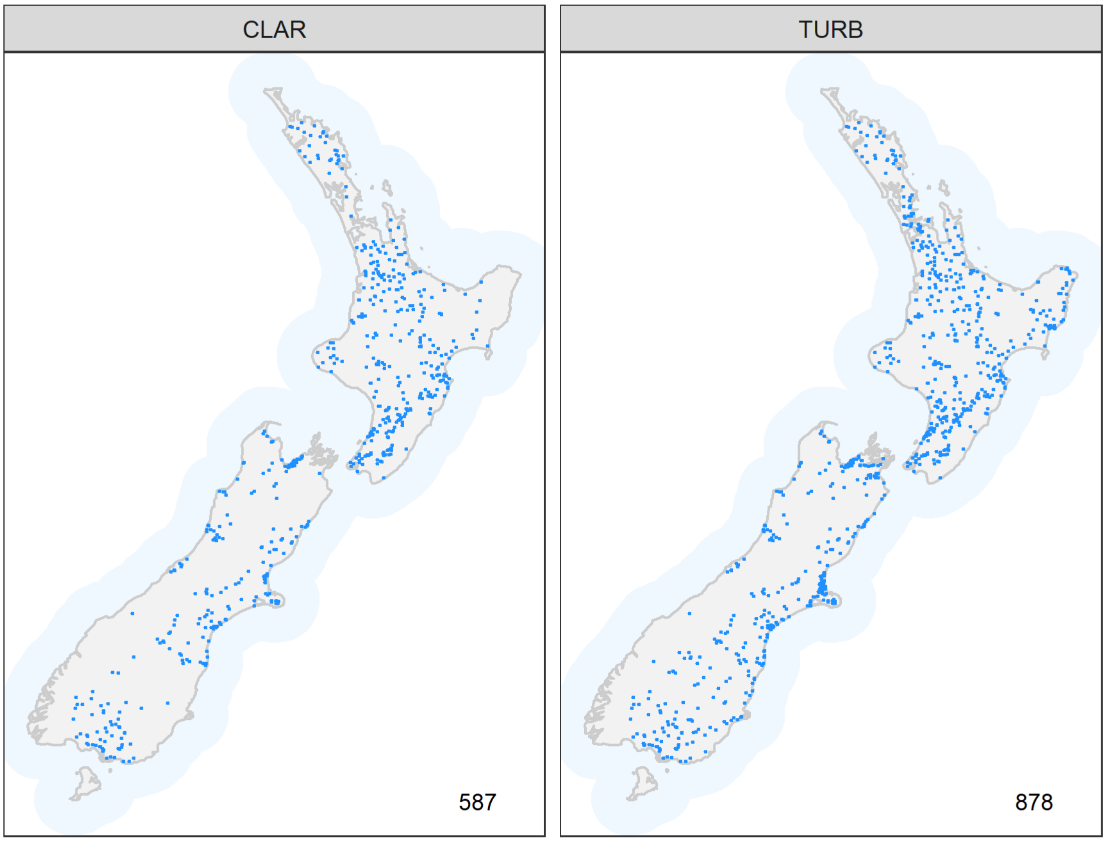

Figure 3-1: Locations of river water quality monitoring sites used for modelling the state of visual clarity (CLAR) and turbidity (TURB). The number in the lower right of each panel corresponds to the number of sites included for each variable (Table 3-3).

Predictor data

The digital river network and catchment boundaries used for Version 2 of the River Environment Classification (REC2)[3] provided the spatial framework for the RF models of turbidity and visual clarity. We selected 33 network attributes from the REC2 database (Table 3-4) for predictor variables. These were largely selected based on previously discovered mechanistic or correlative relationships with water quality (Unwin et al. 2010; Larned et al. 2016; Whitehead 2019). However, given the application of these models to management of river sediment loads, we also included a sediment load related predictor, expressed as a suspended sediment concentration (SSC) since both turbidity and visual clarity are known to be related to SSC. The predictor is the mean annual suspended sediment load passing the observation site divided by the mean annual water discharge, so is a discharge-weighted mean SSC. The mean annual sediment load was derived from the recent national model developed for MfE by Hicks et al. (2019). The mean annual water discharge was derived from the model of Woods et al. (2006).

[3] The national digital stream network has been recently updated to correct errors and to improve its representation of rivers nationally. The REC geodatabase with the updated network is referred to as REC2 (version 2.4). There are approximately 590,000 segments and their corresponding catchments in the REC2 digital network. Each segment in the digital network has a unique identifier, the nzsegment number. REC2 contains spatial data layers describing the climate, topography, geology, vegetation, infrastructure and hydrology of the segment and its catchment. Catchment land cover in REC2 is derived from the national Land Cover Database-4 (LCDB4) which differentiates 32 categories based on analysis of satellite imagery from 2012 (lris.scinfo.org.nz). Descriptions of catchment regolith are derived from the Land Resources Inventory (LRI), including interpretations of the LRI categories made by Leathwick et al. (2003). Additional variables for each segment have been derived from national-scale hydrological modelling (e.g., Booker & Snelder 2012).

3.3.2 Modelling methods

Random forest models

We separately modelled median clarity and turbidity as functions of the predictor variables using RF models (Breiman 1984, 2001; Cutler et al. 2007), with all variables log-transformed (i.e., the $\log_{10}$ of the median of the untransformed raw data). A detailed explanation of the RF modelling approach is provided in Appendix A.

All calculations were performed in the R statistical computing environment (R Core Team 2017) using the *randomForest* package (Liaw & Wiener 2002) and other specialised packages.

Model performance

Model performance was assessed by comparing observations with independent predictions (i.e., sites that were not used in fitting the model), which were obtained from the out-of-bag (OOB) samples. We summarised the models using four statistics: regression R^2**,** Nash-Sutcliffe Efficiencies (NSE), bias, and root-mean-square-deviation (RMSD)[4].

Table 3-4: Predictor variables used in random forest models of median turbidity and median water clarity. *Geological variables are based on regolith, using averages of ordinal values assigned to LRI top-rock categories by Leathwick et al. (2003). The variables usHard and usPsize characterise physical regolith conditions; usPhos and usCalc characterise regolith fertility.

Predictor variable class	Predictor variable description	Abbreviation	Unit
Sediment	Discharge-weighted sediment concentration (the mean annual suspended sediment load divided by the mean annual water discharge)	sedConc	$\log_{10}(g/m^3)$
Geography & topography	Catchment area	usArea	m^2
	Segment mean elevation	segElev	m ASL
	Percentage of catchment occupied by lakes	usLakePerc	%
	Mean catchment elevation	usElev	m ASL
	Mean catchment slope	usSlope	degrees
	Distance to the coast	DistToCoast	m
	Mean segment slope	SegSlope	degrees
	Segment sinuosity (segment length divided by the straight line distance between endpoints)	Sinuosity	unitless
	Distance to furthest headwater segment	DistToHead	m
Climate & flow	Mean segment June air temperature	segTmin	degrees C x 10
	Mean segment January air temperature.	segTwarm	degrees C x 10
	Mean catchment June air temperature	usTmin	degrees C x 10
	Mean catchment January air temperature	usTwarm	degrees C x 10
	Mean annual catchment rainfall	usRain	mm
	Mean catchment coefficient of variation of annual rainfall	usRainvar	mm/yr
	Mean catchment rain days > 10 mm	usRainDays10	days/mo
	Mean catchment rain days > 20 mm	usRainDays20	days/mo
	Mean catchment rain days > 100 mm	usRainDays100	days/mo
	Mean annual catchment potential evapotranspiration	usPET	mm/yr
	Estimated mean flow	MeanFlow	m^3/s

[4] Refer to glossary (Section 7) for explanations of these performance metrics.

Geology*	Mean catchment induration (hardness) of regolith	usHard	Ordinal
	Mean catchment phosphorous content of regolith	usPhos	Ordinal
	Mean catchment particle size of regolith	usPsize	Ordinal
	Mean catchment calcium content of regolith	usCalc	Ordinal
Land cover	Proportion of catchment occupied by combination of high producing exotic grassland, short-rotation cropland, orchard, vineyard and other perennial crops (LCDB4 classes 40, 30, 33)	usIntensiveAg	%
	Proportion of catchment in low producing grassland (LCDB4 class 41)	usPastoralLight	%
	Proportion of catchment in native forest (LCDB4 class 69)	usNativeForest	%
	Proportion of catchment in built-up areas, urban parkland, surface mines, dumps and transport infrastructure (LCDB4 classes 1,2,6,5)	usUrban	%
	Proportion of catchment in scrub and shrub cover (LCDB4 classes 50, 51, 52, 54, 55, 56, 58)	usScrub	%
	Proportion of catchment occupied by lake and pond, river and estuarine open water (LCDB4 classes 20, 21, 22)	usWetland	%
	Proportion of catchment in exotic forest (LCDB3 class 71)	usExoticForest	%
	Proportion of catchment occupied in bare or lightly-vegetated cover (LCDB4 classes 10, 12, 14, 15, 16)	usBare	%

Model predictions

Predictions are made with RF models by "running" new cases down every tree in the fitted forest and averaging the predictions made by each tree (Cutler et al. 2007). The models in this study were fitted to $\log_{10}$-transformed data. When these models are back-transformed, the model error term no longer has a mean of zero. Ignoring this results in retransformation bias (i.e., predictions that systematically underestimate the response). We corrected the retransformation bias using the smearing estimator (S) developed by Duan (1983):

$$S = \frac{1}{n}\sum_{i=1}^{n} 10^{\hat{\varepsilon}_i} \qquad (9)$$

where $\hat{\varepsilon}_i$ are the residuals of a RF model. The predictions were back-transformed by raising them to the power of 10, then corrected for retransformation bias by multiplying by S.

3.3.3 Results

Model performance

The RF models for clarity and turbidity performed well, as indicated by the following statistics: $R^2 > 0.6$, NSE > 0.5, and RMSD < 0.3 for both variables (Table 3-5). Bias in the RF models was low as indicated by the close match between the line representing the regression of the observed versus predicted values (red dashed line in Figure 3-2) and the one-to-one line (blue solid line in Figure 3-2), although there is a tendency for both models to slightly underestimate at high values and overestimate at low values[5]. Based on NSE values, the models for clarity and turbidity had similar performance. The log-space residuals (i.e., predicted – observed log values) were normally distributed at the 5% significance level (K-S test).

[5] We have no explanation as to why this slight misalignment between observed and predicted values occurred. In the following sub-section we apply an empirical adjustment to align the predicted and observed values.

Table 3-5: Performance of median turbidity and visual clarity models. Performance was determined using independent predictions (i.e., sites that were not used in fitting the models) generated from the out-of-bag observations. Regression R^2 = coefficient of determination, NSE = Nash-Sutcliffe efficiency, RSR = relative root-mean-square-error, RMSD = root mean square deviation, and SFE = untransformed standard factorial error associated with the RMSD. Units for RMSD and bias are the $\log_{10}$ transformed units of the respective variable.

Variable	Number of sites	Regression R^2	NSE	Bias	RSR	RMSD	SFE
Visual clarity	587	0.61	0.59	0.003	0.64	0.20	1.58
Turbidity	876	0.57	0.56	-0.004	0.66	0.29	1.95

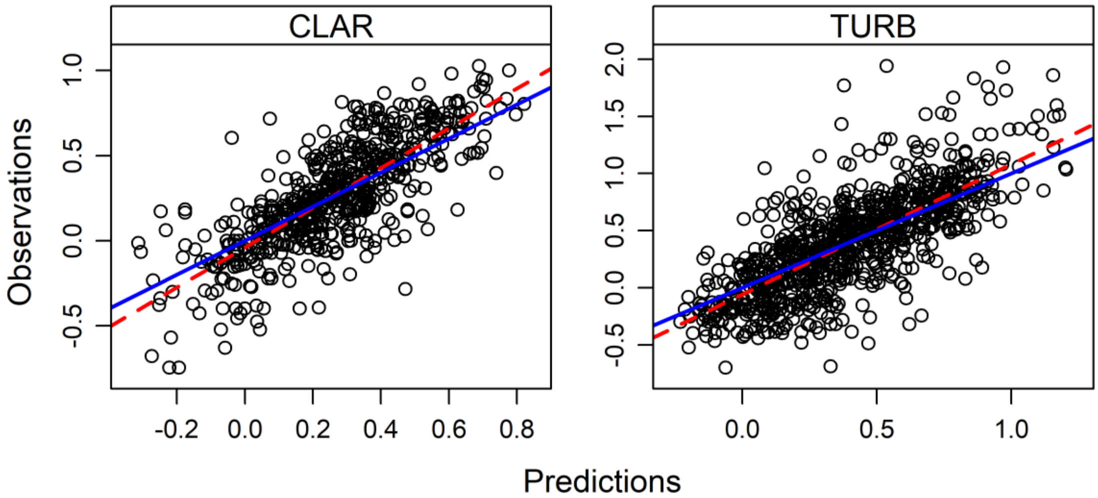

Figure 3-2: Comparison of observed visual clarity (CLAR) and turbidity (TURB) values versus values predicted by the random forest models. Red dashed line is the best fit linear regression of the observed and predicted values. Blue solid line is the one-to-one line. Data are the $\log_{10}$ values.

The predictor variables with high importance in the clarity and turbidity RF models reflected strong associations with land cover and catchment topography, with climate and flow also important. As expected theoretically (since clarity decreases as turbidity increases), important predictors typically showed opposite relationships for clarity and turbidity (e.g., clarity decreased with increasing bare ground in the upstream catchment while turbidity increased; Figure 3-3). Sediment concentration was the 2nd most important predictor of clarity and turbidity (Table 3-6), with clarity decreasing and turbidity increasing with increasing sediment concentration. This is reassuring.

Three variables reflecting the proportion of different land cover types in the upstream catchment were amongst the top twelve most important predictor variables across both models (Table 3-6). The partial plots (Figure 3-3) indicate that clarity decreases and turbidity increases with increasing usIntensiveAg and usBare (ranked 3rd and 4th, respectively). In comparison, clarity increased with increasing usNativeForest (the proportion of late-successional native forest, ranked 12th), while turbidity decreased. These patterns are consistent with previous correlations between land cover and water quality state (Larned et al. 2004, 2016; Whitehead 2019).

Figure 3-4 highlights model performance in relation to the dominant land cover upstream of the training sites, and it also displays how the typical ranges of median turbidity and clarity vary with land cover. The predicted vs observed data points mostly show no land cover related bias. For example, with urban land cover there is similar scatter of turbidity data points above and below the (blue) 1:1 line, and the predicted vs observed trend-line for just the urban points (broken black line) aligns closely with the trend-line for all data points (red line). However, the tussock sites show model under-prediction at high values of both turbidity and clarity (indicated by the tussock data trend-line pivoting counter-clockwise from the 1:1 line), while for the bare ground sites the models tend to consistently over-predict clarity and under-predict turbidity (indicated by the bare ground trend-lines for clarity and turbidity lying, respectively, below and above the 1:1 lines).

Figure 3-4 also shows broad ranges in observed median clarity and turbidity and in the scatter about the 1:1 lines across all land covers (except for wetlands, where there was only one data point). For example, with the urban sites there is a factor-of-80[6] range of observed median turbidity, while the residuals of the observed vs predicted urban data range up to a factor-of-5.

Predictors describing catchment rainfall (usRainvar, usRainDays20) were the 1st and 11th most important predictors for both the clarity and turbidity models (Table 3-6), with clarity increasing and turbidity decreasing as variation and intensity of rainfall increased (Figure 3-3). In addition, reach-scale temperatures (segTwarm, segTmin) were the 4th and 7th most important variables.

Predictors describing catchment slope (usSlope) and upstream area (usArea) ranked 9th and 10th overall (Table 3-6). The partial plots indicated that clarity and MCI increased with increasing values of usSlope and usElev, while the values of all other water quality variables decreased (Figure 3-3).

The predictor usHard ranked 6thoverall (Table 3-6). Clarity increased and TURB decreased with increasing values of usHard (Figure 3-3), indicating the regolith of the catchment is influential on clarity and turbidity, probably through the availability of fine-grained clay minerals (which dominate the optical signature of suspended sediment).

Table 3-6: Rank order of importance of predictor variables retained in the random forest models for at least one water quality variable. Blank cells indicate that the predictor was not included in the reduced model. The predictor variables in the first column are listed in descending order of the median of the rank importance over the two models.

Predictor	Clarity	Turbidity
usRainvar	1	1
sedConc	2	2
usIntensiveAg	4	5
segTwarm	5	6
usBare	6	8
usHard	13	3
segTmin	15	4
meanFlow	10	-
usArea	-	10
usSlope	3	17
usRainDays20	7	14
usNativeForest	12	9
usElev	-	11

[6] The plotted data show a $\log_{10}$ range of 1.9, which transforms to a factorial range of $10^{1.9} = 80$.

Predictor	Clarity	Turbidity
usTmin	16	7
distToCoast	-	12
usPET	11	13
usRain	9	16
usTwarm	8	19
usPhos	-	15
usRainDays100	14	21
distToHead	-	18
usRainDays10	-	20
segElev	-	22
usPsize	-	23

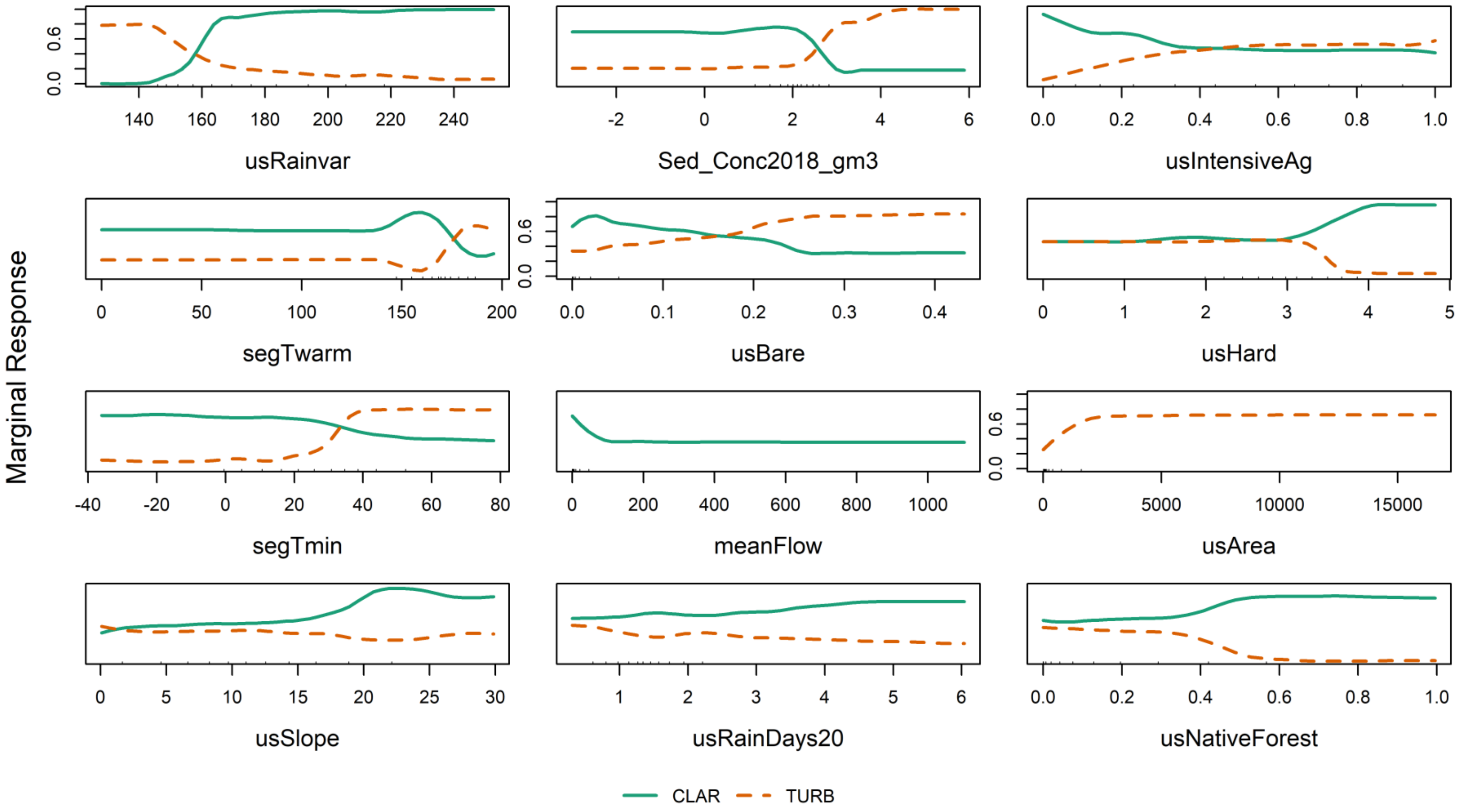

Figure 3-3: Partial plots for the twelve most important predictor variables in random forest models of visual clarity (CLAR) and turbidity (TURB). Each panel corresponds to one predictor, with predictor variables ordered by overall importance from most (top left) to least (bottom right) important. Y-axis scales represent the standardised value of the marginal response for each of the modelled response variables. In each case, the original marginal responses over all twelve predictors were standardised to have a range between zero and one. Plot amplitude (the range of the marginal response on the Y-axis) is directly related to a predictor variable's importance; amplitude is large for predictor variables with high importance.

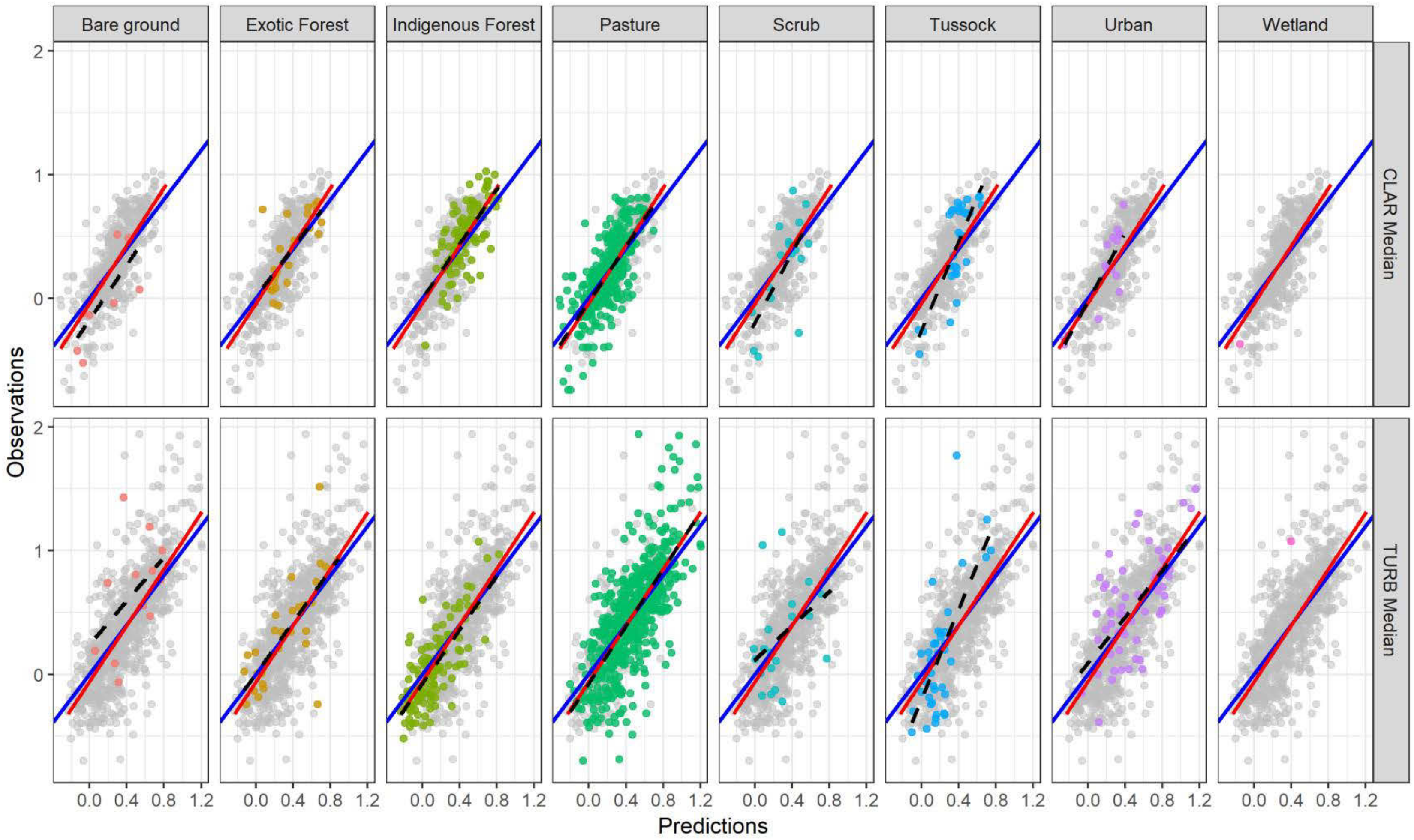

Figure 3-4: Observed visual clarity (CLAR) and turbidity (TURB) versus out-of-bag values predicted by the random forest models identified by dominant upstream land cover. Colours identify sites of specified land cover; grey points show the remainder of the model-training sites (as in Figure 3-2). Red line is the best-fit linear regression of all observed and predicted values; black broken line is best-fit to the land cover specific points; blue line is the one-to-one line. Data are $\log_{10}$ values.

Model predictions

The mapped predictions for clarity and turbidity have similar coarse-scale spatial patterns, with relatively poor values (i.e., low clarity and high turbidity) in low-elevation areas on the east coast of the North, in the Waikato, Auckland and Northland Regions and on the Southland Plains (Figure 3-5, Figure 3-6). In contrast, predicted clarity is relatively high and turbidity generally low in mountainous areas, the Department of Conservation estate, and other areas dominated by native forest land cover. However, high turbidity was predicted across the Southern Alps, which is likely due to summer snow/glacial melt.

Note that the maps in Figure 3-5 and Figure 3-6 consist of nzsegments of Order 3 and above, and some extensive lowland areas are dominated by low order streams (e.g., eastern Auckland, Tauranga). Steep coastal areas of the Marlborough Sounds, Fiordland, Coromandel and Banks Peninsulas and offshore islands are also dominated by low order streams. The predicted clarity and turbidity in low order streams in these areas is not shown on the maps in Figure 3-5 and Figure 3-6.

Plots of predicted vs observed values showed that both RF models tended to slightly under-predict at high values and over-predict at low values (compare red and blue lines on Figure 3-2). Therefore, for the final models used for the sediment load reduction assessments, we applied empirical correction functions that removed these high- and low-range biases. The correction functions were derived from linear regression fits between the log-transformed observed values and the log-transformed values predicted by the random forest models trained to the full datasets. These functions are:

$$CLAR_{corrected} = 0.875CLAR_{RF}^{1.191} \quad (10)$$

$$TURB_{corrected} = 0.794TURB_{RF}^{1.194} \quad (11)$$

We note that these corrections amount to “fine-tuning” that is well inside the model prediction errors (as indicated by the standard factorial errors of 1.58 and 1.95 for the clarity and turbidity models, respectively, listed in Table 3-5).

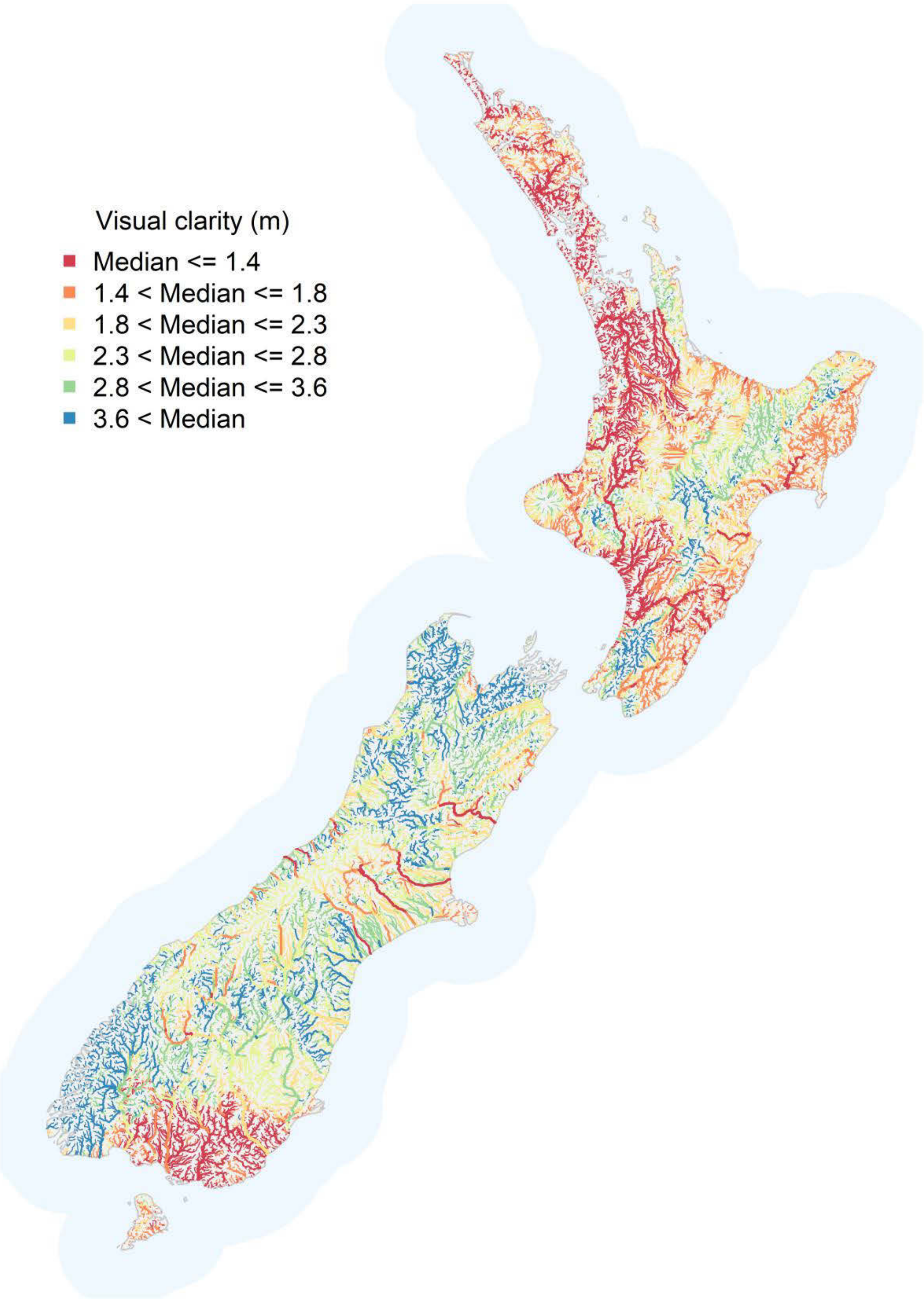

Figure 3-5: Predicted median visual clarity in New Zealand rivers. Map shows all nzsegments of Order 3 and higher. Smaller rivers are omitted to make river networks distinguishable.

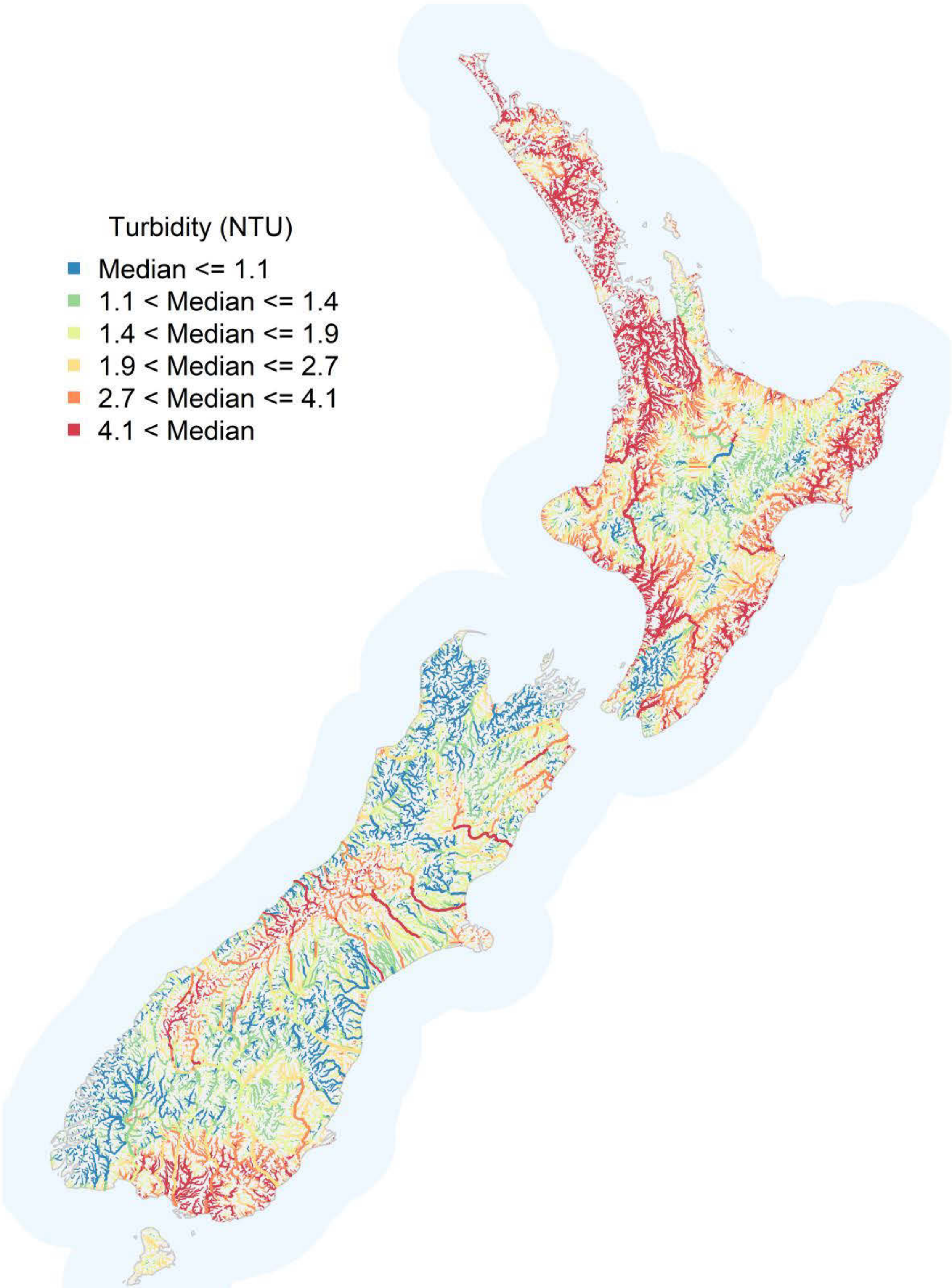

Figure 3-6: Predicted median turbidity in New Zealand rivers. Map shows all nzsegments of Order 3 and higher. Smaller rivers are omitted to make river networks distinguishable.

3.3.4 Discussion

Comparison with previous studies

The median visual clarity and turbidity models developed in this study are similar to those reported by Whitehead (2019) but include sediment concentration as an additional predictor variable. Our results are generally consistent with those of Whitehead (2019), with similar model structures (as indicated by the relative importance of predictor variables and directions of partial plots) and model performance. However, we observed some spatial differences in the predicted values of clarity and turbidity between the two studies. We predicted lower clarity along the Southern Alps and parts of East Cape but higher clarity on the Canterbury Plains, across North Otago, and in the Marlborough/Tasman Districts. The current study predicted higher turbidity on the Southland Plains, Rangitikei-Manawatu coastal plain, Waikato, Auckland and Northland, but lower turbidity in Fiordland, North Otago, Kahurangi and the Hutt Valley. To a fair degree, the clarity pattern is the inverse of the turbidity pattern – which is to be expected given that visual clarity and turbidity are inversely related. Also, to a reasonable degree – but not everywhere – the difference in patterns align with spatial variation in specific sediment yields (compare Figure 3-7 and Figure 3-8), which would stem from inclusion of the mean sediment concentration variable in our new models.

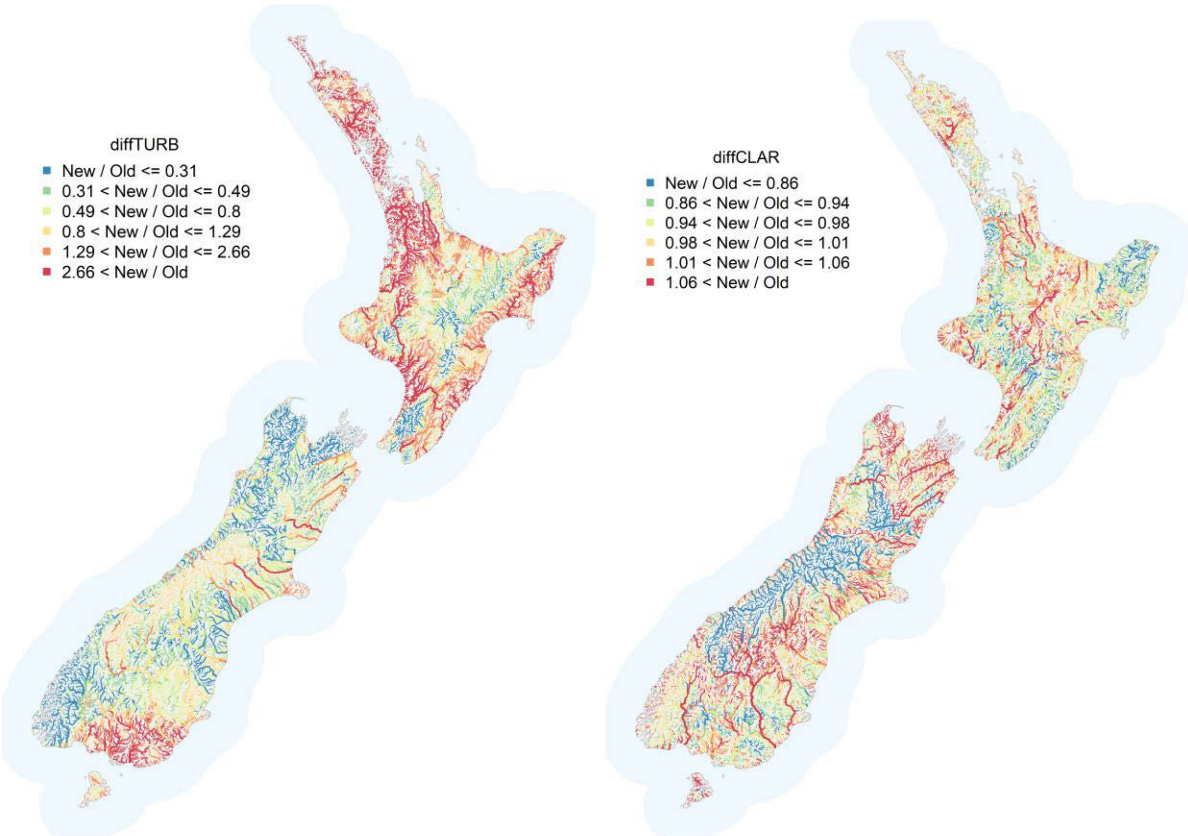

Figure 3-7: Maps of the ratios of median turbidity and clarity predicted by the current (New) and previous (Old) RF models.

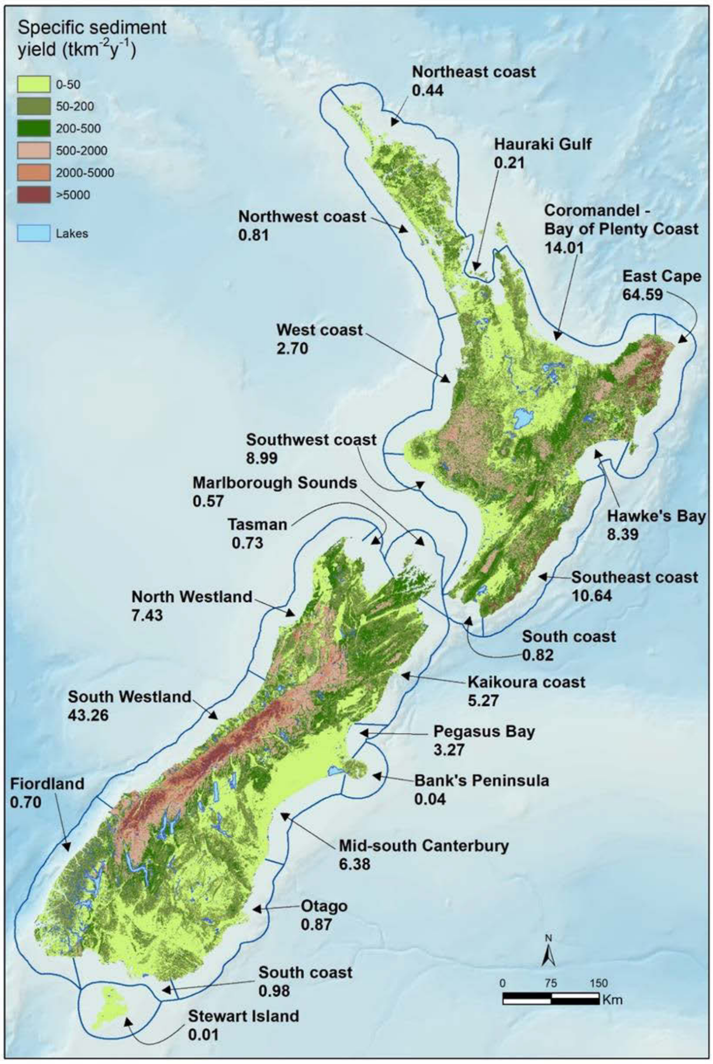

Figure 3-8: Specific suspended sediment yield (t/km^2/yr) predicted by model developed by Hicks et al. (2019) for contemporary land cover. Bold numbers around coast show mean annual sediment load (Mt/yr) discharged to regional spans of the coast. From Hicks et al. (2019).

Model performance check in urban stream segments

MfE have expressed particular interest in model reliability in urban stream segments. As shown in Figure 3-4, our predictive models showed no bias with respect to predicting median clarity and turbidity at the training sites with dominantly urban catchments. A semi-independent check on this was possible by comparing our model predictions with median clarity and turbidity data compiled from urban streams by Gadd et al. (2019). While sourced largely from the same original datasets as used herein, the Gadd et al. dataset used 3-year median values covering different time-spans compared with the 5-year medians used herein, and it contained several more sites. Thus, while it is not completely independent of our training dataset it contains different numbers. Figure 3-9, comparing predicted vs observed values for the Gadd et al. dataset, shows good concordance for the turbidity data but a tendency to under-predict clarity (although not beyond the general scatter of the data). We conclude that there is no indication that our median turbidity model produces biased predictions in stream segments with urban dominated catchments.

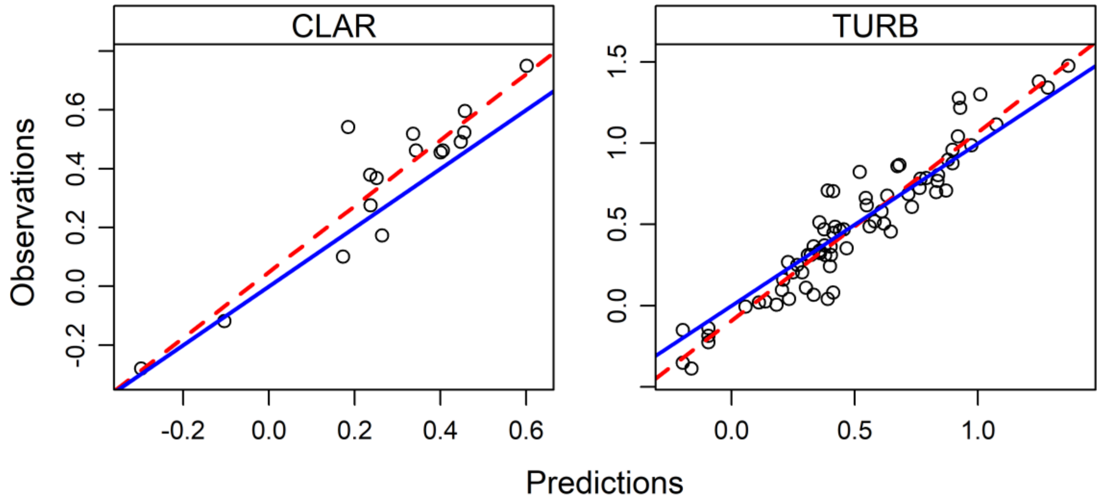

Figure 3-9: Model predicted median clarity and turbidity vs observed values in urban streams from the dataset of Gadd et al. (2019). Blue line is 1:1 line; red broken line is best-fit linear regression line. Log_{10} values of the data are plotted.

The Gadd et al. (2019) data also confirm the wide spread (×80 range) in median turbidity observed in streams with urban catchments.

Model uncertainty

The 95% confidence intervals for median turbidity and visual clarity values predicted by our models for individual stream/river segments can be obtained using

$$95\%\ CI = 10^{[\log_{10}(x)\ \pm\ 1.96 \times RMSD]} \tag{12}$$

where x is the estimated value in the original units, RMSD is the corrected-model error in $\log_{10}$ space (from Table 3-5), and 1.96 is the standard normal deviate or Z-score for probability (0.025 ≤ Z ≥ 0.975).

Equation (12) is appropriate for calculating confidence intervals for variables that were log transformed prior to model fitting, where the prediction uncertainty (RMSD) values have been reported in log space, and where the predicted vs observed values in log space are normally distributed. The prediction confidence intervals for the $\log_{10}$-transformed variables, when expressed in the original units of the variables, are asymmetric and their values vary in proportion to the predicted value. For example, if we let x be a predicted value for clarity of 1.0 m, using RMSD = 0.20 (from Table 3-5), then the lower and upper 95% confidence intervals are 0.4 m and 2.4 m, respectively.

Alternative modelling methods

The RF method that we used here to develop the median visual clarity and turbidity models is well-suited to data from monitoring sites that represent a wide range of environmental conditions. However, it is not the only method available. Alternative statistical models include generalised additive models (GAMs; Hastie et al. 2001), artificial neural networks (e.g., Joy and Death 2001), boosted regression trees (e.g., Leathwick et al. 2005), and optimisation (e.g., Hicks et al 2019). In addition, models that incorporate biophysical processes (e.g., CLUES; Alexander et al. 2002) are available. Refer to Whitehead (2019) for a detailed discussion on why RF models may be considered an appropriate tool for predicting river water quality state variables including turbidity and visual clarity.

3.4 Suspended load reduction required to meet turbidity and visual clarity bottom lines

3.4.1 Overview

This section details:

- the assessment of which river segments around New Zealand have median turbidity and median clarity exceeding the C/D band thresholds presented in Section 2
- the estimated reduction in mean annual up-catchment suspended sediment load required to achieve these thresholds - at the segment and catchment scale
- the uncertainties associated with the estimated reductions in sediment load, and
- a list of outputs passed on to Manaaki Whenua Landcare Research (MWLR).

3.4.2 Which river segments exceed the C/D band thresholds for turbidity and clarity?

Threshold-exceeding river segments around the country were identified by comparing the provided threshold median values (Table 2-1) with the predicted median clarity and turbidity values (from the national models developed in Section 3.3).

Over all 593,551 segments:

- 18.4% exceed the clarity threshold
- 16.4% exceed the turbidity threshold
- 10.3% exceeded both thresholds, and
- 24.5% exceeded one or other or both thresholds.

These exceedances are shown broken-down by Climate-Topography-Geology class in Figure 3-10. Classes 1, 9, and 12 dominate the segment totals, with Classes 1 and 9 showing good concordance between turbidity and clarity (i.e., a large proportion of segments have both turbidity and clarity thresholds exceeded – as indicated by the relative heights of the striped bars) but Class 12 does not. Of the minor classes, 5, 7, 10, and 11 show reasonable concordance but 3 and 8 are dominated by turbidity while 2 and 4 are dominated by clarity. The proportion of all segments exceeding either threshold varies widely by CTG class: ranging from <1% for Class 8 to 66% for Class 2. Classes 1, 2, 3, 9, 10, 11, and 12 have more segments exceeding either threshold than the overall average of 24.5%, while Classes 4, 5, 6, 7, and 8 are below this average.

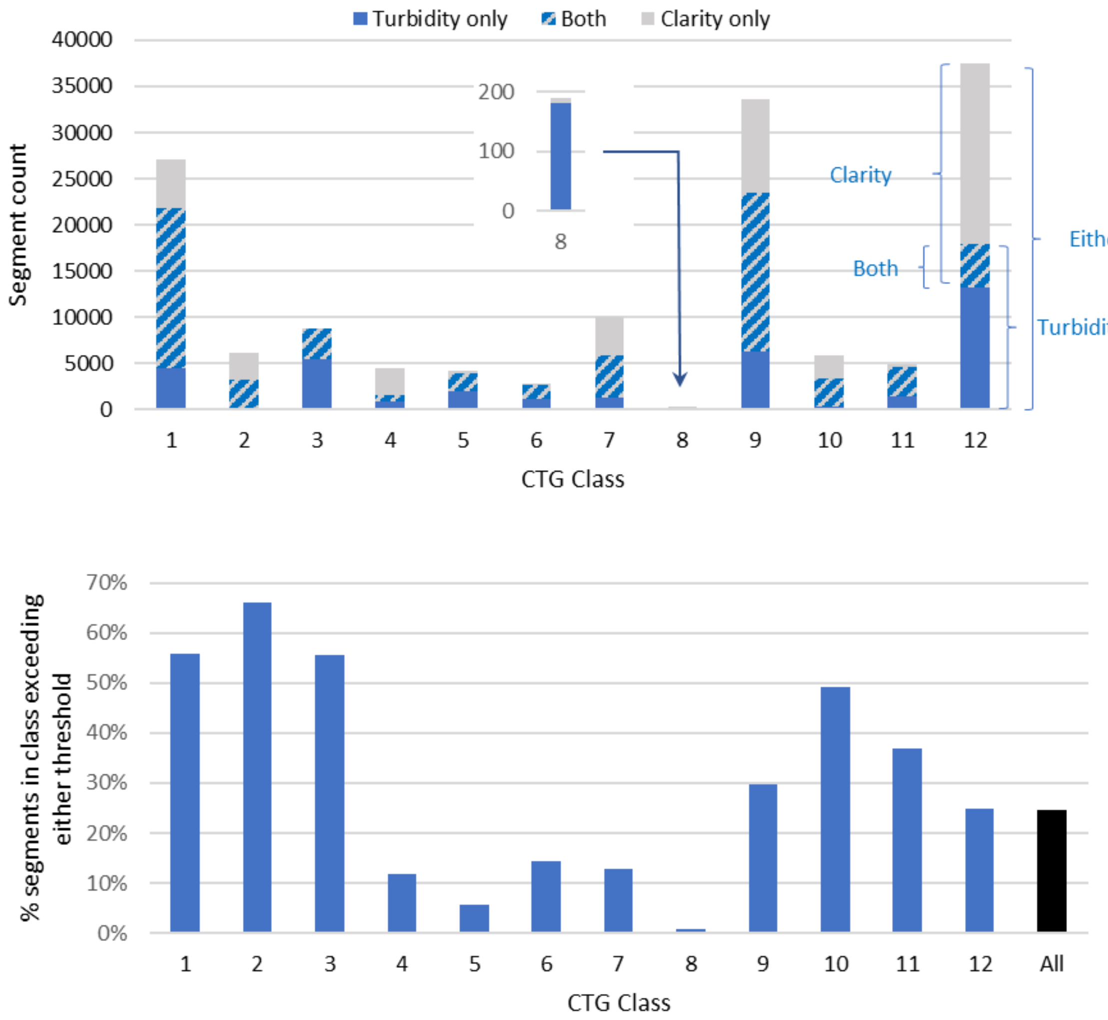

Figure 3-10: Count and percentages of stream segments exceeding clarity and turbidity thresholds by CTG class. On top plot, the total bar height gives the number of segments exceeding either or both thresholds; the striped bar indicates where both are exceeded. Class 8 has only 190 exceedances for either turbidity or clarity and is shown with an expanded scale. On bottom plot, the bar height gives the proportion of all segments in the class that exceed either of the turbidity or clarity thresholds.

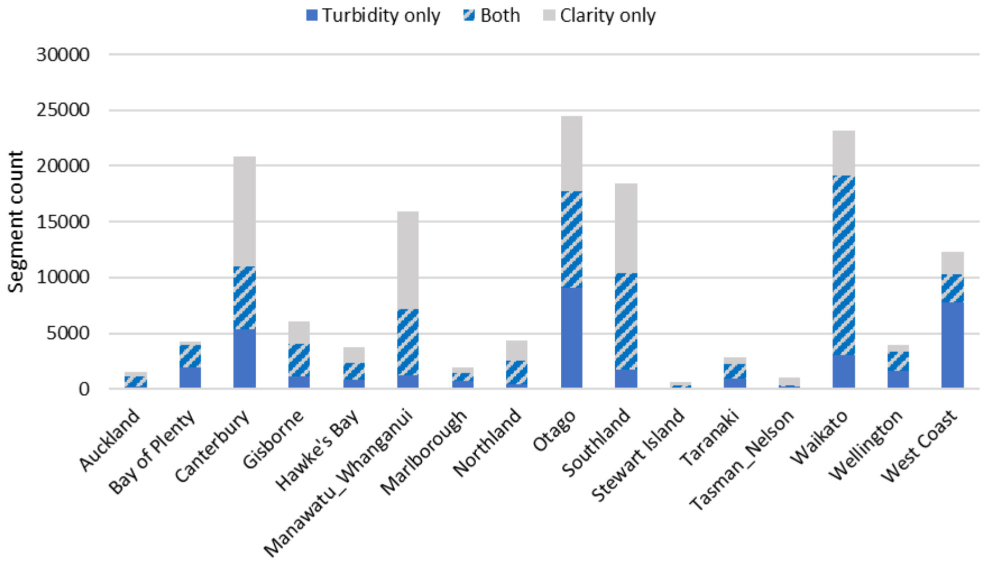

Figure 3-11: Count of stream segments exceeding clarity and turbidity thresholds by region. Total bar height gives the number of segments exceeding either or both thresholds; the striped bar indicates where both are exceeded.

Figure 3-11 shows the equivalent breakdown by region. Canterbury, Manawatu-Whanganui, Southland, Otago, Waikato, and the West Coast show the greatest numbers of exceedances by segment count. Canterbury, Otago, and Westland (and to a lesser extent Southland) all contain rivers with Glacial-Mountain sources-of-flow[7] and which are likely to have relatively higher baseflow turbidity (and lower clarity) due to the presence of fine "glacial flour". The concordance between turbidity and clarity exceedance is reasonable in most regions, except for Canterbury and Tasman-Nelson (where clarity prevails) and the West Coast (where turbidity prevails).

Figure 3-12 shows the proportional breakdown by stream order. The proportions of segments exceeding either threshold are similar across all stream orders, while the concordance increases as stream order increases (e.g., ~ 40% of order 1 segments exceeding any threshold also exceed both thresholds, while 80% of order 8 segments exceeding any threshold also exceed both thresholds).

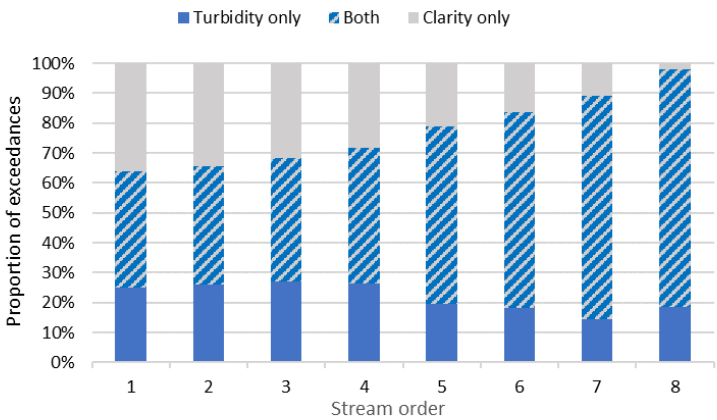

Figure 3-12: Proportion of stream segments exceeding clarity and turbidity thresholds by stream order. The striped bar indicates where both are exceeded.

[7] Based on the source-of-flow REC classification developed by Snelder et al. (2005).

Figure 3-13 shows the threshold exceedance breakdowns by land cover. The top plot shows that the overwhelming majority of threshold exceedances (blue/grey bars) occur in segments with pasture dominated catchments, which largely reflects the prevalence of pasture land cover across the country (yellow bar). The lower plot shows that indigenous forest (notably), exotic forest, scrub, and tussock dominated catchments all have lower threshold exceedance percentages by class compared to the national average of 24.5% (i.e., exceeding either threshold). In contrast, bare ground, miscellaneous/mangroves, urban, and wetland dominated catchments have proportional exceedances higher than the national average. Turbidity threshold exceedances prevail in bare ground, urban, and wetland catchments, while clarity thresholds prevail under the other land covers.

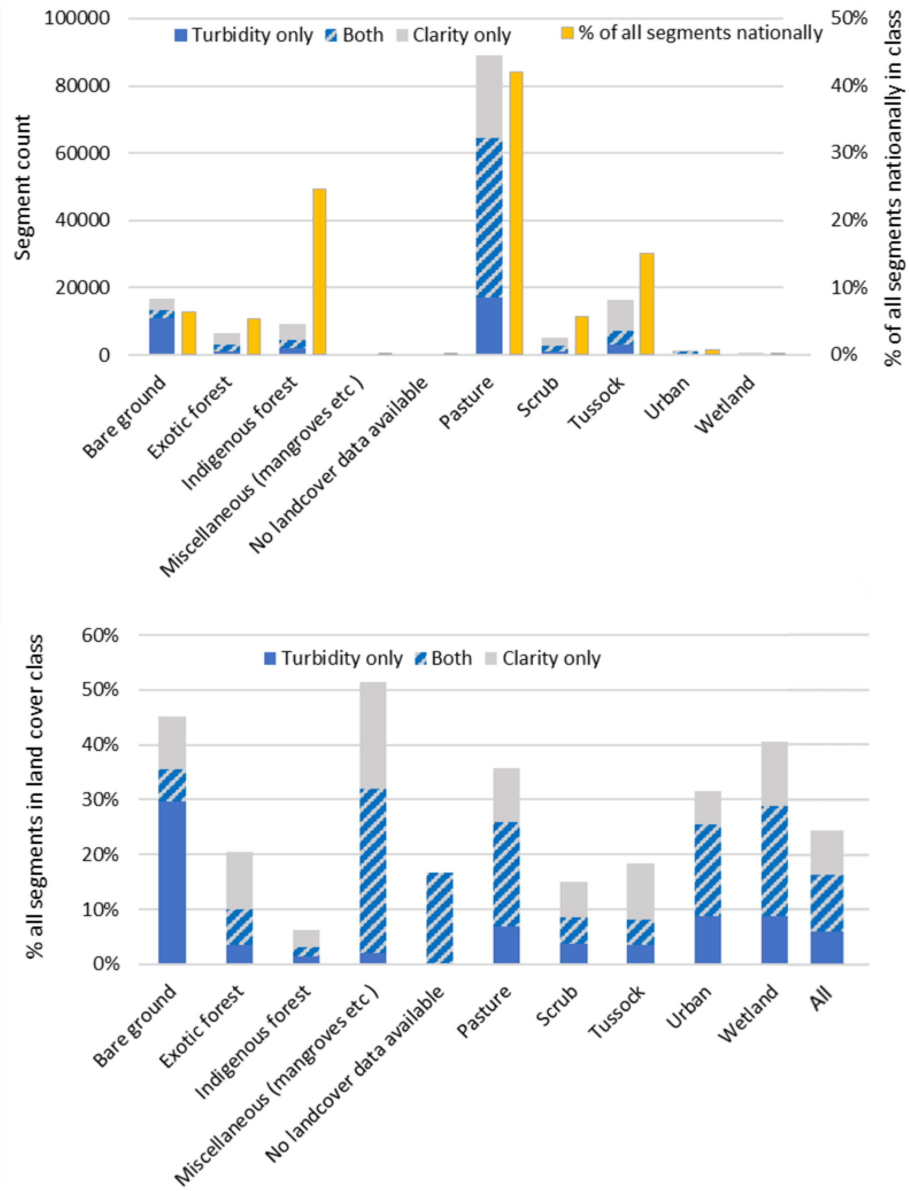

Figure 3-13: Count and proportion of stream segments exceeding clarity and turbidity thresholds by land cover. On top plot: the total blue/grey bar height gives the number of segments exceeding either or both thresholds; the striped bars indicate where both are exceeded; the yellow bars show the national proportions by land cover of all segments. On bottom plot: the blue/grey bar heights give the proportions of all segments in the land cover class that exceed the turbidity or clarity thresholds.

3.4.3 What load reduction is required to achieve the C/D band thresholds for turbidity and clarity?

Equations (7) and (8) from Section 3.1 were used to estimate the load reduction factors for turbidity and visual clarity (R_v and R_t) across all 593551 segments of the REC2 digital network using:

- the provided threshold median values (T_{t50}, V_{t50} from Table 2-1)
- the predicted median clarity and turbidity values (from the national models developed in Section 3.3), and
- the *d* and *f* exponents in the visual clarity vs SSC and turbidity vs SSC relationships developed in Section 3.2 (i.e., -0.76 and 0.98, respectively).

We note that, as defined in Section 3.1, $R = (L - L_t)/L$, where L is the actual sediment load and L_t is the target sediment load that just meets the threshold.

In preliminary assessments, the maximum of the clarity- and turbidity-derived load reduction factors (designated R_{max}) was assigned to each segment. The logic in taking the maximum value was to provide a conservative, worst-case, result. However, following concerns from MfE around the reliability of some of the clarity-based thresholds (which were supported in part by our findings in Section 2), we only used the turbidity-based R_t values in developing the final outputs.

The calculated R_t-values were grouped into 5 classes:

- $0<R_t<0.2$
- $0.2<R_t<0.4$
- $0.4<R_t<0.6$
- $0.6<R_t<0.8$, and
- $0.8<R_t<1$.

Figure 3-14 (left plot) shows the overall, national distribution of R_t values. Note that we have included for completeness negative R_t values (i.e., at segments not requiring load reduction to meet the turbidity threshold. The overall distribution is log-normal, with most segments having R<0 (signalling no need for load reduction) and only 16.4% of segments exceeding the threshold, having R_t >0 and requiring load reduction. Of the latter segments, the right plot shows that 70% have R_t values less than 0.4.

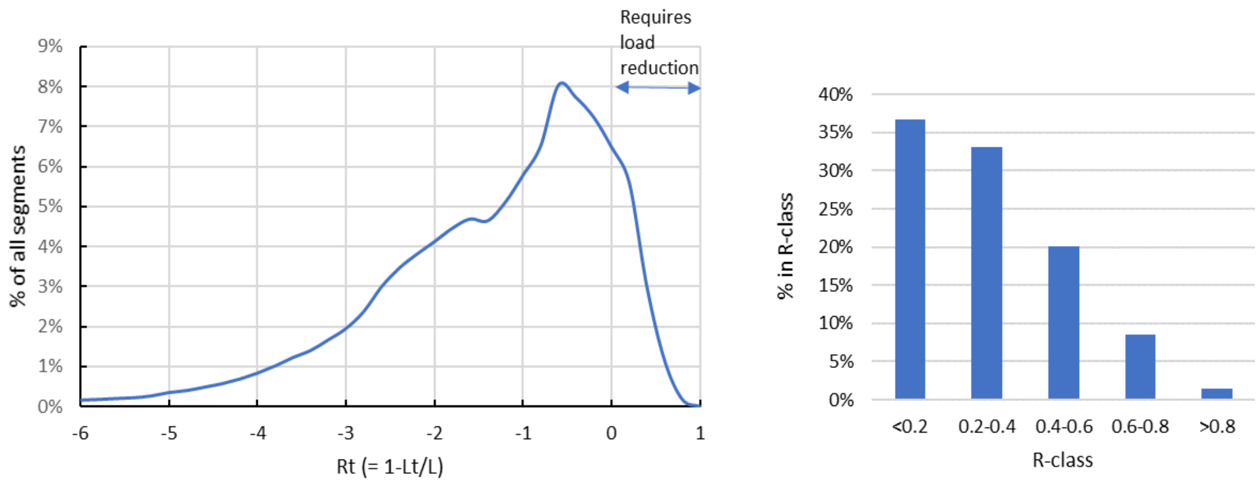

Figure 3-14: Frequency distributions of load reduction factor (R_t) over all segments. Only R_t values > 0 require load reduction to meet turbidity threshold. Plot on right shows proportion of segments for which R_t>0 in each R-class band.

Figure 3-15 shows the distributions of the R_t classes by stream order, CTG class, region, and dominant land cover:

- By stream order: lower order segments (order 1-4) are dominated by relatively low R-values (< 0.4); order 5-7 segments have increasing proportions of higher R-values (R>0.4); while order 8 segments have mainly low R-values but also have the highest proportion of very high R-values (R>0.8).
- By CTG class: low R-values (< 0.4) dominate all CTG classes (particularly 4, 7, 9, and 12) except 1 and 3.
- By region: low R-values (< 0.4) dominate all regions (particularly 4, 7, 9, and 12) except for Auckland and Waikato.
- By land cover: low R-values (<0.4) dominate all land covers except for urban and pasture. Urban catchments have the highest proportion of high R-values (45% higher than 0.4).

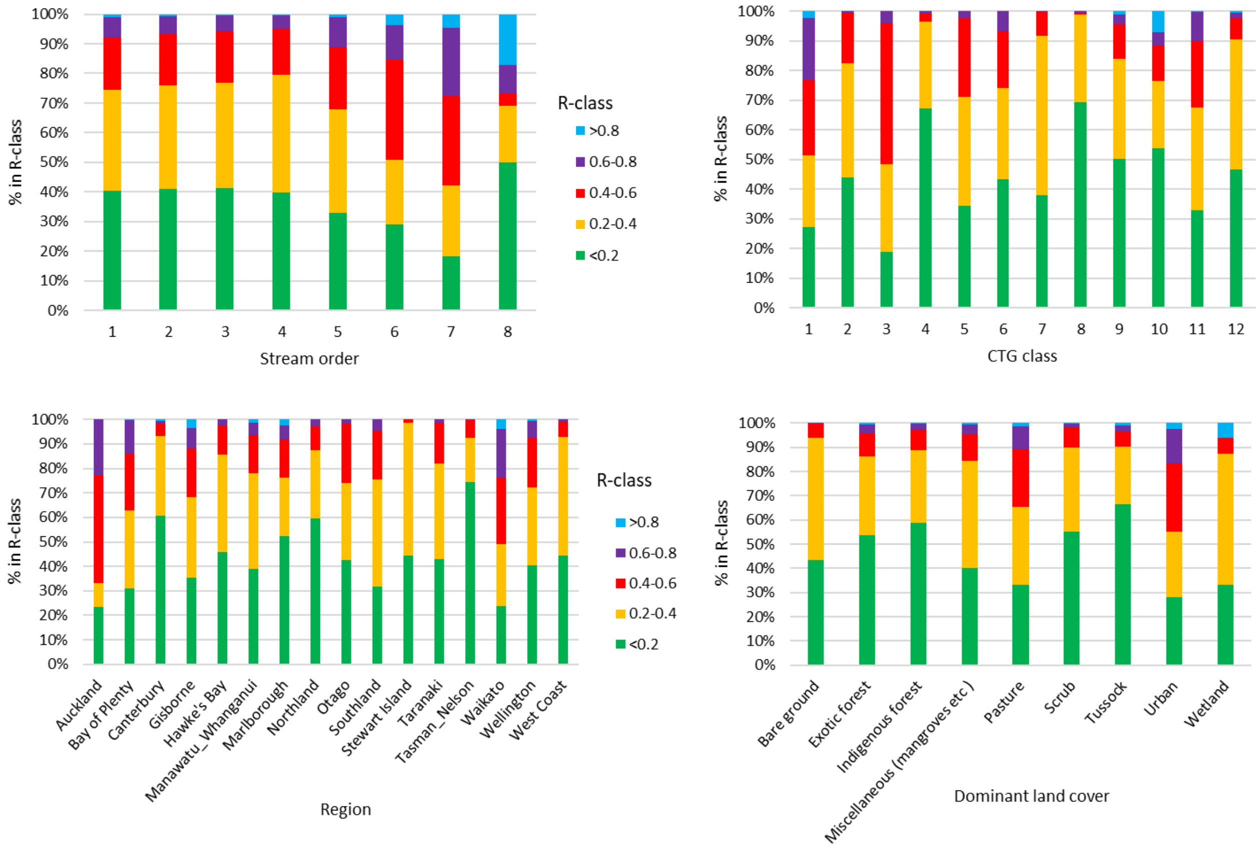

Figure 3-15: R_t-class breakdown by stream order, Climate-Topography-Geology (CTG) class, region, and dominant land cover. R_t-class relates to turbidity threshold. Note that the absolute count of segments with R_t>0 is shown by the combined height of the solid-blue and blue-striped bars in Figures 3-10 to 3-13.

3.4.4 Why the relatively high count and size of R-values in urban catchments?

As shown in Section 3.4.2, stream segments with catchments under dominantly urban land cover exceed the turbidity C/D band threshold more often than the national average (25% compared with 16.4% for all land covers, Figure 3-13). Moreover, as shown in Section 3.4.3, the R-values tend to be higher in urban catchments compared to those in other land covers (45% have R-values > 0.4, Figure 3-15).

In Section 3.3.4, we found no evidence that our median turbidity model produces biased predictions in stream segments with urban dominated catchments. So, why should urban catchments have these relatively high turbidity-based threshold-exceedance percentages and R-values, even though many typically have limited sediment sources? We suggest several reasons:

- As noted by Hicks et al. (2019), urban catchments can have variable sediment yields. High yields can occur from earthworks associated with urban development and roading if erosion control and storm water treatment is inefficient, and these can produce elevated turbidity. Conversely, low yields occur in mature, impervious urban catchments, and these typically produce low turbidity values, particularly on recessions due to sediment exhaustion. This high range is confirmed by the factor-of-80 range in observed median turbidity values for urban sites in the dataset used to train our turbidity predictive-model (Section 3.3.3).

- Many urban streams are low gradient and tidal: low gradient streams tend to remain turbid for longer after runoff events (because of low velocities through pools), while flow-reversing tidal streams exchanging water with mud-rich estuaries are kept turbid on a daily basis by wave and current resuspension over tidal flats.
- 38% of urban stream segments were assigned relatively low C/D Band thresholds (< 4 NTU) – which will combine with a high range of observed turbidity values to produce a relatively high risk of threshold exceedance.

3.4.5 Comparison with R results using observed data

Figure 3-16 compares, for the 847 segments with observed values of median turbidity (that were used to train the national predictive model), the R-class distribution where R_t has been calculated using the observed median turbidity and the R-class distribution using the predicted median turbidity. Of these 847 segments, 265 (31.2%) exceeded the turbidity threshold (i.e., $R_t > 0$), while in the matching predicted dataset 258 (30.4%) exceeded this threshold – a close agreement. The matching predicted and observed distributions were also very similar across the R-classes, which provides reassurance around the model predictions. Compared with the observed-segment dataset, the predicted dataset for the whole country had relatively more segments with low R-values and relatively fewer segments with high R-values (Figure 3-16).

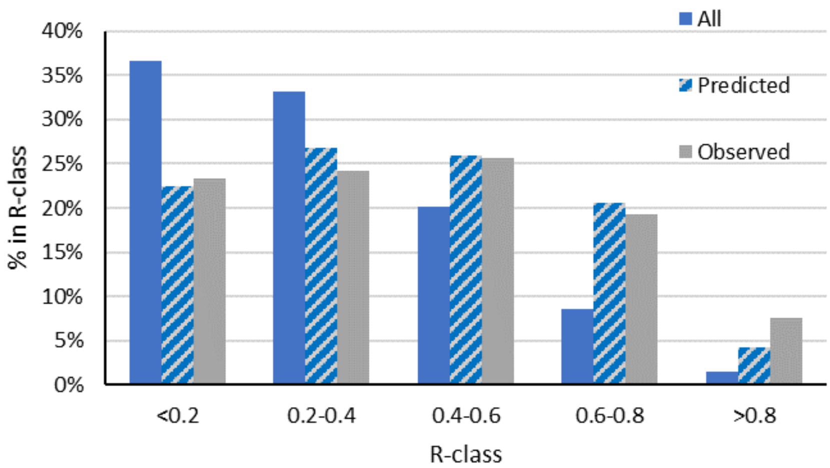

Figure 3-16: Distribution of segments with R_t> 0 by R_t class for all predicted segments, segments with R calculated off observed turbidity data, and matching segments with R calculated off predicted turbidity.

New Zealand's water-quality sites are typically over-represented in environments characterised by low catchment elevations and low catchment slopes and under-represented in catchments with high proportions of native forest land cover and low proportions of intensive agricultural land cover (Whitehead, 2019). Similar patterns of over- and under-representation were observed in this study, with monitored sites for clarity and turbidity over-represented in pastural and urban land covers and under-represented in natural land covers (e.g., indigenous forest, tussock, scrub, Figure 3-17). Similarly, the monitored sites were over-represented in CTG classes 1 and 7, while they were under-represented in Class 12. There were some small differences between the representativeness of the observed clarity and turbidity sites, with clarity under-represented in CTG class 5 while turbidity is slightly over-represented. These differences arise because the two variables are not always monitored at the same sites (Figure 3-1).

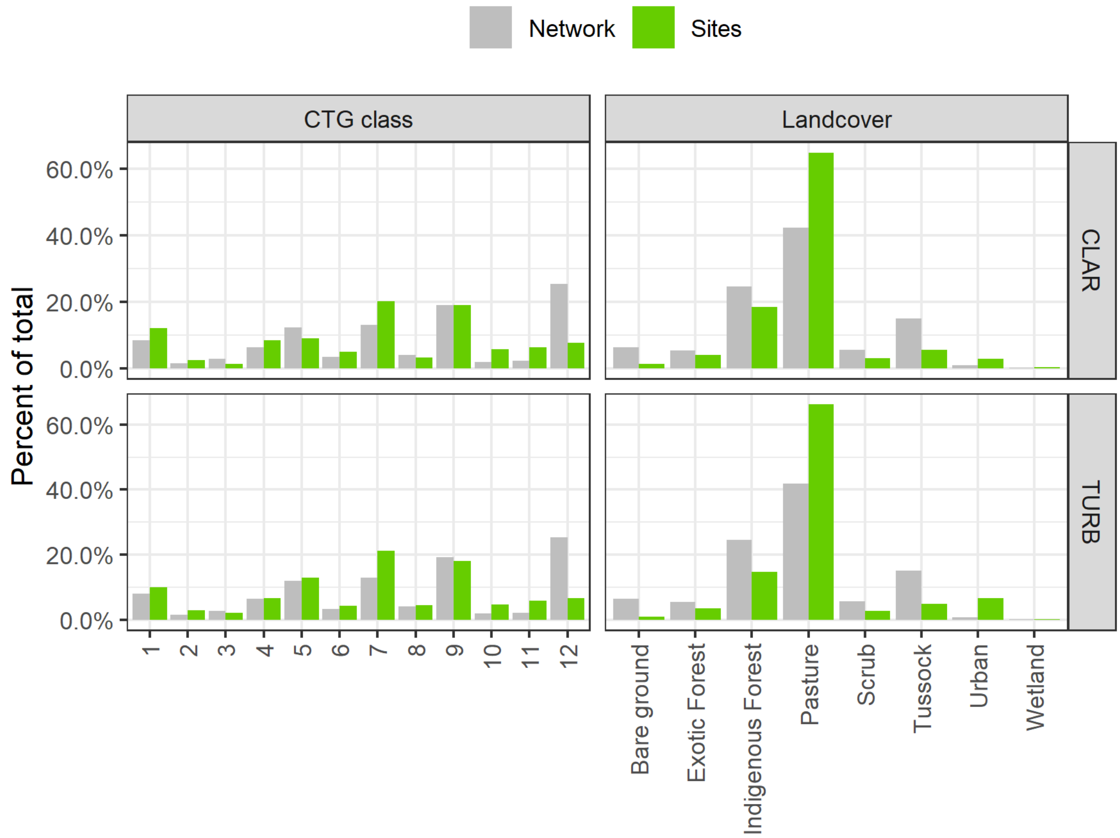

Figure 3-17: The distributions of CTG and land cover class across all segments in the digital river network and at monitored sites. Similarities in the distributions shown in the two histograms in each panel provide an indication of the degree to which environmental variation across the monitoring sites represents environmental variation across the New Zealand river network; complete representativeness would be indicated by exact matches between the histograms.

3.4.6 Uncertainty analysis

We investigated two types of uncertainty:

- uncertainty around whether the median turbidity in a segment exceeds the threshold turbidity, due to uncertainty in the predicted median turbidity, and
- the uncertainty in the calculated values of the load reduction factor, R.

Uncertainty around threshold exceedance

An important question is whether the segments tagged as requiring load reduction (i.e., with R > 0) have been correctly tagged. That is, we are interested in the probability p_{fp} that the true segment R is ≤ 0 when our calculated R is > 0 (the “false positive” case), and the probability p_{fn} that the true segment R is > 0 when our calculated R is ≤ 0 (the “false negative” case) – as illustrated on Figure 3-18.

We addressed this by first noting (from Equation 8) that $R = (1 - T_{t50}/T_{50})^{1/f}$ and that f (=0.98) is approximately equal to 1. Thus, R > 0 equates to $T_{50}/T_{t50} > 1$, and so $\log(T_{t50}/T_{50}) > 0$[8]. By assuming that the threshold turbidity has zero uncertainty (because it is a defined parameter), the uncertainty in $\log(T_{t50}/T_{50})$ stems only from the uncertainty in $\log T_{50}$. The error distribution in the predicted $\log T_{50}$ at any segment is defined by the OOB-RMSE (0.29) derived for the turbidity prediction model (Section 3.3.3) and may be assumed to be normally distributed (because the log-space residuals of this model were normally distributed). On this basis we calculated p-values associated with the calculated range of T_{t50}/T_{50} values (and associated R values) using the t-statistic $t = (\log(T_{t50}/T_{50}) - 0)/0.29$ with 875 degrees of freedom (where 875 is the number of sites used to develop the turbidity prediction model minus 1).

The results (Figure 3-19) show that the case of a false positive R remains significant ($p_{fp} > 0.1$) until the calculated R exceeds 0.56 (and $T_{t50}/T_{50} < 0.44$). Similarly, the chance of a false negative remains significant ($p_{fn} > 0.1$) until $T_{t50}/T_{50} > 2.3$. Table 3-7 shows the false positive probability ranges associated with the five R-band classes.

Figure 3-20 provides a means to gauge the overall importance of false inferences. It shows the proportion of segments associated with a given confidence level that a false inference has not been made. For example: we can be at least 50% confident that 16.4% of all segments actually have R>0 (blue curve); we can be at least 90% confident that 10% of those segments with calculated R>0 actually have R>0 (orange curve).

Table 3-7: False-positive probability range by R-class. Example: if the calculated R for a segment exceeds 0.8, then there is a less than 0.8% chance of that segment actually having R ≤ 0.

R-class	False-positive probability range
0 - 0.2	0.5 – 0.37
0.2 – 0.4	0.37 – 0.22
0.4 – 0.6	0.22 – 0.08
0.6 – 0.8	0.08 – 0.008
0.8 - 1	< 0.008

[8] The log symbol designates a base 10 logarithm.

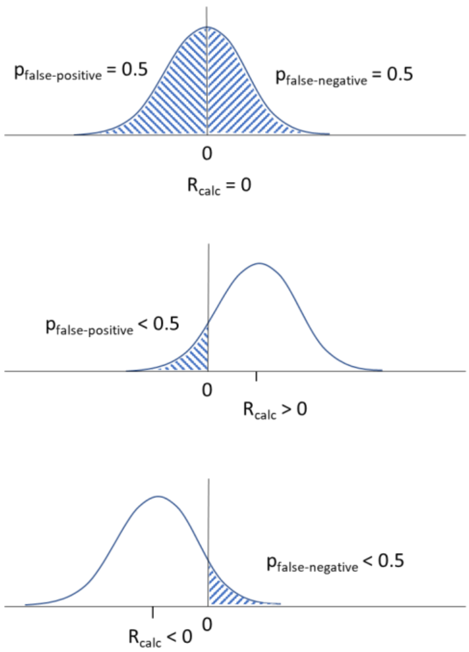

Figure 3-18: Sketch showing the probability of a false inference on the load reduction factor, R. When the calculated R = 0, there is a 0.5 probability that the true R value is either > 0 or < 0. When the calculated R is progressively > 0, there is a lessening probability that the true R value is < 0 (a false-positive). When the calculated R is progressively < 0, there is a lessening probability that the true R value is > 0 (a false-negative).

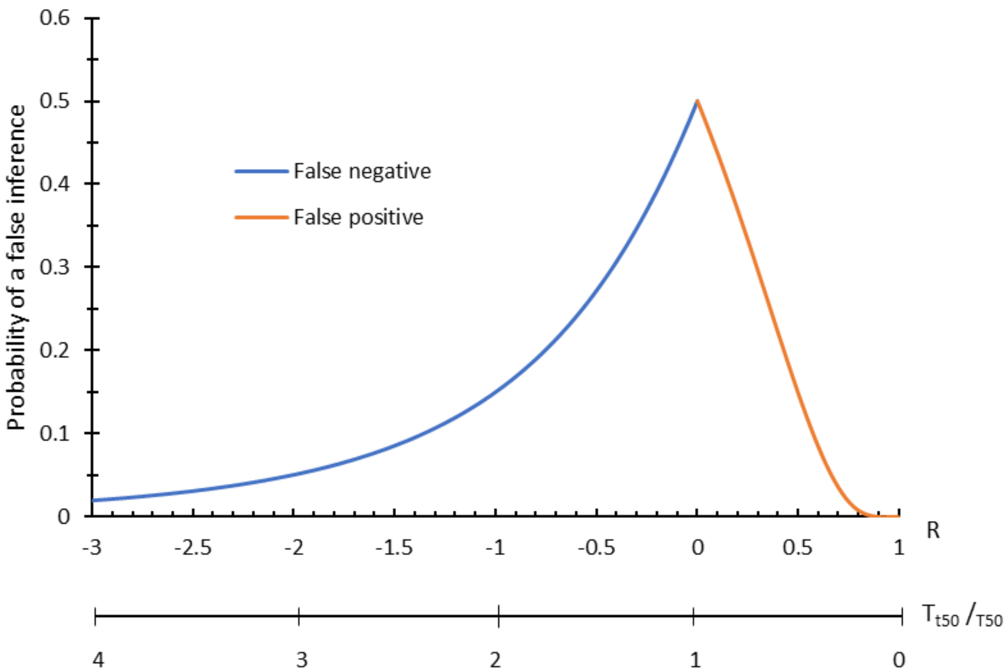

Figure 3-19: Probability of a false inference on whether R is greater than zero. Orange line gives probability that R at a segment may actually be less than 0 despite its calculated value being greater than zero (false positive case). Blue line gives probability that R at a segment may actually be greater than 0 despite its calculated value being less than zero (false negative case). Lower scale shows the equivalent ratio of threshold (T_{t50}) and actual median turbidity (T_{50}). Note R = 0 when $T_{t50} = T_{50}$.

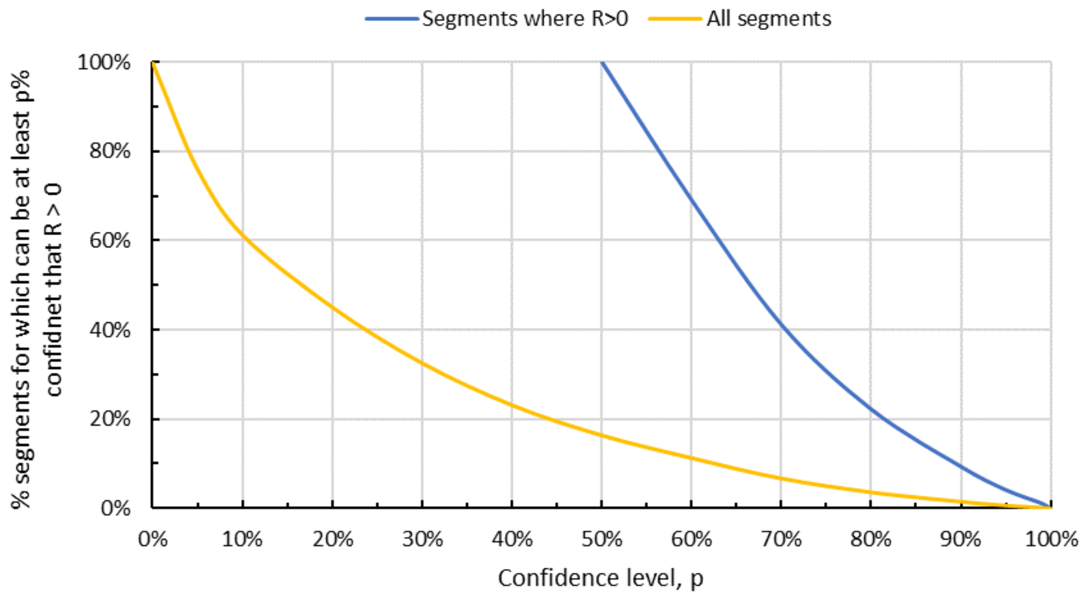

Figure 3-20: Proportion of segments for which calculated R > 0 at a given confidence level. Orange curve is for all REC2 segments; blue curve is only for the 97427 segments (16.4% of total) where calculated R > 0.Example: We can be at least 50% confident that 16.4% of all segments actually have R>0 and that 100% of all those segments with calculated R>0 actually have R>0.

The key conclusion is that – due mainly to the substantial uncertainty in estimating median turbidity – there is a significant risk that many of the segments calculated as requiring load reduction may not need it.

Uncertainty in calculated R-values

If it is accepted that segments where the median turbidity exceeds the threshold have been correctly identified, then the uncertainty in the calculated load reduction factor can be estimated by propagating the standard error in the component variables in Equation (8).

In general, if $R = 1 - (T_{t50}/T_{50})^{1/f}$ and T_{t50}, T_{50}, and f all carry uncertainty, then the error in R (ΔR) is

$$\Delta R = [(dR/dT_{t50} \cdot \Delta T_{t50})^2 + (dR/dT_{50} \cdot \Delta T_{50})^2 + (dR/df \cdot \Delta f)^2]^{0.5} \qquad (13)$$

where dR/dT_{t50}, dR/dT_{50}, and dR/df are the partial differentials of Equation (8) with respect to T_{t50}, T_{50}, and f, and ΔT_{t50}, ΔT_{50}, and Δf are the respective errors on T_{t50}, T_{50}, and f.

Equation (11) can be solved to provide

$$\Delta R/R = (1/f)\ [(\Delta T_{t50}/T_{t50})^2 + (\Delta T_{50}/T_{50})^2 + (1/f)^2 \ln(T_{t50}/T_{50}) \cdot (\Delta f/f)^2]^{0.5} \qquad (14)$$

where $\Delta T_{t50}/T_{t50}$, $\Delta T_{50}/T_{50}$, and $\Delta f/f$ are the proportional errors in the predicted turbidity, the turbidity threshold, and the coefficient f on the relation between turbidity and SSC.

Equation (14) was applied to all segments for which R > 0, using values of 0 for $\Delta T_{t50}/T_{t50}$ (i.e., the thresholds were assumed fixed and without error), 0.95 (= 1- 10^{RMSE} = 1 -$10^{0.29}$) for $\Delta T_{50}/T_{50}$, and 0.17 for $\Delta f/f$ (from Table 3-1). The resulting errors on R were large, equating to a factorial error of ×/÷

1.97, and were dominated by the large (1.95 factorial error) uncertainty in the predicted median turbidity values.

This error pertains to estimates of the load reduction factor at individual segments. When spatially averaging or totalling load reduction results and estimating the error in those averages or totals, a reasonable, conservative assumption might be that such errors are systematic[9] at the catchment scale but vary randomly between catchments. This is because the median turbidity values at stream segments are estimated off upstream catchment characteristics, and so, within a catchment, linked segments share common upstream characteristics – thus the turbidity estimate (and its error) at one segment will not be independent of the estimates at segments linked to it. Conversely, the median turbidity estimates among different catchments (even adjacent ones) will be independent, since there is no physical overlap. It follows that separate tributaries within the same larger catchment will also have independent errors.

Following this logic, we estimate that the relative error on the mean R-value over all 627 pour-point catchments (see below) would be ± 2.7%[10].

3.4.7 Converting proportional load reduction to absolute loads

The actual load reduction (L_r, t/yr) required for any stream segment is derived from

$$L_r = (L - L_t) = R*L \tag{15}$$

where L_t is the load at the threshold and L is the actual load delivered to the segment. L can be estimated from the suspended sediment load estimator developed by Hicks et al. (2019).

In brief, the load estimator includes a raster-type model that relates local sediment yield over a national grid to the local mean annual rainfall, slope, land cover, and erosion terrain. Catchment sediment yields are routed and accumulated downstream to the coast, adjusting for interception by lakes and reservoirs. The model was calibrated using an updated national dataset of measured river suspended sediment loads. The standard factorial error on predicted catchment loads was ×/÷ 1.9 (equating to a $\log_{10}$-RMSE of ±0.28). A second version of the model was produced that scaled the sediment yields across the catchments of the calibration sites so that the model-predicted load at the calibration sites matched the observed loads. This version of the model is called the "corrected" model and is recommended for use in this workstream (because it should be more accurate across catchments where the sediment load has been measured).

[9] "Systematic" means the error at one segment will be similar to the error of linked segments upstream and downstream.
[10] Derived by dividing the pour-point R-value error in $\log_{10}$ space by the square root of the number of catchments, then de-transforming from $\log_{10}$ space. This approach relies on the assumption that the errors of the R-values of the pour-point catchments are independent.

3.4.8 Catchment-scale results from national models

Requirements

The general procedure detailed above provides R_t-values by segment at the national scale. MWLR requested some identification of catchment-scale areas wherein the environmental thresholds were typically exceeded. This was required to identify areas of the country where erosion mitigation would likely be focussed and so where MWLR could assess the cost of mitigation.

"Pour-point" catchments

A series of teleconferences involving MfE, MWLR, and NIWA staff discussed how these "priority" areas could be defined. This produced the concept of "pour-point" catchments, which were defined as those parts of a coast-draining catchment upstream from the first segment where a $R_t > 0$ value was encountered. These segments are termed "pour-points", and their catchments are "pour-point catchments". A conceptual designation is shown in Figure 3-21. Only stream segments of order 3 and higher were used to define the pour-points. Figure 3-22 and Figure 3-23 show worked examples of small and large pour-point catchments, respectively.

Pour-point segments were identified using GIS upstream-tracing algorithms, while their catchment boundaries were created by "dissolving" the boundaries of all the REC2 network sub-catchments upstream from the pour-point. 627 pour-point catchments were so identified. For each pour-point catchment, the count and proportion-by-count and proportion-by-area of the enclosed segments having R_t>0 were determined, and the average R_t values (including zero values) within the catchment were calculated.

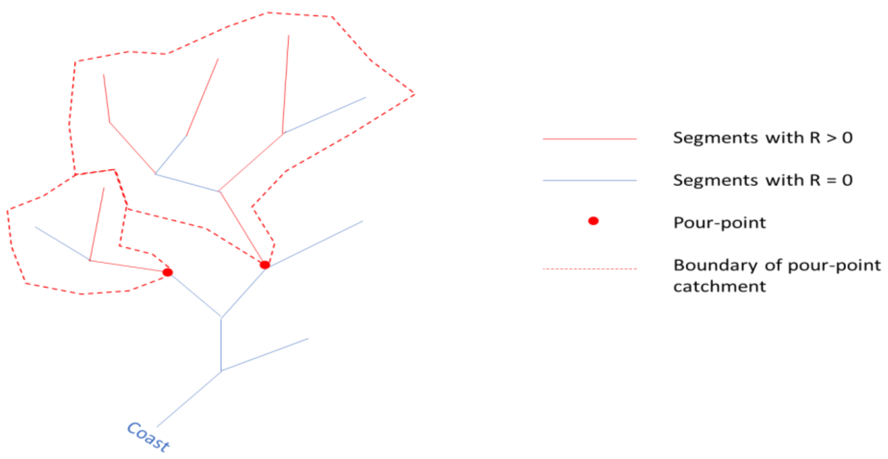

Figure 3-21: Definition sketch of pour-points and their catchments. Pour-points are located at the first occurrence of a segment where the load reduction factor (R) exceeds zero when tracing up a branching network from the coast. The pour-point catchment becomes the whole of the catchment upstream from the pour-point (even if this contains segments with R = 0). This example shows two pour-points and their catchments: one halfway up the mainstem and one on the second tributary.

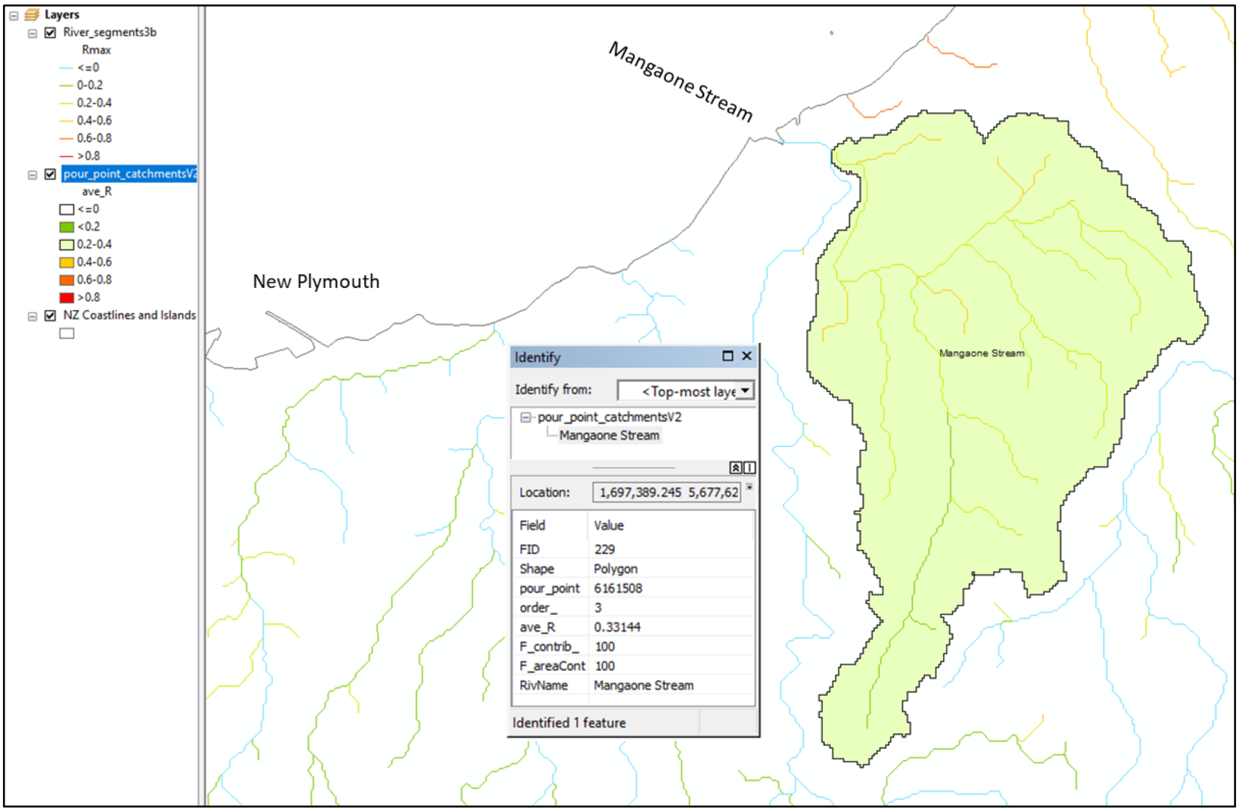

Figure 3-22: Example of a small coastal pour-point catchment: Mangaone Stream, a 3rd-order stream near New Plymouth. The pour-point is located at the first R>0 segment encountered when tracing upstream from the coast. The average segment R-value within the pour-point's catchment is 0.33 (so the catchment is infilled with the pale-green colour of the R=0.2-0.4 band in the legend). 100% of segments within the pour-point catchment (as indicated by the F_contrib value in the information box) have R>0 (as indicated by their colour coding), thus 100% of the catchment area has R>0 segments as well (as indicated by the F_areaCont value). Note that the left-branch of Mangaone Stream has no pour-points tagged because its R>0 segments are only 1st order. Similarly, small nearby catchments having R>0 segments have not been tagged with pour-points because none of these exceed 2nd order. Image extracted from ArcMAP.

Excluded areas

MfE identified two areas that were to be excluded from the MWLR analysis:

- the Department of Conservation (DOC) "conservation estate", and
- catchments in which the dominant source-of-flow (SOF) is classified as "Glacial-Mountain".

The DOC estate was excluded to preserve the natural state. It was defined by a widely available GIS layer (which cuts across catchment boundaries). Catchments with SOF = Glacial-Mountain were identified using the SOF level in the River Environment Classification developed by Snelder et al. (2005). This used a set of rules to define the dominant SOF in every reach of the REC1 digital network[11]. This classification was mapped onto the segments of the REC2 network for this study. The SOF = Glacial-Mountain catchments were defined using an upstream tracing algorithm similar to that

[11] In the Snelder et al. (2005) classification, Glacial-Mountain steams were those that receive more than 50% of their total annual rainfall volume above an elevation of 1000 m and have greater than 2% of their catchment area occupied by permanent ice.

used to define the pour-point catchments. Tributaries with SOF other than Glacial-Mountain were excluded (even if they joined a mainstem where the Glacial-Mountain SOF dominated (Figure 3-24).

We created ArcGIS “shape-files” of both exclusion areas, with the expectation that they would be used as simple visual masks on maps showing the pour-point catchments and/or the stream segments classified by R_t-value.

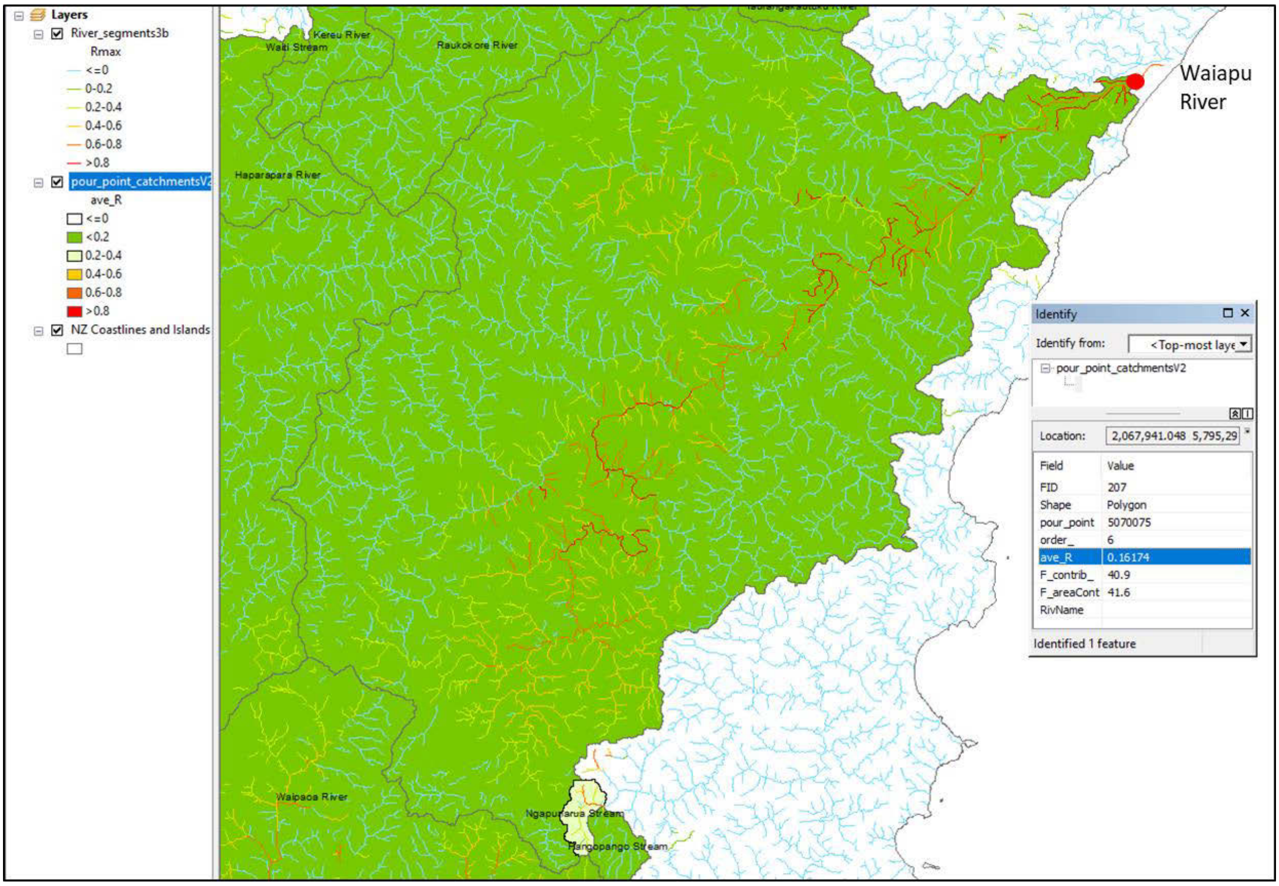

Figure 3-23: Example of a large pour-point catchment: the 6th-order Waiapu River catchment at East Cape. The pour-point is located at the first R>0 segment encountered upstream from the coast (red ball), which in this case occurs at the terminal coastal segment. The pour-point’s catchment covers all of the Waiapu catchment upstream, even though many of its segments have R=0 (blue segments). The yellow-red toned segments highlight the tributaries where the turbidity threshold is predicted to be exceeded. The average segment R-value within the catchment, including the R=0 segments, is 0.16 (so the catchment is infilled with the mid-green colour of the R=0.0-0.2 band in the legend). 40.9% of segments within the pour-point catchment (as indicated by the F_contrib value in the information box) have R>0, and these cover 41.6% of the catchment area (as indicated by the F_areaCont value). Note that the adjoining Raukokere, Haparapara, and Waipaoa Catchments are also pour-point catchments, while the smaller coastal catchments to the east are not. Image extracted from ArcMAP.

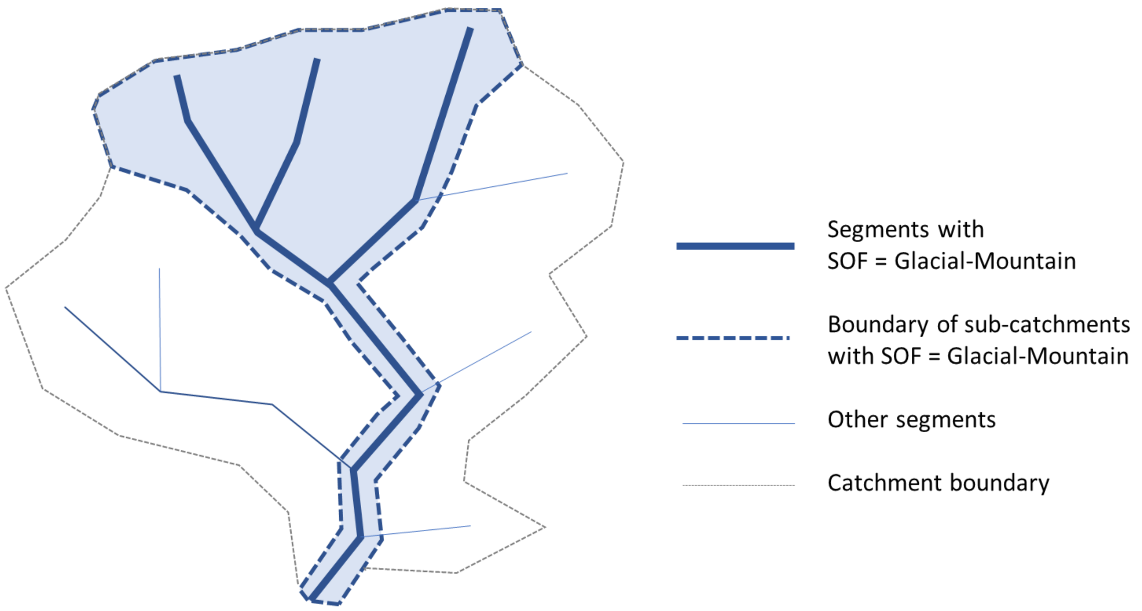

Figure 3-24: Sketch showing how Glacial-Mountain catchments were defined.

3.4.9 Results forwarded

The result-files forwarded to MfE and MWLR include:

- shape-files of the pour-point catchments, including their boundaries and summary statistics in an attribute table (including count and proportion of R-values > 0, and average R-values)
- shape-files of all REC2 segments, with an attribute table including values of R_t, all values of predicted median turbidity and turbidity thresholds used in the calculation of R_t, and estimated mean annual suspended sediment load from the upstream catchment (as derived from the "corrected" version of the Hicks et al. 2019 model)
- shape-files of the DOC conservation estate, and
- shape-files of the SOF = Glacial-Mountain catchments.

3.4.10 Value of continuous turbidity monitoring to inform on sediment load reduction requirements

MfE has queried how continuous turbidity monitoring could help inform future modelling of sediment load reduction requirements in catchment-specific contexts. High-frequency continuous monitoring with in-situ turbidity recorders would help by:

- Providing a more precise measure of median turbidity (compared with the estimate provided by the median of monthly sampling).
- Providing a surrogate for suspended sediment concentration for measuring sediment load. After appropriate calibration with field samples (i.e., to derive the coefficients *e* and *f* in Equation 2) and with a nearby water discharge record, a turbidity record can

be used to measure the suspended sediment load and so more precisely inform on the absolute reduction of sediment load required to achieve a turbidity threshold. When managing catchment sediment exports, it will also help quantify the sediment load reduction effected by erosion mitigation works and the impact of those on the median turbidity.

- Informing on the timing of sediment delivery from a catchment and hence on the location of sediment sources that influence the median turbidity (which occurs during base flows or well-along the recessions of runoff events).
- Providing a proxy record of visual clarity (after a phase of field calibration measurements). Instruments for routine, in situ monitoring of visual clarity are not yet available but turbidity can also be used as a clarity surrogate. In turn, this could assist with managing visual clarity.

Two cautions with using in-situ turbidity monitoring are:

- Different instruments may provide different readings for natural suspended sediment mixtures even when the instruments are all calibrated to a reference suspension such as Formazin. In part this is because different instrument brands use different turbidity measurement protocols (with different light wavelengths and back-scatter detection angles – see, for example, NEMS 2016), but even when the same protocol is used turbidity readings in natural sediment mixtures can vary between instrument brands (and with some brands even between instruments of the same brand). This should not be an issue if the instrument is being used as a surrogate for sediment concentration or visual clarity and is calibrated appropriately, but it may lead to inaccuracies when turbidity is the end result.
- In-situ turbidity sensor lenses are prone to biofouling (e.g. by algae), which elevates the apparent turbidity. Therefore, they must generally be equipped with anti-fouling devices such as miniature wipers, and their records need careful checking for signs of fouling.

4 Deposited fine sediment

4.1 Overview

As discussed in Section 1, the definition of deposited fine sediment (DFS) favoured by MfE for use in the Impacts workstream is the Instream Visual Areal coverage (IVAC), which is a visually-assessed measure of the proportion of the wetted streambed area covered with sediment finer than 2 mm (equating to the SAM2 method of Clapcott et al. 2011 and what we refer to hereafter as the fine sediment cover, FSC). The aim is to derive a model that predicts median (over time) FSC to acceptable accuracy at the national scale so that (like for clarity and turbidity) (i) stream segments where the FSC exceeds the “bottom-line” threshold can be identified, and (ii) the reduction in sediment load required to meet the threshold can be estimated (it follows that to service (ii), the prediction model should also be sensitive to sediment load). Previous attempts to develop a predictive model of FSC with these attributes from New Zealand datasets using machine-learning type models have been found wanting, both through limited performance and a lack of dependence on catchment sediment load. Here, we attempt an alternative, more physically-based modelling approach. In the following sub-sections, we:

- review previous DFS models
- explain the concepts behind a physically-based model, and
- develop and evaluate such a model.

Given the high risk that the physically-based model would perform no better, a conditional workstream was planned that would assess the use of a national model that predicted the Suspendable Inorganic Solids (SIS) measure of DFS coupled with a model relating IVAC to SIS.

4.2 Review of previous models

Hicks et al. (2016) used a boosted regression tree (BRT) approach to develop predictive models for FSC, SIS, and the Shuffle Index from catchment and reach characteristics including suspended sediment load. The FSC dataset pooled data collected across the SAM1, SAM2, and SAM3 protocols of Clapcott et al. (2011) and across varying mesohabitats (i.e., riffles, pools, runs), so a reduced dataset was made for data from runs alone. Multiple data values from the same site were averaged. Sediment load was incorporated in two predictor variables: sediment yield (load per unit catchment area) and segment-load/stream-power (where the segment-load is the load only from the local catchment adjacent to the segment, not from upstream, and stream-power relates to the product of segment slope and discharge).

With FSC, a weak *inverse* relationship was observed with catchment sediment yield but there was a weak positive relationship with segment-load/stream-power. The BRT model with the pooled dataset explained only 22.7% of the variance in the observed FSC, but restricting the model to only run data explained 55.7% of the variance and had a respectable RMSD of 10.2%. However, while performing reasonably statistically, the model showed only weak sensitivity to the sediment load. Moreover, the negative relationship with the load from the catchment upstream, and even the apparent positive link with the load from the local catchment, conflicts with the notion of managing DFS by reducing up-catchment loads.

On the positive side, the Hicks et al. (2016) modelling did provide some learnings, showing that local stream-power made an important contribution to the DFS model (causing a sensible reduction in DFS with higher stream-power), and they suggested that managing local sediment sources (e.g., eroding banks) might be more important that managing the up-catchment sediment load, most of which is transported during high-flow events. The explanatory importance of stream-power as a control on DFS was confirmed for streams in the UK by Naden et al. (2016), who also showed that sediment yield made little contribution to variation in DFS.

Clapcott and Goodwin (2017) extracted from the dataset compiled by Hicks et al. (2016) a "%coverB" subset of FSC records, comprising observations using the SAM1, SAM2, Rapid-habitat-assessment methods for run-habitat combined with records from the NZ Freshwater Fish database that did not identify mesohabitat type. They used a BRT approach to predict %coverB, relating it in two stages to land cover then catchment/site environmental variables. Their model explained 26% of the variance in the logit-transformed %coverB dataset and had a RMSE of 1.1, which equates to a standard error of ± 25% FSC when the estimated FSC is 50%[12]. Their "Pastoral heavy" variable was the most influential land cover variable, while elevation was the most important physical variable. The overall result was, however, no better than the Hicks et al. (2016) model.

4.3 Physically-based model

4.3.1 Conceptual model

Building from the learnings in Hicks et al. (2016), we propose that the time-averaged (or median) FSC at any stream segment should relate to four general factors:

FSC = [Sediment supply] × [Sediment trap-efficiency] × [Probability-of-occurrence] x [Space]

The sediment supply factor relates to flux of settling sediment, which is the product of the sediment fall speed and the SSC under flow conditions that promote sediment deposition. Deposition can only occur during flow recessions or at baseflows – since during rising flows any sediment depositing will be immediately re-entrained. Thus, it is the supply and concentration of suspended sediment under these waning/base flow conditions that is important. In turn, this will depend on the relative phasing of the sediment and water delivered to the stream network upstream from the segment of interest. For example, if the runoff is supplied (by rain) uniformly over the catchment, but the main sediment sources are at the catchment headwaters, then the SSC will peak on the flow recession. Erosion processes that operate on recessions (e.g., bank failure due to unbalanced pore water pressure in newly exposed banks and earth flows) will also contribute to the supply factor, as will processes that deliver sediment at base-flows (e.g., glacial melt and stock wading in streams). Thus, while one might expect a broad relationship with the long-term average sediment load (e.g., if there is zero load, there will be no deposited sediment), the key factor for DFS is when this sediment arrives and where it comes from[13] - which should relate to factors such as the erosion terrain, land-cover and land-use.

Sediment fall speed is controlled largely by grainsize, which, in turn, is controlled by catchment lithology and soil characteristics.

[12] The logit transformation means that their model absolute errors varied with the predicted FSC, getting smaller toward 0% and 100% cover.

[13] This concept explains why poor correlations are experienced with the long-term average load, most of which is discharged under flood peaks, and suggests that targeting erosion sites/processes that deliver their sediment loads late during runoff events or at base-flows should effect greater improvements in DFS than mitigating all erosion sites uniformly.

The supply factor will also relate inversely to the “flashiness” of the runoff regime, since this will control the time-period while elevated SSCs occur during recessions, and this will depend on catchment steepness.

The <u>sediment trap-efficiency</u> factor relates to the local hydraulic conditions at the reach (as indexed by the local shear stress, since this determines the intensity of flow turbulence which hinders sediment settling) and, again, the suspended sediment fall speed.

The <u>probability-of-occurrence</u> factor sets the likelihood that deposited sediment will be observed during a monitoring program. It should be appreciated that DFS is a transient feature of most streambeds – it deposits during recessions/base-flows and is re-entrained and flushed during high-flow events – so the likelihood of randomly encountering DFS should increase as the frequency of floods decreases (and the time-span between floods increases).

The <u>space</u> factor applies at the channel scale and depends inversely on the size of the streambed framework material. It concerns the space available for DFS to build up between coarse bed-material clasts. For example, with a boulder bed, considerable deposition is required to infill the gaps between boulders before they are all covered, whereas a fine gravel bed will be covered by a much smaller volume of fine sediment.

4.3.2 Model form

We developed this conceptual model into the following solvable form, using appropriate variables where available or readily created off the REC database:

$$FSC = \underbrace{\overbrace{[T]^{Ct}}^{\text{Recession SSC}}\ \overbrace{[C]^{Cc}[LC][ET][\tfrac{Xs}{Xw}]^{Cx}/[USslope]^{Cs}}^{\text{Factors controlling phasing \& duration of sediment delivery}}}_{\text{Supply factor}}\ \overbrace{\left[\frac{u*}{W_s}\right]^{Cu}}^{\text{Trap efficiency}}\ \underbrace{[FRE3]^{Cf}}_{\text{Probability-of-occurrence}}\ \overbrace{[D65_s]^{Cd}}^{\text{Space factor}} \tag{16}$$

where: *FSC* is the temporal median proportion of the streambed covered with fine sediment; *LC* is a land cover factor weighted by area and land-cover-specific coefficients, *CLi*; *ET* is an erosion terrain factor weighted by area and erosion-terrain-specific coefficients, *CEj; Ct*, *Cc*, *Cx, Cd, Cs, Cf, and Cu* are other coefficients to be solved for; and the other terms are the predictor variables which are defined in Table 4-1 and their derivation is detailed in Appendix B.

Table 4-1: **Variables included as predictors in DFS model.**

Variable abbreviation	Description	Units	Rationale/ explanation	Source
T	Predicted median turbidity	NTU	Turbidity is directly related to SSC and the median value should be reasonably representative of that during flood recessions.	Median turbidity derived by model reported in Section 3.
FSC	Aerial proportion of stream bed covered by fine sediment	-	-	-
X_s/X_w	Ratio of the distance from the segment of interest to the centre-of-mass of the upstream sediment supply (X_s) and the distance to the centre-of-mass of the upstream runoff (X_w).	-	Distance of the source of sediment and water discharge from the reach of interest influence the transport of fine sediment at different flood sizes and also different stages of the hydrograph.	Segment local-catchment sediment loads from Hicks et al. (2019) and mean discharge (existing REC2 variable).
C	Load-weighted suspended sediment concentration, equal to the mean annual suspended sediment load divided by the mean annual water discharge.	g/m^3	An index of the SSC at times when sediment is depositing; is directly linked to the long-term average sediment load.	Segment upstream-catchment sediment loads from Hicks et al. (2019) and mean discharge (existing REC2 variable).
W_s	Suspended sediment fall velocity.	m/s	Along with concentration, determines the rate at which sediment settles from a suspension.	Not activated (set to 1), as no model predicting suspended sediment size available.
Af, Ap, At, Ao	Aerial proportions of upstream catchment in land cover groups (forest, pasture, tussock, other).	-	Different land covers generate different suspended sediment loads and therefore influence the supply of fine sediment.	LCDB3 land cover class groupings as extracted in Hicks et al. (2019).

Variable abbreviation	Description	Units	Rationale/ explanation	Source
AEj	Aerial proportions of upstream catchment in *j*th erosion terrain group.	-	Suspended sediment load generation, as source of fine sediment deposition, highly depends on the erodibility of the upstream catchment lithologies.	MWLR Erosion Terrain class groupings as extracted in Hicks et al. (2019).
$D65_s$	Segment bed surface grainsize, represented by the size for which 65% of material is finer.	m	The bed framework size influences the volume of fine sediment required to cover a given area of streambed.	Predicted for REC2 by Haddadchi et al. (2018).
UsSlope	Average slope of the upstream catchment.	-	Catchment slope influences fine sediment deposition mainly through the direct effect on travel time of the runoff to the river reach during rainfall events.	Existing REC2 calculated field.
FRE3	Average frequency of high flow with peaks-flow > 3 x median flow.	Events/yr	Number of events higher than 3 median flows represent hydrologic flashiness which directly influence transport of fine sediment deposition from the river bed.	Existing REC2 predicted field.
$u*$	Shear velocity – a measure of shear stress and an index of turbulence intensity.	m/s	Total stress as a combination of flow and reach slope influence the transport of sediment from upstream reach and also recycle of deposited sediment within the river reach.	Calculated off existing segment slope calculated REC2 field, predicted mean discharge REC2 field, and Jowett (1998) hydraulic geometry relations between mean depth and discharge.

Variable abbreviation	Description	Units	Rationale/ explanation	Source
SOF	Source of flow	-	Source of flow was used to differentiate mountain, hill, lowland, and lake-fed rivers.	Existing REC2 calculated field.

To optimise the coefficients in the FSC model, Equation (16) was $\log_{10}$ transformed (indicated by "log" throughout)[14]. The resulting transformed equation is:

$$\begin{aligned}\log(FSC) = {} & Cbase + Ct\ \log(T) + Cc\log(C) + Cx\ \log\left(\frac{X_s}{X_w}\right) \\ & + \log(CLf \times Af + CLp \times Ap + CLt \times At + CLo \times Ao) \\ & + \log\left(\sum_{j=1}^{12} CEj \times AEj\right) - Cd\ \log(D65s) - Cs\ \log(USslope) \\ & - Cf\log(FRE3) - Cu\ \log\left(\frac{U*}{W_s}\right)\end{aligned} \tag{17}$$

Lastly, we considered that the REC source of flow class (SOF) might also influence fine sediment deposition, for example by indexing sediment supply processes during base flow conditions – for example, through diurnal snow melt in mountains. To explore for this effect, we calibrated additional models from training datasets partitioned into four different dominant flow sources: mountain, hill, lowland, and lake-fed (as defined and classified by Snelder et al. 2005).

4.3.3 Response dataset

The FSC predictor model was calibrated against a national dataset compiled from sites where data had been measured by the Instream Visual (SAM2) method in run habitat[15]. The details of these data are reported in Clapcott and Goodwin (2017). The full dataset represented 560 sites. Where more than one observation was available at the same site these were averaged. The number of observations at each site varied from 1 to 12, but 67% of sites had only 1 observation and 89% had 3 or less. Across the sites the FSC values were skewed towards lower values (Figure 4-1): most (80%) showed an average FSC < 40%. 7.3% had FSC = 0% and 6.8% had FSC = 100%. We omitted the 41 sites with 0% FSC from the analysis in order to undertake a log-transformation, while a further 52 sites were not used because of difficulties obtaining values for some predictor variables. Of the 467 remaining sites, 177 had dominantly hill SOF, 268 had dominantly lowland SOF, 18 had dominantly mountain and glacial mountain SOF, and 4 had lake-fed SOF.

[14] This transformation required omitting the 7.3% of sites where the FSC value was zero (Section 4.3.3).
[15] An updated dataset was supplied to this study by Dr Joanne Clapcott from Cawthron Institute in January 2019. In this dataset, site locations were assigned to REC1 reach numbers. For this study, we re-mapped these onto REC2 segments.

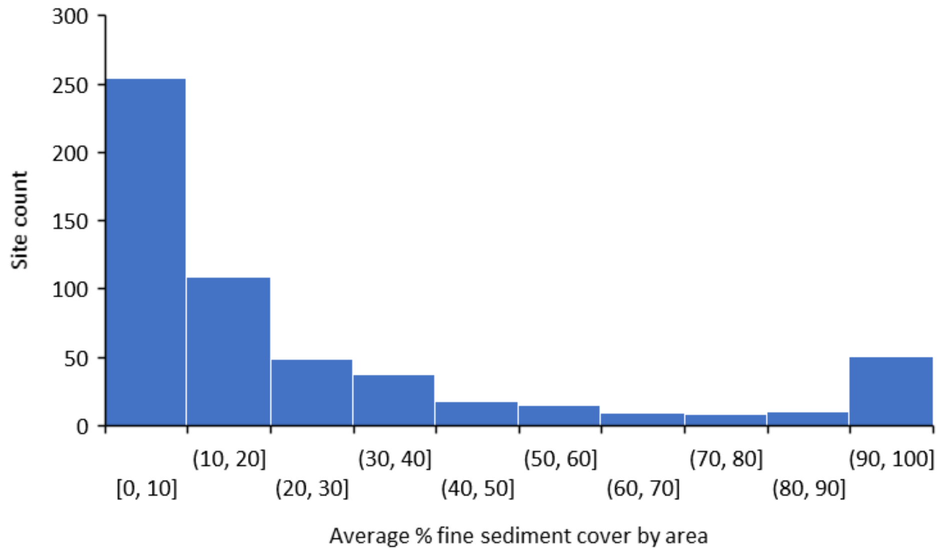

Figure 4-1: Histogram of average % fine sediment cover by area across 560 sites. Fine sediment cover assessed by Instream Visual (SAM2) method.

An alternative dataset was available for 606 sites where FSC had been observed using the Bankside Visual method (SAM1). In the first instance we chose not to work with this dataset because it is regarded as less reliable than the Instream Visual method (Clapcott et al. 2011). We left this open to revisit if the results from the Instream visual dataset proved sufficiently promising. For similar reasons, we avoided using datasets where the type of mesohabitat was mixed or unrecorded.

4.3.4 Correlation analysis

We used a correlation matrix (Table 4-2) to explore the interdependency between the predictor variables (listed in Table 4-1) and also their relationship with FSC.

Confirming our conceptual model, fine sediment cover inversely correlates with the upstream catchment slope (R=-0.28), *FRE3* (R=-0.24), *u** (R=-0.19), and the substrate framework *D65* size (-0.18). Moreover, a reasonable positive correlation (R=0.25) is shown with the lagging sediment delivery parameter X_s/X_w, high values of which indicate more of a catchment's sediment load should be carried on recessions. Inverse correlation of FSC with forest land cover (R=-0.19) but a positive correlation with other land covers is consistent with observations of early-during-event sediment exhaustion from forest catchments but lingering sediment supplies under non-forest land covers (e.g., Haddadchi and Hicks, 2019). Weak positive correlations with the loess and tephra, gullied argillite, and alpine schist erosion terrains suggest a tendency of those terrains to sustain sediment supplies during events. The FSC correlation with median turbidity is weak but positive, while the weak negative correlation with the load-weighted SSC suggest the overall load is of less importance than when it arrives.

Upstream slope, *u**, *FRE3*, land cover, and erosion terrain all show some degree of inter-correlation (as quantified by the correlation coefficient, R); load-weighted SSC correlates with median turbidity and Erosion Terrain 3 (with erodible mudstone lithologies); but median turbidity relates more to the presence of pasture than any particular erosion terrain.

We conclude that most of the essential elements of the conceptual model are supported by the correlation matrix.

Table 4-2: Correlation coefficients between predictor variables and DFS parameter. The correlation coefficients, R, are Pearson coefficients fitted to the actual data.

	T	*C*	*X* X_s/X_w	*Forest*	*Pasture*	*Tussock*	*Other*	*ET1*	*ET2*	*ET3*	*ET4*	*ET5*	*ET6*	*ET7*	*ET8*	*ET9*	*ET10*	*ET11*	*ET12*	*FRE3*	*usSlope*	*D65*	*u**
T		0.36	-0.14	-0.24	0.24	-0.04	0.00	0.12	-0.03	-0.01	0.00	0.00	-0.05	-0.09	-0.07	-0.05	-0.02	0.09	-0.01	0.08	-0.08	-0.06	0.00
C			-0.13	-0.07	0.12	-0.05	-0.07	-0.01	-0.09	0.23	0.00	0.00	-0.06	0.03	-0.04	-0.02	-0.01	0.01	-0.07	0.03	0.09	-0.06	0.12
X_s/X_w				0.19	-0.25	-0.01	0.20	-0.20	-0.06	-0.05	0.00	0.00	0.07	0.29	0.06	-0.05	-0.03	-0.08	0.20	-0.13	-0.06	0.04	-0.18
Forest					-0.77	-0.13	-0.11	-0.26	-0.06	0.08	0.00	-0.02	0.21	0.26	0.12	0.15	0.09	-0.09	-0.08	0.25	0.49	0.48	0.40
Pasture						-0.37	-0.30	0.31	0.05	-0.01	0.00	0.03	-0.18	-0.22	-0.12	-0.12	-0.07	0.02	-0.23	-0.09	-0.40	-0.39	-0.28
Tussock							-0.05	-0.07	0.07	-0.07	0.00	-0.01	0.02	0.03	0.06	-0.02	-0.01	0.07	-0.05	-0.11	0.09	0.11	0.01
Other								-0.11	-0.10	-0.06	0.00	-0.01	-0.03	-0.08	-0.04	-0.02	-0.01	0.04	0.77	-0.19	-0.19	-0.22	-0.17
ET1									-0.45	-0.28	0.00	-0.05	-0.16	-0.38	-0.17	-0.08	-0.05	-0.16	-0.19	-0.10	-0.30	-0.36	-0.47
ET2										-0.13	0.00	-0.02	-0.07	-0.18	-0.08	-0.04	-0.02	-0.07	-0.09	0.02	-0.04	0.23	0.14
ET3											0.00	-0.01	-0.05	-0.11	-0.05	-0.02	-0.01	-0.05	-0.06	0.06	0.13	0.01	0.25
ET4												0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ET5													-0.01	-0.02	-0.01	0.00	0.00	-0.01	-0.01	-0.01	-0.01	0.00	0.05
ET6														-0.06	-0.03	-0.01	-0.01	-0.03	-0.03	0.08	0.03	0.16	0.19
ET7															-0.07	-0.03	-0.02	-0.06	-0.08	0.04	0.38	0.23	0.27
ET8																-0.01	-0.01	-0.03	-0.03	0.10	0.07	0.10	0.06
ET9																	0.00	-0.01	-0.02	0.21	0.11	0.15	0.19
ET10																		-0.01	-0.01	0.05	0.05	0.04	0.12
ET11																			-0.03	-0.04	0.04	-0.05	-0.06

	T	***C***	***X_s/X_w***	***Forest***	***Pasture***	***Tussock***	***Other***	***ET1***	***ET2***	***ET3***	***ET4***	***ET5***	***ET6***	***ET7***	***ET8***	***ET9***	***ET10***	***ET11***	***ET12***	***FRE3***	***usSlope***	***D65***	***u******
ET12																				-0.15	0.13	0.50	0.42
FRE3																					0.37	0.58	0.29
usSlope																						0.50	0.59
D65																							0.42
*u**																							
FSC	0.08	-0.09	**0.25**	-0.19	0.13	-0.11	**0.22**	-0.05	0.10	-0.04	0.00	0.04	-0.03	-0.11	0.04	-0.06	-0.03	-0.03	0.21	-0.24	-0.28	-0.18	-0.19

4.3.5 Model fitting

Genetic algorithm optimisation was used to solve the 24 coefficients of the FSC model given by Equation (17). The Optquest algorithm in Oracle's Crystal Ball software (Oracle 2015) was used with 10,000 iterations, and the Latin Hypercube sampling method with 2500 iterations was applied. The model coefficients were first optimised using all the data, then coefficients were determined separately using different sources of flow. This includes 268 sites with lowland flow sources, 177 with hill, 18 with mountain and glaciated mountain, and 4 lake-fed.

The optimisation technique sought to minimise the root-mean-square-deviation calculated for the residuals between the log of the observed and modelled FSC values (log-RMSD)[16]:

$$log - RMSD = \sqrt{\frac{\sum_{k=1}^{m}\left(\log FSC_{o,k} - \log FSC_{p,k}\right)^2}{m}} \tag{18}$$

where m is the number of sites and $FSC_{o,k}$ and $FSC_{p,k}$ are the observed and predicted DFS values, respectively.

The final optimised coefficients for the different FSC model runs (with and without the source of flows partitions) are presented in Table 4-3. For the direct and inverse variable types (which enter Equation (16) as exponents), the magnitude of the coefficient is a measure of the sensitivity of FSC to the variable (e.g., a value close to 1 indicates a near-linear response). For the land cover and erosion terrain variables, the coefficients demonstrate the relative sensitivity within each group.

Considering first the model with all the data, we see:

- Positive values for the *Ct, Cx, Cs, Cf,* and *Cd* coefficients – indicating direct responses to turbidity and the sediment lag factor and inverse responses to flood frequency and catchment slope, as expected from the conceptual model.
- Near-linear sensitivity to flood intermittency (*Cf* = 0.82).
- Moderate sensitivity to median turbidity (*Ct* = 0.49) but not the load-weighted SSC (Cc = 0), suggesting greater influence by SSC on recessions than during floods when the bulk of the suspended load is delivered.
- Relatively weak sensitivity to the sediment lag factor (*Cx* = 0.13), the upstream slope (*Cs* = 0.13), and the bed-material *D65* (*Cd* = 0.08).
- Not sensitive to the re-entrainment index $u*$ - perhaps because of the large uncertainty in predicting this (due to uncertainty in the hydraulic geometry relations and segment slope[17]).
- Relatively greater sensitivity to pasture land cover and "other" land cover (which includes bare ground and snow and ice) and less with forest and tussock.

[16] The calibration approach used log-transformed FSC to reduce the skew due to the variation in percentage of FSC.
[17] We note that segment slope is at best the averaged slope of a long reach and need not reflect the slope of runs within that reach. This is because the reach-averaged slope averages the slopes of component riffles, runs and pools and will depend on the relative proportions of these meso-habitats as well as their typical slopes.

- Relatively greater sensitivity (*CE* coefficients > 1.7) for the Tephra and loess, Intensely gullied crushed argillite and greywacke from the East Cape region, Volcanic rocks (lavas, rhyolite etc group), and Coarse plutonics and metamorphics erosion terrain groups. This may reflect the effects of lithology on grainsize (e.g., the coarse plutonic and volcanic terrains tend to erode to coarse sediment particles, which will be more inclined to settle) or on erosion process and sediment delivery (e.g., the East Cape gullied terrain tends to deliver high suspended sediment concentrations on event recessions – Hicks et al. 2004).

With the SOF-partitioned models:

- The SOF = Lowland model showed a similar pattern of coefficients to the all-dataset model, with strong sensitivity to turbidity (*Ct* = 0.69) and inverse sensitivity to *FRE3* (*Cf* = 0.94).
- The SOF = Hill model responded in the anticipated way to all of the direct/inverse variables but was generally less sensitive to any of these.
- The SOF = Mountain model appeared to be at best only moderately sensitive to turbidity and a few erosion terrains, but the number of sites (15) was very limited.
- The SOF = Lake-fed model showed strong sensitivity to *FRE3* (*Cf* = 1.3) and forest cover, but with only four sites little should be drawn from this.

Table 4-3: Predictor variable coefficients for all-data FSC model and models with data partitioned by source of flow (SOF). Variables are types as “direct” if a positive response with FSC is expected (numerator of Equation (16)) or “inverse” if an inverse response is expected (denominator of Equation (16)). Base coefficient is defined in Equation (17). Refer Appendix B for detail on land cover and erosion terrane groups.

Type	Variable	Coefficient	All data	SOF =Hill	SOF = Mountain	SOF = Lowland	SOF = Lake-fed
	Base	Cbase	2.25	2.11	1.57	2.20	1.75
Direct	*T*	Ct	0.49	0.24	0.30	0.69	0.10
	C	Cc	0.00	0.02	0.00	0.00	0.00
	Xs/Xw	Cx	0.13	0.19	0.01	0.15	0.01
Inverse	*UsSlope*	Cs	0.14	0.13	0.01	0.03	0.01
	FRE3	Cf	0.82	0.43	0.01	0.94	1.30
	*u**	Cu	0.00	0.13	0.00	0.00	0.00
	D65	Cd	0.08	0.01	0.01	0.16	0.01
Land cover aerial proportion	Forest	CLf	1.06	0.64	0.81	1.27	1.86
	Pasture	CLp	1.26	1.01	0.34	1.51	0.30
	Tussock	CLt	0.92	0.82	0.20	1.20	0.01
	Other	Clo	1.55	0.93	0.73	1.94	0.65
Erosion terrain aerial proportion	Sand country, flood plains, fans and terraces, peat	CE1	0.99	0.77	0.57	1.18	1.53
	Tephra and loess	CE2	1.70	1.31	1.49	2.23	1.30
	Tertiary mudstone, sandstone and soft limestone	CE3	1.29	1.68	0.71	1.28	0.33
	Intensely gullied crushed argillite and greywacke	CE4	1.48	1.66	1.11	1.75	1.80
	East Cape - Intensely gullied crushed argillite and greywacke	CE5	2.63	1.95	2.38	3.36	2.40
	Lavas, rhyolite, volcanic slopes	CE6	1.93	1.33	1.61	2.42	0.84
	Greywacke, argillite and hard limestone	CE7	1.12	1.05	0.84	1.70	1.57
	Schist and South Island greywacke (incl. alpine, ice and snow)	CE8	1.47	2.63	0.63	1.66	0.73
	Coarse crystalline plutonics and metamorphics	CE9	2.49	3.44	1.56	2.99	2.08
	Deeply weathered plutonics	CE10	0.66	0.74	0.01	0.66	0.07
	Water	CE11	1.43	1.89	1.02	2.21	0.56
	Other	CE12	1.22	0.65	0.29	1.34	0.68

4.3.6 Model performance

As well as by the log-RMSD optimisation function, model performance was evaluated by five other metrics: the root-mean-square-deviation of the actual (not log-transformed) values (RMSD), the Nash-Sutcliffe efficiency (NSE), the percent bias (PBIAS), Lin’s concordance correlation coefficient, and the R^2 (see glossary in Section 7 for explanations). These metrics are listed in Table 4-4, while the correspondence between measured and modelled FSC values is plotted in Figure 4-2.

Table 4-4: Performance metrics for all-data model and models with data partitioned by dominant source of flow. SOF column shows overall results from a composite model using the source of flow models with source of flow partitioning.

	All-data	Hill	Lowland	Lake-fed	Mountain	SOF
Sites	467	177	268	4	15	467
log-RMSD	0.51	0.51	0.49	0.45	0.27	0.49
SFE	3.23	3.23	3.09	2.82	1.86	3.09
RMSD (%)	24.80	18.36	27.85	3.57	8.74	24.00
NSE	0.15	0.64	0.49	1.00	0.99	0.20
PBIAS (%)	0.00	2.52	-0.91	-1.93	-5.45	-0.08
Concordance	0.48	0.38	0.47	0.99	0.60	0.49
R^2	0.24	0.19	0.23	0.99	0.39	0.25

The log-RMSD using all datasets was 0.51 (which equates to a standard factorial error, SFE, of ×/÷ $10^{0.51}$ = 3.23); after optimizing the model using different source of flows the log-RMSD was slightly reduced to 0.49 (SFE = ×/÷ 3.09). Standard errors exceeding a factor of 3 would have to rate as poor.

The absolute RMSD on predicted FSC is ~ 24%, which again signifies poor accuracy – particularly if the predicted FSC is less than 50%.

The near zero percent bias metric for the all-data and combined SOF models indicate that the general results from these models were unbiased overall. However, the typically ~ 0.5 values obtained with for the concordance correlation coefficient (which indicates how far the predictions deviate from the 1:1 line on a predicted vs measured plot) indicate only moderate concordance. This is very apparent on Figure 4-2, which shows the models over-predicting at FSC < 10% except for the SOF = Mountain and SOF = Lake-fed models, which have few data points and are likely “over-fitted”.

The NSE value for the all-data model (at 0.15) shows that it has predictive power only a little better than a simple model represented by the dataset mean[18]. While the NSE values are better for the SOF-partitioned models (particularly for the lake-fed and mountain models, but only because these have only a small number of data points), the NSE of the combined SOF model (0.2) is only a marginal improvement.

Similarly, the low R^2 values for the all-data and combined SOF models show that they only explain 24-25% of the variance in the measured dataset.

Our overall conclusion is that while FSC responds to most of the predictor variables as expected from our conceptual model, the accuracy of the FSC predictions is poor (standard factorial error > 3), predictions tend to be overestimated when FSC < 10%, and using different models according to dominant source of flow provides minimal improvement overall.

[18] The 467 sites have a mean FSC of 23% and a standard deviation of 26%.

Thus, at least in terms of statistical performance, this model offers no improvement over previous attempts at modelling FSC (Hicks et al. 2016; Clapcott and Goodwin 2017; Section 4.2). However, it is more consistent with the observations.

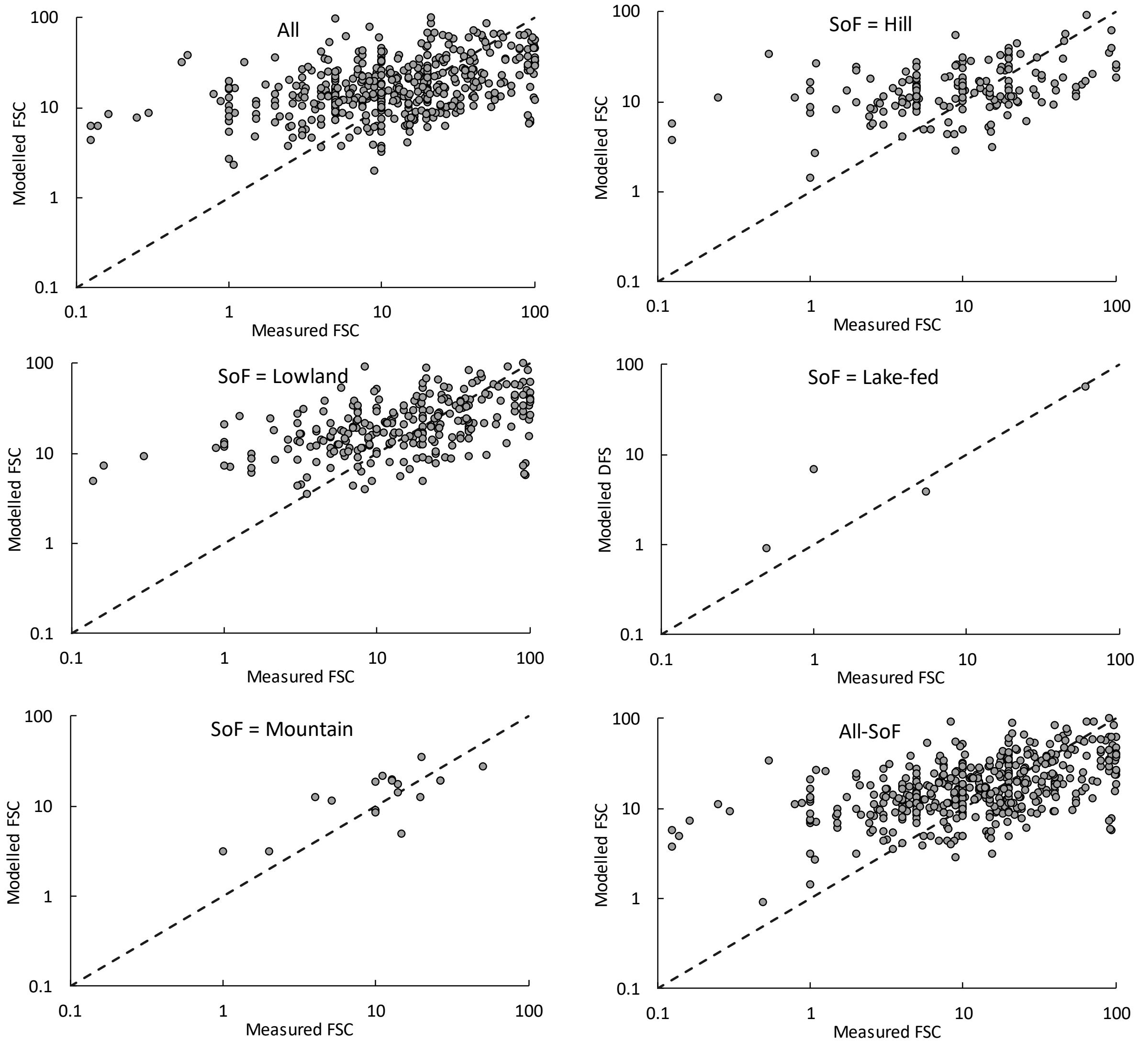

Figure 4-2: Measured vs modelled FSC for models using all data, different source of flow (SOF), and source of flow models combined. The FSC values are %-values.

4.4 Discussion

4.4.1 Why is the FSC model performance poor?

We suggest a number of reasons why the performance of the FSC prediction model is poor:

- Measurement error. Hicks et al. (2016) considered that significant operator variability occurs around the country when making FSC measurements.
- Large imprecision in mean FSC due to sampling error. As expected from our conceptual model and as observed at the relatively few sites in the training dataset with multiple observations, FSC shows considerable temporal variation at-a-site. This variation is shown on Figure 4-3 by the averaged-across-sites standard deviation, standard error of the mean, and range of the observed values. For example, at the one site with 12 FSC observations, the FSC standard deviation was 20% while the FSC range covered 66%. Most of the training sites had very few observations (67% had only 1 observation, and 89% had 3 or less), therefore, the precision of the estimated mean FSC at these sites will be poor. Based on the data in Figure 4-3, standard errors of ± 20-30% are likely associated with the mean FSC at those sites where there was only one observation.
- Uncertainty in predictor variables. Several of the predictor variables used in the FSC model are themselves estimated using model that are associated with significant uncertainty, e.g., median turbidity (SFE[19] = 1.95), load-weighted SSC (SFE at least 1.9[20]), D65 (SFE ≈ 1.4-1.8[21]), FRE3 (RMSE ≈ 4 events/year[22]), and shear velocity (error includes that from mean depth predicted from hydraulic geometry relation and calculated local channel slope, which is only approximated by the segment slope which has its own uncertainty).
- Missing key variables. A representative suspended sediment grain size is expected to be a key control on FSC because it determines the sediment fall velocity; however, there is no national predictor available for this. Suspended sediment grain size is represented in the current model only by the influence of erosion terrain and perhaps land cover.
- Process complexity and variability. As described in the conceptual model (Section 4.3.1), fine sediment deposition on streambeds reflects the balance between settling and re-entrainment which depends on sediment supply, sediment characteristics, local hydraulic conditions, and antecedent flow conditions – which are all difficult to represent in a national-scale, annual-average modelling framework.

Most of these issues (apart from the last one) could be resolved for future modelling by monitoring FSC:

- on a regular basis (e.g., monthly, quarterly) for long enough to adequately average-out temporal variability

---

[19] SFE = standard factorial error = $10^{\text{log-RMSE}}$.
[20] Hicks et al. (2019).
[21] Haddadchi et al. (2018).
[22] Booker (2013).

- at network of sites covering a gradient of stream physical characteristics and catchment erosion terrain
- by robustly following FSC measurement procedures
- by collecting continuous records of turbidity and water discharge, and
- by collecting site-representative data on channel hydraulics (velocity, depth, channel slope), bed-material size grading, and suspended sediment size grading.

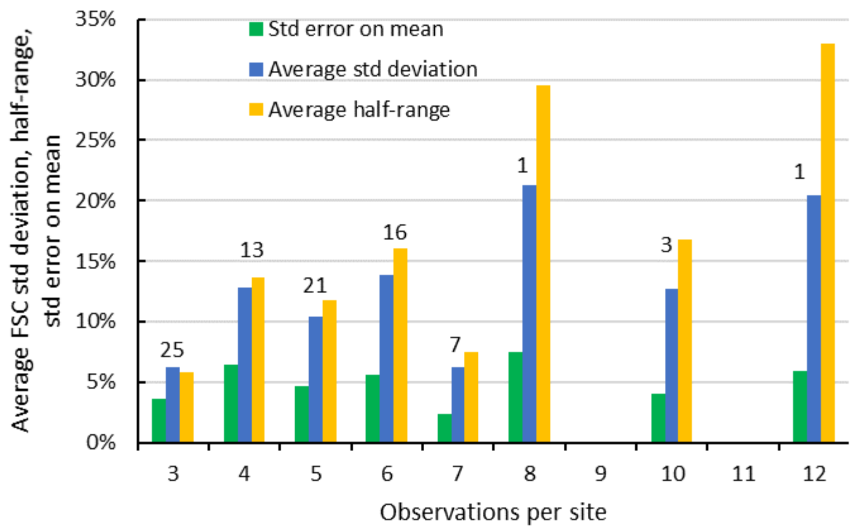

Figure 4-3: Standard deviations, standard error on mean, and ranges of observed FSC values, distributed by number of observations per site. The half-range is the maximum and minimum FSC observed at a site divided by two. Results have been averaged over sites in each number-of-observations class (the number of sites is labelled above the bars). For example, 3 sites had 10 observations, and the statistics are averaged for those 3 sites.

4.4.2 Managing DFS by regulating catchment sediment loads

The above work presents two problems for managing DFS by regulating mean annual catchment sediment loads, both for the impacts assessment workstream and for future policy application:

- In the context of the Band C/D thresholds proposed for FSC, which range between 21% and 97% (Table 2-1), the model's predictive capability is too inaccurate to adequately resolve which stream segments are over- or under-threshold – except perhaps for CTG classes 1, 5, and 11 (where the threshold exceeds 89%).
- Even if the predictive accuracy was high (e.g., to ± a few %), this and previous modelling has shown minimal dependence on catchment mean annual sediment load – that is, there is no function to use to calculate what the load reduction factor would need to be to meet any FSC threshold.

Based on the thinking outlined in our conceptual model for DFS (Section 4.3.1), a possible way forward is to abandon the focus on managing overall catchment average sediment loads but, instead, manage the supply of sediment to the late stages of flood recessions and during baseflows – which is when fine sediment deposition in baseflow channels occurs (until it is flushed by the next competent high-flow event). Likely sediment sources to focus on include:

- Bank erosion sites – where banks exposed on recessions collapse due to pore-water-pressure imbalances, and/or where blocks of bank-material have fallen into the baseflow stream and are gradually winnowed.
- Terrain where erosion features such as earthflows and gullies "bleed" sediment into the stream network.
- Headwater sources in long river networks, where the sediment load lags the flood wave.
- Land-use activities, such as stock-access to channels or earth-moving, that mobilise riparian sediment at baseflows or during small runoff events.
- Natural processes, such as diurnal glacial melt, which contribute a fine suspended load during baseflow periods.

Identifying these will require catchment-scale assessment of sediment source locations and the timing of their delivery to the stream sediment during runoff events – at a level more detailed than afforded by models that operate at national and mean-annual scales. The temporal aspect of sediment delivery, in particular, remains in the research domain but is now featured in the complementary Managing Mud (NIWA) and Smarter Targeting of Erosion Control (MWLR) research programs.

A key addition to these existing research programs would be to integrate catchment- and event-scale monitoring and modelling of fine sediment delivery and transfer with monitoring and physically-based modelling of streambed fine sediment deposition. This integrated erosion-transfer-deposition research should ideally include catchments featuring contrasting erosion processes and sediment-source distributions.

4.5 Alternative, linked-model approach

The initial workplan was that if our attempts at developing a satisfactory predictive model of FSC failed, then an alternative approach would be to look at coupling existing relationships between FSC and SIS with a national predictive model of SIS.

However, after discussion with MfE, this option was abandoned. The reason being that the existing SIS-prediction BRT model developed by Hicks et al. (2016) using data from 362 sites in run habitat also did not meet the criteria for accuracy and dependence on catchment sediment load. While the model did explain 39% of the deviance in the log-transformed SIS values, it also showed an apparent negative response to sediment load variables – which implies sediment load should be increased to reduce DFS!

5 Conclusions

The main conclusions of this study are:

- The sediment load reduction required to meet visual clarity and turbidity thresholds may be estimated as simple functions of the existing and target median clarity and turbidity and the relations of suspended sediment concentration with visual clarity and turbidity.
- The relations of suspended sediment concentration with visual clarity and turbidity are reasonably well estimated by national average values.
- Existing median visual clarity and turbidity were able to be predicted (to tolerable uncertainty) across the national drainage network by random forest models informed by catchment physical and hydrological characteristics, with mean annual sediment load the most important predictor variable, confirming the underpinning approach.
- Over the national stream network, 18.4% of segments exceeded the supplied clarity thresholds, 16.4% exceeded the turbidity thresholds, 10.3% exceeded both thresholds (indicating moderate concordance between the two thresholds), and 24.5% exceeded one or other or both thresholds. Thus, most stream segments across New Zealand require no sediment load reduction to meet the turbidity and visual clarity bottom lines.
- With sediment load reduction focussed on that required to meet the turbidity thresholds, of the 16.4% of segments nationally exceeding the turbidity thresholds, 70% required a load reduction factor (R) less than 0.4.
- The results for the load reduction factor are associated with significant uncertainty, largely due to the uncertainty on the modelled median turbidity values. This produces a significant risk that many of the segments calculated as requiring load reduction may not need it (while some that do have been missed). These large uncertainties pertain to R estimates at individual segments. Such errors are likely to be systematic at the catchment scale but vary randomly between catchments, so the uncertainty on R estimates averaged over multiple catchments, regionally, and nationally will be relatively smaller.
- Areas where erosion mitigation would likely be focussed can be identified from stream network maps colour-coding the load reduction factor in each segment or by mapping catchments upstream from the first segment where R exceeds zero on an upstream trace.
- Previous attempts to model deposited fine sediment from New Zealand datasets have been found wanting, both through limited performance and a lack of dependence on catchment sediment load.
- A more physically-based modelling approach considered that the time-averaged fine sediment cover should relate to four factors: sediment supply, sediment trap efficiency, probability of occurrence, and space. Using predictor variables to represent these factors, fine sediment cover models calibrated using Genetic Algorithm

Optimisation confirmed the importance of most of the factors described above, but model performance remained relatively poor (standard factorial error exceeding ×/÷ 3). Probable reasons for the disappointing performance included measurement error, sampling error, uncertainty in the predictor variables (several of which were predicted themselves), missing key variables, and process complexity and variability.

- Two problems arise for the impacts assessment workstream and for future policy application: the fine sediment cover model's predictive capability is generally too inaccurate to adequately resolve which stream segments are over- or under-threshold, and there is minimal dependence on catchment average annual sediment load.
- A possible way forward is to abandon the focus on managing catchment average sediment load but, instead, manage the supply of sediment to the late stages of flood recessions and during baseflows – which is when fine sediment deposition in baseflow channels occurs. Research that confirmed and calibrated the conceptual model presented herein would increase confidence in this suggestion. Such research would include monitoring and modelling fine sediment delivery to the stream network and associated streambed fine sediment deposition at catchment and event scales, ideally covering catchments featuring contrasting erosion processes and sediment-source distributions.

6 Acknowledgements

We thank:

- Dr Joanne Clapcott (Cawthron Institute) for supplying data on deposited fine sediment.
- Dr Les Basher (MWLR) and Stephen Fragaszy (MfE) for discussions around the concept of pour-point catchments.
- Dr Doug Booker (NIWA) for advice on statistics and modelling.
- Stephen Fragaszy (MfE), Heather Martindale (GNS), and Dr Ton Snelder (LWP Ltd) for comments on an early draft of this report.

7 Glossary of abbreviations and terms

Bias	Model bias measures the average tendency of the predicted values of water quality variables to be larger or smaller than the observed values. Positive values indicate underestimation bias and negative values indicate overestimation bias (Moriasi et al. 2007).
BRT	Boosted Regression Tree: a type of regression modelling that uses machine learning.
Concordance correlation coefficient	Measures the agreement between two variables in terms of how close their relationship approaches the 1:1 line. A value of 1 overlays the 1:1 line.
DFS	Deposited Fine Sediment: sediment of grain size finer than 2 mm that is deposited on or within the coarse framework of a streambed.
DOC	Department of Conservation.
ESV	Environment State Variable: a variable that captures an aspect of the state of the physical, chemical, or ecological environment.
FDC	Flow Duration Curve: graphical relationship between water discharge and the % time that discharge is exceeded.
FSC	Fine sediment cover as an aerial proportion of the streambed.
Hysteresis	A “loop” in a relationship between two variables, e.g., when SSC is higher at a given discharge on the rising stages of a flood hydrograph compared to the falling stages.
IVAC	In-stream Visual Aerial Coverage: a proportion-of-area based measure of fine sediment deposited on streambeds.
Lithology	Rock-type.
Ln	Natural base logarithm.
Log	Base 10 logarithm.
MfE	Ministry for the Environment.
MWLR	Manaaki Whenua Landcare Research.
NRWQN	National River Water Quality Network. A monitoring network of 77 river sites run by NIWA since 1989, with an aggregate catchment about 50% of NZ’s land area (Davies-Colley et al. 2011).
NSE	Nash-Sutcliffe Efficiency: a measure of the fit between observed values and model predictions, it determines the relative magnitude of the residual variance in the estimated yield compared to the measured yield variance. NSE ranges from $-\infty$ to 1, with 1 indicating a perfect match to predictions, 0 indicating that predictions are as accurate as the mean of the observed data, and negative values indicating that the observed mean is a better predictor than the model (Nash and Sutcliffe 1970).

NZSegment	Individual river segment within REC2, with associated environmental information available. Segment boundaries occur at confluences.
Out-of-Bag R^2 (OOB R^2)	The average proportion of the total variance explained by a random forest predictive model developed from n data records when the model is re-calculated n times, each time removing 1 record in turn from the derivation. Provides an estimate of the predictive power of the model for new cases.
Partial dependence plots	Show the marginal contribution of a predictor to the response (i.e., the response as a function of the predictor when the other predictors are held at their mean value) in a RF model.
Power function	A curvilinear function of the form: $Y = aX^b$, where X and Y are variables and *a* and *b* are fitting parameters.
R^2	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).
REC1	River Environment Classification version 1.
REC2	River Environment Classification version 2.
RF	Random Forest. A flexible regression technique in which final predictions are based on averages across an ensemble of regression trees.
RMSD	Root-Mean-Square-Deviation. A measure of the absolute precision of fit between observed values and model predictions. A lower RMSD indicates a better fit between observed and predicted values. RMSD can be used to evaluate the confidence intervals of model predictions.
RSR	Relative root Mean Square Error: The ratio of the root-mean-square-deviation to the standard deviation of the observed data. A dimensionless measure of the precision of fit between observed values and model predictions. A lower RSR indicates a better fit between predicted and observed values.
SAM1	Sediment Assessment Method 1: Bankside visual estimate of % sediment cover. Rapid qualitative assessment of the surface area of the streambed covered by sediment.
SAM2	Sediment Assessment Method 2: In-stream visual estimate of % sediment cover. Semi-quantitative assessment of the surface area of the streambed covered by sediment. At least 20 readings are made within a single habitat.
SAM4 (also SIS)	Sediment Assessment Method 4: Resuspendable sediment (Quorer method). Quantitative measure of total suspendable solids deposited on the streambed. Six samples are collected from a single habitat. Samples are processed in the laboratory for Total Inorganic/Organic Sediment by areal mass and/or Suspendable Benthic Solids by Volume. Also termed Suspendable Inorganic Solids measurement (SIS).
Sediment load	The mass flux of sediment delivered from a catchment (typically in t/yr).
Sediment yield	The sediment load per unit catchment area (typically in $t/km^2/yr$).

SMA (regression)	Standard major axis regression – minimizes the variance of both the X and Y variables, in contrast to ordinary least squares regression which minimizes variance only in Y. Useful when there is no particular reason to treat either one of X and Y as the 'independent' variable.
SOF	Source of Flow category from REC2 (derived for REC1).
SS	Suspended Sediment.
SSC (also C)	Suspended Sediment Concentration: mass of sediment suspended per unit volume of water (units of mg/l or g/m^3 are equivalent), measured by filtration of the *whole* a water sample, in contrast to TSS which is measured by filtration of a subsample.
Strahler stream order	The number of times a channel branches going upstream from a point minus 1.
T (also TURB)	Turbidity of water.
TSS	Total suspended sediment (concentration) – measured by filtration of a *subsample* of a water sample, in contrast to SSC which is measured by filtration of the whole sample. Ideally TSS would equal SSC, but if the subsampling is not representative, typically owing to rapid settling sand, TSS may differ (and be biased).
V (also CLAR)	Visual water clarity – quantified by the black disc visibility (in the horizontal direction).

8 References

Booker, DJ., Snelder, TH. (2012) Comparing methods for estimating flow duration curves at ungauged sites. Journal of Hydrology **434**:78–94.

Booker, D. (2013) Spatial and temporal patterns in the frequency of events exceeding three times the median flow (FRE3) across New Zealand. Journal of Hydrology (NZ) **52** (1): 1-25.

Breiman, L. (1984) Classification and regression trees. Wadsworth International Group, Belmont, Calif.

Breiman, L. (2001) Random Forests. Machine Learning **45**:5–32. Available from https://doi.org/10.1023/A:1010933404324.

Chow, VT., Maidment, DR., Mays, LW. (1988) Applied Hydrology. McGraw-Hill International Editions, New York, 572 p.

Clapcott, JE., Young, RG., Harding, JS., Matthaei, CD., Quinn, JM., Death, R.G. (2011) Sediment Assessment Methods: Protocols and guidelines for assessing the effects of deposited fine sediment on in-stream values. Nelson, New Zealand, Cawthron Institute.

Clapcott, JE., Goodwin, E. (2017) Technical report on developing a deposited sediment classification for New Zealand streams. Prepared for Ministry for the Environment. Cawthron Report No. 2994. 36 p. plus appendices.

Cutler, DR., Edwards, TC., Beard, KH., Cutler, A., Hess, KT., Gibson, J., Lawler, JJ. (2007) Random forests for classification in ecology. Ecology **88**:2783–2792. Available from http://doi.wiley.com/10.1890/07-0539.1.

Davies-Colley, RJ., Hicks, DM., Hughes, AO., Clapcott, J., Kelly, D., Wagenhoff, A. (2015) Fine sediment effects on freshwaters, and the relationship of environmental state to sediment load – a literature review, NIWA, NIWA Client Report HAM2015-104.

Duan, N. (1983) Smearing Estimate: A Nonparametric Retransformation Method. Journal of the American Statistical Association **78**:605–610. Available from http://www.tandfonline.com/doi/abs/10.1080/01621459.1983.10478017.

Dymond, JR., Davies-Colley, RJ., Hughes, AO., Matthaei, CD. (2017) Predicting improved optical water quality in rivers resulting from soil conservation actions on land. Science of the Total Environment 603–604: 584-592.

Franklin, P., Stoffells, R., Clapcott, J., Booker, D., Wagenhoff, A., Hickey, C. (2019) Deriving potential fine sediment attribute thresholds for the National Objectives Framework. NIWA Client Report No: 2019039HN, prepared for Ministry for the Environment, February 2019.

Gadd, J., Snelder, TH., Fraser, C., Whitehead, AL. (2019) Urban river and stream water quality state and trends 2008-2017. NIWA Client Report 2018328AK prepared for Ministry for the Environment. NIWA, Auckland.

Haddadchi, A., Hicks, DM. (2019) Understanding the effect of catchment characteristics on the dynamics and yield of suspended sediments. Geophysical Research Abstracts Vol. **21**, 2019 EGU General Assembly, Vienna, April 2019.

Haddadchi ,A., Booker, D., Measures, R. (2018) Predicting river bed substrate cover proportions across New Zealand. Catena **163**: 130-146.

Hastie, TJ., Tibshirani, R., Friedman, JH. (2001) The elements of statistical learning, data mining, inference and prediction, 2nd edition. Springer-Verlag, New York.

Hicks, DM., Clapcott, J., Davies-Colley, R., Dymond, J., Greenwood, M., Hughes, A., Shankar, U., Walter, K. (2016) Sediment Attributes Stage 1. NIWA Client Report No: CHC2016-058:189 prepared for Ministry for the Environment.

Hicks, M., Semadeni-Davies, A., Haddadchi, A., Shankar, U., Plew, D. (2019) Updated sediment load estimator for New Zealand. NIWA Client Report No. 2018341CH, prepared for Ministry for the Environment, March 2019.

Hicks, DM., Gomez, B., Trustrum, NA. (2004) Event suspended sediment characteristics and the generation of hyperpycnal plumes at river mouths: East Coast continental margin, North Island, New Zealand. Journal of Geology **112**: 471-485.

Larned, ST., Scarsbrook, MR., Snelder, TH., Norton, NJ., Biggs, BJF. (2004) Water quality in low-elevation streams and rivers of New Zealand: Recent state and trends in contrasting land-cover classes. New Zealand Journal of Marine and Freshwater Research **38**:347–366.

Larned, ST., Snelder, TH., Unwin, MJ. (2016) Water Quality in New Zealand Rivers Modelled Water Quality State. NIWA Client Report CHC2016-070 prepared for New Zealand Ministry for the Environment. Christchurch.

Larned, ST., Whitehead, AL., Snelder, TH., Fraser, C., Yang, J. (2018) Water quality state and trends in New Zealand rivers: analyses of national data ending in 2017. NIWA Client Report 2018347CH prepared for the Ministry for the Environment. Christchurch.

Leathwick, JR., Overton, JM., McLeod, M. (2003) An environmental domain classification of New Zealand and its use as a tool for biodiversity management. Conservation Biology **17**:1612–1623. Available from http://doi.wiley.com/10.1111/j.1523-1739.2003.00469.x.

Liaw, A., Wiener, M. (2002) Classification and Regression by randomForest. R News **2**:18–22.

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. Available from https://www.ucn.dk/Files/Billeder/ucn/Uddannelser/Multimediedesigner/Curriculum-Multimedia-Design-and-Communication-National-part-UCN.pdf.

Nash, JE., Sutcliffe, JV. (1970) River flow forecasting through conceptual models part I — A discussion of principles. Journal of Hydrology **10**:282–290. Available from http://linkinghub.elsevier.com/retrieve/pii/0022169470902556.

NEMS (2016) Turbidity Recording: Measurement, processing and archiving turbidity data, Version 1, issued October 2016. National Environmental Monitoring Standards Document, Wellington.

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

R Core Team. (2017) R: A Language and Environment for Statistical Computing. Vienna, Austria. Available from https://www.r-project.org/.

Snelder T, Biggs, BJF., Woods RA. (2005) Improved eco-hydrological classification of rivers. River Research and Applications **21**: 609-628.

Svetnik, V., Liaw, A., Tong, C., Wang, T. (2004) Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. Pages 334–343 in F. Roli, J. Kittler, and T. Windeatt, editors. Multiple Classifier Systems. Springer Berlin Heidelberg, Berlin, Heidelberg.

Unwin, MJ., Snelder, TH., Booker, DJ., Ballantine, D., Lessard, J. (2010) Predicting water quality in New Zealand rivers from catchment-scale physical, hydrological and land cover descriptors using Random Forest models. NIWA Client Report CHC2010-037. Christchurch.

Whitehead, AL. (2019) Spatial modelling of river water-quality state: incorporating monitoring data from 2013 to 2017. NIWA Client Report 2018360CH prepared for the Ministry for the Environment. Christchurch.

Appendix A Random forest models

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

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

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

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

RF models can include any of the original set of predictor variables that are chosen during the model fitting process. Inclusion of marginally important and correlated predictor variables does not degrade the performance of the RF models. However, these predictor variables may be redundant (i.e., their removal does not affect model performance) and their inclusion can complicate model interpretation. We used a backward elimination procedure to remove redundant predictor variables from the initial 'saturated' models (i.e., models that included any of the original predictor variables). The procedure first assesses the model error (MSE) using a 10-fold cross validation process. The predictions made to the hold out observations during cross validation are used to estimate the MSE and its standard error. The model's least important predictor variables are then removed in order, with the MSE and its standard error being assessed for each for each successive model. The final, 'reduced' model is defined as the model with the fewest predictor variables whose error is within one standard error of the best model (i.e., the model with the lowest cross validated MSE). This is equivalent to the "one standard error rule" used for cross validation of classification trees (Breiman 1984).

An alternative approach is to choose the model with the smallest error. We used the former procedure as it retains fewer predictor variables than the latter procedure, while achieving an error rate that is not different, within sampling error, from the "best solution". Importance levels for predictor variables were not recalculated at each reduction step to avoid over-fitting (Svetnik et al. 2004).

We note that, because fitting a RF model involves randomly selecting observations and predictor variables throughout the fitting process, successive models fitted to the same data set will exhibit subtle differences in structure and diagnostics such as total explained deviance, MSE, partial dependence plots, and the order of predictor importance. In the current study, the variability in model error between individual fits of the model for each water quality variable were within the reported model performance (see Section 3.2).

Appendix B Derivation of predictor variables used in fine sediment cover model

The physically based model predicting fine sediment cover is

$$FSC = \underbrace{\overbrace{[T]^{Ct}\,[C]^{Cc}}^{\text{Recession SSC}}\,\overbrace{[LC][ET][\tfrac{Xs}{Xw}]^{Cx}/[USslope]^{Cs}}^{\text{Factors controlling phasing \& duration of sediment delivery}}}_{\text{Supply factor}}\;\overbrace{\left[\frac{u*}{W_s}\right]^{Cu}}^{\text{Trap efficiency}}\;\underbrace{[FRE3]^{Cf}}_{\text{Probability-of-occurrence}}\;\overbrace{[D65_s]^{Cd}}^{\text{Space factor}}$$

(B-1)

where: *FSC* is the temporal median proportion of the streambed covered with fine sediment; *Ct, Cc, Cx, Cd, Cs, Cf,* and *Cu* are coefficients to be solved for; and the other terms are the predictor variables. The derivation of these predictor variables is detailed below.

In Equation (B-1), the supply factor comprises a recession-SSC factor (influenced both by the long-term average load-weighted SSC, *C*, and the median turbidity, *T*) and several factors influencing the phasing and duration of sediment on hydrograph recessions, including the land cover and erosion terrain (which influence the phasing of runoff and the type of erosion process), the relative distance upstream to the centre-of-mass of erosion and of runoff (X_s/X_w), and the runoff "flashiness" (indexed by the average upstream catchment slope, *usSlope*)[23].

The load-weighted suspended sediment concentration (*C*) is the long-term average suspended load divided by the mean annual water discharge. While direct use of the mean annual load is important in the context of the current study, we also considered it likely that recession concentrations might better correlate with the temporal-median turbidity – since the temporal-median turbidity is most likely to occur on a recession – hence we used these two potential terms to index the recession SSC.

The *LC* term in Equation (B-1) incorporates land cover on an area-weighted basis, using a simplified land cover grouping of four classes (*pasture, forest, tussock*, and *other* - as developed by Hicks et al. 2019), thus:

$$LC = CLf \times Af + CLp \times Ap + CLt \times At + CLo \times Ao \qquad \text{(B-2)}$$

Similarly, the *ET* term in Equation (B-1) incorporates erosion terrain on an area-weighted basis, using a simplified 12-class grouping of MWLR's erosion terrain classification (Table B-1, as developed by Hicks et al. 2019), thus:

$$ET = \textstyle\sum_{j=1}^{12} CEj \times AEj \qquad \text{(B-3)}$$

[23] Runoff "flashiness" can be indexed by the "time to peak", T_p, of rainstorm runoff. A common estimator (Chow et al. 1988) relates T_p to the average channel slope upstream, catchment land cover and soil type, and upstream channel length.

Table B-1: Erosion terrain groups used in the FSC model.

Associated model parameter	Erosion terrain group
ET1	Sand country, flood plains, fans and terraces, peat
ET2	Tephra and loess
ET3	Tertiary mudstone, sandstone and soft limestone
ET4	Intensely gullied crushed argillite and greywacke
ET5	East Cape - Intensely gullied crushed argillite and greywacke
ET6	Lavas, rhyolite, volcanic slopes
ET7	Greywacke, argillite and hard limestone
ET8	Schist and South Island greywacke (incl. alpine, ice and snow)
ET9	Coarse crystalline plutonics and metamorphics
ET10	Deeply weathered plutonics
ET11	Water
ET12	Other

For the X_s/X_w parameter, X_s was calculated as:

$$X_s = \sum_{i=1}^{n}(x_i L_i) \,/\, \sum_{i=1}^{n} L_i \tag{B-4}$$

where L_i is the locally-supplied sediment mean annual sediment load to the ith segment upstream from the target segment and x_i is the distance upstream to the ith segment from the target segment.

Similarly, X_w was calculated as:

$$X_w = \sum_{i=1}^{n}(x_i Q_i) \,/\, \sum_{i=1}^{n} Q_i \tag{B-5}$$

where Q_i is the locally-supplied mean discharge to the ith segment upstream from the target segment and x_i is the distance upstream to the ith segment from the target segment.

The <u>trap-efficiency</u> factor, u^*/W_s (often called the "Rouse Number"), is effectively a measure of the relative magnitude of upward turbulent velocity fluctuations (which are indexed by the shear velocity, u^*) and the fall velocity of sediment in suspension (W_s). We calculated u^* as:

$$u^* = 1.91 {Q_{mean}}^{0.12} {S_{local}}^{0.5} \tag{B-6}$$

which is based on $u^* = (gYS_{local})^{0.5}$, where g is the gravitational acceleration, S_{local} is the local segment slope, and Y is mean depth. Equation (B-6) was solved at mean flow conditions using Jowett's (1998) downstream hydraulic geometry function for mean depth at mean flow. We had no national model available with which to predict W_s, therefore we assigned it a "neutral" value of 1 m/s; however, we expect that the dominant erosion terrain parameter should capture this (since the erosion terrain groups are essentially lithology based) – as was observed by Hicks et al. (2016) from a relatively small dataset of measured suspended sediment size gradings.

The probability-of-occurrence factor was indexed by the *FRE3* parameter, which is the average number of runoff events per year that exceed 3x the median flow and has been shown by several studies to be a useful predictor of event based environmental processes (e.g., Booker 2016; Hoyle et al. 2017). Note that the inverse of this indexes the average time between such events – or event intermittency.

We represented the space factor by the 65th percentile bed-surface material size, which was predicted nationally by Haddadchi et al. (2018).

Appendix C Results using alternative turbidity and clarity bottom lines

Introduction

An updated set of C/D band thresholds for median clarity and turbidity (Table C-1) was provided by MfE in May 2019. This appendix assesses which river segments have median turbidity and median clarity exceeding these new thresholds and updates the estimated reduction in mean annual up-catchment suspended sediment load required to achieve the turbidity thresholds.

Note that these thresholds are generally more “relaxed” than those used previously (i.e., the turbidity thresholds are higher, the clarity thresholds are lower).

Table C-1: C/D band thresholds for median turbidity and visual clarity for Level 4 suspended sediment classification, issued May 2019. Source: MfE, adapted from Franklin et al (2019).

Class	Turbidity threshold (NTU)	Visual clarity threshold (m)
1	5.51	0.9
2	5.51	0.9
3	7.75	0.71
4	7.75	0.71
5	12.27	0.45
6	12.27	0.45
7	4.09	1.16
8	12.27	0.45
9	4.29	1.27
10	4.09	1.16
11	4.09	1.16
12	5.51	0.9

Which river segments exceed the new C/D band thresholds for turbidity and clarity?

Threshold-exceeding river segments around the country were identified by comparing the new threshold values (Table C-1Table C-1: C/D band thresholds for median turbidity and visual clarity for Level 4 suspended sediment classification, issued May 2019.) with the predicted median clarity and turbidity values from the national models developed in Section 3.3.

Over all 593551 segments:

- 4.0% exceed the clarity threshold
- 4.7% exceed the turbidity threshold
- 2.7% exceed both thresholds, and

- 6.1% exceed one or other or both thresholds.

Note that these percentages are approximately one quarter of those obtained using the earlier set of thresholds – so with the updated thresholds, substantially fewer segments require sediment load reduction.

These exceedances are shown broken-down by suspended sediment CTG class in Figure C-1. Classes 1, 2, 5 and 7 dominate the segment totals. Classes 1 and 10 show good concordance between turbidity and clarity (i.e., a large proportion of segments have both turbidity and clarity thresholds exceeded – as indicated by the relative height of the striped bars in the lower plot). Classes 2, 3 , and 6 are dominated by the turbidity thresholds, while Classes 4 and 8 are dominated by the clarity thresholds. Note that Class 8 has only one segment exceeding the turbidity threshold and 13 exceeding the clarity threshold, while Class 3 has only 16 segments exceeding the turbidity threshold and none exceeding the clarity threshold.

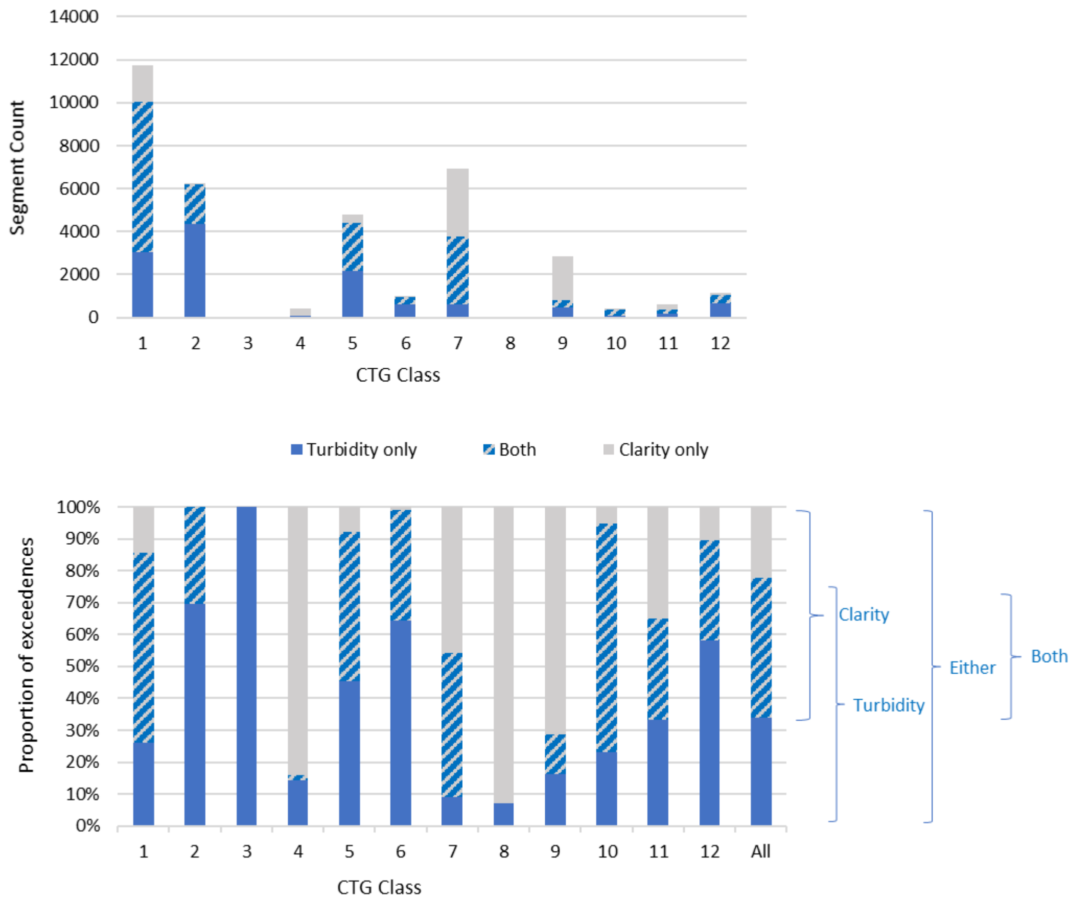

Figure C-1: Count and proportion of stream segments exceeding new clarity and turbidity thresholds by CTG class. The total bar height gives the number or proportion of segments exceeding either threshold; the striped bar indicates where both are exceeded.

Figure C-2 shows the threshold exceedance breakdowns by land cover. The top plot shows that the overwhelming majority of threshold exceedances (blue/grey bars) occur in segments with pasture dominated catchments. This partly reflects the prevalence of pasture land cover across the country (yellow bars) and partly the proportion of pasture segments exceeding the threshold (lower plot). The lower plot shows that only very small proportions (<1%) of the indigenous forest, tussock, scrub, and (surprisingly) bare ground segments exceed the thresholds, which is why they barely show on the top plot despite having significant national coverage (yellow bars). These land covers, along with exotic forest, all have lower threshold exceedance percentages by class compared to the national average of 6.1% (i.e., exceeding either threshold). In contrast, pasture, miscellaneous/mangroves, urban, and wetland dominated catchments have proportional exceedances higher than the national average. Turbidity threshold exceedances prevail in most land covers (i.e., blue bars are higher than the grey bars in lower plot in Figure C-2).

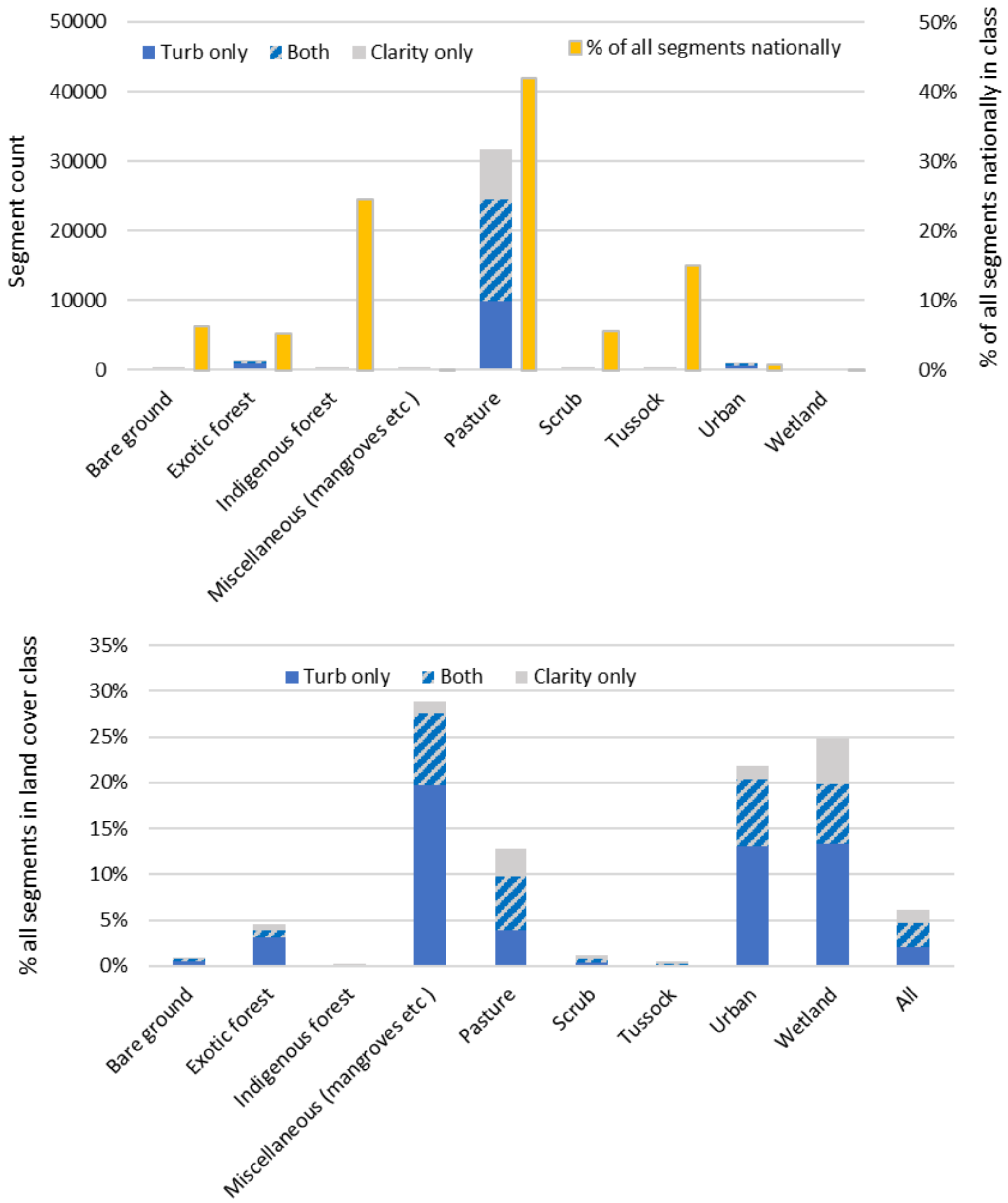

Figure C-2: Count and proportion of stream segments exceeding new clarity and turbidity thresholds by land cover. On top plot: the total blue/grey bar height gives the number of segments exceeding either or both thresholds; the striped bars indicate where both are exceeded; the yellow bars show the national proportions by land cover of all segments. On bottom plot: the blue/grey bar heights give the proportions of all segments in the land cover class that exceed the turbidity or clarity thresholds.

What load reduction is required to achieve the new C/D band thresholds for turbidity?

We estimated the load reduction factors for turbidity (R_t) across all 593551 segments of the REC2 digital network as detailed in Section 3.4.3. We note that $R_t = (L – L_{tt})/L$, where L is the actual sediment load and L_{tt} is the target sediment load that just meets the turbidity threshold.

As before, the calculated R_t-values were grouped into 5 classes:

- $0<R_t<0.2$
- $0.2<R_t<0.4$
- $0.4<R_t<0.6$
- $0.6<R_t<0.8$, and
- $0.8<R_t<1$.

Figure C-3 shows the distributions of the R_t classes by stream order, CTG class, region, and dominant land cover. In summary:

- By stream order, most orders (order 1-7) are dominated by relatively low R-values (< 0.4); order 8 segments have a higher proportion of moderate R-values (0.4<R<0.6) but fewer low R-values and no high R-values (R>0.6).
- By CTG class, low R-values (< 0.4) dominate most CTG classes except 2 and 10.
- By region, low R-values (< 0.4) dominate most regions except for Manawatu-Whanganui, Marlborough, Waikato and Northland. Note no turbidity threshold exceedances in Tasman-Nelson and Stewart Island.
- By dominant land cover, low R-values (< 0.4) prevail except in the miscellaneous class, which includes mangroves.

Comparison with R results using observed data

Figure C-4 compares the R-class distribution of segments where R_t has been calculated off the observed median turbidity and off the predicted median turbidity, as well as the distribution from all predicted segments. The observed dataset (which is the one used to train the predictive model) has 847 segments. Of these, 128 (15.1%) exceeded the new turbidity thresholds (i.e., $R_t >0$), while in the matching predicted dataset 124 (14.6%) exceeded these thresholds – a close agreement. The matching predicted and observed distributions were reasonably similar across the R-classes, which provides reassurance around the model predictions. Compared with the observed dataset, the full predicted dataset had relatively similar proportions of segments with low R-values but relatively more segments with high R-values.

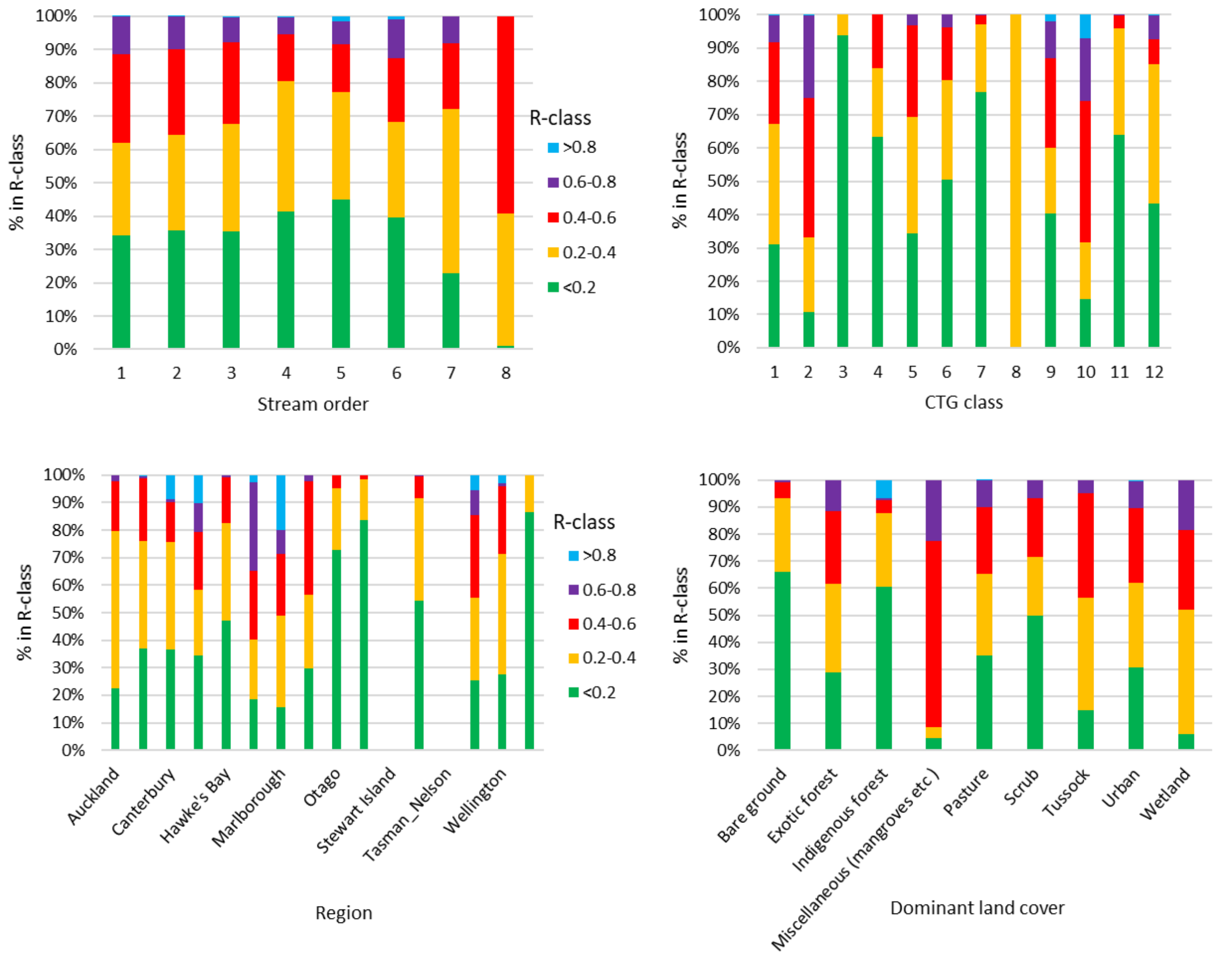

Figure C-3: R_t-class breakdown by stream order, Climate-Topography-Geology (CTG) class, region, and dominant land cover.

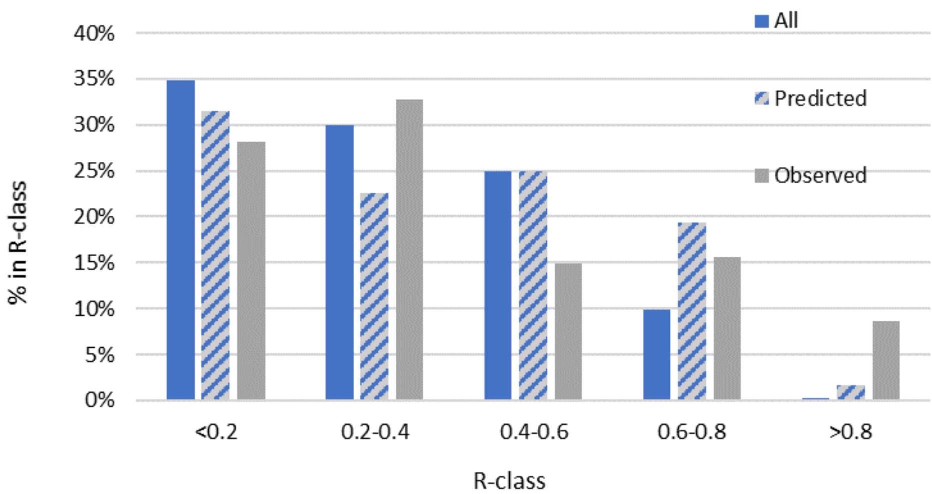

Figure C-4: Distribution of segments with R_t> 0 by R_t-class for all predicted segments, segments with R calculated off observed turbidity data, and matching segments with R calculated off predicted turbidity.

Catchment-scale results from national models

As in Section 3.4.7, pour-point catchments (defined as those parts of a coast-draining catchment upstream from the first segment where a $R_t > 0$ value was encountered) were mapped for the updated set of turbidity thresholds.

Result-files forwarded to MfE included:

- shape-files of the new set of pour-point catchments, including their boundaries and summary statistics in an attribute table (including count and proportion of R-values > 0, and average R-values)
- shape-files of all REC2 segments, with an attribute table including values of R computed off turbidity (R_t), computed off clarity (R_c),and computed as the maximum of R_t and R_c (R_{max}).