MC10

Assessment of recent reductions in *E. coli* and sediment in rivers of the Manawatū-Whanganui Region

Including associations between water quality trends and management interventions

February 2018

Prepared By:

Ton Snelder

For any information regarding this report please contact:

Ton Snelder

Phone: 03 377 3755

Email: ton@lwp.nz

LWP Ltd
PO Box 70
Lyttelton 8092
New Zealand

LWP Client Report Number:

Report Date: January 2018

LWP Project: LWP Client Report 2017-06

Quality Assurance Statement

[Click here and type text]

Version	Reviewed By	
1	John Quinn (NIWA)	

Table of Contents

Abstract

This study investigated state and trends in river water quality measures that indicate human health risk and sediment contamination in rivers of the Manawatū-Whanganui region. The water quality measures included one indicator of human health risk; the concentration of *Escherichia coli* (*E. coli*), and three measures of sediment contamination; visual clarity suspended solids concentration (SSC) and turbidity. The water quality data was derived from monitoring carried out over the past decade by Horizons Regional Council (HRC) at river sites distributed throughout the region.

The study investigated the following key areas:

1. Comparison of the national swimming map, which was derived from a large national dataset of water quality monitoring sites, with a regional swimming map derived from a smaller number of regional state of environment (SoE) sites.
2. Assessment of trends in three *E. coli* statistics (median and the proportion of samples exceeding 260 and 540 *E. coli* $100mL^{-1}$; referred to as the G260 and G540), and the three sediment related water quality variables at SoE sites over the 10 and seven-year time-periods ending in 2016.
3. The associations between the water quality trends and several interventions that HRC have initiated aimed at improving water quality including upgrading point source discharges and land management initiatives.

The spatial patterns in swimming grades defined by the national and regional swimming maps were reasonably consistent. The national map indicated that 45% of the region's large rivers (i.e., river segments or order 4 and greater) are swimmable (grade 'fair' or better). The regional swimming map indicated that 38% of large rivers are swimmable. When all rivers (small and large) were considered, the national and regional swimming maps indicated 36% and 37% of rivers are swimmable respectively. The study showed that the models that underlie the swimming maps are sensitive to the input data and that the swimming grades at individual sites are sensitive to the time-period of analysis. Therefore, swimming maps and the estimated quantity of swimmable rivers should be regarded as indicative.

The proportion of 10-year trends at SoE sites that were at least as likely as not to be improving were 65%, 81% and 80% for median *E. coli*, G260 and G540 respectively. For the seven-year time-period, the proportion of trends at SoE sites that were at least as likely as not to be improving were 72%, 91%, 81%, 78%, 99% and 95% for median *E. coli*, G260, G540, clarity, SSC and turbidity respectively. Thus, the trend analyses provide strong statistical evidence of regional improvement in the water quality measures over the past decade.

Several independent analyses undertaken by the study found weak but statistically significant positive associations between improving trends for all water quality variables and the proportion of catchment area involved in sustainable land use initiative (SLUI) farm plans. The associations between trends and HRC riparian planting and new fencing initiatives were less clear with some analyses indicating the initiatives were associated with degradation and others with improvement. These results cannot prove that the interventions caused the water quality changes and may be confounded by other factors such as changes in land use intensity and climate. Overall, the results provide weak evidence that a package of mitigation measures implemented at many locations across a region can produce regional scale water quality improvements. By contrast, the analysis provided strong statistical evidence of water quality improvements associated with upgrading point source discharges throughout the region.

Executive Summary

This study investigated changes in river water quality measures that indicate human health risk and sediment contamination in rivers of the Manawatū-Whanganui region. The water quality measures included one indicator of human health risk; the concentration of *Escherichia coli* (*E. coli*), and three measures of sediment contamination; visual clarity as measured by black disc (clarity) suspended solids concentration (SSC) and turbidity. The water quality data was derived from river water quality monitoring carried out over the past decade by Horizons Regional Council (HRC) at sites distributed throughout the region.

Over the past 10 years HRC have initiated a range of interventions aimed at improving water quality in the region including upgrading discharges and land management initiatives. A geographic database describing the location of these interventions had been maintained by HRC and this data was also used in this study.

The contamination of freshwater with pathogens and sediment is a national water quality issue due to the impacts on human use and ecological values. The National Objectives Framework (NOF) of the National Policy Statement – Freshwater Management (NPS-FM) defines 'swimming grades' from excellent (A) to poor (E). In addition, 'national swimming maps', showing estimated swimming grades for all large rivers in New Zealand, were used to set targets for improving the total length of 'swimmable' freshwaters nationally. The NOF does not currently define objectives related to sediment contamination. However, sediment attributes are under development. In addition, management of sediment contamination of freshwaters is a requirement of the NPS-FM irrespective of the existence of specific sediment-related NOF attributes due to sediment's impacts on compulsory values. Assessment of the state of freshwaters with respect to human health risk and sediment impacts and evaluation of interventions aimed at improving water quality are therefore of interest both nationally and regionally.

This regional case study investigated the following areas:

1. Comparison of the national swimming map, which was derived from a large national dataset of water quality monitoring sites, with a regional swimming map derived from a smaller number of regional state of environment (SoE) sites. The comparison included differences between the swimming maps that include only large rivers (order 4 and greater) and including smaller rivers and streams (order 1, 2 and 3).

2. Comparison of regional swimming maps derived using year-round data compared to maps derived using data pertaining to the 'bathing season' (November to March).

3. Trends in *E. coli* and three sediment related water quality variables (visual clarity, suspended sediment concentration and turbidity) over the 10 and seven-year time-periods ending in 2016.

4. The associations between water quality interventions and water quality trends.

The majority of these investigations were conducted for a subset of the monitoring sites that can be regarded as state of environment (SoE) sites. Water quality measured at SoE sites is broadly representative of conditions in the region and have been regularly measured for a number of water quality variables. For the 10-year time-period, 69 sites had adequate data and were included in the analysis and for the 7-year time period up to 86 sites were included. The main outcomes of these investigations are summarised below.

1. <u>National versus regional swimming maps</u>

The national swimming map indicated that 45% of the region's large rivers (i.e., river segments or order 4 and greater) are swimmable (grade 'fair' or better). The regional swimming maps indicated that 38% of large rivers are swimmable. When all rivers (small and large) were considered, the national swimming maps indicated 36% of rivers are swimmable and the regional maps indicated 37% are swimmable.

The spatial patterns in swimming grades defined by the national and regional swimming maps were reasonably consistent. Differences between the swimming maps indicate that the underlying spatial models are sensitive to the input data including the numbers and combinations of sites that are included. In addition, it was shown that the swimming grade for a site can differ depending on the time-period of the assessment. All swimming maps are therefore only indicative of the general pattern of human health risk associated with rivers. Modelled swimming grades shown for specific locations on the maps are uncertain. The most reliable grades are those derived for specific SoE sites from monitoring data. The estimated quantity of swimmable rivers or rivers belonging to a specific grade as represented by swimming maps should be regarded as indicative.

2. Year-round versus bathing season data

Swimming grades for individual SoE sites derived from 10 years of year-round data were compared with grades defined from the same time-period but restricted to data for the bathing season (summer months). These comparisons indicated that a larger proportion of sites were swimmable (grade 'fair' or better) during the bathing season than year-round (59% versus 55%). However, small headwater rivers (order 1, 2 & 3) tended to have lower swimming grades (i.e., less suitable) in the bathing season compared to their year-round grades. By contrast, large rivers (order 4 and greater) tended to have better swimming grades in the bathing season compared to their year-round grades. The reasons for these differences were not investigated.

Because of the generally poorer swimming grades for small rivers during the bathing season, the bathing season swimming map indicates that only 17% of all rivers (by length) are swimmable. However, the bathing season swimming map indicates that 36% of large rivers are swimmable, which is consistent with the year-round regional swimming map.

3. Water quality trends

The study analysed trends for the seven and 10-year time-period in three *E. coli* statistics that are used to define swimming grades: the median and the proportion of samples exceeding 260 and 540 *E. coli* $100mL^{-1}$ (referred to as the G260 and G540). The trend in the median value was evaluated by conventional trend analysis, which assessed the change through time of monthly *E. coli* samples. The trend in G260 and G540 was evaluated by calculating the annual values in these statistics for each site and assessing the change in these values through time. Because the G260 and G540 trends were based on annualised values, the analyses had lower statistical power than the assessments of trends in the median. Additional trend analyses were performed for the seven-year dataset on measured monthly data describing three water quality variables that reflect sediment contamination: visual clarity, suspended sediment concentration and turbidity.

When considered on an individual site basis, a large proportion of trends in all four variables for both time-periods were uncertain (i.e., trend direction misclassification error risk > 5%). For example, median *E. coli* trends for the 69 sites analysed over 10-year time-period, were uncertain at 75% of sites and were improving and degrading at only 14% and 10% of sites respectively (i.e., trend direction misclassification error risks < 5%). Consistent with the lower statistical power for the G260 and G540 measures of *E. coli*, the 10-year period trends were improving at 10% and 6% of site respectively and were degrading at 0% and 4% of sites

respectively. There were similarly high proportions of uncertain trends for all variables for the seven-year time-period.

When the traditional confidence level of 95% was relaxed there was a clear pattern of improving trends across all variables for both time periods. For example, the proportion of 10-year trends that were at least as likely as not to be improving were 65%, 81% and 80% for median *E. coli*, G260 and G540 respectively. For the seven-year time-period, the proportion of trends that were at least as likely as not to be improving were 72%, 91%, 81%, 78%, 99% and 95% for median *E. coli*, G260, G540, clarity, SSC and turbidity respectively. Thus, the trend analyses of SoE sites provide strong evidence of regional improvement in the water quality measures over the past decade.

Spatial modelling was used to determine associations between trend direction and catchment characteristics. The direction of trends at all SoE sites was used to inform these models irrespective of the misclassification error risk. The logic for this is that over many sites, incorrect classifications of direction cancel (i.e., as many sites will be misclassified as increasing, as sites misclassified as decreasing). The spatial models for all water quality variables over both time periods were consistent in associating the highest probability of improving trends with hill-country catchments of moderate size (area ~500 km^2) and dominated by soft sedimentary geology and pastoral land cover. This is evidence for positive benefits associated with HRC's interventions because farms in the types of catchments that the association describes have been targeted for interventions aimed primarily at reducing erosion by the sustainable land use initiative (SLUI). Since 2006, 683 whole farm plans (on 'SLUI farms') have been developed, with approximately 80–85% having implemented some on-the-ground works to control or mitigate erosion and sediment loss.

To estimate the change in swimming grades across all rivers through time, the spatial models of trend direction were combined with the spatial models that underlie the regional swimming maps. The combination of models indicated that over the 10-year period the region's rivers in the swimmable category (grades 'fair' to 'excellent') increased from 35% to 40% (a 5% improvement). Large rivers (order 4 and above) increased from 33% to 43% (a 10% improvement). For the seven-year period, modelling indicated the region's rivers in the swimmable category increased from 35% swimmable to 40%, (a 5% improvement). Over the seven-year period, large rivers (order 4 and above) increased from 36% to 44% (an 8% improvement).

All spatial modelling carried out by this study was associated with large site-scale uncertainties. In addition, the predicted increases in swimmable rivers was based on the combination of two sets of spatial models, for which the combined error could not be quantified. Therefore, the swimming grade maps and changes in swimming grades produced by this study should be regarded as indicative.

4. <u>Association between water quality trends and interventions</u>

HRC have implemented a range of water quality interventions aimed at improving water quality in the region since 2004. This study investigated the relationship between those interventions and trends at SoE sites in median *E. coli*, G260 and G540 for the 10-year period and median *E. coli*, G260, G540, clarity, SSC and turbidity for the seven-year period. The analysis of associations was possible because HRC had maintained records of the actions that included the geographic location. This highlights the value of not only water quality monitoring, but also monitoring and recording management actions.

The results showed weak but statistically significant associations between improving trends for all water quality variables and the proportion of catchment involved in SLUI farm plans.

There were also significant associations between improving water quality and additional HRC initiatives associated with riparian planting and new fencing.

The statistical models that were used to test the association between water quality improvement and interventions controlled for the land areas that were subject to erosion in 2004. This indicates that improving water quality is not only associated with natural processes of recovery from the events in 2004.

HRC water quality monitoring data includes sites that represent effluent discharges directly (discharge sites) and the in-river impact of these discharges some distance downstream of the discharge point (impact sites). Analysis of trends at paired discharge-impact sites indicated that trend directions at impact sites were significantly associated with trend directions at discharge sites. In addition, the analysis indicated a strong regional pattern of water quality improvement associated with point source discharges. Overall, the analysis provided strong statistical evidence of water quality improvements associated with upgrading point source discharges throughout the region.

The analysis of the associations between trends and interventions is based on correlations and cannot prove that the interventions caused the water quality changes. The associations may be confounded by other factors such as changes in land use intensity and climate (which the study showed had varied over the last decade). However, several independent analyses undertaken by the study found associations between trends and interventions and that the water quality changes are consistent with expectations. The study therefore provides weak but positive evidence that HRC's interventions have contributed to the observed improvements and suggests that a package of mitigation measures implemented at many locations across a region can produce regional scale water quality improvements.

1 Introduction

The contamination of freshwater with pathogens and sediment is a national water quality issue due to the impacts on human use and ecological values. The National Policy Statement – Freshwater Management (NPS-FM (Ministry for the Environment, 2017a) has changed the regulations associated with managing human health values in freshwater. The changes include introducing a new attribute in the National Objectives Framework (NOF) that defines 'swimming grades' from excellent (A) to poor (E). In addition, 'national swimming maps', showing estimated swimming grades for all large rivers in New Zealand, were used to set targets for improving the total length of 'swimmable' freshwaters nationally. The NOF does not currently contain attributes related to sediment contamination but potential sediment attributes have been under active development since 2014 (Ministry for the Environment, 2015). Due to its impact on compulsory values, sediment contamination of freshwaters is a requirement of the NPS-FM irrespective of the existence of specific sediment-related NOF attributes. Assessment of the state of freshwaters with respect to human health risk and sediment impacts and evaluation of interventions aimed at improving water quality are therefore of interest both nationally and regionally.

River water quality is measured at long term state of environment (SoE) monitoring sites at approximately 800 locations nationally. The measurements are used to quantify the current state and changes in state over time (trends) (e.g., Larned *et al.*, 2015). Human health risk is indicated at these sites by the concentration of *Escherichia coli* (*E. coli*). *E. coli* is an indicator of human or animal faecal contamination and the risk of infectious human disease from waterborne pathogens. Statistics that are derived from *E. coli* data are used to define swimming grades, which indicate the level of risk to humans who use the river for recreation (e.g., swimming) and other activities (e.g., kai gathering). The concentration of suspended sediment determines several aspects of water quality that have implications for both human use and ecological values. Suspended sediment can affect organisms living in the water directly by clogging gills and causing abrasion. In addition, suspended sediment alters the optical properties of water resulting in changes in colour, light penetration and visual clarity. The state of water quality with respect to sediment contamination at SoE sites is generally indicated by measurements of suspended sediment concentration (SSC), visual clarity (clarity) and turbidity. The state and trends indicated by *E. coli*, clarity, SSC and turbidity at a site are relevant to water quality management, including understanding the impact of land use intensification and improvement in conditions that may arise from interventions (i.e., mitigations).

Horizons Regional Council (HRC) undertakes river water quality monitoring at 224 sites in the Manawatū-Whanganui region. In addition, over the past 10 years HRC have initiated a range of interventions aimed at improving water quality in the region including upgrading discharges and land management initiatives. It is expected that these interventions will have improved water quality including reducing *E. coli* concentrations and reducing sediment discharges. The combination of ongoing monitoring and the interventions provided the opportunity to quantify the changes to water quality and asses the effectiveness of management. The information is of interest to the region and can also provide insights for national water management

This study investigated improvements in river water quality measures that indicate sediment contamination (clarity, SSC and turbidity) and human health risk (*E. coli* concentration). There were three key aspects to the work. First, the state of the rivers with respect to the four water quality measures was assessed. Spatial modelling was undertaken that allowed the length of

rivers in each swimming grade to be quantified. This type of spatial model was used to produce national swimming maps that informed the policies and national targets that are now part of the NPS-FM (Ministry for the Environment, 2017a). Use of modelling to transform site scale water quality measurements to regional estimates of state will be integral to reporting on current conditions in the future and assessing progress toward targets. This study has considered and further developed the modelling techniques that were used in the production of the national swimming maps and has made some specific recommendations concerning the assessment of swimming grades at the regional scale. There were three questions concerning the national swimming maps that this study aimed to answer:

- Do estimates of swimming grades in the region made using the national swimming maps agree with data and models that are specific to the Region?
- How do results differ for rivers of order four and greater compared to rivers of all orders?
- How do the river swimming grades calculated from monitoring data collected in all seasons compare to grades that are assessed using only data pertaining to the summer 'swimming' season (November to March)?

The second aspect of the work involved using trend analyses to assess changes in the river water quality measures that indicate sediment contamination (clarity, SSC and turbidity) and human health risk (*E. coli* concentration) in the Manawatū-Whanganui region. The analyses extended the use of statistical spatial modelling to the spatial modelling of trends and used this to develop methods to estimate spatial changes in water quality across the Region. In addition, this study considered issues associated with flow adjustment of water quality data and made some specific recommendations concerning future trend analyses.

The third aspect of the work involved analyses of the association between river water quality trends and interventions that have occurred in the region over the last decade.

2 National swimming grades and maps

2.1 Swimming grades

The NOF attribute table for *E. coli* in the NPS-FM (Ministry for the Environment, 2017a) defines swimming grades for water quality monitoring sites. The swimming grade provides an assessment of the average level of risk to human health associated with swimming (immersion) at a site. The concentration of *E. coli* has been linked to the risk of infection by the pathogen *Campylobacter* through a quantitative microbial risk assessment (QMRA: (McBride, 2016; Ministry for the Environment, 2017b). The actual level of risk on a particular day is quantifiable by a sample on that occasion. This NPS-FM establishes the requirement to carry out surveillance monitoring of this risk at primary contact sites. The NOF does not define attributes for any water quality measures that are related to sediment; however, NOF attributes for sediment are currently being developed (Ministry for the Environment, 2015). Although there are guideline values for sediment related water quality measures (e.g., (MFE, 1994), these do not have the statutory significance of NOF attributes. Therefore, in this study observed or modelled values of clarity, SSC or turbidity have not been expressed as grades.

The NPS-FM human health for recreation attribute table defines the swimming grade at a site based on four statistics derived from *E. coli* measurements: median, percentage of exceedances over 540 *E. coli* $100mL^{-1}$, percentage of exceedances over 260 *E. coli* $100mL^{-1}$, and the 95th percentile. Thresholds for each statistic are associated with a category from A (Excellent) to E (Poor) (Table 1). These thresholds are associated with the level of risk of Campylobacter infection (). The swimming grade for a site is the lowest grade indicated by the individual statistics. Each grade indicates the site's average level of risk (Table 1; (Ministry for the Environment, 2017b).

The 95th percentile is estimated with lower precision than the other three statistics. This imprecision cannot be reduced because it is inherent to the available data and varies between sites in association with the level of variability in the individual *E. coli* measurements. The imprecision affects the robustness of swimming grade assessments, particularly those that use spatial models to estimate grades at unmonitored locations (Stats NZ, 2017). A precisely measured 95th percentile value is consistent with the average level of risk indicated by the other three statistics, but an imprecise measurement may result in an erroneous allocation of a site to a swimming grade. Because this study used spatial modelling to make region-wide assessments of swimming grades, the 95th percentile statistic was not used to assess swimming grades.

Table 1. The statistical measures used to define the swimming grades in this study. The grade that applies at a site is the lowest category over the three statistics. Note that the fourth NPS-FM criteria (95th percentile) was excluded.

Category	Colour	Median *E. coli* 100 mL^{-1}	Exceedance of *260 E. coli* 100 mL^{-1} (G260)	Exceedance 540 *E. coli* 100 mL^{-1} (G540)	Average risk of campylobacter infection
A (Excellent)	Blue	≤130	≤0.2	< 0.05	<1%
B (Good)	Green	≤130	0.2 – 0.3	0.05-0.1	<2%
C (Fair)	Yellow	≤130 or less	0.2 – 0.34	0.1 - 0.2	<3%
D (Intermittent)	Orange	> 130	0.3 – 0.5	0.2 – 0.3	>7%
E (Poor)	Red	> 260	> 0.5	> 0.3	>12%

2.2 Definition of national swimming maps

National swimming maps for rivers are based on the *E. coli* regression modelling approach outlined in (Snelder *et al.*, 2016a) and in section 4.5.1 of this report. Separate models were constructed for each of the three statistics outlined in Table 1 and were used to predict the values of each statistic for each segment of a digital representation of the national river network[1]. The grades shown on the swimming maps were derived by categorising segments for each of three statistics according to the criteria in Table 1 and assigning the segment's swimming grade as the lowest category. Some adjustments were made to the assessed grades at segments that were associated with monitoring sites to ensure the map was consistent with measured values at those locations (see MFE, 2017, Swimming Committed Work Report for details).

[1] River Environment Classification version 1

3 Data

3.1 Water quality data

Of the 231 water quality monitoring sites operated by HRC, 143 can be regarded as state of environment (SoE) sites, meaning their water quality is broadly representative of conditions in the region and have been regularly measured for a number of water quality variables (Figure 1). Discharge sites represent treated wastewater prior to it being discharged to the river. Discharges are monitored at 35 sites across the region (Figure 1). River water quality downstream of point sources is monitored at 53 'impact' sites (Figure 1). Many of the discharge sites are upstream of (and therefore linked to) impact sites.

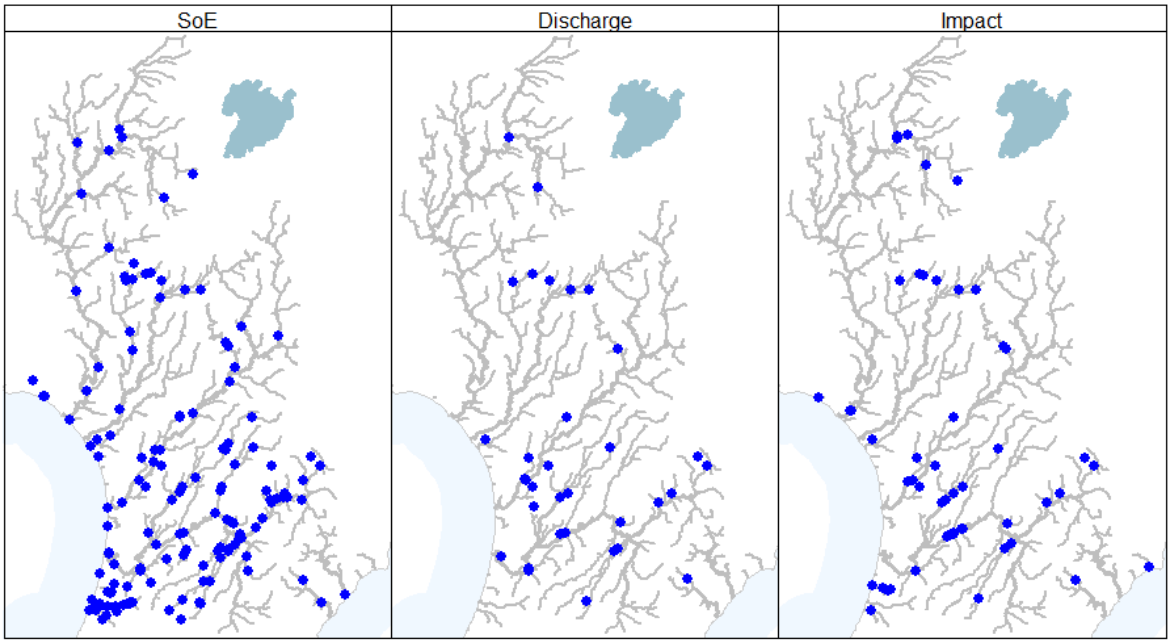

Figure 1. Map showing location of the river water quality monitoring sites in the Region. Grey lines represent main stem rivers (stream order of 4 or greater).

E. coli clarity (measured in the field using a black disc), turbidity, and suspended sediment concentrations (SSC) have been measured at130 SoE, 30 discharge and 48 impact sites in the region (Table 2). All *E. coli* measurements in the dataset used in this study were analysed using the maximum probably number (MPN) method. Turbidity was measured in the laboratory using three types of method; EPA (26,000 data points), ISO FNU (8,507 data points) and ISO NTU (8,343 data points). In this study it has been assumed that all turbidity measurements are commensurate. The laboratory analyses of SSC have varied over time and include the following methods; APHA (2005) 2540 D, APHA 21st Edition Method 2540 D modified and ASTM D3977-97. In this study it has been assumed that all SSC measurements are commensurate.

Data for these sites were obtained from the HRC water quality database. *E. coli* was included in this study because it is a measure of human health risk and statistics derived from *E. coli* measurements are the basis for swimming grades. Clarity, turbidity and suspended sediment concentration (SSC) are included primarily because they represent the effects of sediment contamination on water quality. Sediment enters freshwater via runoff processes in a similar way to *E. coli* (McDowell *et al.*, 2008). Therefore, inclusion of the sediment-related variables provides additional evidence of water quality changes associated with human health risk and

their association with management (including interventions). The sediment-related variables also have relevance to the suitability of water for human contact recreation, however they are not used in the assessment of swimming grades.

Table 2. Water quality variables included in this study.

Variable	Abbreviation	Units
Escherichia coli	*E. coli*	MPN 100 mL^{-1}
Visual clarity (black disc)	Clarity	m
Turbidity	Turbidity	NTU
Suspended sediment concentration	SSC	g m^{-3}

The water quality data included the site name, date, water quality variable and the measured value. For all variables, some true values were too low or too high for laboratory or field-based methods to measure with precision. These measurements are called censored values. For very low values of a variable, the minimum measurement with acceptable precision corresponds to the “detection limit”. For very high values of a variable, the maximum measurement with acceptable precision corresponds to the “reporting limit”. The data included measurements that were below detection and above reporting limits for *E. coli*, clarity and turbidity. Cases in the HRC dataset where values of variables were below the detection limit or above the reporting limit were indicated by the prefix to recorded values “<” and “>” respectively.

The duration of sampling across all SoE sites and variables in this study varied between 1 and 25 years (Figure 2). Sampling start years were variable between sites. Sampling of clarity and *E. coli* started at many sites in 2005 and 2006, but most turbidity and SSC sampling did not start until 2010 (Figure 2). The total number of samples varied between SoE sites, partly reflecting variation in the number of years that the variables had been measured and partly in association with differences in sampling frequency. Although SoE sites tended to have monthly measurements for most variables, there were often multiple measurements of variables at a site within months. Most SoE sites had a low proportion of samples having censored values for turbidity and *E. coli* (Figure 2). The proportion of censored values was generally higher for clarity and SSC for which there were sites with more than 10% of values being censored (Figure 2).

River water quality can be strongly associated with flow, and this can influence trend analysis. The effect of flow can be accounted for in analysis of trends by flow adjustment (see Section 4.4.4). Most of the 130 SoE water quality monitoring sites (between 97 and 130 sites depending on variable) had flow records for at least some sample occasions; these data were obtained from HRC. Of the 130 SoE sites, between 14 and 25 had flow measurements corresponding to every sample, and between 28 and 40 sites had flow data for >80% of samples, depending on the variable. Flows for each water quality monitoring site were either measured at a gauge that was located at, or close to, the monitoring sites or at a ‘proxy’ gauge that may be located some distance away.

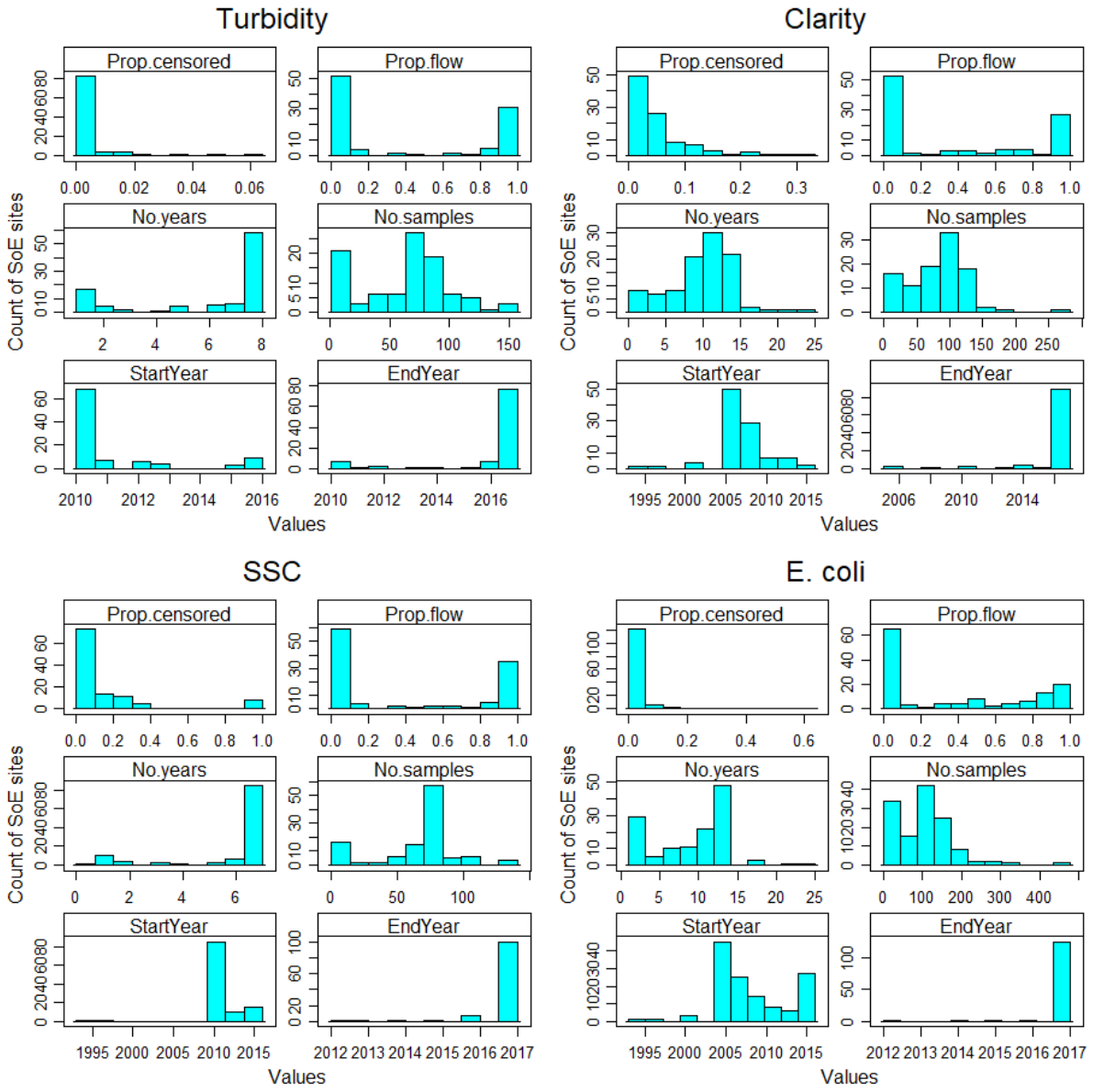

Figure 2. Histograms describing the available data for the 130 SoE river water quality monitoring sites.The histograms describe, for each variable, the number of measurements of each variable (No.Samples), the proportion of censored values (Prop.censored), the proportion of samples with associated measurement of flow (Prop.flow), the duration of the sampling period (No.years), the start and end year of the samples (StartYear, EndYear).

3.2 Data describing interventions and potential covariates

3.2.1 Sustainable land use initiative

The Manawatū-Whanganui region experienced a major storm event in February 2004. Rainfall exceeded a 150-year return period over the mid-catchment hill country areas of much of the region resulting in widespread erosion and flooding (Schierlitz and Dymond, 2006). The event prompted the Sustainable Land Use Initiative (SLUI), which seeks to implement whole farm plans specifically tailored to manage areas of highly erodible land. The SLUI programme has four objectives: reduce hill country erosion rates to natural levels; increase the resilience of the regional economy to future major storm events; protect lowland communities from the

impacts of erosion; and to improve water quality. Actions that are undertaken on SLUI farms include: retirement of land, fencing, and erosion planting. Modelling indicated that implementing farm plans on 500 of the most erosion-prone farms would reduce the current regional sediment load by 47% with average reductions in sediment over all catchments of 27% by 2040 (Schierlitz and Dymond, 2006). The reduction in sediment loads can be expected to result in improving trends in the four water quality variables considered by this study (*E. coli*, clarity, turbidity and SSC).

Since 2006, 683 whole farm plans (on 'SLUI farms') have been developed, with approximately 80–85% having implemented some on-the-ground works to control or mitigate erosion and sediment losses (Manderson *et al.*, 2015). The SLUI farms (Figure 3) cover a total area of 493,650 ha (comprising approximately 22% of the region). Areas of the region that were categorised as subject to erosion after 2004 covered 16,937 ha. These areas may also be associated with water quality trends because healing of erosion scars may have contributed to improvements in water quality over the intervening period. GIS data were obtained from HRC that mapped the SLUI farms and land areas that were designated as subject to erosion in 2004 (Figure 3 and Figure 4).

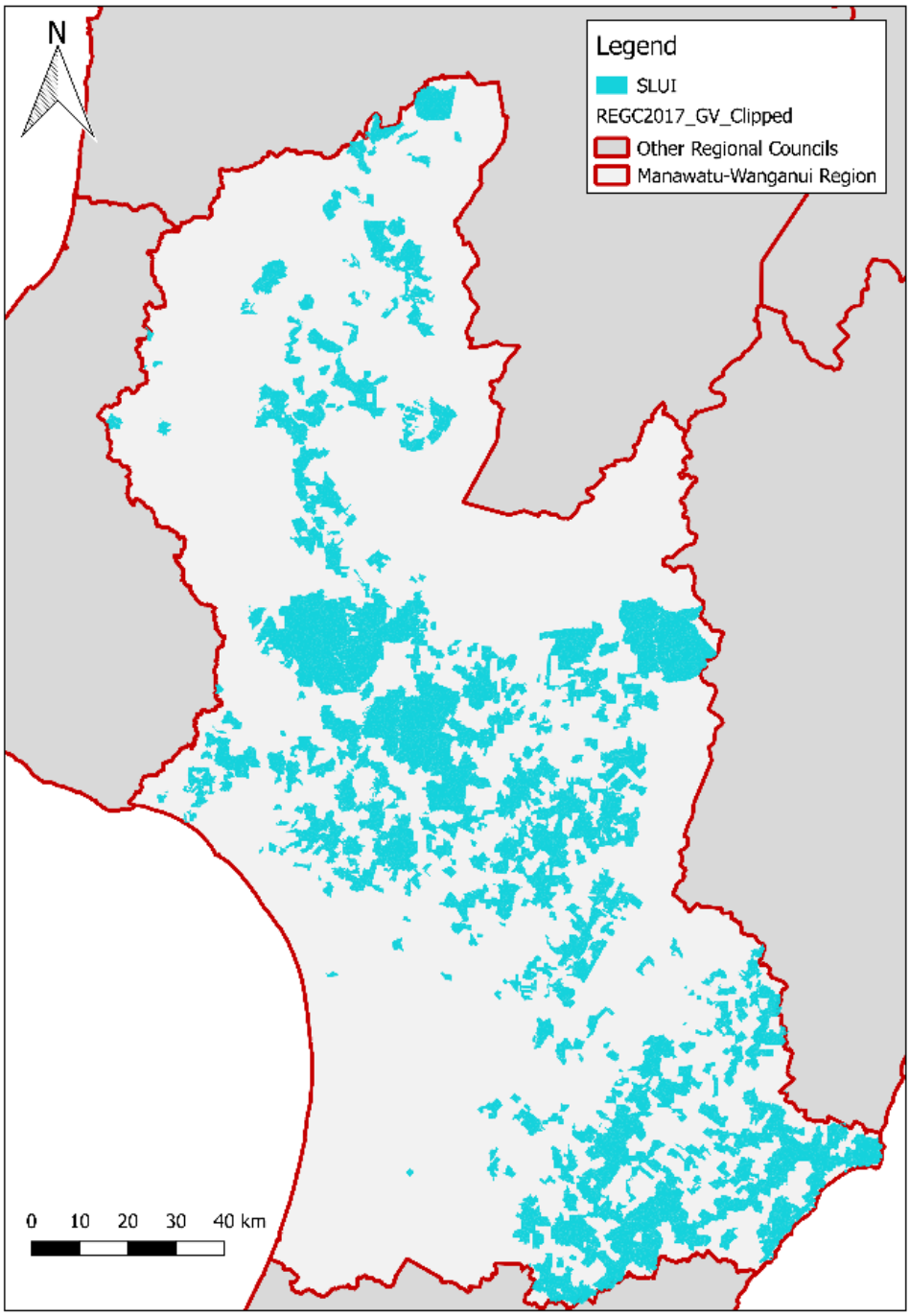

Figure 3. Location of SLUI farms throughout the region. The map indicates 683 'SLUI farms' for which whole farm plans have been developed.

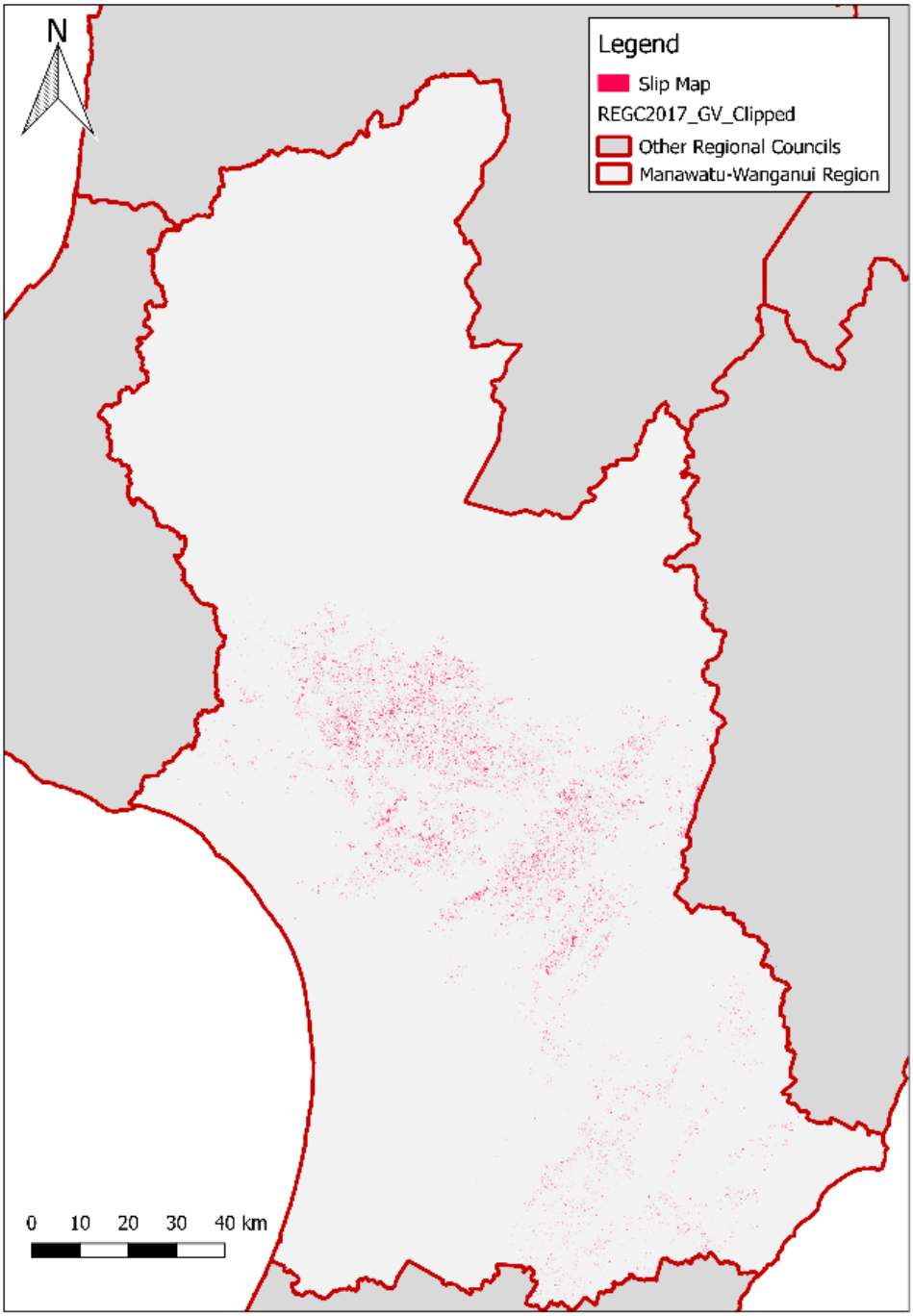

Figure 4. Area subject to erosion in 2004.

3.2.2 Fencing and planting initiatives

Since 2010, HRC has funded and supported a freshwater environmental grant programme that promotes fencing and riparian planting on streams on agricultural land. HRC targets freshwater intervention work programmes based on consideration of a range of factors, including the likely vulnerability of catchments and receiving environments to pressures such as the proportion of land under intensive land use, erosion, hydroelectric schemes, existing fencing and planting, nutrient loss through leaching and run-off, wastewater discharges and contaminated sites such as landfills. Catchments are targeted if current water quality and

biomonitoring indicates either degrading trends or degraded state. In addition, catchments and receiving environments are targeted if they have been identified as having specific community values (as identified in the One Plan, community engagement, Manawatū River Leaders' Accord and Lake Horowhenua Accord Action Plans), or are catchments subject to nutrient management under the One Plan. Finally, because these interventions are generally non-regulatory, catchments and receiving environments are targeted if the community is willing to invest in interventions.

HRC has maintained a geospatial database recording the location and extent of fencing (linear metres) and planting (area). A copy of this database was provided for this study from HRC (Figure 5 and Figure 6).

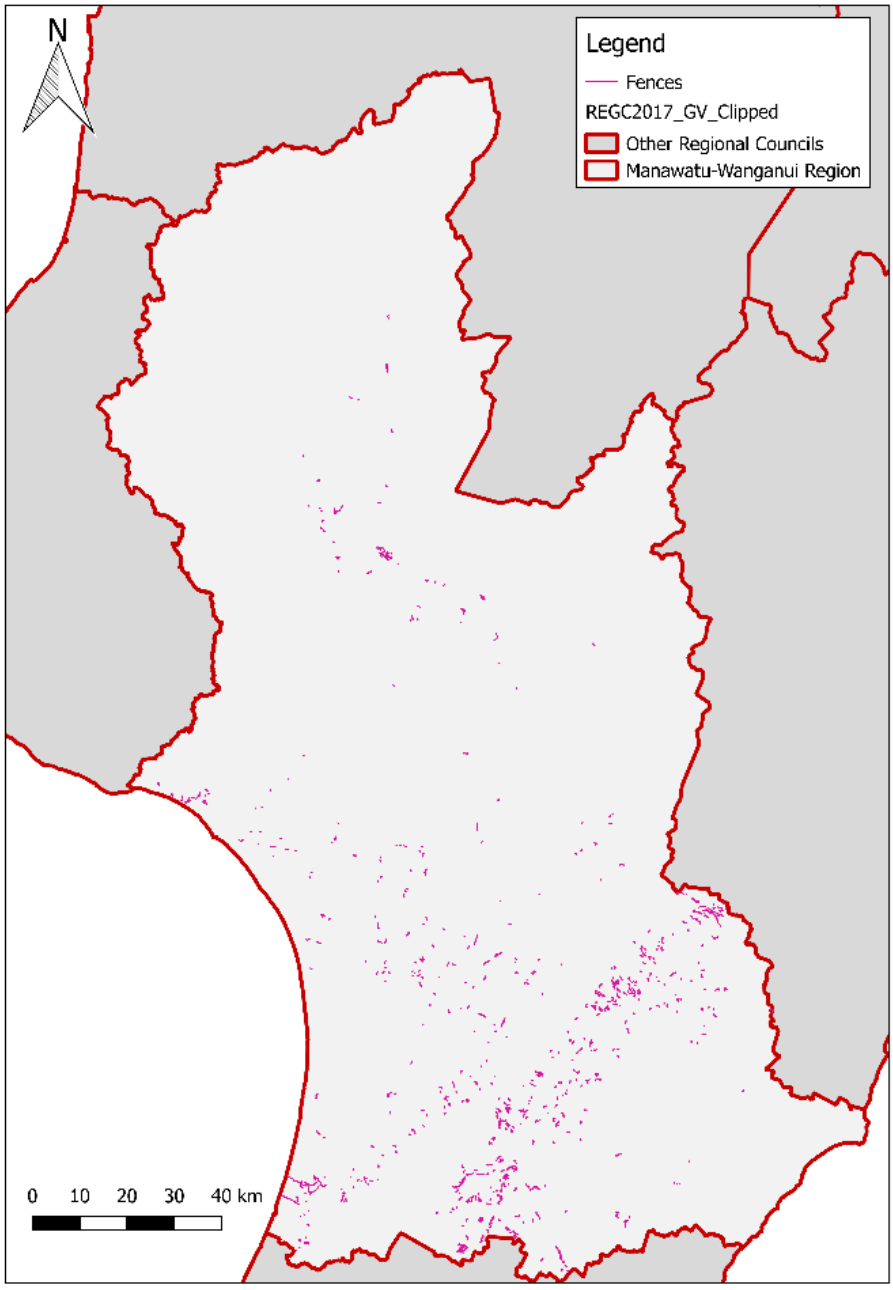

Figure 5. Location of fencing work throughout the region.

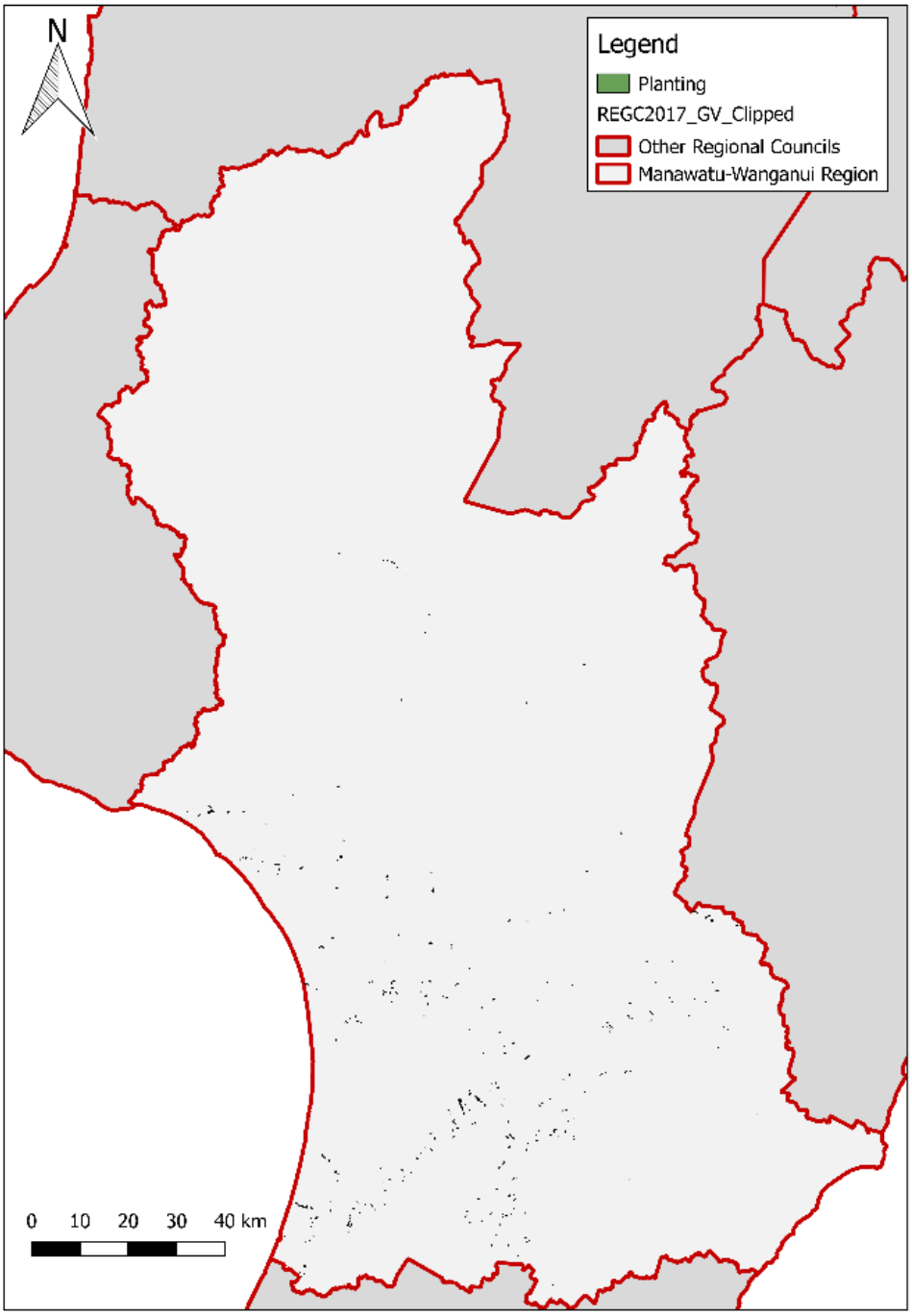

Figure 6. Location of planting work throughout the region.

3.2.3 Climate and flow

Trends in water quality variables may be at least partially attributable to climate effects. Climate may affect the water quality variables of concern in this study in a variety of ways. The main mechanism of mobilisation and transport of *E. coli* and sediment is associated with episodic runoff and high flows (McDowell *et al.*, 2014). These processes are complex, and it was beyond the scope of this study to undertake a detailed analysis of the association between trends and climate or flows. However, trends in mean annual rainfall and mean annual flows were analysed for representative stations through the same periods as the trend analyses

performed on the water quality variables. The results of these analyses provided an indication of the extent to which climate and flows varied through the periods and allowed a qualitative evaluation of extent to which these natural processes may be associated with trends.

Annual rainfall and daily flow data were obtained for 13 and 15 long term climate and flow recording stations, distributed throughout the region from HRC, respectively (Figure 7). The mean annual daily flows were converted to mean flow for the 15 flow recording stations in each year of record to make them comparable to the annual rainfall statistics.

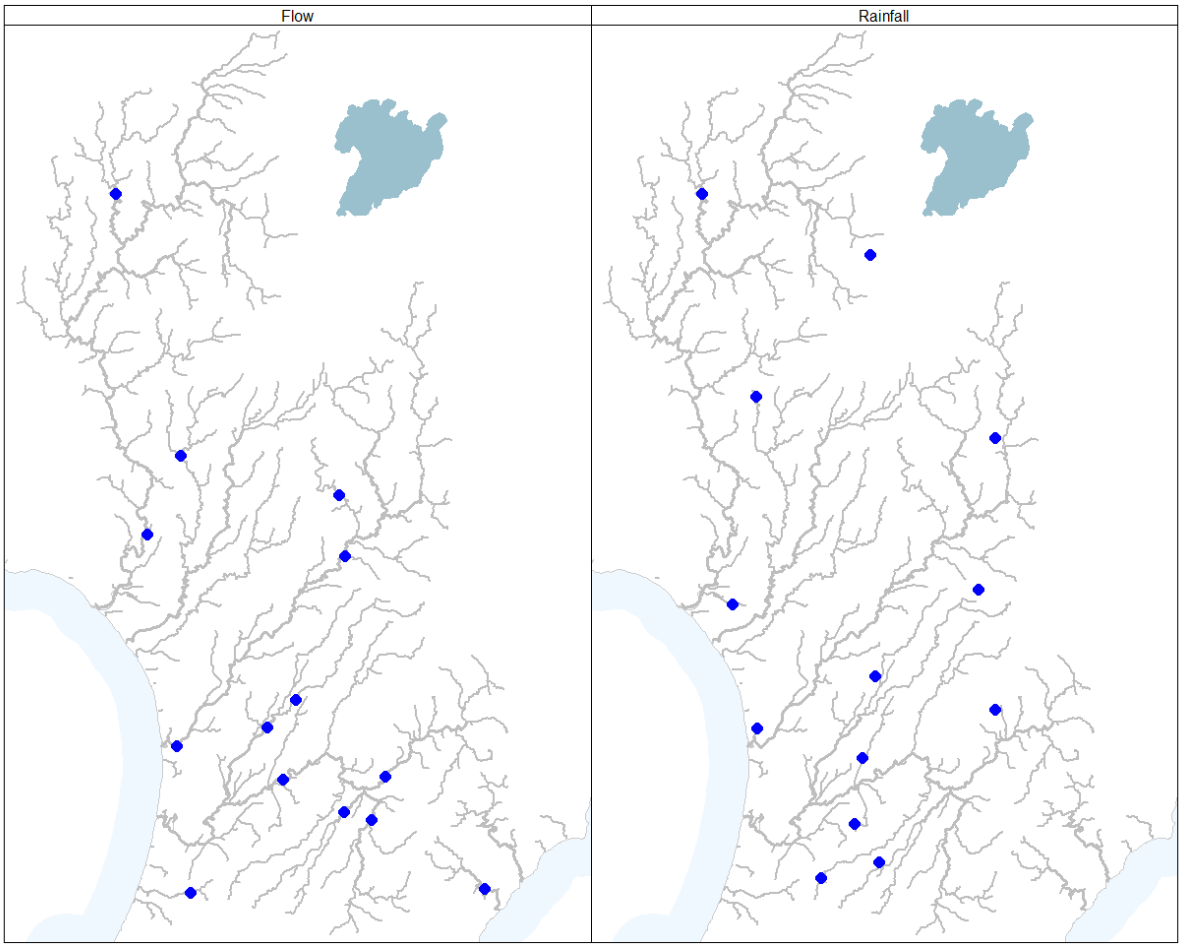

Figure 7. Map showing location of the river flow recorder and climate stations in the Region. Grey lines represent main stem rivers (stream order of 4 or greater).

4 Methods

4.1 Categorisation of sites

Sites were categorised into three types: discharge, impact and state of environment (SoE) sites (Figure 1). Discharge and impact sites represent specific point source discharges or locations downstream of significant and specific point sources, respectively. SoE sites are located at points in the river network that are not significantly affected by point source discharges so that they reflect water quality arising from diffuse and widespread sources of contaminants.

The three categories of sites were used and analysed differently in the study. SoE sites were assumed to be representative of general regional water quality conditions. State and trends were evaluated at SoE sites and were used to make inferences about water quality conditions across the entire region including estimating the length of rivers in different swimming grades (see Section 4.2). Trends were evaluated at discharge and impact sites to provide information about the association between river water quality trends and interventions that have occurred in the region over the last decade (see Section 4.7). Neither state nor swimming grades were evaluated at the discharge or impact sites because these sites are atypical of general conditions.

4.2 Sampling dates and time-periods for analyses

The analyses that follow concern two characterisations of water quality: state and trend. For each variable, the state at SoE sites was characterised by statistics that were calculated from the samples, for example the median value. The trend at all sites (SoE, discharge and impact) was characterised by the rate of change of the variable (or a statistic estimated from the variable; see Section 4.4) through time.

Because water quality changes through time, both the state and trend depend on the time-period over which the state and trends are assessed (e.g., Ballantine *et al.*, 2010; Larned *et al.*, 2016a). Therefore, state and trend assessments are specific for a given period of analysis. In this study, state and trends were characterised for two time-periods. The statistical robustness of determinations of water quality state and trends depends on the variability in the measurements through the time-period and on sample size (i.e., the number of sampling dates). As a general rule, the rate at which confidence increases for estimates of population statistics levels off above a sample size greater than 30 (i.e., above this size there are diminishing returns on increasing confidence with increasing sample size; McBride, 2005). Because water quality data tends to be seasonal, it is also important that each season is well represented over a period of record. In this study, seasons were generally represented by months because most sites have been sampled monthly over the past decade. However, because formerly quarterly was a common sampling interval and because some sites have been sampled less frequently, some of the analyses presented below represent seasons by quarters.

The dataset had variable starting and ending dates, variable sampling frequencies, and variable numbers of missing values. Because the analyses that follow are concerned with assessing regional patterns in state and trends, it was important to maintain the maximum number of sites over the longest time-period, while ensuring the characterisation of state and trends for those sites were as robust and precise as possible. Filtering rules were therefore used to achieve a reasonable trade-off between length of time-period, sample size and numbers of sites. In addition, analyses performed on the grouped results of site trends used

all results irrespective of the level of certainty of the individual trends (see Section 4.6 and 4.7.2). The filtering rules ensured that a reasonable number of samples informed all trends.

In a recent national analysis, Larned *et al.* (2015) used filtering rules that restricted site and variable combinations that were analysed for trends in a given time-period such that there were measurements for at least 90% of the years and at least 90% of seasons. This study adopted similar filtering rules but with criteria relaxed to 80% of years and sample seasons, as suggested by Helsel and Hirsch (1992). All site by variable combinations that did not comply with these filtering rules were excluded from the analysis.

The two time periods evaluated by this study were determined by examining the trade-off between the number of qualifying sites (i.e., sites that met the filtering rules concerning missing measurements) and the time-period. The trade-off between length of time-period and numbers of sites was assessed based on treating the data as both monthly and quarterly samples.

4.3 Analysis of state and swimming grades

For each time-period, the state at each SoE site was characterised using statistics calculated from the water quality measurements. Three *E. coli* statistics were calculated for each site: the median *E. coli* concentration and the proportion of samples that exceeded 260 and 540 *E. coli* 100 mL^{-1} (referred to as G260 and G540, respectively[2]). For clarity, turbidity and SSC the state at each site was characterised by the median of the measurements.

For each time-period, the swimming grade for each site was assessed using the three *E. coli* statistics (i.e., median, G260 and G540). For each site, a grade was calculated for each statistic based on the thresholds shown in Table 1. A final grade was then assigned to each site as the lowest grade achieved by the three statistics. Hence, if a site's grades for the median, G260 and G540 were "Good", "Fair" and "Good", respectively, then the final site grade was "Fair".

The swimming grade at each SoE site was also characterised using the three *E. coli* statistics (i.e., median, G260 and G540) calculated from samples pertaining to the summer 'swimming' season (1st November to 31st March). To ensure maximum precision, the summer season *E. coli* statistics were calculated for the longest time-period only. The swimming grade for each site was calculated from the three statistics based on the thresholds shown in Table 1. The grade at each site for the summer season was compared to the grade derived from the whole record for the longest time-period.

4.4 Analysis of trends

4.4.1 Input data

The trend analyses conducted in this study determined the rate of change in the central tendency of the measurements through the time-period. The underlying model considers monotonic change and the analysis is non-parametric so the measured trend represents rate of change of the median of the data. Trends were evaluated for median *E. coli*, clarity, turbidity and SSC at SoE, discharge and impact sites by conducting analyses based on all samples

[2] This nomenclature follows from (Elliott and Whitehead, 2016) and is based on the use of F to signify the cumulative frequency distribution (F(x) is proportion of time that the variable is less than x). G(x) is used here to represent the proportion of the time that the variable is greater than x. (Snelder *et al.*, 2016a) used the abbreviations PropGT260 and PropGT540 to represent G260 and G540.

through each time-period. The results were interpreted as the annual rate of change of the median value.

The *E. coli* statistics G260 and G540 were only analysed at SoE sites. Trends in the G260 and G540 were evaluated by first calculating the annual values of G260 and G540 as the number of samples that exceeded 260 and 540 *E. coli* 100mL^{-1} in each calendar year divided by the number of samples in that year. The trend in the time series of annual values was analysed for each site and statistic and interpreted the results as the annual rate of change of G260 and G540 through the period.

4.4.2 Missing data and censored values

Trends are most robust when there are few censored values in the time-period of analysis. It has been common to substitute the censored values with 0.5×detection limit and 1.1×reporting limit. Although common, replacement of censored values with constant multiples of the detection and reporting limits can result in misleading results when statistical tests are subsequently applied to those data (Helsel, 2012).

In a recent national analysis, Larned *et al.* (2015) substituted censored values with values that were imputed from the data. In that study, the effect of censored values and missing data on the evaluated trend magnitude was minimal because sites and variable combinations were restricted to those for which the number of censored values was < 15% of the total number of observations. Imputation of censored values is an accepted method for obtaining sample statistics (e.g., mean values and standard deviations). The use of imputed values in trend analysis by Larned et al. (2015) was not strictly correct because the imputation process cannot account for the time order of samples. However, the restriction rules avoided making incorrect determinations of trend magnitude because this quantity is unaffected by censoring when fewer than 15% of the data are censored values.

A different approach to dealing with censored values and missing data to that of Larned *et al.* (2015) was adopted in this study. The approach was based on recent consideration of how to handle censored values in trend analysis, a procedure that has been implemented within the TimeTrends software (Jowett, 2017). The approach does not restrict analysis of sites based on censored values. Instead, it allows the analysis to indicate if there is sufficient data to produce a conclusive result. This is possible because the implementation of trend analysis by the TimeTrends software does not impute[3] replacement values for censored values. In general, the TimeTrends software treats censored values as unknown and does not use them in trend analyses. This means that when there are many censored values (and missing data), the analysis returns an inconclusive result. In cases where there are many missing or censored values TimeTrends will not analyse the data. These cases are reported as "not analysed" in the results (see Section 4.4.5).

Some of the analyses that follow use the evaluated trend slope directions and magnitudes irrespective of the statistical confidence in these evaluations (see Section 4.6). The use of all the evaluated trend assessments (irrespective of whether the analyses were certain regarding individual trend directions) is robust due to the filtering rules (described in section 4.2) applied to each site and variable combination.

[3] However, note that TimeTrends does impute replacements for censored value for calculation of descriptive statistics. The default option applies robust "regression on order statistics" (ROS) to left censored values and Kaplan–Meier (K-M) to right censored values for these calculations.

4.4.3 Trend analysis

The method used for statistical trend analyses in this study differs from the approach used in previous analyses of water quality data (e.g., Ballantine *et al.*, 2010). In the previous studies, the non-parametric Sen slope estimator was used with the Kendall trend test, and trends were determined to be statistically "significant" or "insignificant". Two key problems with the previous ("traditional") approach were identified by Larned *et al.* (2015):

1) conclusions about the significance of trends are strongly influenced by sample size in addition to trend magnitude;

2) the failure to reject the null hypothesis is often incorrectly treated as evidence that there is no trend (e.g., that water quality conditions are "stable" or "being maintained").

To overcome these problems, Larned *et al.* (2015) developed a new trend assessment method. Briefly, confidence intervals are used to draw inferences about trend direction; if a symmetric confidence interval around the trend (estimated using the Sen slope estimator) does not contain zero, then the trend direction (either positive of negative) is "established with confidence". If it does contain zero, it is concluded that there are insufficient data to determine the trend direction and the assessment is that the trend is "uncertain".

In this study, trend assessments for all variables were based on a Sen slope estimator (SSE), which expresses trends in units reflecting the change in the variable per year. When there is seasonal variation in the observations, the seasonal Sen slope estimator (SSSE) should be used (Hirsch *et al.*, 1982). Because the seasonal estimator has lower statistical power than the non-seasonal estimator (due to smaller sample sizes), it is important to establish whether data are seasonally varying. Therefore, where the trend in the median value was being assessed, the trend analysis commenced by testing for the effect of season (i.e., month) on each site and variable combination using a Kruskal Wallis test. When there was a statistically significant effect ($p \leq 0.05$) of season on the value of a variable, the SSSE was evaluated (the Seasonal Kendall analysis in the TimeTrends software). For the two annual values (G260 and G540), and where the trend in the median was being assessed but the effect of season was not significant (Kruskal Wallis $p > 0.05$), the SSE was evaluated (the Mann-Kendall analysis in the TimeTrends software).

To help understand the implications of shifting from the traditional trend analysis procedures to the procedures used in this study, trends were also tested using the traditional Kendall test of correlation. The results of the Kendall tests for all sites and variables are included in the supplementary files for trend analyses.

All trends (i.e., SSE and SSSE) were quantified by relative Sen slope (RSS), which is the annual rate of change divided by the median value of the variable over the time-period (% year^{-1}). The standardisation associated with the RSS value allows trends for statistics measured on different scales, such as median *E. coli,* G260 and G540, to be compared.

Sen slopes and their confidence intervals were calculated with the TimeTrends software (Version 6.1; http://www.jowettconsulting.co.nz/home/software). Subsequent processing and plotting of the TimeTrends output was undertaken using the R statistical software (http://www.r-project.org).

4.4.4 Flow adjustment of river water quality variables

Flow rate at the time that a river water quality measurement is made can affect the observed values because many water quality variables are subject to either dilution (decreasing

concentration with increasing flow) or wash-off (increasing concentration with increasing flow) (Smith *et al.*, 1996). Different mechanisms may dominate at different sites so that the same water quality variable (e.g., *E. coli*) can exhibit positive or negative relationships with flow (Snelder *et al.*, 2016b).

Removing the effect of flow (or any covariate) decreases variation and increases statistical power (i.e., increases the likelihood of detecting a trend with certainty; Helsel and Hirsch, 1992). In addition, a trend in a water quality variable may arise because there is a relationship between time and flow on sample occasion (i.e., a trend in the flow on sample occasion such as increasing or decreasing flow with time). Removing the effect of flow may change this trend's direction and/or magnitude. Previous studies have often provided trend analyses based on both flow adjusted and raw data (e.g., Ballantine *et al.*, 2010; Larned *et al.*, 2015). The appropriate interpretation of the two sets of results by previous studies has been unclear (e.g., Ballantine, 2012).

Flow adjustment requires that water quality samples are associated with the flow at the time of sampling. In this study, flow data was not available for all sites, or for all sample occasions at most sites. To be consistent, trends were all evaluated from raw (i.e., non-flow adjusted) data for all sites. Tests of whether conclusions would have differed substantially if trends had been evaluated using flow adjusted data were carried out by examining differences between raw and flow adjusted trends for a subset of sites and variables for which flow data was available for at least 80% of sample occasions. These tests and further consideration of flow adjustment are detailed in Appendix B. The tests for the subset data indicated that differences in trend directions and magnitudes derived from raw and flow adjusted data were not large. It was concluded that the overall findings of this study would not be appreciably different were the analysis to be performed using flow adjusted data.

4.4.5 Interpretation of trends

Outputs from the TimeTrends analysis were processed to classify the trend obtained for each site and variable combination. Trends were classified into four categories: increasing, decreasing, uncertain and not analysed. An increasing or decreasing trend category was assigned when the 90% confidence interval did not contain zero and the Sen slope was positive or negative, respectively (i.e., the trend direction is established with confidence; Larned *et al.*, 2016b). An uncertain trend category was assigned when the 90% confidence interval contained zero. It is noted that if the 90% confidence interval does not contain zero, the trend direction is established with 95% confidence[4]. Trends will be classified as "not analysed" for three reasons, only the second of which occurred in this study:

1) Trends cannot be assessed when a large proportion of the values (approximately >80%) are censored. This arises because trend analysis is based on examining differences in the value of the variable under consideration between all pairs of sample occasions. When a value is censored, it cannot be compared with any other value and the comparison is treated as a "tie" (i.e., there is no change in the variable between the two sample occasions). When there are many ties there is little information content in the data and a meaningful statistic cannot be calculated.

[4] To achieve a 95% confidence level the procedure uses symmetric intervals at the 90% level, *not* 95%. The rationale is fully explained in Appendix 1 of Larned et al. (2015). Briefly, this arises because the new direction-testing procedure uses a two one-sided ("TOST") methodology, rather than the traditional single "two-sided" method. A notable and beneficial feature of this is that it makes the test more powerful than the traditional two-sided approach.

2) Trends cannot be assessed when there is no, or very little, variation in the data because this also results in ties. The annual values for G260 and G540 had little variation or a large proportion of zero values at some sites.

3) The laboratory analysis of some variables has low precision (i.e., values have few or no significant figures). In this case, many samples have the same value and this then also results in ties.

The evaluation of the trend direction by the new confidence interval approach facilitates a more nuanced inference rather than the 'yes/no' output for the chosen acceptable misclassification error rate of 5%. The approach produces the probability that a trend has a given direction. Trends are declared to be "confidently" detected when direction is established with 95% certainty (following the traditional alpha value of 0.05). This means that the acceptable misclassification rate is 5%. However, if there is insufficient data to infer the direction of a variable's trend at a minimum of 95% confidence, the direction can be determined with lower levels of confidence and a categorisation can be used to convey that information. This study has used the approach to presenting levels of confidence of the Intergovernmental Panel on Climate Change (IPCC; Stocker *et al.*, 2014) to convey the certainty of trend directions (Table 3). The categorical levels of confidence were used to express the likelihood that water quality was improving for each site and variable.

Table 3. Level of confidence categories used to convey the likelihood that water quality was improving (Stocker et al., 2014).

Categorical level of confidence	**Probability (%)**
Virtually certain	99–100
Extremely likely	95–99
Very likely	90–95
Likely	67–90
About as likely as not	33–67
Unlikely	10–33
Very unlikely	5–10
Extremely unlikely	1–5
Exceptionally unlikely	0–1

4.5 Spatial modelling of current water quality state and swimming grades

4.5.1 Modelling approach

The current state of all river segments in the region was modelled for both the analysed time periods and the *E. coli* statistics were also modelled for the summer swimming season (defined as 1st November to 31st March). Modelling used the same approach as those used to generate the national swimming maps (Snelder *et al.*, 2016a) and other predictions of water quality at regional to national scales (e.g., Larned *et al*., 2016; Unwin *et al.*, 2010). The approach combines the SoE site statistics with a spatial framework provided by a database representing the national river network (see Snelder *et al.* (2016a) for details). The database contains a range of variables that represent the characteristics of the catchments upstream of every segment of the river network (Wild *et al.*, 2005). The statistical spatial models used the same catchment characteristics as Larned *et al*. (2016) and Snelder *et al.* (2016a) as

predictors in the models (Table 4). The catchment characteristics obtained from the database were then used to predict water quality state for all river segments in the Region.

Table 4. Predictor variables used in spatial models.

Predictor	Abbreviation	Description	Unit
Geography and topography	usArea	Catchment area	m^2
	usLake	Proportion of upstream catchment occupied by lakes	%
	usCatElev	Catchment mean elevation	m ASL
	usAveSlope	Catchment mean slope	degrees
	segAveElev	Segment mean elevation	degrees
Climate and flow	usAvTWarm	Catchment averaged summer air temperature	degrees C x 10
	usAvTCold	Catchment averaged winter air temperature	degrees C x 10
	usAnRainVar	Catchment average coefficient of variation of annual rainfall	mm y^{-1}r
	usRainDays10	Catchment average frequency of rainfall > 10 mm	days $month^{-1}$
	usRainDays20	Catchment average frequency of rainfall > 20 mm	days $month^{-1}$
	usRainDays100	Catchment average frequency of rainfall > 100 mm	days $month^{-1}$
	segAveTCold	Segment mean minimum winter air temperature	degrees C x 10
	usFlow	Estimated mean flow	m^3 s^{-1}
Geology*	usHard	Catchment average induration or hardness value	Ordinal*
	usPhos	Catchment average phosphorous	Ordinal*
	usParticleSize	Catchment average particle size	Ordinal*
Land cover	usPastoral	Proportion of catchment occupied by combination of high producing exotic grassland, short-rotation cropland, orchard, vineyard and other perennial crops (LCDB3 classes 40, 30, 31, 33)	Proportion
	usIndigForest	Proportion of catchment occupied by indigenous forest (LCDB3 class 69)	Proportion
	usUrban	Proportion of catchment occupied by built-up area, urban parkland, surface mine, dump and transport infrastructure (LCDB3 classes 1,2,6,5)	Proportion
	usScrub	Proportion of catchment occupied by scrub and shrub land cover (LCDB3 classes 50, 51, 52, 54, 55, 56, 58)	Proportion
	usWetland	Proportion of catchment occupied by lake and pond, river and estuarine open water (LCDB3 classes 20, 21, 22)	Proportion
	usBare	Proportion of catchment occupied by bare ground (LCDB3 classes 10, 11, 12,13,14, 15)	Proportion
	usExoticForest	Proportion of catchment occupied by exotic forest (LCDB3 class 71)	Proportion
	usGlacial	Proportion of catchment occupied by ice (LCDB3 classes 14)	Proportion

Relationships were fitted using random forest (RF) statistical models. RF models are a machine learning technique that automatically detect and fit non-linear relationships and high order interactions, both of which can be expected when modelling relationships between water quality and catchment conditions over large environmental gradients (Unwin *et al.*, 2010).

Determining and specifying non-linearities and interactions in more traditional statistical models such as linear models or general linear models requires significant skill and insight by the modeller into the relationships being modelled. Because RF models automatically detect and fit these complex relationships, it is more likely that results generated by different modellers will be comparable. In addition, RF models achieve higher accuracy than more traditional statistical models. High predictive performance is achieved by basing predictions on an ensemble of single regression trees (a forest) (Breiman, 2001). Detailed descriptions of RF models and their diagnostic tools are described in detail in Breiman (2001) and Cutler *et al.* (2007).

Although RF models do not depend on distributional assumptions, transformation of the response variable to an approximately symmetric distribution can improve model performance. The effect of various transformations of the water quality (i.e., response) variables on the model performance was investigated. Where performance was improved, transformations were adopted. Details of the investigations of transformations are explained in Appendix A.

Fitted models were used to make predictions of the water quality variables for all 49,000 segments representing the rivers of the Region. The predicted values of the *E. coli* statistics were used to calculate the swimming grade for all the Region's river segments based on the criteria shown in Table 1. This assessment of swimming grades was compared with assessments made using the national predictions of the same *E. coli* statistics made by Snelder *et al.* (2016a), which underlie the national swimming maps.

4.5.2 Model performance

RF models produce a set of predictions for all cases in the training dataset that is independent of the fitting process and that can be used to evaluate model performance. Model performance was quantified by comparing the independent predictions with the observations and expressing the degree of agreement using five statistics: Nash-Sutcliffe efficiency (NSE; (Nash and Sutcliffe, 1970), bias, percent bias (PBIAS), the relative root mean square error (RSR) and the root mean square deviation (RMSD). NSE indicates how closely the observations coincide with the model predictions. NSE values range from $-\infty$ to 1. A NSE of 1 corresponds to a perfect match between predictions and the observations, values greater than 0 indicate the model has some predictive skill and values greater than 0.5 are commonly considered to indicate good model performance (Moriasi *et al.*, 2007). Bias measures the average tendency of the predicted values to be larger or smaller than the observed values. Optimal bias is zero, positive values indicate underestimation bias and negative values indicate overestimation bias (Piñeiro *et al.*, 2008). PBIAS is computed as the sum of the differences between the observations and predictions divided by the sum of the observations (Moriasi *et al.*, 2007). RSR is a measure of the characteristic model uncertainty and is estimated as the mean deviation of predicted values with respect to the observed values divided by the standard deviation of the observations (Moriasi *et al.*, 2007). RSR varies from the optimal value of 0, which indicates zero RMSE or residual variation and therefore perfect predictions, to a large positive value. RMSD is the mean deviation of predicted values with respect to the observed values and quantifies the characteristic uncertainty of the predictions (Moriasi *et al.*, 2007). The normalization associated with PBIAS and RSR allowed the performance of models to be compared across all the modelled water quality variables. Categorical descriptions of the quality of the predictions for different values of the normalised performance measures that are accepted 'rules of thumb' are shown in Table 5 (Moriasi *et al.*, 2007).

Table 5: Quality of predictions based on performance measure values. The categorical descriptions of the quality of the predictions are 'rules of thumb' (Moriasi et al.*, 2007).*

Prediction Quality	NSE	RSR	PBIAS
Poor	x <0.5	x > 0.7	|x| > 55%
Satisfactory	0.5< x < 0.65	0.6< x < 0.7	30%< |x| <55%
Good	0.65< x < 0.75	0.5< x < 0.6	15%< |x| <30%
Very Good	x >0.75	x <0.5	|x| <15%

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

Water quality data is frequently right skewed. Logarithmic or other transformations that compress the larger values have been used in previous studies to improve model performance (Larned *et al.*, 2016; Snelder *et al.*, 2005; Unwin *et al.*, 2010). Transformations of the modelled response variables were similarly used in this study. Because these transformations are non-linear the model predictions need to be corrected for re-transformation bias. In this study, the smearing coefficient of Duan (1983) was used to correct for back-transformation bias when variables were $\log_{10}$-transformed prior to model fitting, following the approach of Larned *et al.* (2016) and Snelder *et al.* (2016a). Consideration of some of the details of the transformation and retransformation are included in Appendix A.

4.6 Spatial modelling of change in state

Assessment of changes in state was made by combining the spatial model predictions of state with a predictive classification model of trend directions that were observed at the SoE sites. The modelling process comprised three steps. First, for each statistic and time-period, a classification model was used to discriminate sites based on observed trend direction (i.e., positive or negative RSS values). The classification model was fitted using RF models; the same type of statistical model as that used to model state but in a classification mode. The fitted classification models were then used to predict whether the trend directions were negative or positive for each variable for every segment of the river network. The predictors for the classification models were the same catchment characteristics used by the state model (Table 4). It is important to note that no predictors used by the classification models represented actual water quality interventions.

Trend directions at all SoE sites were used to train the classification model, irrespective of the level of confidence in direction. This decision is justifiable on the basis that the confidence intervals are used to judge confidence in trend direction at individual sites. The choice of confidence interval based on the acceptable risk of making incorrect inferences about trend direction at individual sites is arbitrary (i.e., alpha value of 0.05 is arbitrary but is generally accepted). The misclassification error risk for individual sites can be disregarded when considering water quality trends globally (i.e., for all sites across the region) because incorrect classifications of direction will cancel each other (i.e., as many sites will be misclassified as increasing as sites misclassified as decreasing). Therefore, the "face value" of each site's trend (i.e., the direction indicated RSS value) was used to train the classification model.

Misclassification rates and receiver operating curves (ROCs) were used to evaluate the performance of the classification models. ROC plots show the true positive rate (sensitivity)

against the false positive rate (1−specificity) as the probability threshold used to classify a case varies from 0 to 1 (Hanley and McNeil, 1982). Good models have high true positive rates and relatively small false positive rates and, therefore, have ROC plots that rise steeply at the origin, and level off near the maximum value of 1. The ROC plot for a poor model lies near the diagonal, where the true positive rate equals the false positive rate for all thresholds. The model performance was quantified using the area under the ROC curve (AUC). AUC is a measure of the performance of a binary classifier, with good models having an AUC near 1, while a poor models will have an AUC near 0.5 (Hanley and McNeil, 1982). The following rules of thumb were used to express the quality of the model indicated by AUC in narrative terms: very good (0.9 – 0.8); good (0.8 - 0.7); satisfactory (0.7 - 0.6); poor (0.6 - 0.5).

The second step was to adjust the spatial predictions of state for each variable (i.e., the three *E. coli* statistics, and median values for clarity, SSC and turbidity) to represent an expected value at the beginning and end of the time-period. The adjustment was made for each statistic by adding and subtracting a trend slope multiplied by half the time-period from the predicted value of the statistic for the entire time-period (i.e., the predicted state). Thus, if the predicted median value was 100 *E. coli* $100mL^{-1}$ and the trend was -2 *E. coli* yr^{-1}, then the median at the start of a seven-year time-period was estimated to be 107 *E. coli* $100mL^{-1}$ and the median at the end of the time-period was estimated to be 93 *E. coli* $100mL^{-1}$. The trend slope used at each segment was equal to the median of all RSS values with either decreasing or increasing trends, depending on whether the classification model prediction for that segment was a decreasing or increasing trend. Thus, segments that were predicted to have increasing trends had their expected values at the beginning and end of the trend year period adjusted such that the rate of increase was uniform over the region and equal to the “average” (i.e., median) of the observed rates of increase. The same logic was applied to segments for which the classification model indicated a decreasing trend. Thus, two sets of spatial predictions of state were derived for each statistic (i.e., median, G260 and G540 of *E. coli* and median clarity, SSC and turbidity) representing the ‘expected value’ of state at the beginning and end of the time-period.

For the *E. coli* statistics, a third step calculated the swimming grade at each segment at the beginning and end of the time-period based on the criteria shown in Table 1. The change in total river length in each swimming grade over the time-period, for all segments and segments of order four and above, were then calculated using the stream orders and segment lengths that are available for the digital river network.

4.7 Association between trends, interventions and other factors

4.7.1 Sustainable land use initiative, fencing and planting.

The strength of the relationship between the direction and magnitudes of trends at SoE sites and mitigation measures associated with three intervention initiatives was assessed. First, the SLUI mitigation initiative was quantified as the proportion of the catchment area upstream of SoE sites that were occupied by SLUI farms (Table 6). Second, the new fencing resulting from the freshwater environmental grant programme was quantified as the proportion of upstream stream length with new fences (Table 6). Note that this variable ranges between zero and 200% because both sides of a stream may be fenced. Third, the new planting resulting from the freshwater environmental grant programme was quantified as the proportion of catchment area subject to new planting. The proportion of the catchment area subject to erosion in 2004 was also included as a covariate in this analysis.

The first step of the analysis was to use the digital river network to accumulate the area of SLUI farms, new fencing, new planting and the area of land designated as subject to erosion in 2004. The accumulation produced a value for all network segments being the proportion of catchment occupied by SLUI farms (Figure 8), the proportion of river length subject to new fences (Figure 9), the proportion of catchment area subject to new planting (Figure 10) and the proportion of catchment area subject to erosion in 2004 (Figure 11). It was hypothesised that these four variables (hereafter SLUI, Fencing, Planting and Erosion) are associated with trend direction and magnitude. In addition, it was hypothesised that the interaction (i.e., multiplicative combination) of SLUI and erosion (Erosion_SLUI) would be associated with trend direction and magnitude on the basis that erosion sites subject to farm plans would show stronger benefits to water quality than either predictor in isolation.

Table 6. Summary of explanatory variables used in assessments of association between trends and management interventions.

Explanatory variable	**Description**	**Units**
SLUI	Proportion of upstream catchment area occupied by farms subject to a farm plan prepared under the SLUI program.	%
Erosion	Proportion of upstream catchment area subject to erosion in 2004.	%
SLUI_Erosion	Interaction (i.e., multiplicative combination) of SLUI and erosion	-
Planting	Proportion of upstream catchment area subject to riparian planting carried out under HRC freshwater environmental grant programme.	%
Fencing	Proportion of upstream riparian river length fencing carried out under HRC freshwater environmental grant programme.	%

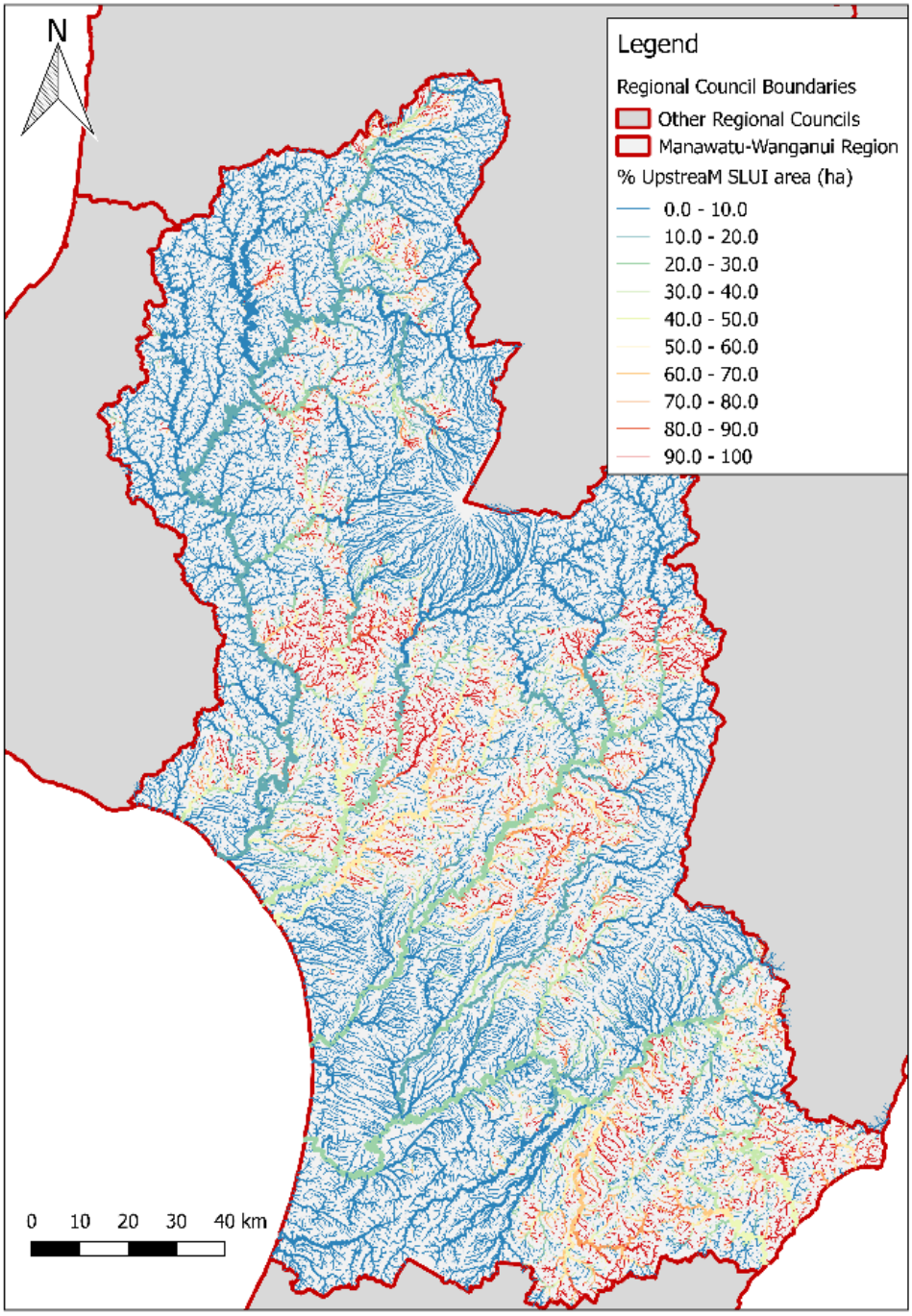

Figure 8. Proportion of catchment occupied by SLUI farms.

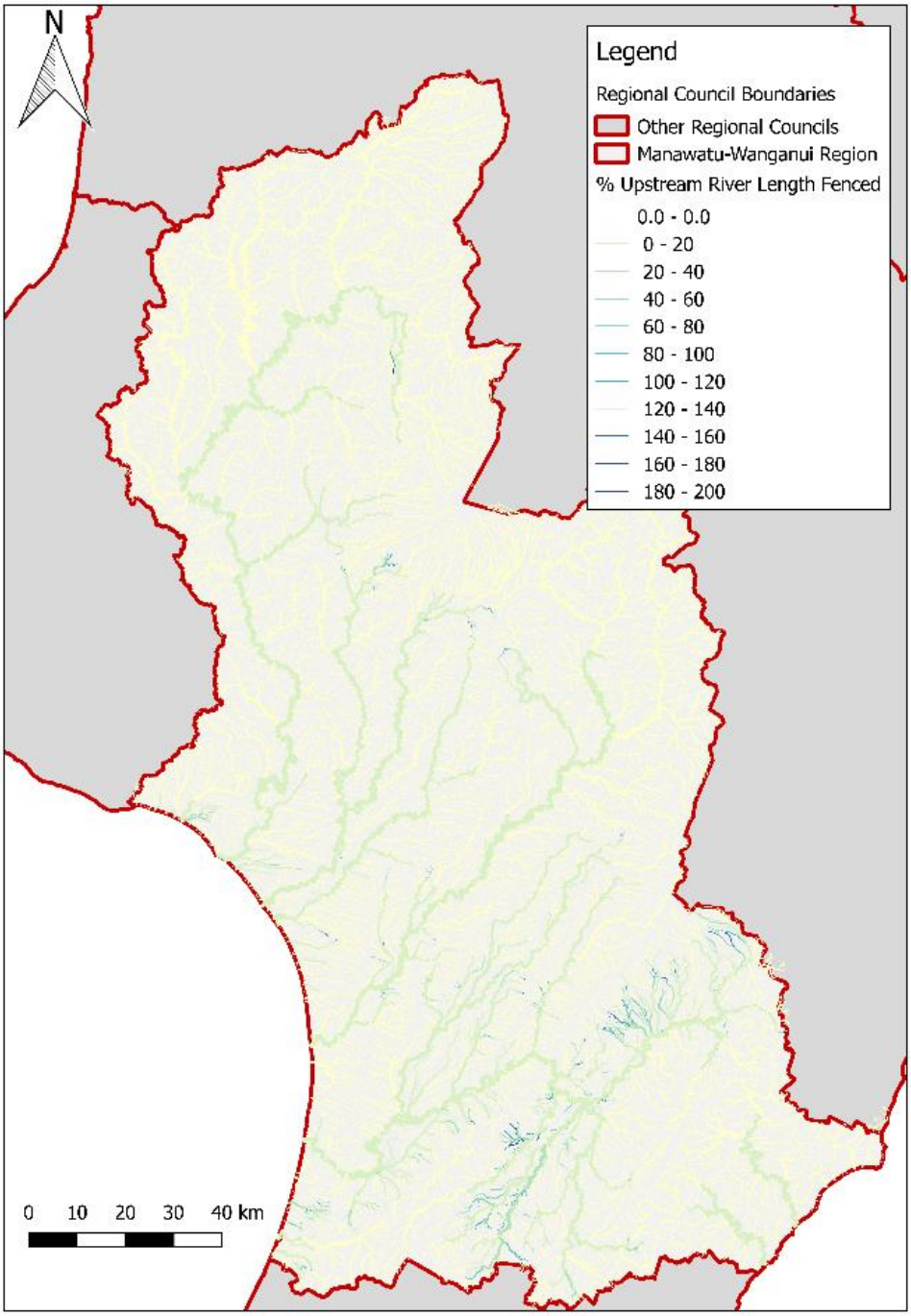

Figure 9. Proportion of total catchment stream length fence.

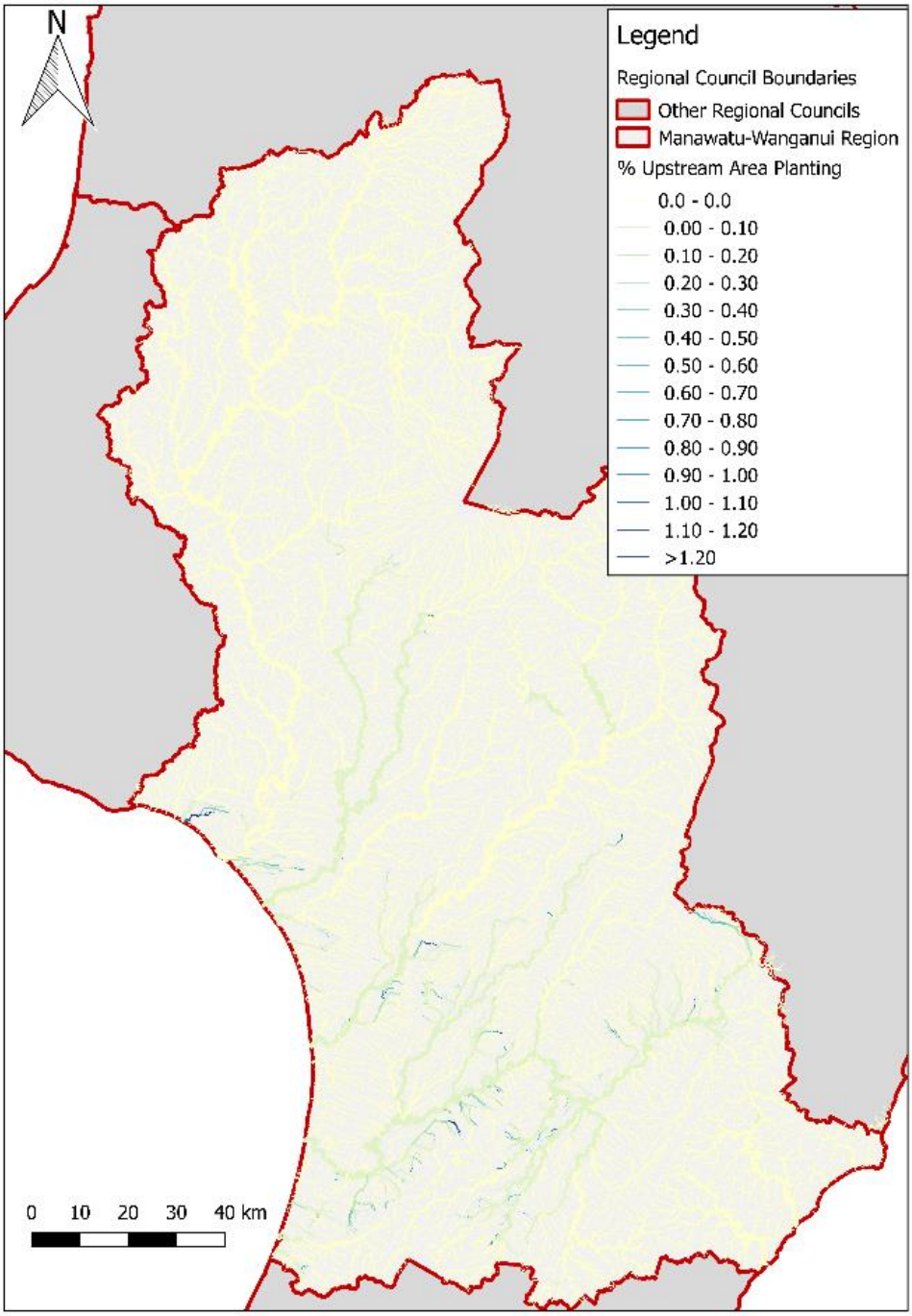

Figure 10. Proportion of total catchment subject to planting works.

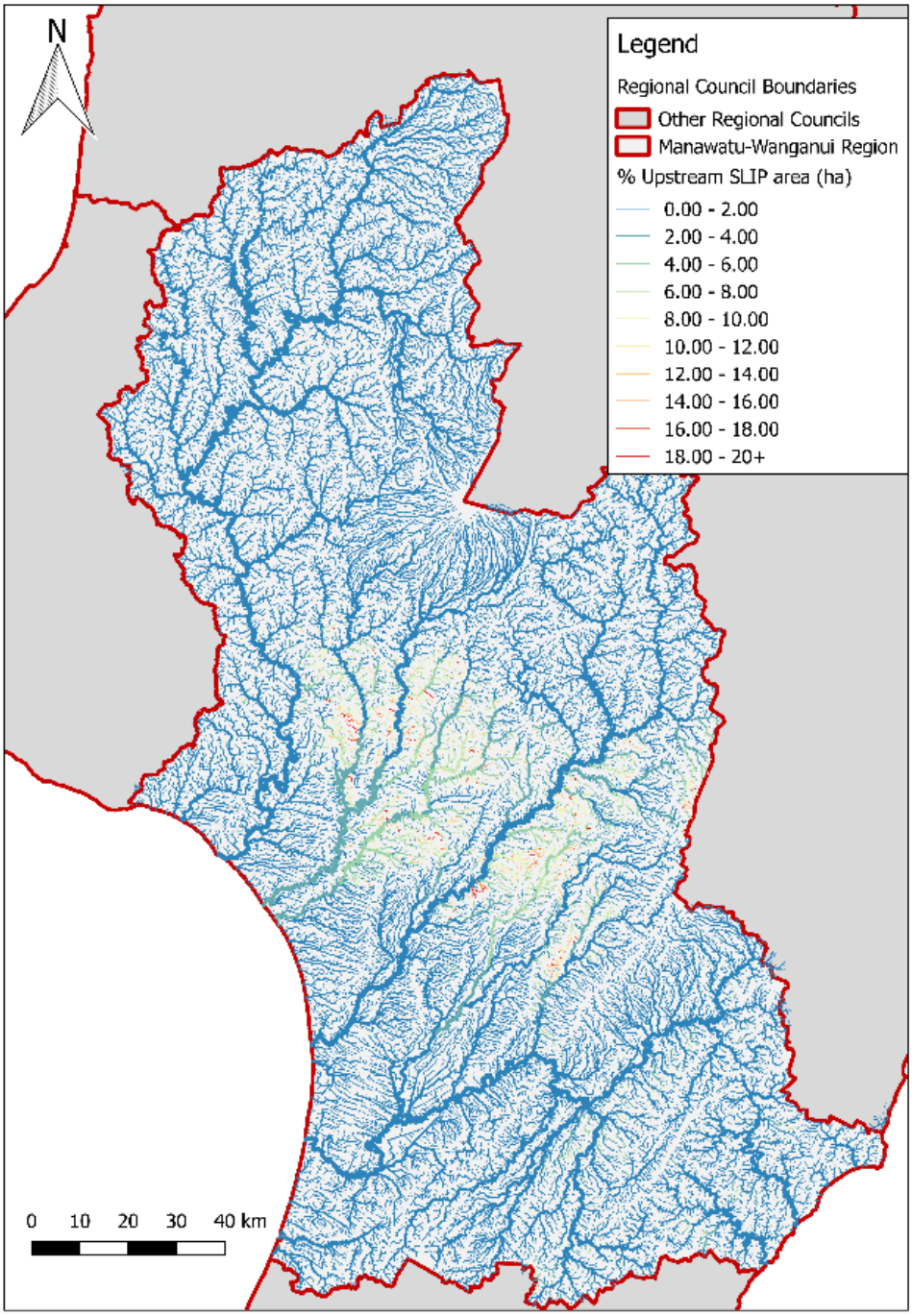

Figure 11. Proportion of catchment subject to erosion in 2004.

The relationships between trend direction and the explanatory variables were tested using a random forest classification model. In these RF models, the categorical dependent variable was the trend direction and the predictors were the continuous independent variables: the proportion of catchment occupied by SLUI farms; the proportion of river length subject to new fences; the proportion of catchment area subject to new planting; and the proportion of catchment area subject erosion in 2004.

A backward elimination procedure was used to remove redundant predictor variables from the initial 'saturated' RF models (i.e., models that included all predictors). The variables retained in the reduced model were interpreted as having significant associations with the response variable. The procedure first assesses the model error (MSE) using a cross validation process (Svetnik *et al.*, 2004). 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 *et al.*, 1984).

The relationships between trend magnitude and the individual explanatory variables were evaluated using stepwise linear regression. A standard forward and backward stepwise model fitting procedure was applied to a saturated model that included all predictors and a term expressing the interaction between erosion and SLUI. In this procedure, the Akaike information criterion (Akaike, 1973) was used to apply a penalised log-likelihood method to evaluate the trade-off between the degrees of freedom and fit of the model as explanatory variables were added or removed (Crawley, 2002). As for the classification RF model, the retained variables were interpreted as having significant associations with the response variable. The performance was expressed as the proportion of variation in the response explained by model (r^2).

4.7.2 Improvements of point source discharges

HRC provided data that linked discharge sites to their closest downstream impact sites. The discharge-impact pairs were used to examine if trend directions at both sites were concordant. Consistent concordance between trends at discharge and impact sites was interpreted as evidence of association between the two sets of trends.

Binomial tests were used to assess the level of association between the paired discharge-impact site trends. It was deemed that there was an association in a certain direction if the number of sites that had concordant trends (i.e., both increasing or both decreasing) were greater than could be expected if increasing and decreasing trends were equally likely. To perform a binomial test, the number of concordant pairs of discharge-impact site trends and their directions were counted. All pairs of trends were included regardless of the confidence in the classification of the trend direction. A 'two-tailed' binomial test was then performed based on the expectation that pairs of sites have a 50% probability of being concordant. If the binomial test *p*-value was less than 0.05, the null hypothesis was rejected, (i.e., it was concluded that there were more concordant trends than could be expected by chance and that there was an association between discharge trends and trends at downstream impact sites). In this case, the overall trend was positive if the proportion of concordant increasing trends was greater than 50% (and the binomial test was significant), or negative if 50% of concordant trends were decreasing (and the binomial test was significant).

4.7.3 Climate and flows

Trends in time series of annual rainfall depths and mean flows were evaluated at the 13 and 15 climate and flow recording stations (Figure 7). Trend directions were quantified by RSS values and confidence in trend direction was expressed as the likelihood that trends were decreasing (Table 3).

Overall ‘regional trends’ in annual rainfall and mean flows were assessed using binomial tests. It was deemed that there was a regional trend in a certain direction if the number of sites that had trends in the same direction (i.e., positive or negative RSS values) were greater than could be expected if increasing and decreasing trends were equally likely. To perform a binomial test, the number of climate and flow recording station trends and their directions were counted. All trends were included regardless of the confidence in the classification of the trend direction. A ‘two-tailed’ binomial test was then performed based on expectation that the trend at a station has a 50% probability of being in a particular direction (e.g., decreasing). If the binomial test *p*-value was less than 0.05, the null hypothesis was rejected, i.e., it was concluded that there were more increasing trends than could be expected by chance and that there was a regional trend. When there was a regional trend, its magnitude was defined as the median of all RSS values.

5 Results

5.1 Analysis of time-periods for SoE sites

The results of the analysis of the number of SoE sites with adequate data versus length of time-period are shown in Figure 12. For the 10-year period ending 2016 there were 69 sites with adequate *E. coli* data i.e., sites that met the filtering rules. There were no sites with 10 years of adequate Turbidity or SSC and only 23 sites with adequate Clarity data. The 10-year period was adopted for analysis of *E. coli* as this represents a reasonable number of sites for spatial modelling (69) and a period over which HRC's initiatives are expected to have improved water quality. Relaxing the frequency of monitoring requirement to quarterly would increase the number of eligible sites for Clarity (Figure 12). However, relaxation of sampling frequency for Clarity would have introduced inconsistency in the filtering rules. Analysis that included spatial modelling of time-periods longer than 10 years was precluded because the number of sites with adequate data was too small to adequately represent regional variability (Figure 12).

The 7-year period ending 2016 was adopted as a second time-period for the analysis. This shorter period was less ideal for analysis of trends but provided a larger number of sites for spatial modelling: 86 *E. coli*, 37 clarity, 76 SSC and 62 turbidity (Figure 12).

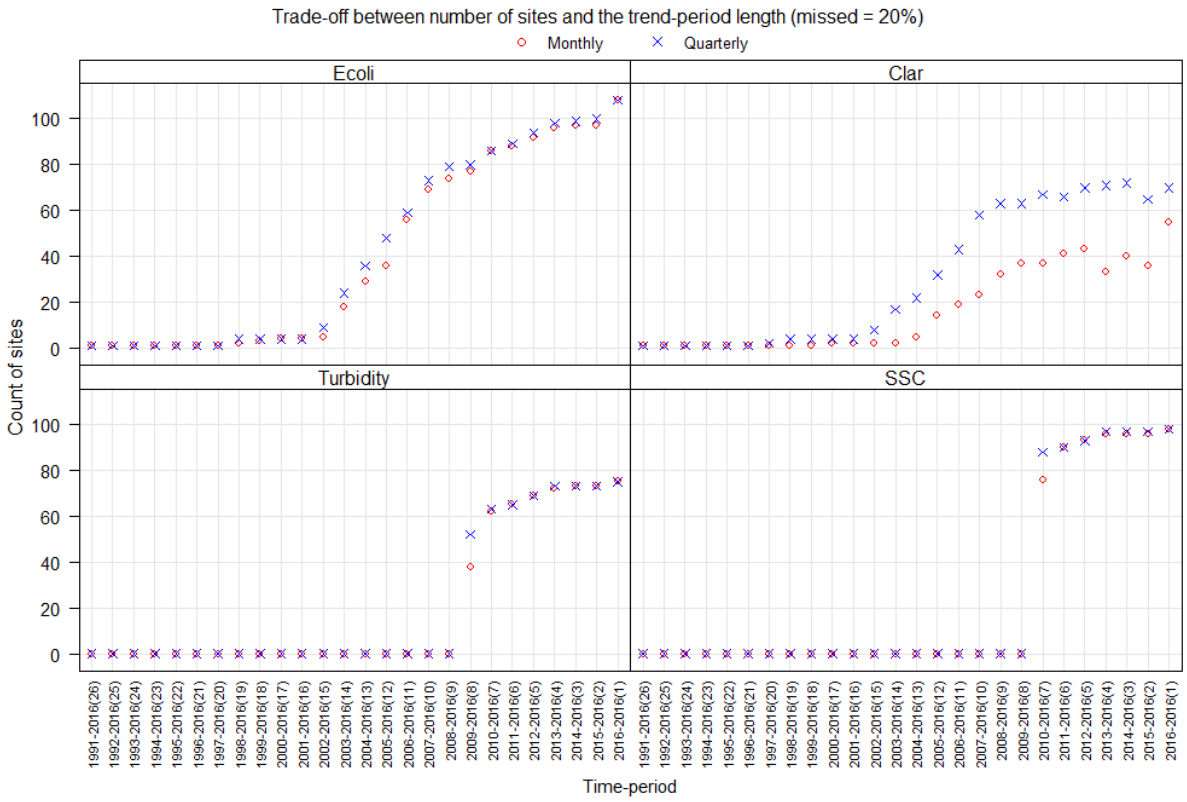

Figure 12. Trade-off between number of SoE sites and the trend-period length. The plots for each variable show the number of sites that comply with the filtering rules when seasons are defined as months or quarters.

5.2 Swimming grades at SoE sites

5.2.1 10-year time-period

Swimming grades for the 69 SoE sites in the 10-year dataset, for both the complete dataset and the summer only data, are summarised in Table 7 and shown in Figure 13[5].

Table 7: Summary of number of SoE sites in each swimming grade.

Grade	No. of sites (annual grade)	No. of sites (summer grade)
A	9	8
B	10	2
C	9	21
D	25	16
E	16	22

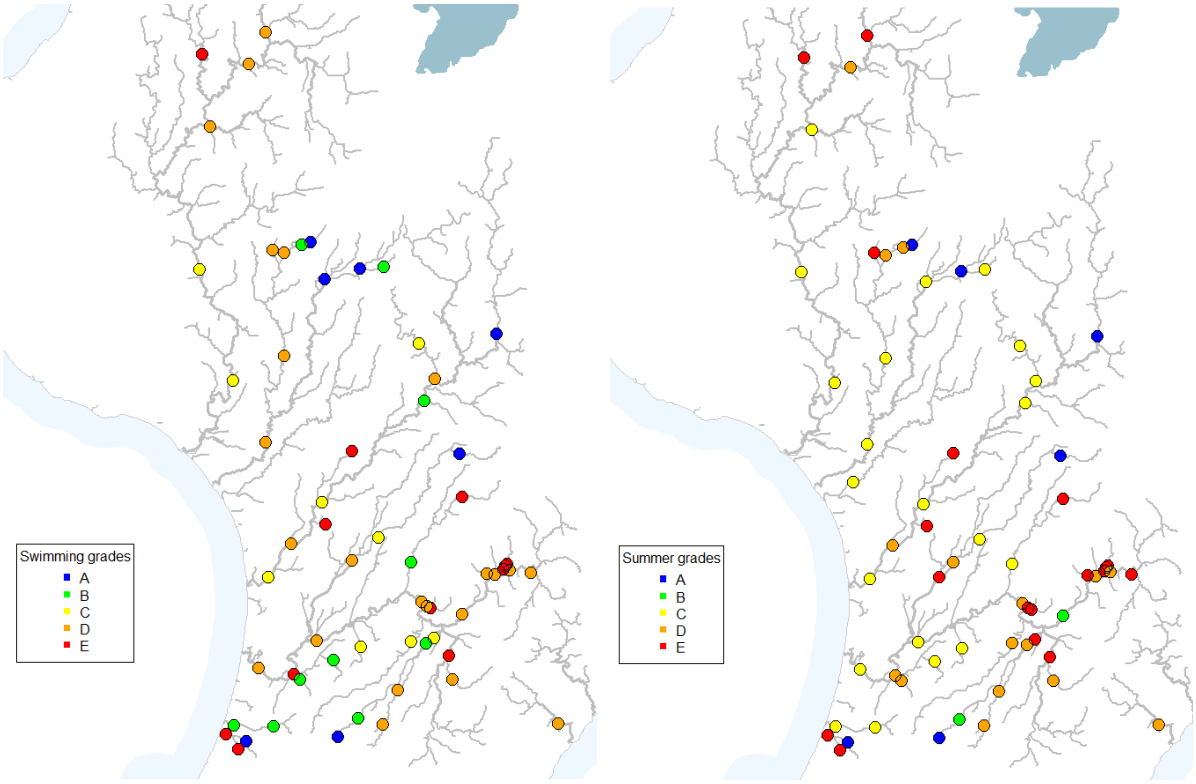

Figure 13. Swimming grades assigned to the 69 SoE sites included in the 10-year time-period dataset for all data (left) and the summer season (right). The "summer" grades were derived from E. coli samples pertaining to the summer swimming season only (1st November to 31st March). The grey lines represent rivers of stream order four or greater.

[5] A complete set of *E. coli* statistics for the 69 SoE sites included in the 10-year dataset is provided as supplementary data in file "SoESite_SwimGrading_10Yr.csv"

Sites that were graded excellent (A) and good (B) tended to be located in headwaters and sites that were graded intermittent (D) and poor (E) tended to be on main stem rivers (Figure 13).

A larger proportion of sites were swimmable (i.e., grade fair or better) for the summer swimming season compared to the all year grades (45% versus 40%). However, fewer sites were in the excellent and good grades for summer (10 sites) compared to all year (19 sites). There was a pattern in the spatial distribution of differences between summer and all year grades (Figure 14). The largest negative differences (i.e., grades being poorer in summer compared to all year) were concentrated in the smaller headwater streams and the largest positive differences (i.e., grades being better in summer compared to all year) tended to occur on mainstem rivers. Consistent with this pattern, there was a small but statistically significant positive relationship between catchment area and the difference between summer and all year grades (r^2 = 8%, p = 0.01).

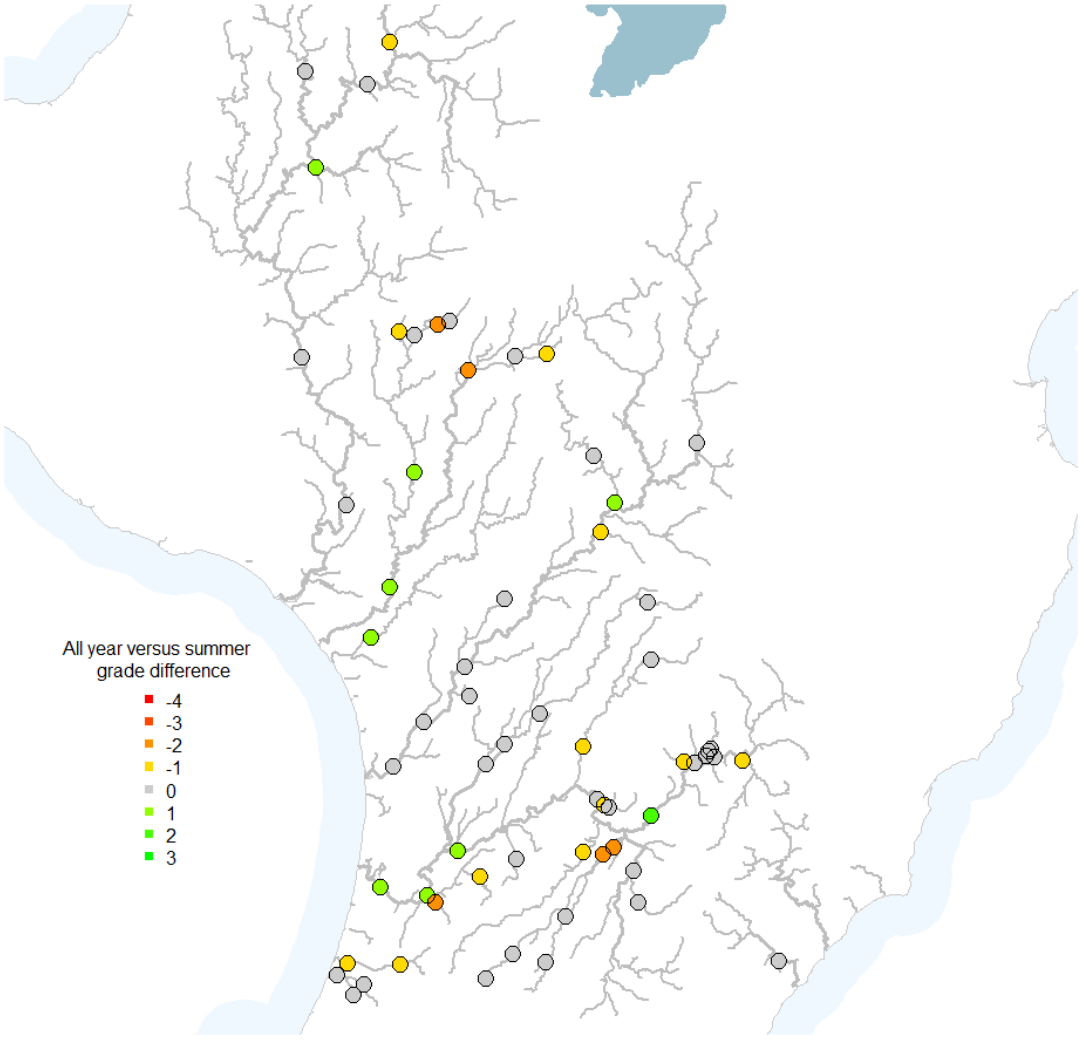

Figure 14. Difference between the all year grades and the summer swimming season grades (1st November to 31st March) for the 10-year time-period dataset. The plotted values indicate the difference in grade between all year and summer. For example, an A for all year and D for summer is a difference of -3. The grey lines represent rivers of stream order four or greater.

5.2.2 Seven-year time-period

A complete set of statistics describing state for *E. coli*, turbidity, SSC and clarity at the 88 SoE sites included in the 7-year dataset is provided as supplementary data[6]. Of the 86 SoE sites in the 7-year dataset with *E. coli* data, 24 sites were graded excellent (A), 7 were graded good (B), 5 were graded fair (C), 30 were graded intermittent (D) and 20 were graded poor (E) (Figure 15). Sites that were graded excellent (A) and good (B) tended to be located in headwaters and sites that were graded intermittent (D) and poor (E) tended to be on main stem rivers (Figure 15).

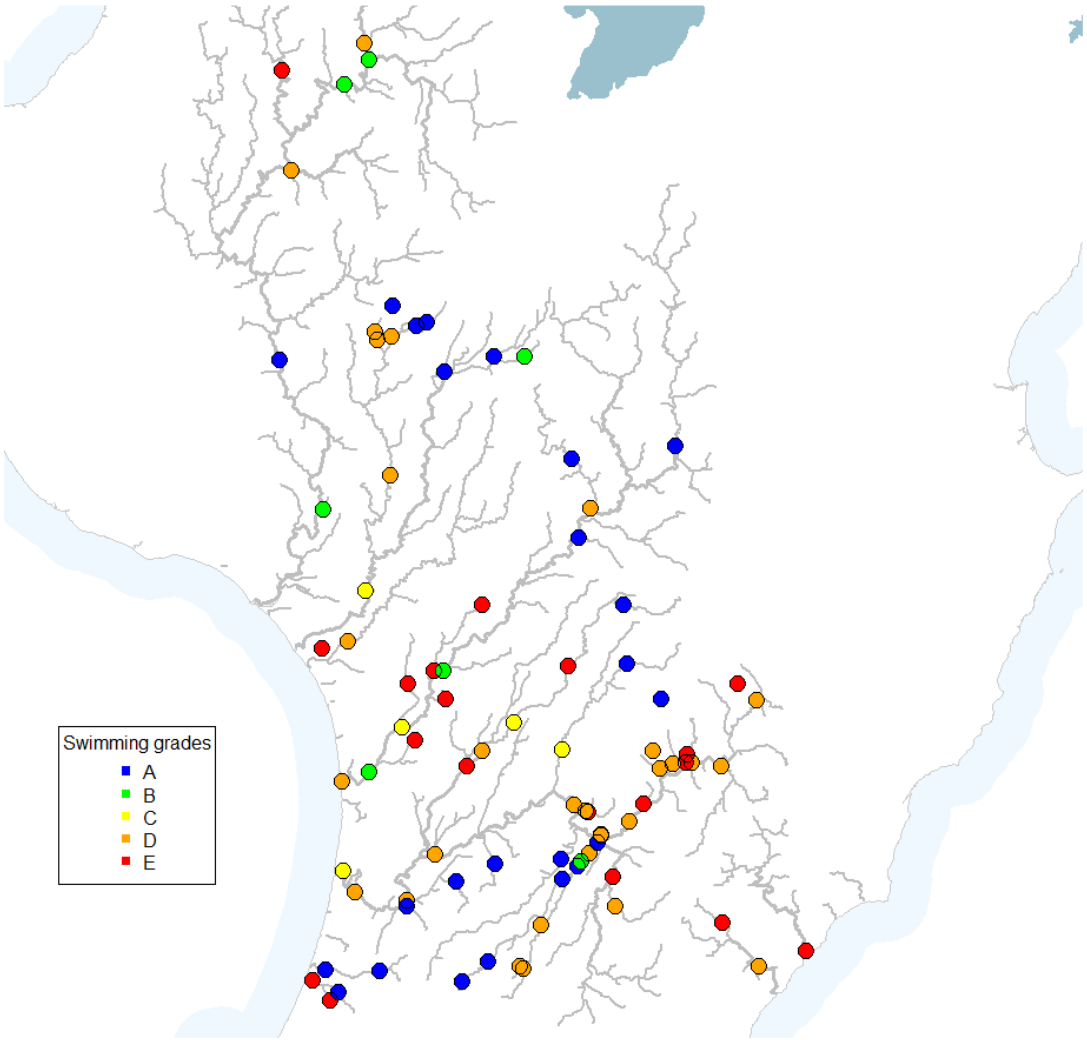

Figure 15. Swimming grades assigned to the 86 SoE sites that had E. coli data in the 7-year time-period dataset. The grey lines represent rivers of stream order four or greater.

Site swimming grades assessed for each time-period were not always the same (Table 8). Of the 69 sites that were in common to the 10-year and seven-year time period, 49 (i.e., 71%) had the same grading for both time periods (Table 8). Sites in the D and E grades had the most stable grades between the two time-periods and sites in the A and B grades were the least stable (Table 8).

[6] SoE_Sites7-Year_StateResults.csv

Table 8. Comparison of swimming grades at 69 sites evaluated for the 10-year and seven-year time-periods. The values in parentheses are the percentage of sites.

		Swimming grades seven-years				
		A	**B**	**C**	**D**	**E**
Swimming grades 10-years	**A**	9 (15)	0 (0)	0 (0)	0 (0)	0 (0)
	B	7 (10)	2 (3)	1 (1)	0 (0)	0 (0)
	C	4 (6)	3 (4)	1 (1)	1 (1)	0 (0)
	D	0 (0)	1 (1)	2 (3)	22 (32)	0 (0)
	E	0 (0)	0 (0)	0 (0)	1 (1)	15 (22)

5.3 Trends at SoE sites

5.3.1 10-year time-period

Of the 69 SoE sites with *E. coli* data included in the 10-year time-period dataset, 18 had flow data for at least 80% of sample occasions. An examination of differences in trend directions and magnitudes for raw and flow adjusted data was carried out for these sites and is described in Appendix B. Based on this examination it was concluded that this study's findings would not be significantly different if flow adjusted trends had been used. All trends reported below for both time-periods are therefore based on analyses performed using raw (i.e., not flow adjusted) data.

A large proportion of *E. coli* trends (raw data) for the 69 SoE sites included in the 10-year time-period dataset was uncertain (Table 9, Figure 16). However, there were more trends that were characterised as likely to be improving than degrading (Figure 17). For median *E. coli*, G260 and G540, 61%, 77% and 78% of site trends were as likely as not to be improving (Figure 17). Trends in G260 and G540 were not analysed for some sites due to high proportions of zeros in the time series of annual values. There was no discernible geographic pattern in the trend status and improving, degrading and uncertain trends occurred throughout the region (Figure 16).

Table 9. Trend analysis results for E. coli at the 69 SoE sites included in the 10-year period dataset. Values in parentheses are proportion of sites (%).

Variable	Decreasing	Increasing	Uncertain	Not Analysed
E. coli	10 (14)	7 (10)	52 (75)	0 (0)
G260	7 (10)	0 (0)	55 (80)	7 (10)
G540	4 (6)	3 (4)	53 (77)	9 (13)

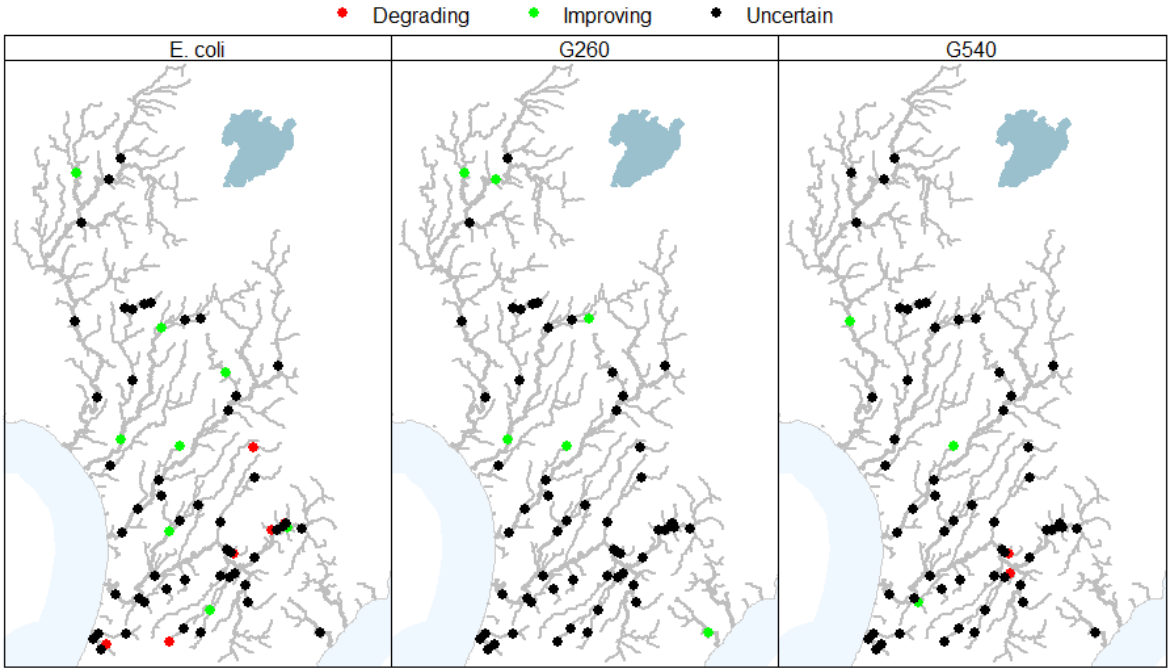

Figure 16. Map of sites classified by their 10-year raw trend descriptions for the three E. coli statistics. Note that trend descriptions indicate degrading and improving (rather than trend direction). Trends are all based on analyses performed using raw (i.e., not flow adjusted) data.

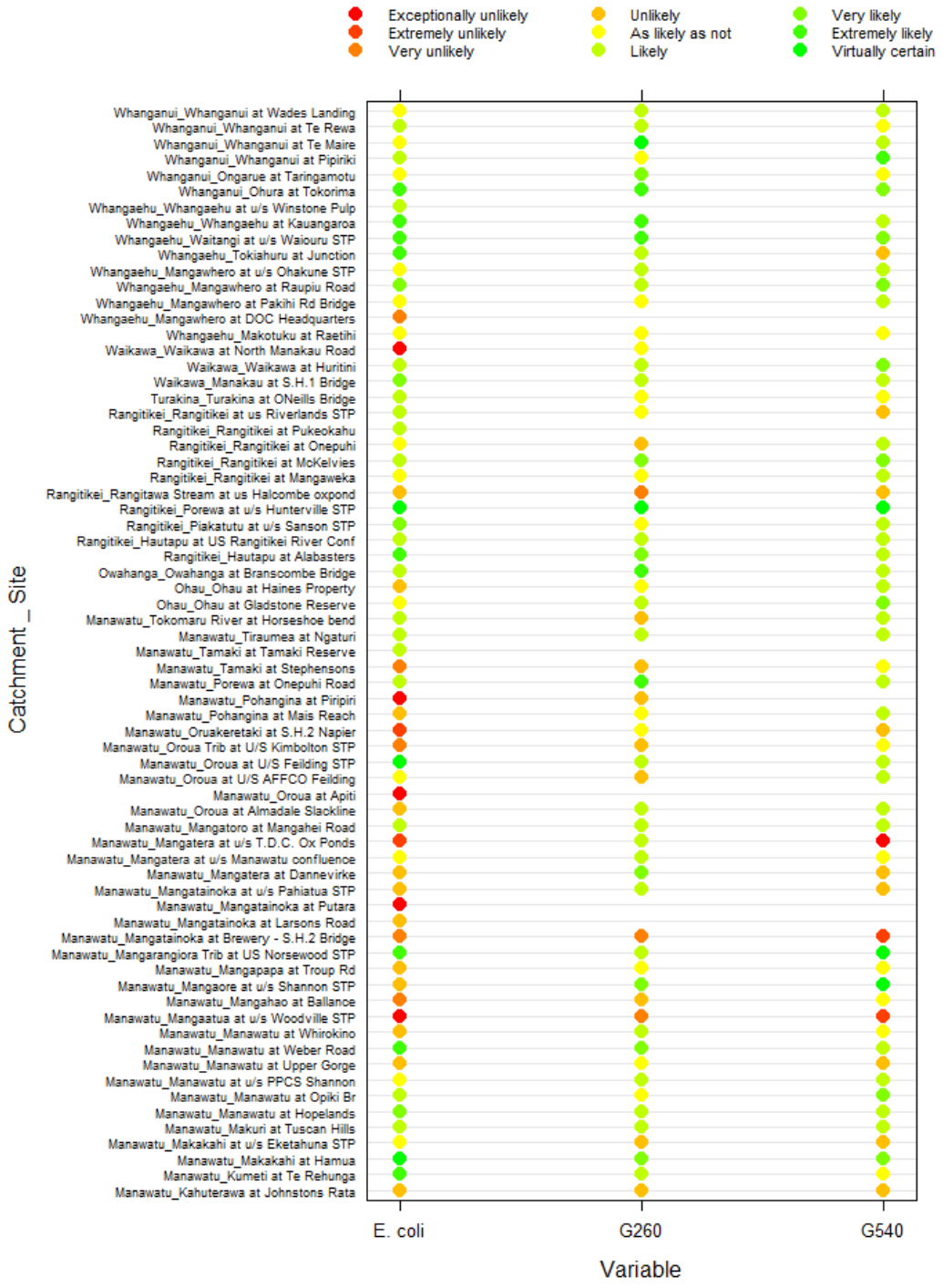

Figure 17. Summary plot of 10-year time-period trend analysis results. The plot shows the level of confidence that water quality was improving at each site and variable. Combinations of site and variable for which data was not available or trends were categorised as not analysed are shown as missing dots. Missing dots indicate the variable was either not monitored or the water quality trend description was 'not analysed'. See Table 3 for details of the confidence categories. Sites are grouped by the sea draining catchment to which they belong and then alphabetical order of the site names (separated by an underscore).

Trend direction (irrespective of confidence) was predominantly improving water quality, particularly for trends in the proportion of samples exceeding the 260 and 540 thresholds. However, trend magnitudes differed greatly between sites (Figure 18).

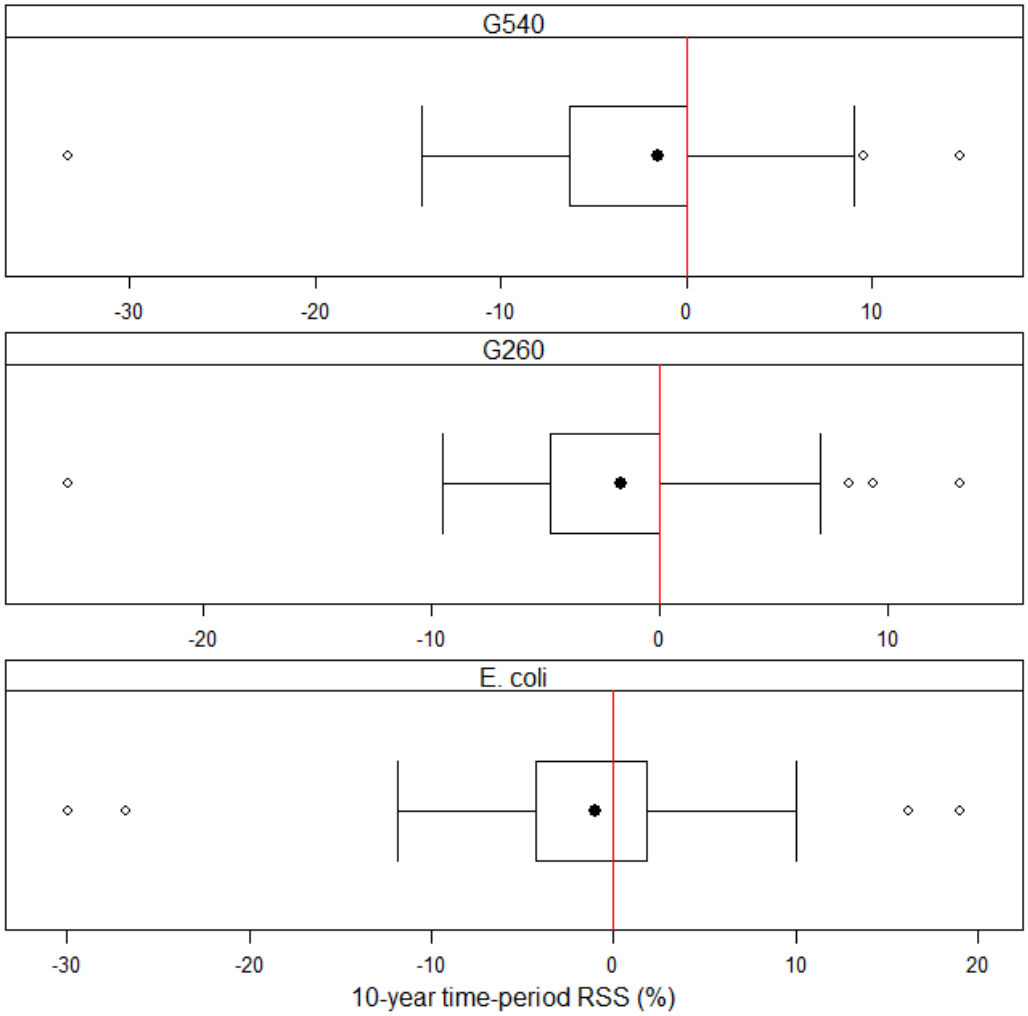

Figure 18. Distribution of trend magnitudes (RSS values) for the E. coli statistics at the 69 SoE sites included in the 10-year time-period dataset.Trends are all based on analyses performed using raw (i.e., not flow adjusted) data. All sites complied with the inclusion rules but their directions were not necessarily established with confidence.

5.3.2 Seven-year time-period

Of the 88 SoE sites retained in the seven-year time-period dataset, 86 had sufficient *E. coli* data for trend analysis, and 37, 76 and 62 had sufficient data for Clarity, SSC and Turbidity respectively. A large proportion of trends were uncertain (Table 10, Figure 19). However, there were more trends that were characterised as likely to be improving than degrading (Figure 20). For *E. coli* median, G260 and G540, 65%, 81% and 78% of site trends were as likely as not to be improving (Figure 20). For Clarity, SSC and Turbidity, 73%, 95% and 90% of site trends were as likely as not to be improving (Figure 20). Most of the certain trends for clarity, SSC and turbidity were improving (Table 10). Trends in G260 and G540 were not analysed for some sites due to high proportions of zeros in the time series of annual values. There was no discernible geographic pattern in the trend status and improving, degrading and uncertain trends occurred throughout the region (Figure 16).

Table 10. Trend analysis results for the 88 SoE sites included in the seven-year period dataset. Values in parentheses are proportion of sites (%). Trends are all based on analyses performed using raw (i.e., not flow adjusted) data.

Variable	Decreasing	Increasing	Not Analysed	Uncertain
E. coli	6 (7)	7 (8)	0 (0)	73 (85)
G260	26 (30)	1 (1)	2 (2)	59 (67)
G540	17 (19)	1 (1)	6 (7)	64 (73)
Clarity	2 (5)	3 (8)	0 (0)	32 (86)
SSC	8 (11)	0 (0)	0 (0)	68 (89)
Turbidity	26 (42)	0 (0)	0 (0)	36 (58)

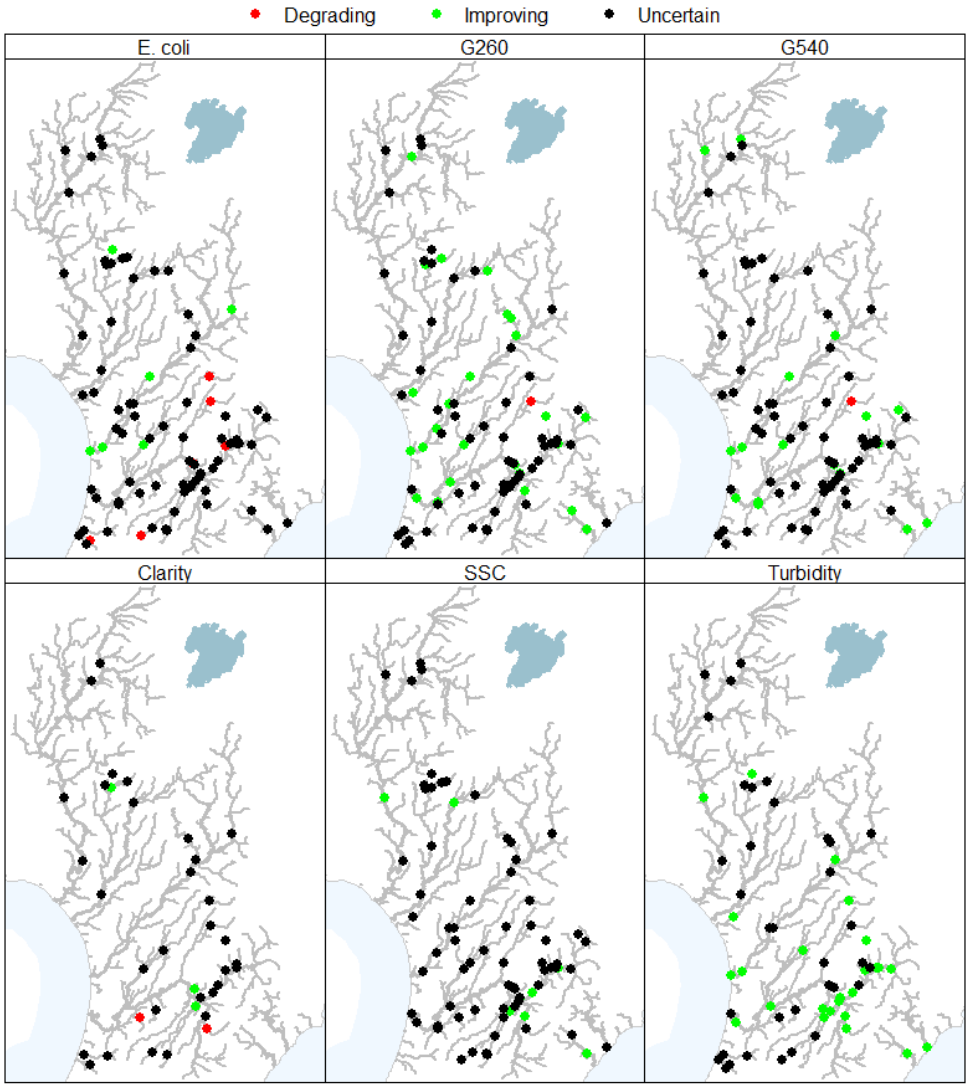

Figure 19. Map of SoE sites classified by their seven-year trend descriptions. Note that trend descriptions indicate degrading and improving (rather than trend direction). Trends are all based on analyses performed using raw (i.e., not flow adjusted) data.

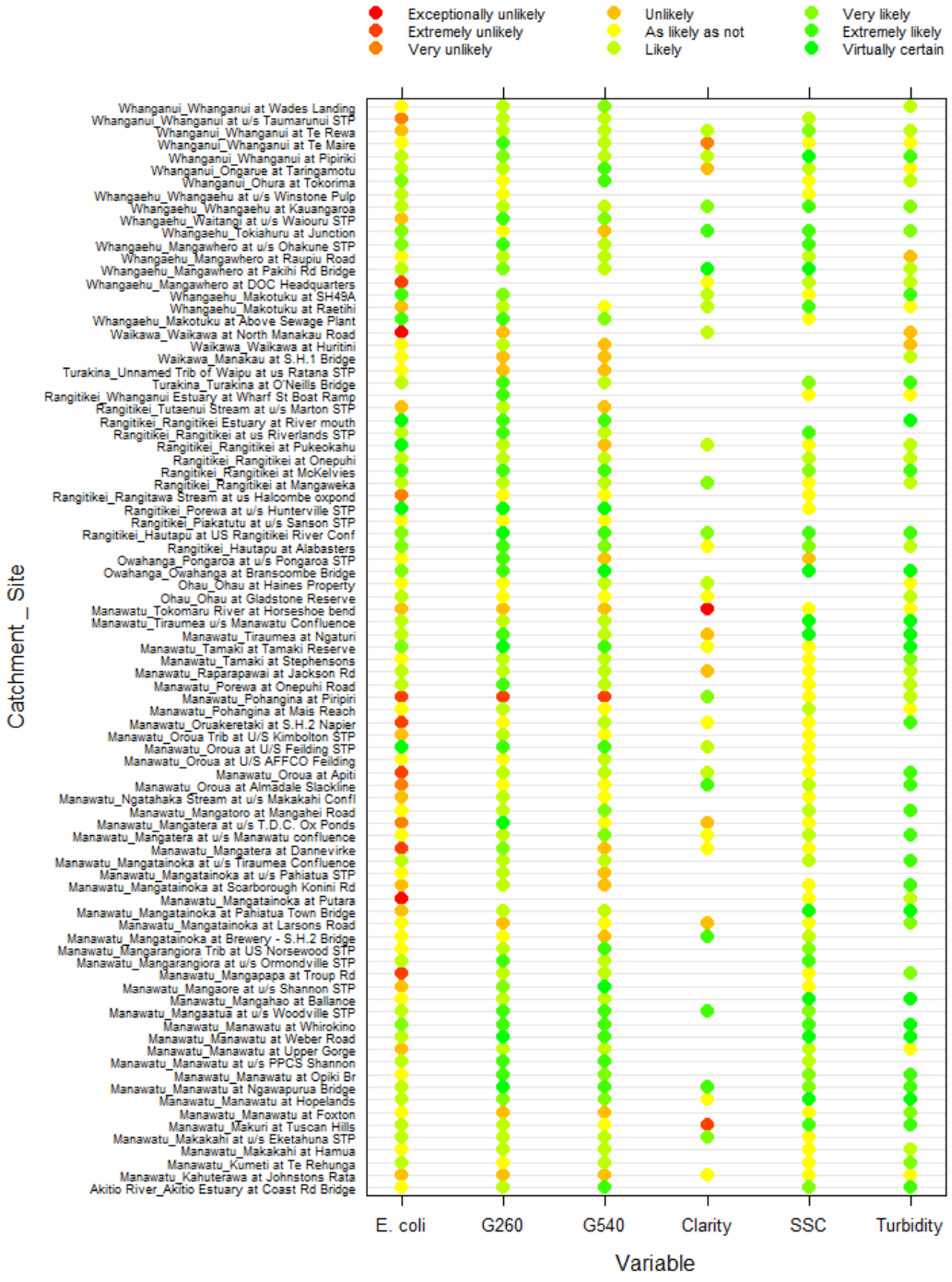

Figure 20. Summary plot of seven-year time-period trend analysis results. The plot shows the level of confidence that water quality was improving at each site and variable. Combinations of site and variable for which data was not available or trends were categorised as not analysed are shown as missing dots. Missing dots indicate the variable was either not monitored or the water quality trend description was 'not analysed'. See Table 3 for details of the confidence categories. Sites are grouped by the sea draining catchment to which they belong and then alphabetical order of the site names (separated by an underscore).

For the seven-year time-period, trend direction (irrespective of confidence) was predominantly improving water quality, particularly for trends in the proportion of samples exceeding the 260 and 540 thresholds. However, trend magnitudes differed greatly between sites (Figure 21).

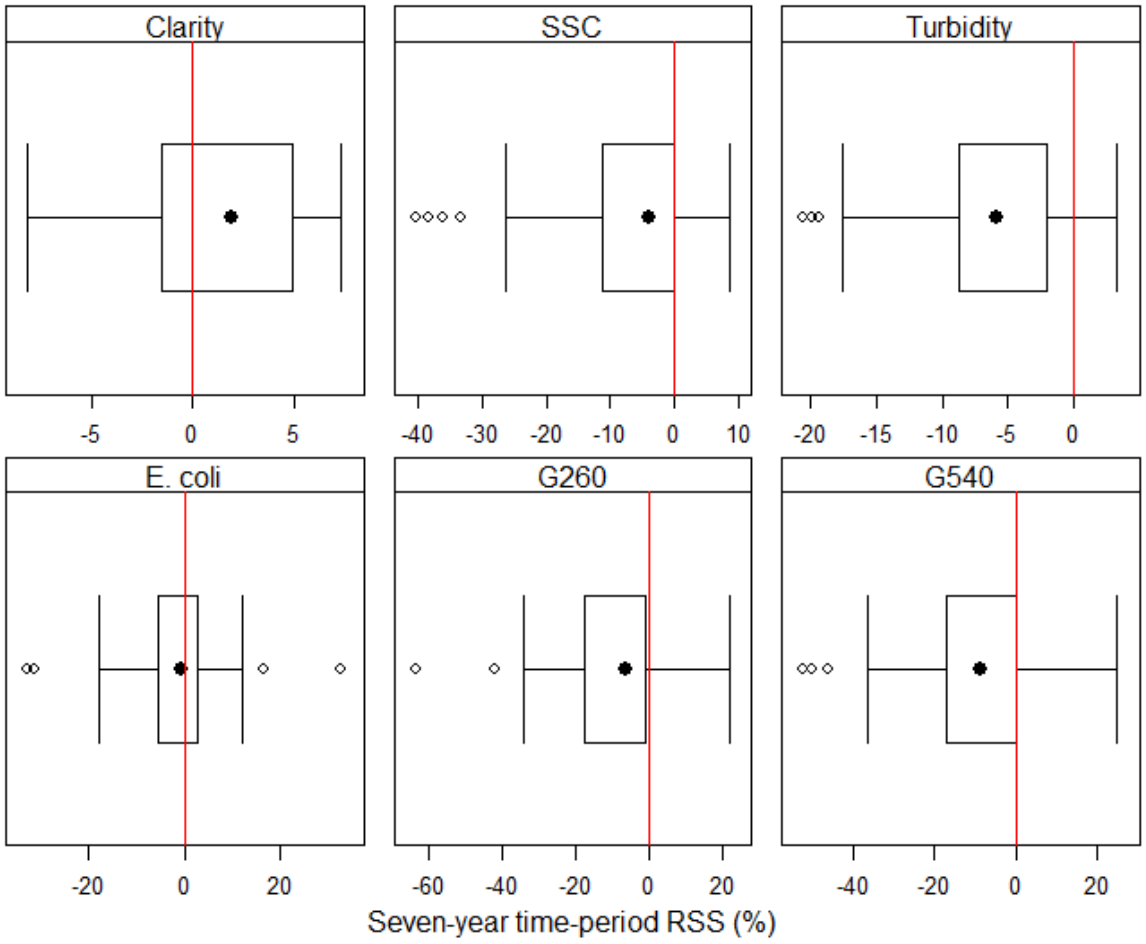

Figure 21. Distribution of trend magnitudes (RSS values) for the water quality variables at the 89 SoE sites included in the seven-year time-period dataset. Trends are all based on analyses performed using raw (i.e., not flow adjusted) data. All sites complied with the inclusion rules but their directions were not necessarily established with confidence.

5.4 Trends at discharge and impact sites

5.4.1 Analysis of time-periods

The results of the analysis of the number of discharge and impact sites with adequate data versus length of time-period are shown in Figure 22. There were no sites that met the filtering rules with 10 years of adequate turbidity or SSC data. There was a reduction in the rate of increase of sites that met the filtering rules for time periods less than seven years. For the seven-year period ending 2016 there was an increase in the total number of impact and discharge sites that met the filtering rules if quarterly rather than monthly data was used for trend analysis. For example, for SSC and clarity, there were 11 and 8 sites that met the filtering rules if monthly data were considered and this increased to 22 and 19 sites respectively if quarterly data were used (Figure 22).

The seven-year period ending 2016 and quarterly data was adopted for analysis of the discharge and impact sites. This represented a reasonable trade-off between number of sites and the duration of the time-period over which HRC's initiatives are expected to have improved water quality. This time period yielded 23, 0, 22 and 0 discharge sites that met the filtering rules for *E. coli*, clarity, SSC and turbidity respectively (Figure 22). The seven-year time period yielded 31, 19, 30 and 4 impact sites that met the filtering rules for *E. coli*, clarity, SSC and turbidity respectively (Figure 22).

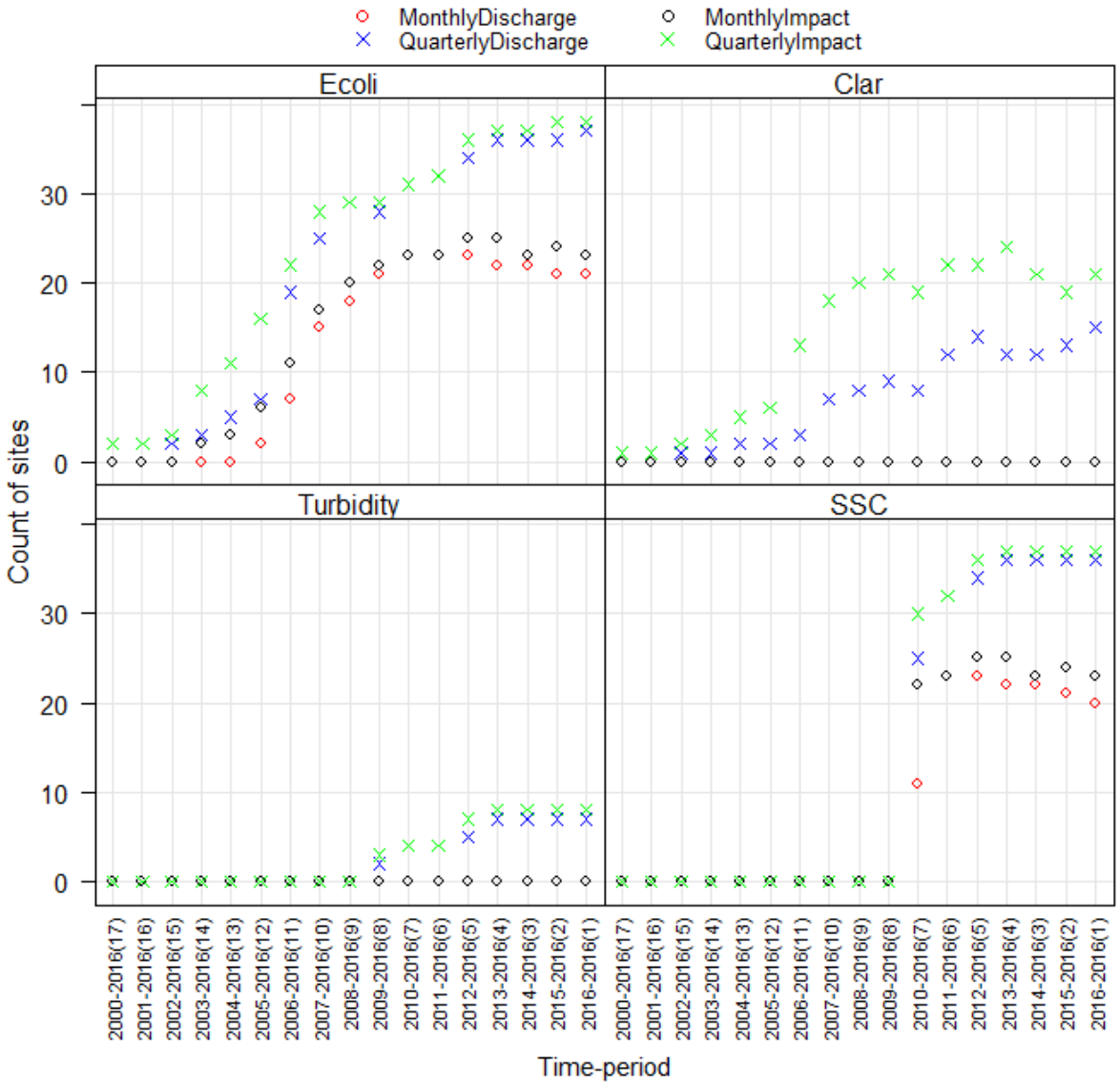

Figure 22. Trade-off between number of impact and discharge sites and the trend-period length. The plots for each variable show the number of sites that comply with the filtering rules when seasons are defined as months or quarters.

5.4.2 Discharge sites

There were 23 discharge sites that complied with the filtering rules for the seven-year time-period of which 23 and 22 had *E. coli* and SSC data respectively[7]. A large proportion of trends were uncertain; however, more of the certain trends for were improving than degrading (Table 11, Figure 23). In addition, there were more trends characterised as likely to be improving than degrading (Figure 24. There was no discernible geographic pattern in the trend status and improving, degrading and uncertain trends occurred at discharge sites throughout the region (Figure 23).

Table 11. Trend analysis results for E. coli, and SSC at the 23 discharge sites included in the seven-year period dataset. Values in parentheses are proportion of sites (%). Trends are all based on analyses performed using raw (i.e., not flow adjusted) data.

Variable	Degrading	Improving	Uncertain
E. coli	3 (13)	6 (26)	14 (61)
SSC	6 (26)	7 (30)	9 (39)

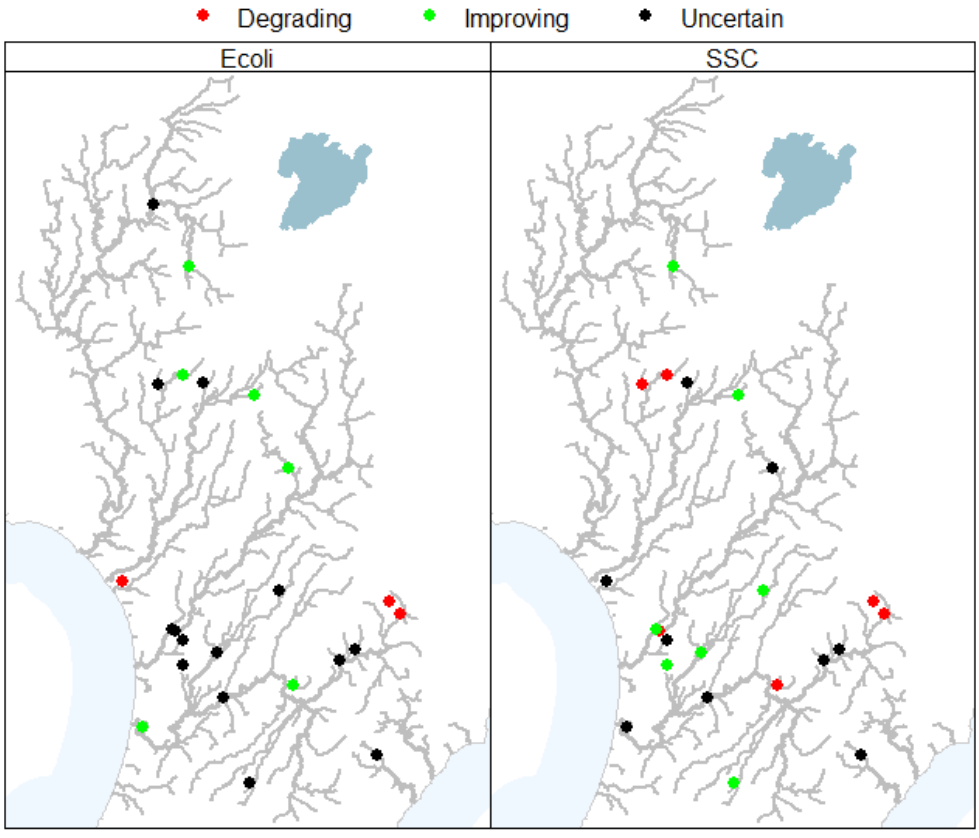

Figure 23. Map of discharge sites classified by their 10-year trend descriptions. Note that trend descriptions indicate degrading and improving (rather than trend direction). Trends are all based on analyses performed using raw (i.e., not flow adjusted) data.

[7] A complete set of trend analysis results for the 23 discharge sites is provided as supplementary data in file "TrendsDischargeSites.csv"

Figure 24. Summary plot of seven-year time-period trend analysis results based on quarterly data for impact sites. The plot shows the level of confidence that water quality was improving at each site and variable. Combinations of site and variable for which data was not available or trends were categorised as not analysed are shown as missing dots. Missing dots indicate the variable was either not monitored or the water quality trend description was 'not analysed'. See Table 3 for details of the confidence categories. Sites are grouped by the sea draining catchment to which they belong and then alphabetical order of the site names (separated by an underscore).

5.4.3 Impact sites

There were 31 impact sites that complied with the filtering rules for the seven-year time-period of which *E. coli*, clarity, SSC and turbidity data was available for at 31, 19, 30 and 4 sites respectively. A large proportion of trends were uncertain but most of the certain trends were improving (Table 12, Figure 25). In addition, there were more trends characterised as likely to be improving than degrading (Figure 26). There was no discernible geographic pattern in the trend status and improving, degrading and uncertain trends occurred throughout the region (Figure 25).

Table 12. Trend analysis results for E. coli, clarity SSC and turbidity at the impact sites included in the 10-year period dataset. Values in parentheses are proportion of sites (%). Trends are all based on analyses performed using raw (i.e., not flow adjusted) data.

Variable	Degrading	Improving	Uncertain
Clarity	0 (0)	5 (26)	14 (74)
E. coli	4 (13)	6 (19)	21 (68)
SSC	1 (3)	2 (6)	27 (91)
Turbidity	0 (0)	0 (0)	4 (100)

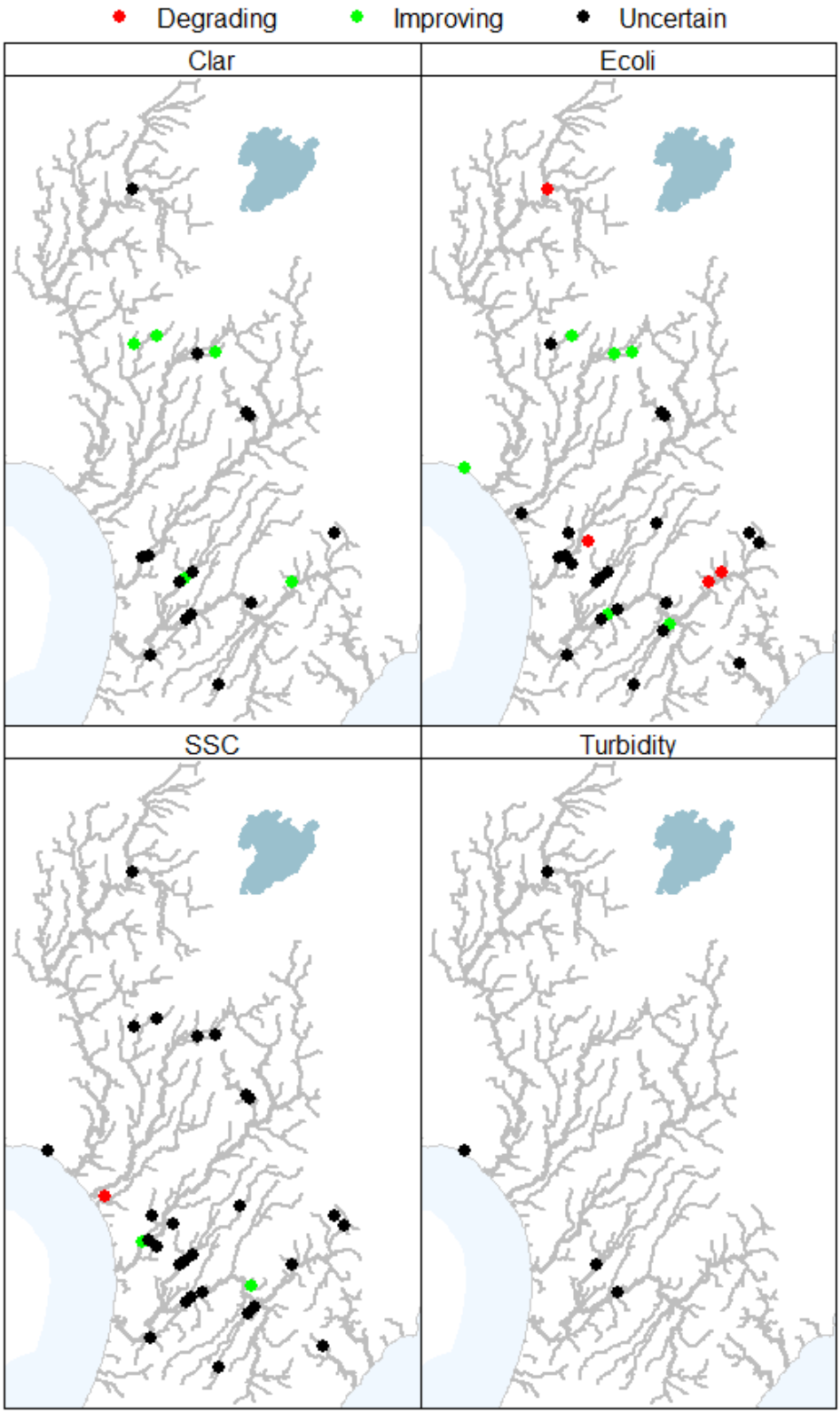

Figure 25. Map of impact sites classified by their 10-year trend descriptions. Note that trend descriptions indicate degrading and improving (rather than trend direction). Trends are all based on analyses performed using raw (i.e., not flow adjusted) data.

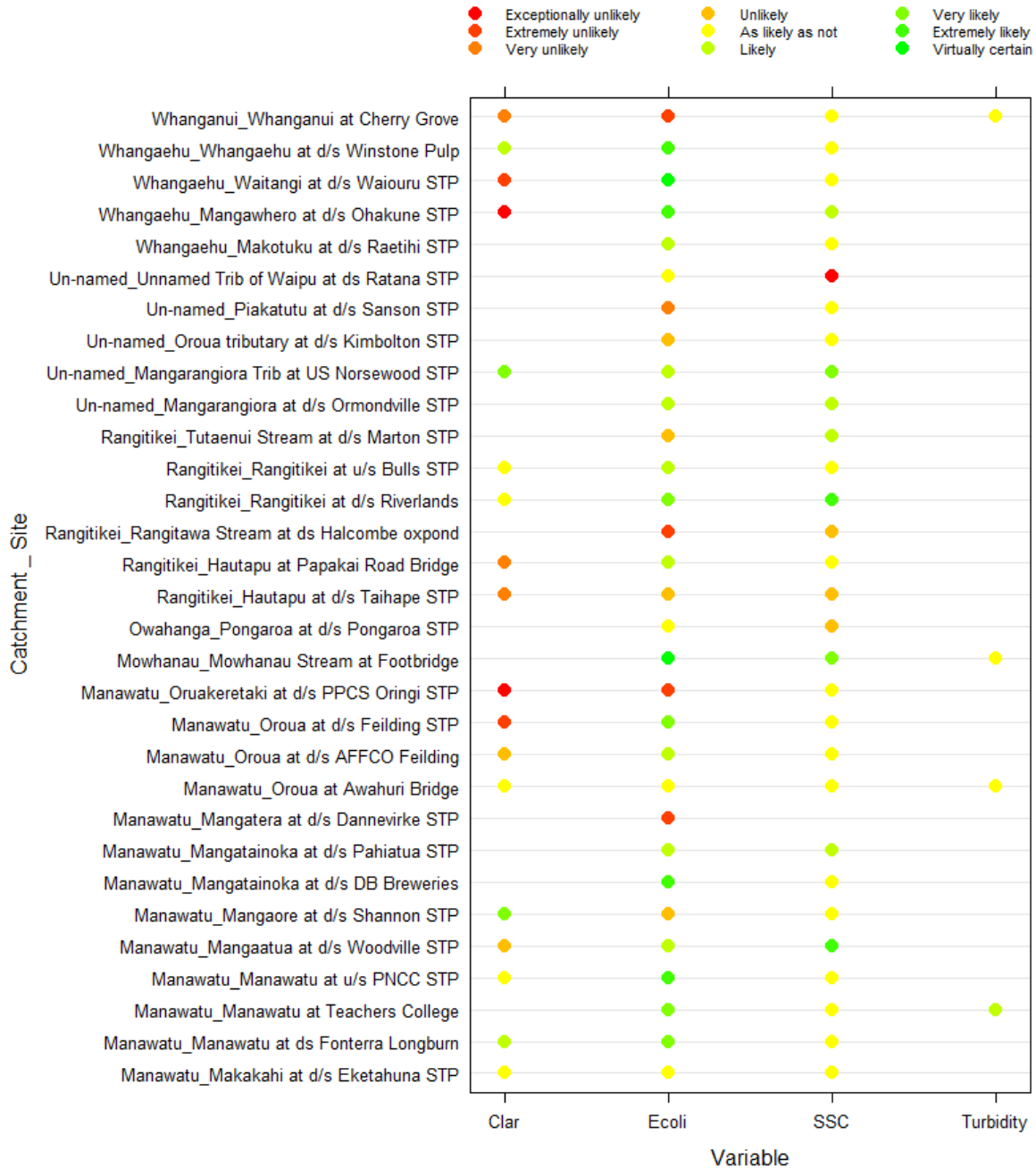

Figure 26. Summary plot of seven-year time-period trend analysis results for impact sites. The plot shows the level of confidence that water quality was improving at each site and variable. Combinations of site and variable for which data was not available or trends were categorised as not analysed are shown as missing dots. Missing dots indicate the variable was either not monitored or the water quality trend description was 'not analysed'. See Table 3 for details of the confidence categories. Sites are grouped by the sea draining catchment to which they belong and then alphabetical order of the site names (separated by an underscore).

5.5 Spatial models of current water quality state

5.5.1 Swimming grades based on 10-year time-period dataset

The performance of the regional spatial models of the *E. coli* statistics (median, G260 and PropGT540) were at or close to adequate as indicated by the following statistics NSE > 0.50, RSR < 0.70, and if PBIAS < ±25% (Table 13, Moriasi *et al.* (2007)). The performance of the regional spatial models was lower than the national models of the same *E. coli* statistics reported by Snelder *et al.* (2016a) (Table 14).

Table 13. Performance of the spatial models of E. coli state based on the 10-year time-period dataset. The fitting dataset comprised 68 regional sites. Performance was determined using independent predictions (i.e., sites that were not used in fitting the models) generated from the out-of-bag observations. NSE = Nash-Sutcliffe efficiency, PBIAS = percent bias (%), RSR = relative root mean square error, RMSD = root mean square deviation. Units for RMSD for Median are log_{10} E. coli 100mL^{-1}. RMSD units for G260 and G540 are logit transformed proportions.

Model (*E. coli* statistic)	NSE	PBIAS	RSR	RMSD
Median	0.48	-1	0.7	0.38
PropGT260	0.51	5	0.7	0.94
PropGT540	0.50	0.4	0.7	0.79

Table 14. Performance of the national spatial models of E. coli state. The fitting dataset comprised the 753 national sites (see Snelder et al., 2016a). Performance was determined using independent predictions (i.e., sites that were not used in fitting the models) generated from the out-of-bag observations. NSE = Nash-Sutcliffe efficiency, PBIAS = percent bias (%), RSR = relative root mean square error, RMSD = root mean square deviation. Units for RMSD for Median are log_{10} E. coli 100mL^{-1}. RMSD units for G260 and G540 are proportions (i.e., original scale [0 – 1])

Model (*E. coli* statistic)	NSE	PBIAS	RSR	RMSD
Median	0.72	-0.40	0.5	0.35
G260	0.67	-1.02	0.6	0.16
G540	0.58	-2.07	0.6	0.13

The relationships between predictors and the three *E. coli* statistics are indicated by partial dependence plots (PDP, Figure 27). The PDPs indicate associations between *E. coli* statistics and the proportion of the catchment occupied by high producing exotic grassland and scrub land (usPastoral and usScrub), catchment elevation and slope (usCatElev, usAveSlope), climatic variables (usAveTCold, usAveTWarn) and segment elevation (segAveElev) (Figure 27). These relationships were consistent between *E. coli* statistics and with expectations. For example, the values of all *E. coli* statistics increased with increasing pastoral land cover and decreased with increasing catchment elevation and slope (Figure 27). The association of the *E. coli* statistics with pastoral land cover and elevation is consistent with the dominant source of faecal contamination being grazing animals and is consistent with recent evaluations of environmental patterns in river water quality (e.g., Larned *et al.* (2016b); Unwin *et al.* (2010)).

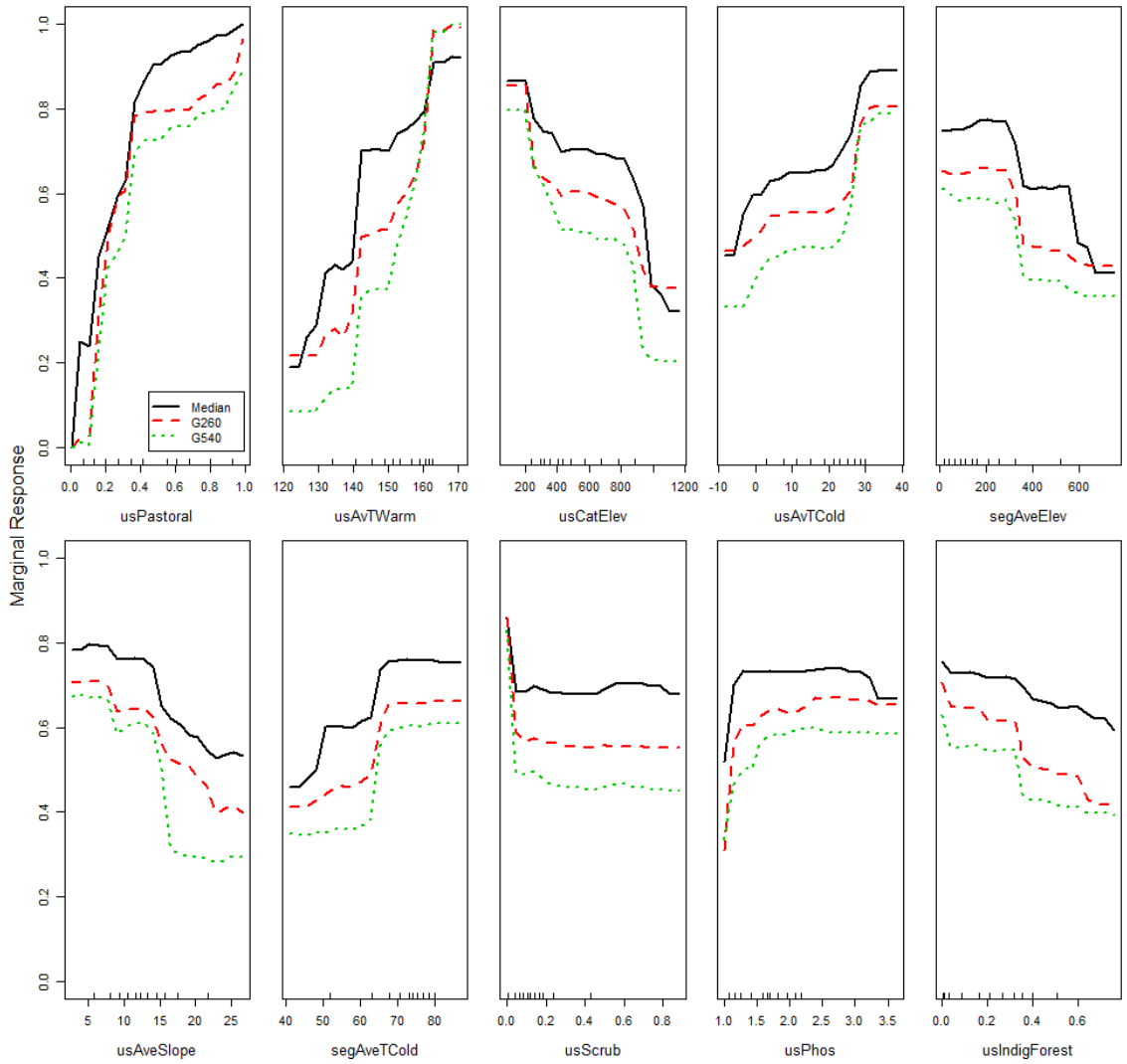

Figure 27. PDPs for the eight most important predictor variables in RF models of the three Escherichia coli statistics based on the 10-year time-period dataset. Each panel corresponds to one predictor. The Y-axis is the standardised value of the marginal response for each the three statistics. In each case, the original marginal responses over all eight 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 predictors with high importance. Legend in top left panel applies to all panels. Predictor variables are defined in Table 4.

Predictions of state and swimming grades assessments based on the regional 10-year *E. coli* models are shown in Figure 28. Assessment of regional river swimming grades, calculated from predictions made by the regional 10-year models and the national models of Snelder *et al.* (2016a), are shown in Table 15. The proportion of river segments in each swimming grade estimated using the national model and the regional model were in reasonable agreement. The national model estimated that 45% of river length of stream order four or greater was swimmable (grade fair or better) whereas the regional model estimate was 38%. When all segments were considered, the national model predicted 36% swimmable and the regional model predicted 37%.

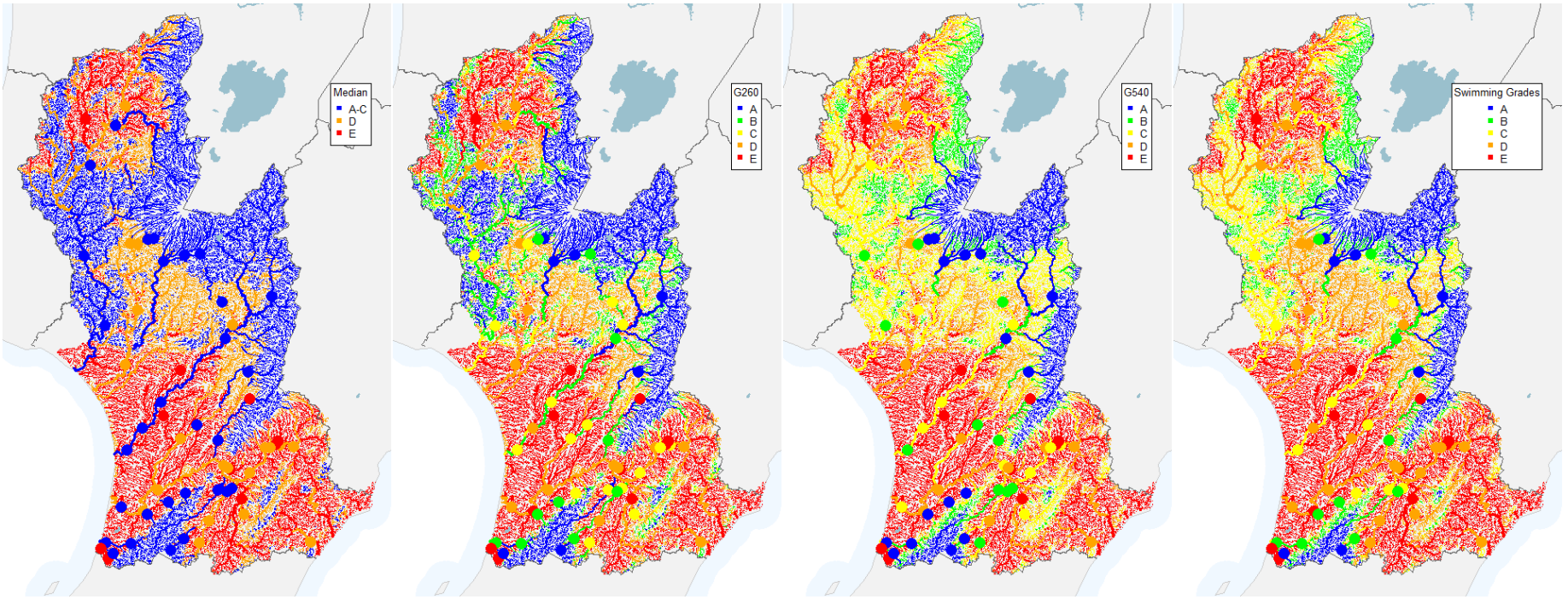

Figure 28. Spatial model predictions made using RF models and transformed response variables for the 69 SoE sites represented in the 10-year dataset. A log_{10} transformation was applied to the median and logit transformations was applied to G260 and G540 prior to fitting the models. The right-hand map represents the predicted swimming grade derived from analysis of the predicted values of the three statistics to the left. SoE sites are shown as dots with the colour representing the observed grade for the site (i.e., not the grade predicted by the model).

Table 15. Swimming grades determined using the regional models based on the 10-year dataset and the national models. Tabulated values are proportions of the river network by length (%) in each swimming grade.

Models	Excellent	Good	Fair	Swimmable	Intermittent	Poor	Not swimmable
National models (Order 4+)	12	12	21	46	40	16	56
Regional model (Order 4+)	12	8	18	38	35	26	61
National models (All segments)	20	10	6	36	20	44	64
Regional model (All segments)	12	10	15	37	22	41	63

5.5.2 Swimming grades based on summer data

The performance of the regional spatial models of the summer *E. coli* statistics (median, G260 and G540) were at or close to adequate as indicated by the following statistics NSE > 0.50, RSR < 0.70, and if PBIAS < ±25% (Table 16, Moriasi *et al.*, 2007) and were very similar to the performance of the models based on statistics derived from the complete dataset (Table 13). The modelled relationships represented by the summer models were very similar to the models derived from the complete dataset (data not shown).

Table 16. Performance of the spatial models of E. coli state based on the summer season statistics for 10-year time-period dataset. The fitting dataset comprised 69 regional sites. Performance was determined using independent predictions (i.e., sites that were not used in fitting the models) generated from the out-of-bag observations. NSE = Nash-Sutcliffe efficiency, PBIAS = percent bias (%), RSR = relative root mean square error, RMSD = root mean square deviation. Units for RMSD for Median are log_{10} E. coli $100mL^{-1}$. RMSD units for G260 and G540 are logit transformed proportions.

Variable (model)	NSE	PBIAS	RSR	RMSD
E. coli	0.48	-1	0.7	0.38
G260	0.47	5	0.7	0.99
G540	0.51	0.5	0.7	0.78

Predictions of state and swimming grades assessments based on the summer data for the 10-year period are shown in Figure 29. The summer model estimated that 36% of river length of stream order four or greater was swimmable (grade fair or better, Table 19), which was consistent with the all-year model estimate (38%, Table 15). By contrast, when all segments were considered, the summer model predicted 17% (Table 17swimmable and the all year model predicted 37% (Table 15). The difference is consistent with the observation that SoE sites representing smaller headwater streams tended to have poorer grades in summer compared to grades derived from the complete dataset, whereas sites on mainstem rivers tended to have better grades in summer compared grades derived from the complete dataset (Figure 14). Because smaller rivers make up the largest contribution to total river length in the

region, the proportion of swimmable rivers is much lower when all rivers are considered compared to just forth order or larger.

Table 17. Swimming grades determined using the summer models based on the 10-year dataset and the national models. Tabulated values are proportions of the river network by length (%) in each swimming grade.

Models	Excellent	Good	Fair	Swimmable	Intermittent	Poor	Not swimmable
Summer models (Order 4+)	8	7	21	36	28	35	63
Summer models (All segments)	2	9	6	17	17	66	83

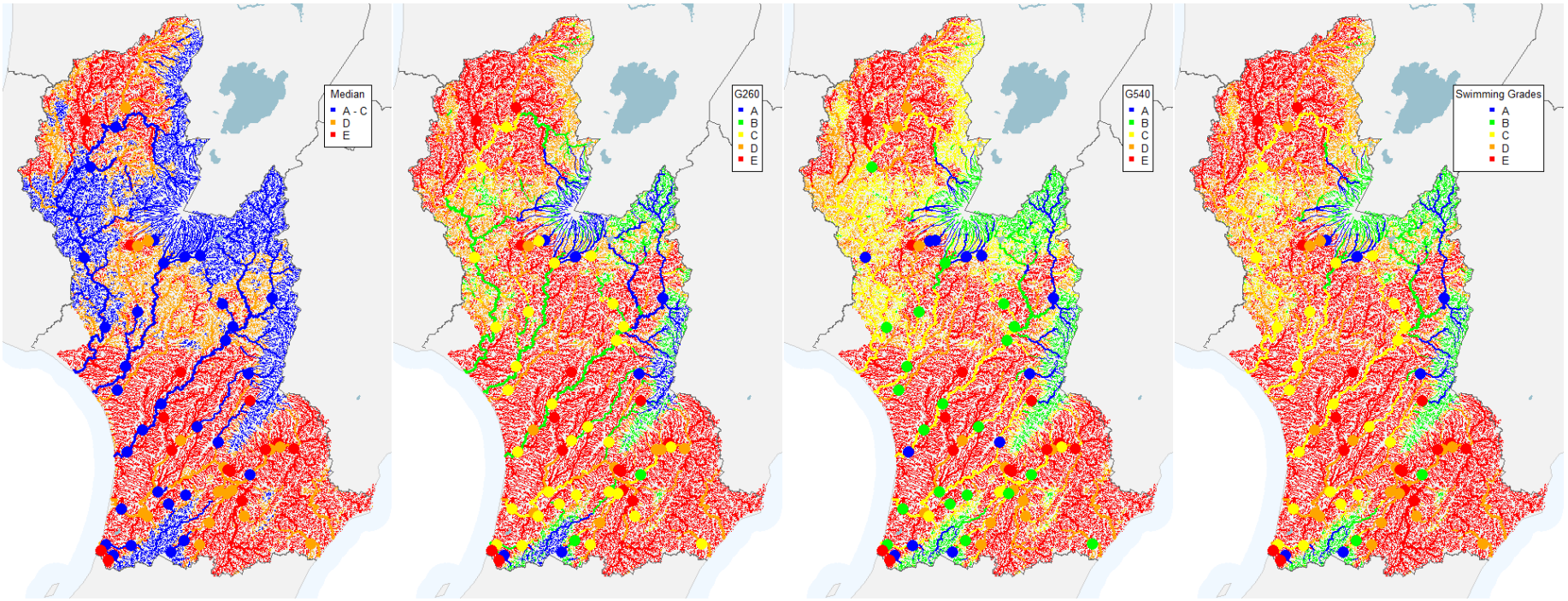

Figure 29. Spatial model predictions made using RF models and transformed response variables for the 69 SoE sites represented in the summer 10-year dataset. A log_{10} transformation was applied to the median and logit transformations was applied to G260 and G540 prior to fitting the models. The right-hand map represents the predicted swimming grade derived from analysis of the predicted values of the three statistics to the left. SoE sites are shown as dots with the colour representing the observed grade for the site (i.e., not the grade predicted by the model).

5.5.3 Swimming grades based on seven-year time-period dataset

The performance of the regional spatial models of the *E. coli* statistics (median, G260 and PropGT540) based on the seven-year time-period dataset was at or close to adequate as indicated by the following statistics NSE > 0.50, RSR < 0.70, and if PBIAS < ±25% (Table 18, Moriasi *et al.* (2007). The performance of the regional spatial models was lower than the national models of the same *E. coli* statistics (Table 14).

Table 18. Performance of the E. coli spatial models based on the seven-year time-period dataset. The fitting dataset comprised 86 sites. Performance was determined using independent predictions (i.e., sites that were not used in fitting the models) generated from the out-of-bag observations. NSE = Nash-Sutcliffe efficiency, PBIAS = percent bias (%), RSR = relative root mean square error, RMSD = root mean square deviation. Units for RMSD for Median are log_{10} E. coli / 100mL. RMSD units for G260 and G540 are logit transformed proportions.

Model (*E. coli* statistic)	**NSE**	**PBIAS**	**RSR**	**RMSD**
Median	0.49	-1.0	72	0.35
G260	0.47	0.5	74	0.65
G540	0.46	0.3	75	0.57

The relationships between predictors with high importance in the RF models and the three *E. coli* statistics are indicated by PDPs (Figure 30). The PDPs indicate associations between the *E. coli* statistics and the proportion of the catchment occupied by high producing exotic grassland and scrub land (usPastoral and usScrub), catchment elevation and slope (usCatElev, usAveSlope), climatic variables (usAveTCold, usAveTWarn) and segment elevation (segAveElev) (Figure 30). As for the state models derived using the 10-year time-period data, these relationships were consistent between *E. coli* statistics and with expectations.

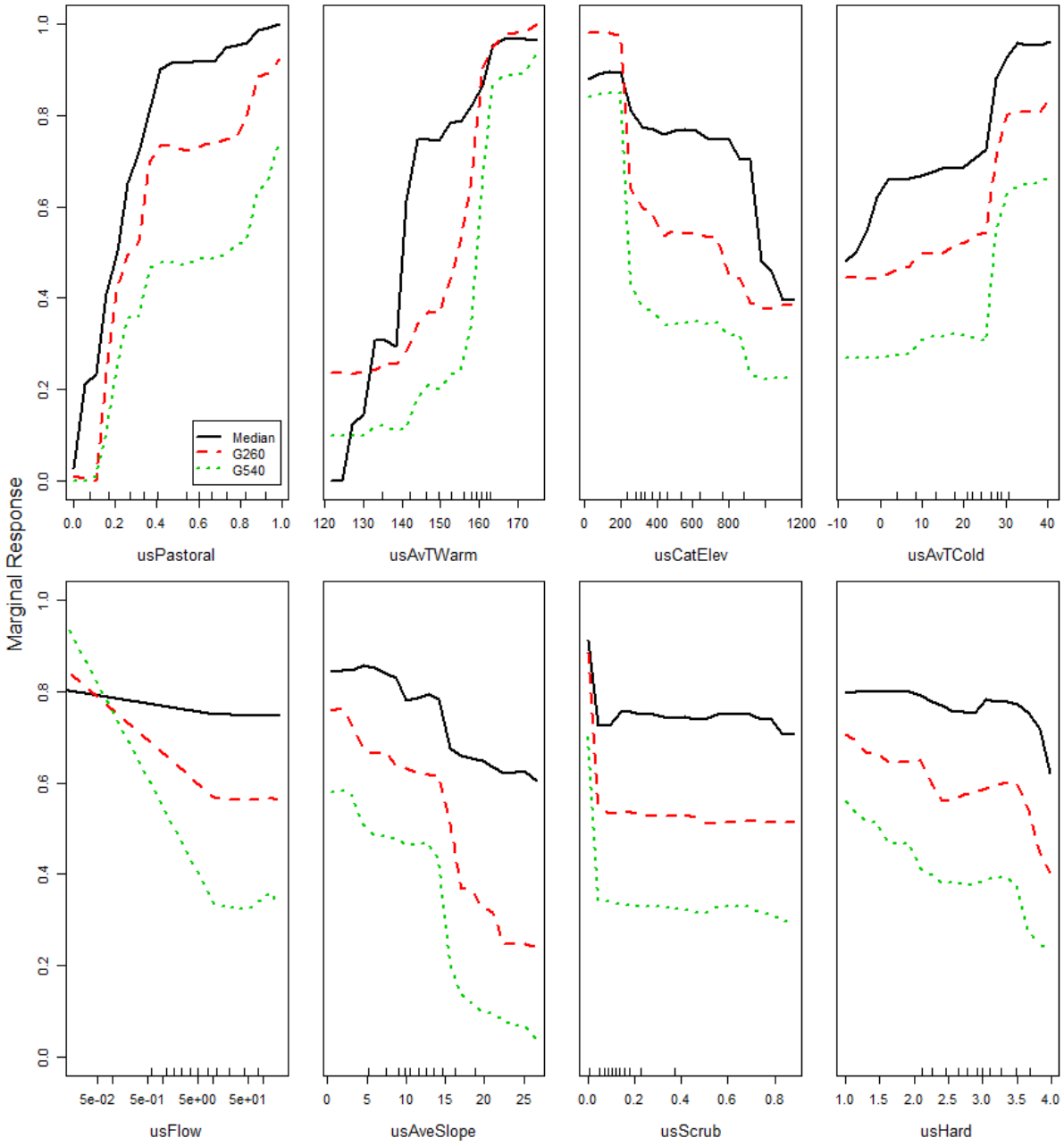

Figure 30. PDPs for the eight most important predictor variables in Random Forest models of the three Escherichia coli statistics for the seven-year time-period. Each panel corresponds to one predictor. The Y-axis is the standardised value of the marginal response for each the three statistics. In each case, the original marginal responses over all eight 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 predictors with high importance. Legend in top left panel applies to all panels. Predictor variables are defined in Table 4.

Predictions of state and swimming grades assessments using the regional seven-year models are shown in Figure 31. Assessment of regional river swimming grades, calculated from these predictions are shown in Table 18. The proportion of river segments in each swimming grade estimated using the seven-year models did not strongly agree with those of the 10-year

dataset or the national models. The seven-year model estimated lower proportions of river length in excellent and fair grades than the 10-year and national models and higher proportions in the good grade (Table 15, Table 19). The seven-year models resulted in very few network segments being allocated the fair grade. However, the proportion of order 4 or higher segments of grade fair or better (i.e., swimmable) estimated using the seven-year models were in close agreement with the 10-year models (41%, 38% respectively) and were in reasonable agreement with the national models (45%). When all segments were considered, the proportion of segments of grade fair or better (i.e., swimmable) estimated using the seven-year, 10-year and national models closely agreed, reporting results of 38%, 37% and 36% respectively (Table 15, Table 19).

Table 19. Swimming grades determined using models of the E. coli statistics based on the seven-year time-period dataset. Tabulated values are proportions of the river network by length (%) in each swimming grade.

Models	Excellent	Good	Fair	Swimmable	Intermittent	Poor	Not swimmable
Regional model (Order 4+)	6	31	4	41	36	24	59
Regional models (All segments)	2	24	12	38	25	37	62

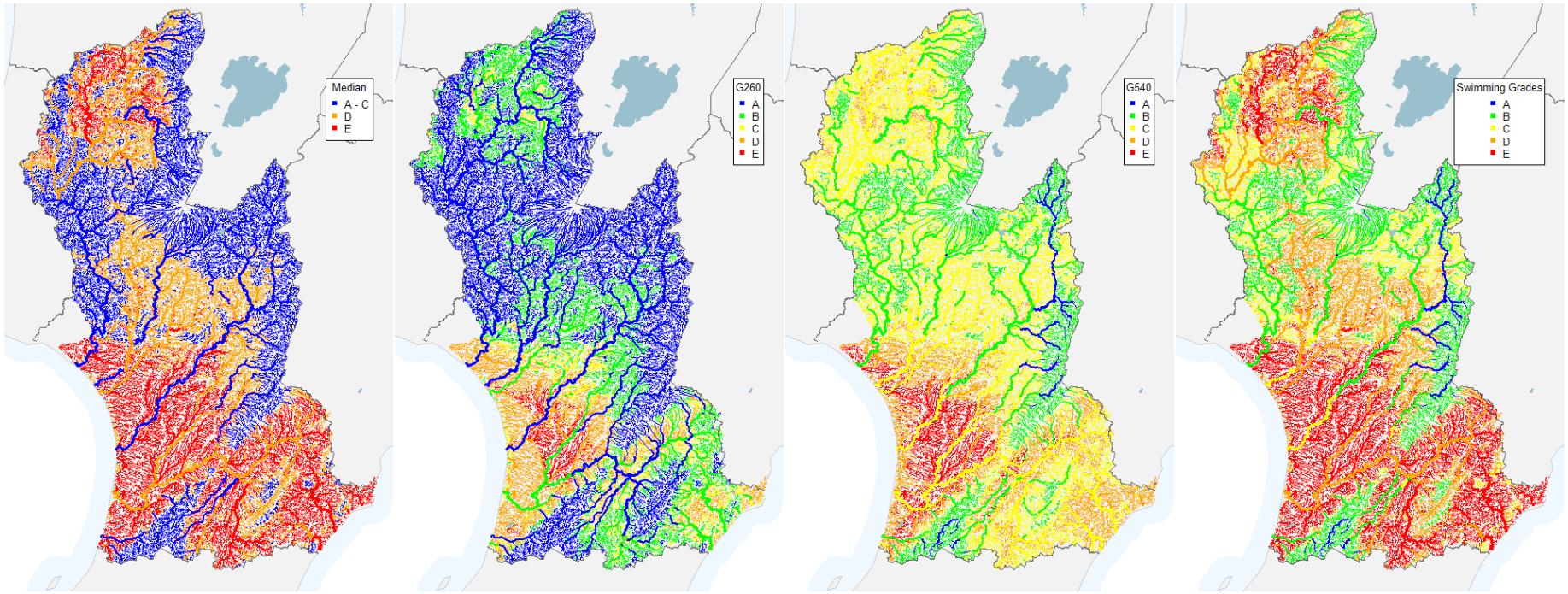

Figure 31. Spatial model predictions made using RF models and transformed response variables for the 86 SoE sites represented in the seven-year dataset. A log_{10} transformation was applied to the median and logit transformations were applied to G260 and G540 prior to fitting the models. The right-hand map represents the predicted swimming grade derived from analysis of the predicted values of the three statistics to the left. SoE sites are shown as dots with the colour representing the observed grade for the site (i.e., not the grade predicted by the model).

5.5.4 State for clarity, SSC and turbidity based on seven-year time-period dataset

The performance of the regional spatial models of the median site values of clarity, SSC and turbidity were satisfactory as indicated by the following statistics NSE > 0.50, RSR < 70, and if PBIAS < ±25% (Table 20, Moriasi *et al.*, 2007).

Table 20. Performance of the spatial models of clarity, SSC and turbidity based on the seven-year time-period dataset. The fitting dataset comprised varying numbers of sites by variable. Performance was determined using independent predictions (i.e., sites that were not used in fitting the models) generated from the out-of-bag observations. NSE = Nash-Sutcliffe efficiency, PBIAS = percent bias (%), RSR = relative root mean square error, RMSD = root mean square deviation. Units for RMSD are log_{10} of the original units for each variable (see Table 2).

Variable (model)	Number of sites	NSE	PBIAS	RSR	RMSD
Clarity	37	0.61	3.28	62.7	0.18
SSC	75	0.57	0.29	65.3	0.33
Turbidity	61	0.66	-1.63	58.4	0.28

The predictor variables with high importance in the RF models of clarity, SSC and turbidity state reflected associations of all three variables with catchment area and river mean flow (usArea and usFlow; Figure 32). Turbidity and SSC increased with area and mean flow and clarity decreased (Figure 32). Turbidity and SSC increased, and clarity decreased, with increasing values of usPhos (Figure 32). This relationship is expected because the predictor usPhos is high in catchments dominated by soft sedimentary geology, which tend to be more erosion-prone than catchments dominated by harder geological material. Turbidity and SSC increased, and clarity decreased, with increasing values of usPastoral and usExoticForest (Figure 32). There were complex relationships between clarity, SSC and turbidity and catchment slope, probably because of interactions with geology, which tends to be more erosion prone in soft sedimentary hill country and less so in steeper headwater catchments. The association of clarity, SSC and turbidity with the predictors are consistent with recent evaluations of environmental patterns in river water quality (e.g., Larned *et al.*, 2016b; Unwin *et al.*, 2010).

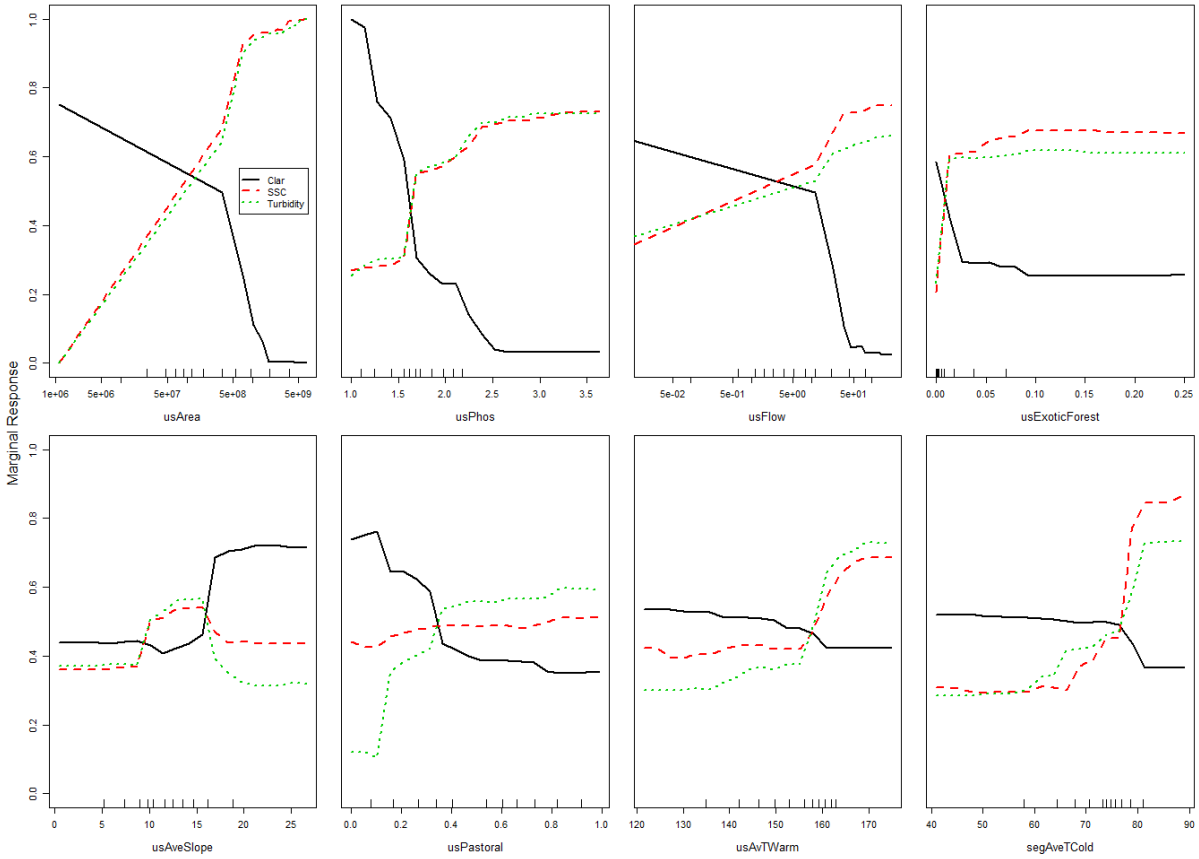

Figure 32. PDPs for the eight most important predictor variables in Random Forest models of clarity, SSC and turbidity based on the 7-year time-period dataset. Each panel corresponds to one predictor. The Y-axis is the standardised value of the marginal response for each the three variables. In each case, the original marginal responses over all eight 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 predictors with high importance. Legend in top left panel applies to all panels. Predictor variables are defined in Table 4.

Predictions of state for clarity, SSC and turbidity based on the seven-year period are shown in Figure 29. Mainstem rivers had lower clarity and higher turbidity and SSC that smaller rivers. Clarity was lower and turbidity and SSC was higher in rivers whose catchments were dominated by pasture and the soft sedimentary hill county areas that occupy the central catchment areas of the large river of the region (Figure 33). The current state predictions has not been compared to targets because HRC do not have targets for SSC and turbidity and the clarity target applies only to flows below the median.

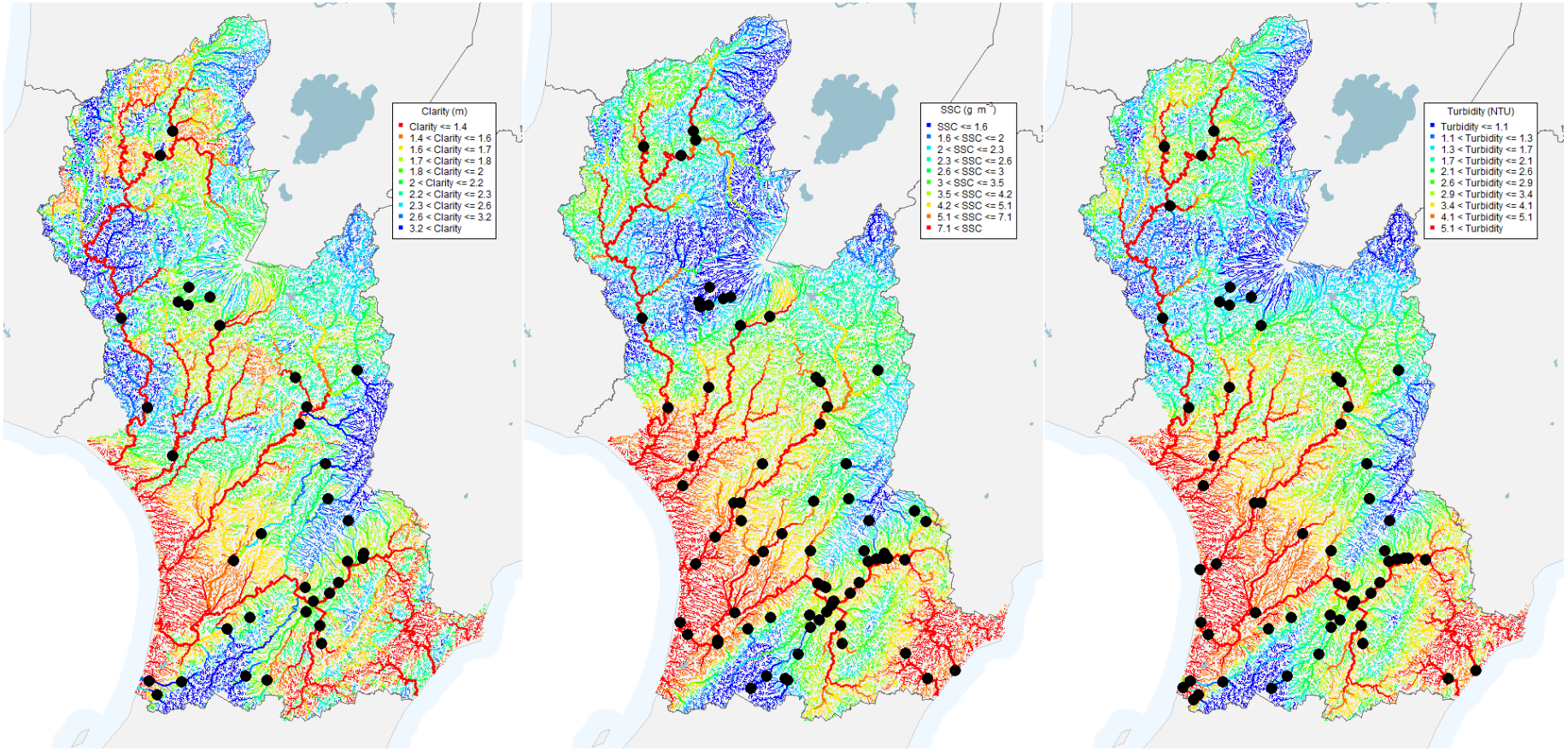

Figure 33. Spatial model predictions of clarity, SSC and turbidity made using RF models and transformed response variables for the SoE sites represented in the seven-year dataset. The fitting dataset comprised varying numbers of sites by variable (see Table 20). The colour scale on these maps is chosen to optimally discriminate the range in values and does not indicate the acceptability of the current state. Site grades are not shown on the map because the NOF does not define attributes for any water quality measures that are related to sediment.

5.6 Spatial model of changes in water quality

5.6.1 Changes in swimming grades for the 10-year time-period

Input data for step one of the assessment of changes in river swimming grades for the 10-year period are the analysis of trend direction and magnitude shown in Table 21. The mix of sites with increasing and decreasing trends was reasonably balanced for the median statistic but the G260 and G540 were dominated by decreasing trends.

Table 21. Sites with increasing and decreasing trends by E. coli statistic for the 10-year time-period. Trends at all sites were included in this analysis irrespective of confidence in trend direction.

***E. coli* statistic**	**No sites decreasing**	**No sites increasing**	**Median of decreasing trends**	**Median of increasing trends**
Median	39	27	-3.1	2.5
G260	46	13	-3.2	4.4
G540	47	10	-2.8	2.9

The random forest models of trend direction for the 10-year time-period had predictive misclassification rates of 36%, 25% and 17% for *E. coli*, G260 and G540 respectively. The AUC statistics for the models were 0.66, 0.72 and 0.61 for *E. coli*, G260 and G540 respectively. These AUC statistics indicate satisfactory performance for the three classification models. Low misclassification rates for G260 and G540 partly reflect the low occurrence of increasing trends (Table 21).

The relationships between trend direction and the model predictors were consistent across the three *E. coli* statistics (Figure 34). Predictions of trend direction for the region also showed a reasonable level of consistency across the three *E. coli* statistics (Figure 35). The model PDPs (Figure 34) and the regional predictions (Figure 35) indicate that:

1. The most important predictor is usPhos, which indicates that there is an association with trend direction and geology. The probability that the trend at a site was decreasing increases with increasing values of usPhos. Sites with catchments that are either volcanic or soft sedimentary geology tend to have high values of usPhos.
2. The partial plot indicates that the probability of a site having a decreasing trend increases as rainfall variability (usAnnRainVar) increases. This variable may not be causative but is correlated, having low values in the central part of the region which has a predominance of sites with increasing trends (Figure 35).
3. The probability of a site having a decreasing trend is maximum for river sites with intermediate sized (i.e., usArea ~500 km^2) catchments.
4. The probability of a site having a decreasing trend is maximum for river sites with intermediate values of usHard and generally increases with increasing values of usParticleSize. This indicates that there is an association with trend direction and catchment geology.
5. The probability of a site having a decreasing trend decreases with increasing indigenous forest cover (usIndigForest) and increasing scrub (usScrub).

6. The relationships with usAveSlope and usCatElev indicate that the probability of a site having a decreasing trend is maximum at intermediate slopes and elevations.

7. The probability of a site having a decreasing trend is maximum for river sites with catchments having intermediate values of high intensity rainfall (i.e., usRainDays10~3-3.5 month^{-1}).

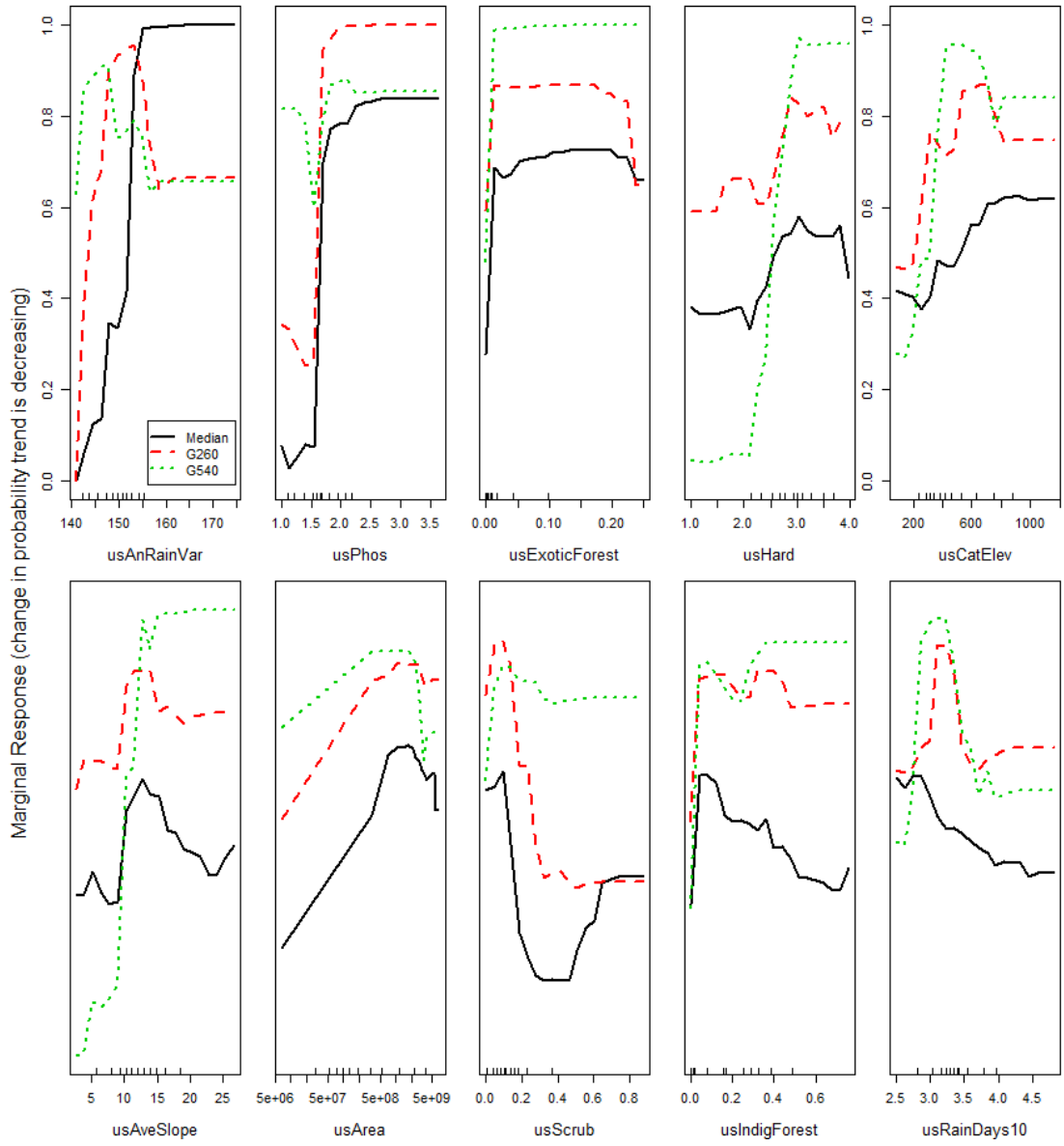

Figure 34. PDPs for the eight most important predictor variables in Random Forest models of the trend direction for the three E. coli statistic for the 10-year dataset. Each panel corresponds to one predictor. The Y-axis is the standardised value of the marginal change in probability the trend is decreasing for each of the eight modelled variables. Note that a decreasing trend in the variables shown indicates water quality improvement. In each case, the original marginal responses over all eight 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. Legend in top left panel applies to all panels. Predictor variables are defined in Table 4.

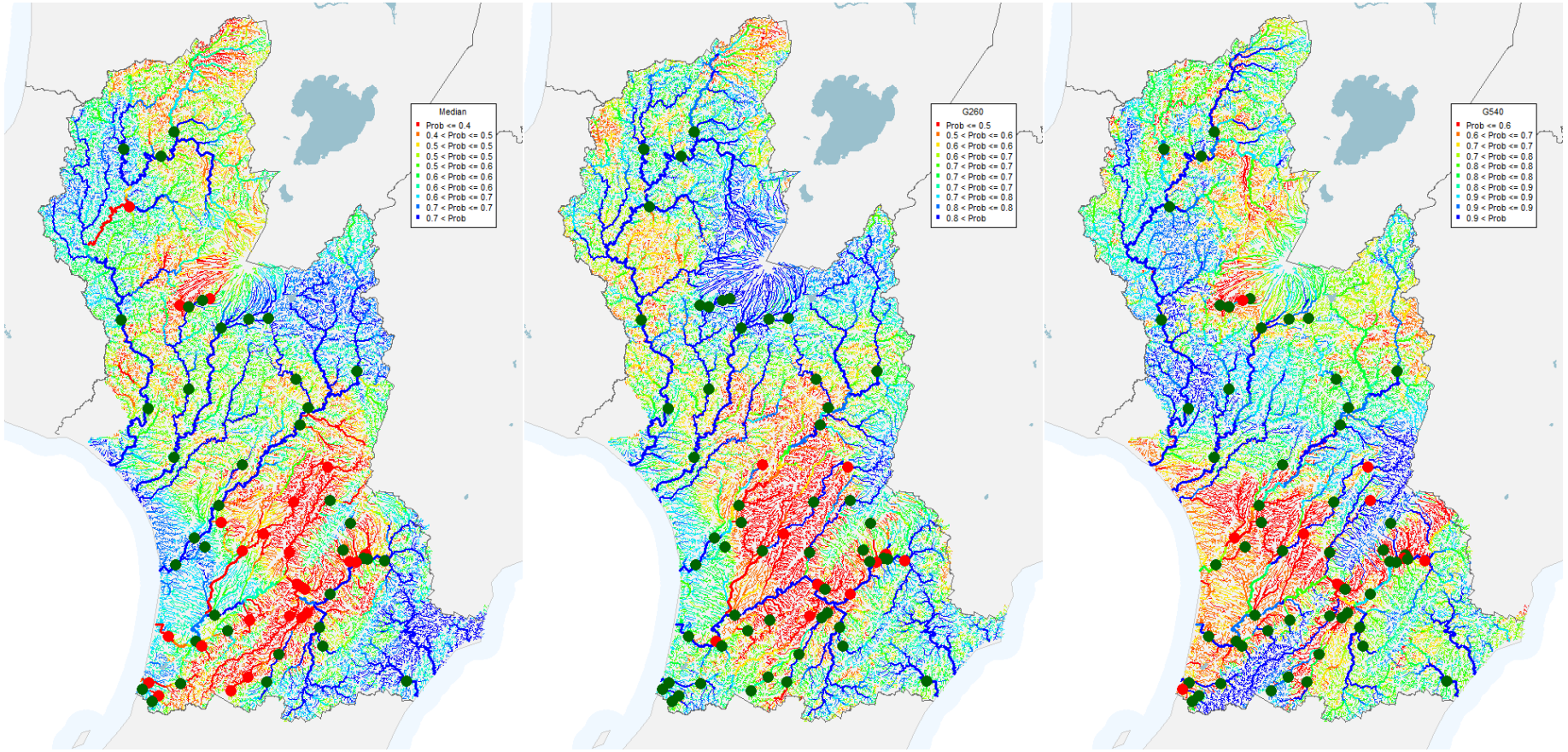

Figure 35. Spatial model predictions made using RF models of trend direction for the 69 SoE sites represented in the 10-year dataset. The plotted colours values represent differences in the probability that trend is decreasing. The red and green points represent the observed trend direction at the 69 SoE monitoring sites (degrading and improving respectively).

Predicted swimming grades at the beginning and end of the 10-year period were produced by combining predictions of trend direction (Figure 35) with the predictions of state (Figure 28) and the median magnitudes over sites grouped by decreasing and increasing trends (Table 21). Maps of swimming grades at the beginning and end of the 10-year period are shown for network segments of order four and greater on Figure 36. The changes in swimming grade over all segments are shown on Figure 37. The predictions shown on Figure 36 and Figure 37 are consistent with the information provided by the partial plots (Figure 34) indicating, for example, that there was improvement trends in many moderate size (main stem) rivers.

The predicted swimming grades at the beginning and end of the 10-year period were used to calculate the increase in river length in the five swimming grades (Table 22). The estimated percentage of all segments with grade of fair or better (grades A-C; Table 1) at start was 35% and at end was 40% (increase of 5%, Table 22). Estimated percentage of segments order 4+ with grade of fair or better (grades A-C; Table 1) at start was 34% and at end was 42% (estimated increase of 8%, Table 22).

Table 22. Predicted proportion of the river network by length in swimming grades at the start and end of the 10-year trend period and changes over the period for all segments and segments of order four and above.

Grade	Excellent	Good	Fair	Swimmable	Intermittent	Poor	Not swimmable
All segments start	10	10	15	35	19	46	65
All segments end	14	12	14	40	23	37	60
Segments order 4+ start	11	6	17	34	32	35	67
Segments order 4+ end	13	12	17	42	37	20	57

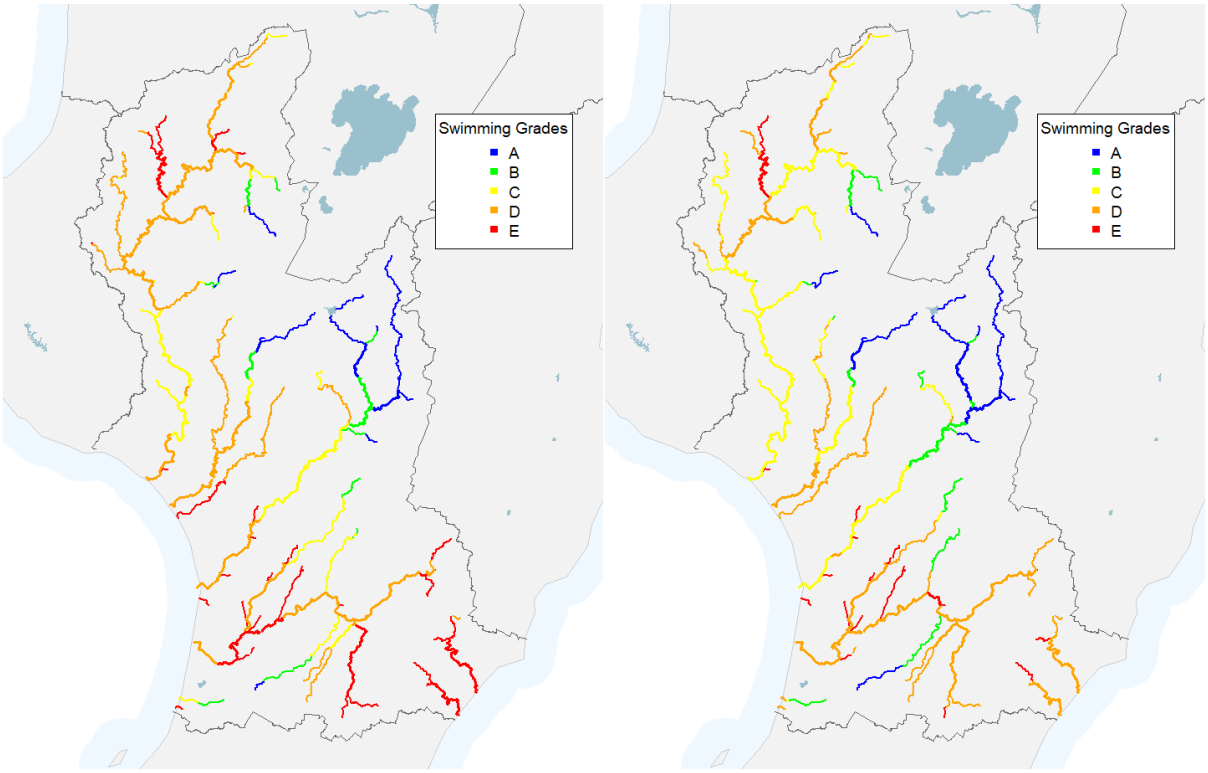

Figure 36. Estimated swimming grades at the beginning (left map) and end (right map) of the 10-year time-period based on spatial modelling for segments of Order 4+.

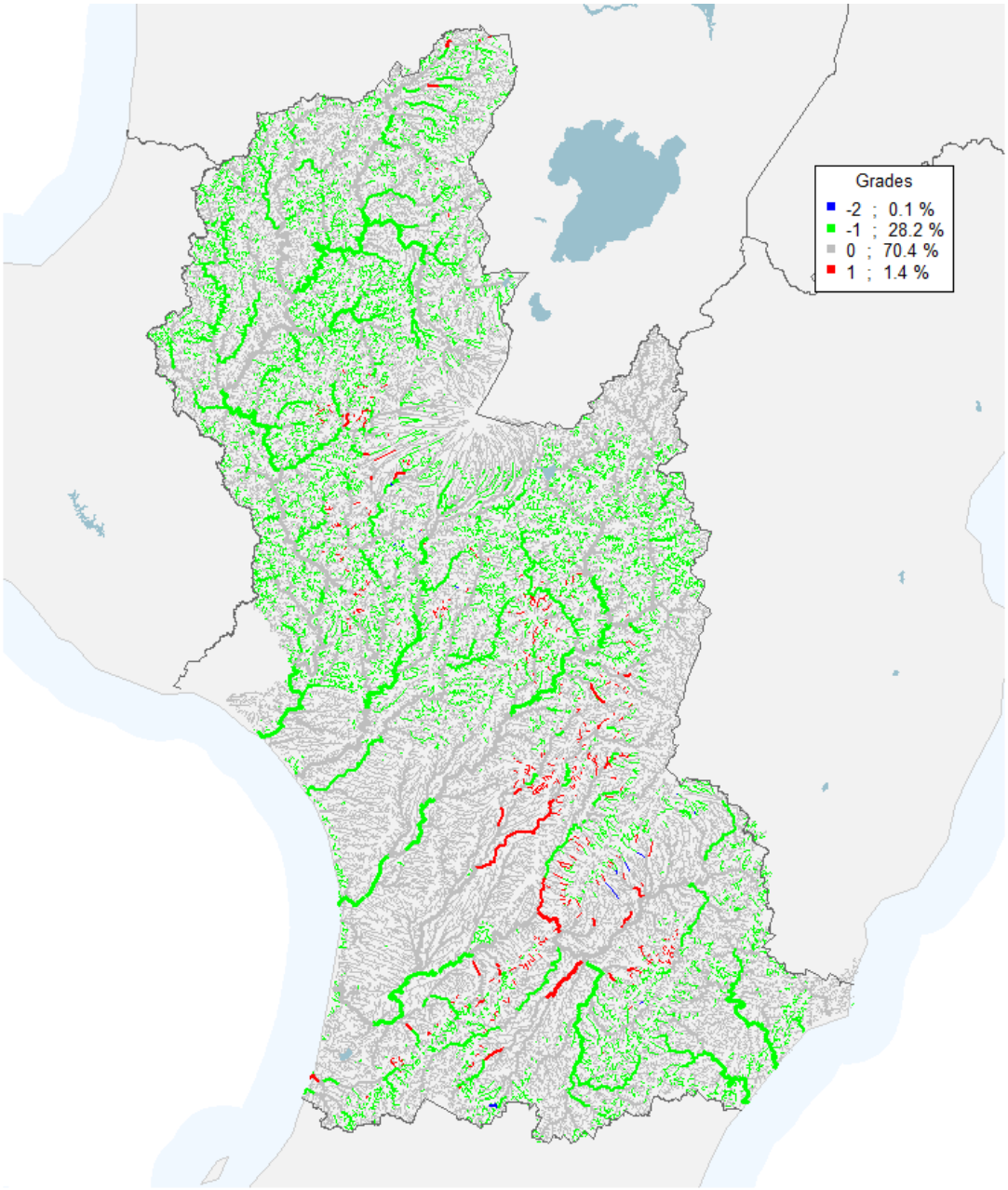

Figure 37. Predicted change in swimming grade for the 10-year time-period. A change in grade of -1 indicates an improvement of one swimming grade, e.g., a change from grade C at the start of the period to grade B at the end.

5.6.2 Changes in swimming grades seven-year time-period

Input data for step one of the assessment of changes in river swimming grades for the seven-year period are the analysis of trend direction and magnitude shown in Table 23. The mix of increasing and decreasing sites was reasonably balanced for the *E. coli* median but the G260 and G540 were dominated by decreasing trends.

Table 23. Sites with increasing and decreasing trends by statistic for the seven-year time-period. Trends at all sites were included in this analysis irrespective of confidence in trend direction.

Variable	No sites decreasing	No sites increasing	Median of decreasing trends	Median of increasing trends
Median	56	30	-3.7	5.5
G260	68	14	-10.2	4.4
G540	61	17	-11.6	7.1

The AUC statistics for the random forest models of trend direction for the seven-year *E. coli* statistics indicated poor performance (0.5 < AUC < 0.6) (Table 24). The misclassification rates for the *E. coli* statistics were higher than the misclassification rates associated with the 10-year dataset. Misclassification rates were low (< 30%) for G260 and G540 despite the poor model performance due to the low occurrence of increasing trends (Table 23).

Table 24. Misclassification rates of the RF models predicting trend directions for the E. coli statistics included in the seven-year dataset.

Variable	Misclassification rate (%)	AUC
Median	41	0.59
G260	28	0.57
G540	29	0.54

The relationship between trend direction and the model predictors are show in the partial plots in Figure 38. The model PDPs (Figure 38) and the regional predictions (Figure 39) indicate that:

1. The was a high level of consistency in the relationships between the response (the probability of trend decreasing) and the predictors for all three *E. coli* statistics. Note that the relationships are generally reversed for clarity for which a decreasing trend indicates degradation.
2. The probability of a site having a decreasing trend increased with catchment area (usArea) and reached a plateau or decreased for catchments greater than ~500 km^2. There was a similar relationship with river mean flow (data not shown).
3. The probability of a site having a decreasing trend decreases with increasing indigenous forest cover (usIndigForest). There was a similar relationship with scrub data not shown.
4. The probability of a site having a decreasing trend decreased as rainfall intensity increases (usRainDays20).
5. The probability of a site having a decreasing trend increased with increasing catchment geological phosphorus (usPhos). This indicates that there is an association with trends and geology. Catchments with soft sedimentary geology tend to have high values of usPhos.

6. The relationships with catchment slope (usAveSlope), catchment elevation (usAveElev) and catchment average temperature (usAvTCold) indicates that the probability of a site having decreasing trends was maximum at intermediate slopes, elevations and temperature.

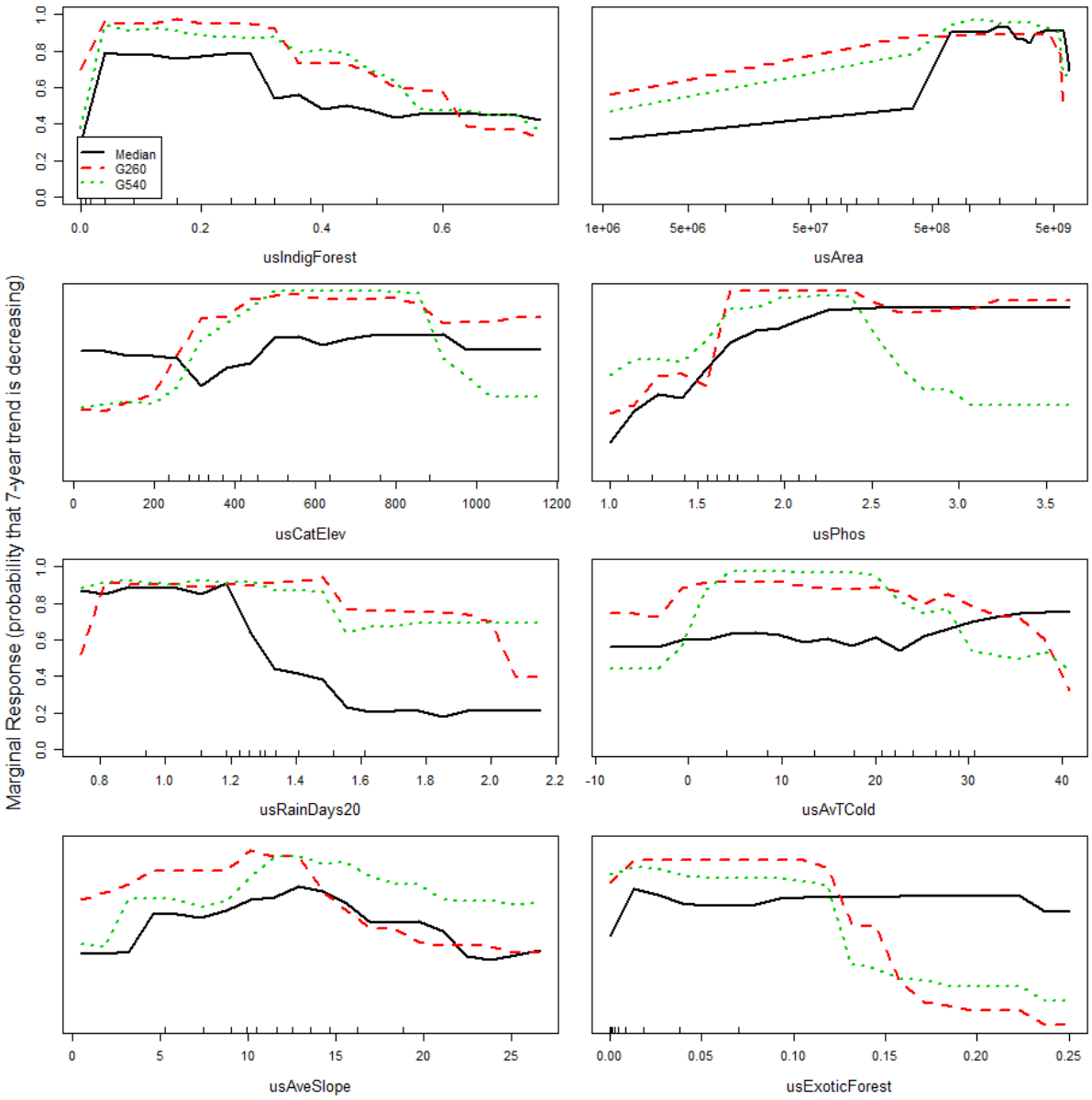

Figure 38. PDPs for the eight most important predictor variables in Random Forest models of the trend direction for the E. coli statistics included in the seven-year dataset. Each panel corresponds to one predictor. The Y-axis is the standardised value of the marginal change in probability the trend is decreasing for each of the eight modelled variables. Note that a decreasing trend in the variables shown indicates water quality improvement. In each case, the original marginal responses over all eight 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. Legend in top left panel applies to all panels. Predictor variables are defined in Table 3.

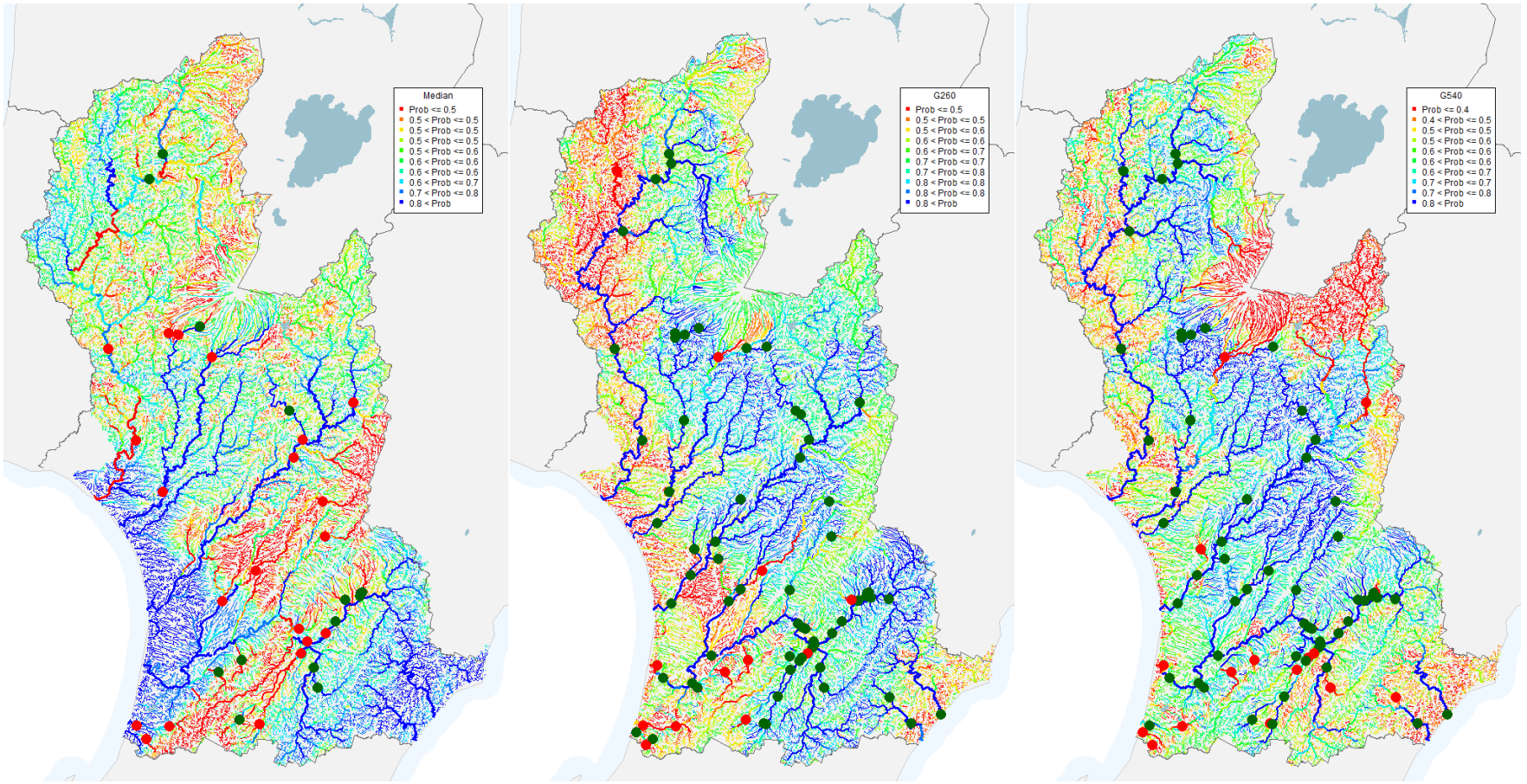

Figure 39. Spatial model predictions made using RF models of trend direction for the 85 SoE sites represented in the seven-year dataset. The plotted colours values represent the probability that trend is decreasing. SoE sites are shown as dots with the colour representing the trend directions (irrespective of confidence (red and green indicate increasing and decreasing trends respectively).

Predictions of trend direction (Figure 39) were combined with the predictions of state (Figure 31) and the median magnitudes of decreasing and increasing trends (Table 23) to produce predicted swimming grades at the beginning and end of the seven-year period. Maps of swimming grades at the beginning and end of the seven-year period for network segments of order four and greater are shown on Figure 40 and the change in swimming grade over all segments is shown in Figure 41. The predictions shown in Figure 40 and Figure 41 are consistent with the information provided by the partial plots (Figure 38) indicating, for example, that there were improving trends in many moderate size (main stem) rivers.

The predicted swimming grades at the beginning and end of the seven-year period were used to calculate the increase in river length in the five swimming grades (Table 25, Figure 40). The estimated segments of all orders with grades of fair or better (grades A - C) at the start was 35% and at the end was 40% (increase of 5%, Table 22). Estimated segments of order 4+ with grades of fair or better (grades A - C) at the start was 36% and at the end was 43% (estimated increase of 7%, Table 22).

Table 25. Predicted proportion of the river network by length in swimming grades at the start and end of the seven-year trend period for all segments and segments of order four and above.

Grade	Excellent	Good	Fair	Swimmable	Intermittent	Poor	Not swimmable
All segments start	3	17	15	35	24	41	65
All segments end	12	20	8	40	27	33	60
Segments order 4+ start	8	15	13	36	36	28	64
Segments order 4+ end	19	23	1	43	39	17	56

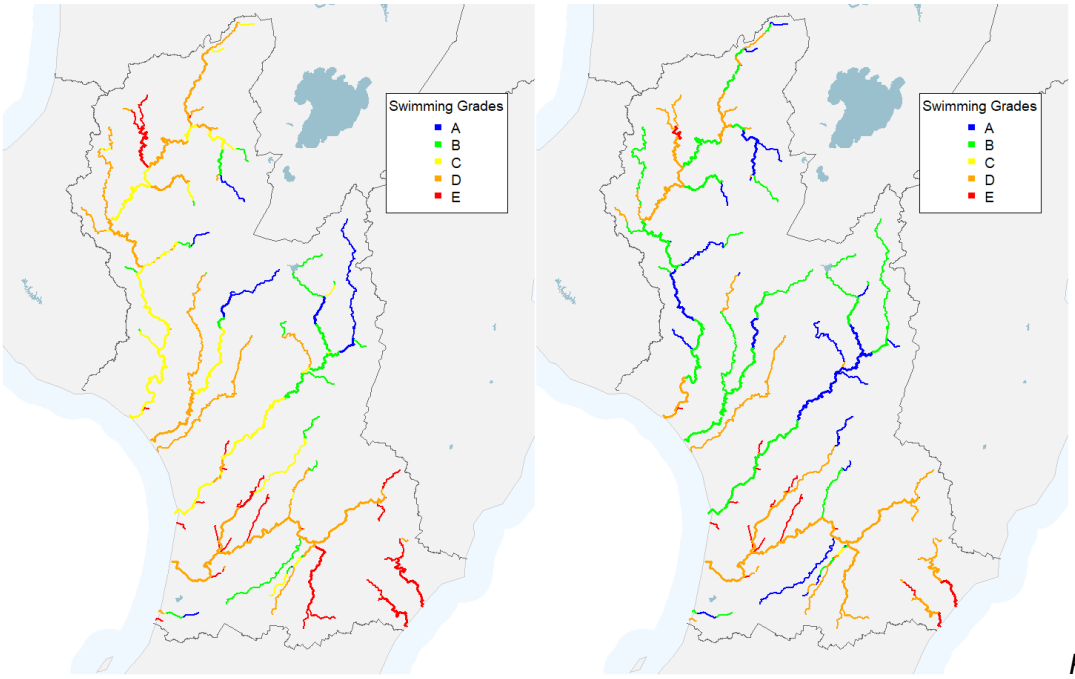

Figure 40. Estimated swimming grades at the beginning (left map) and end (right map) of the seven-year time-period based on spatial modelling for segments of Order 4+.

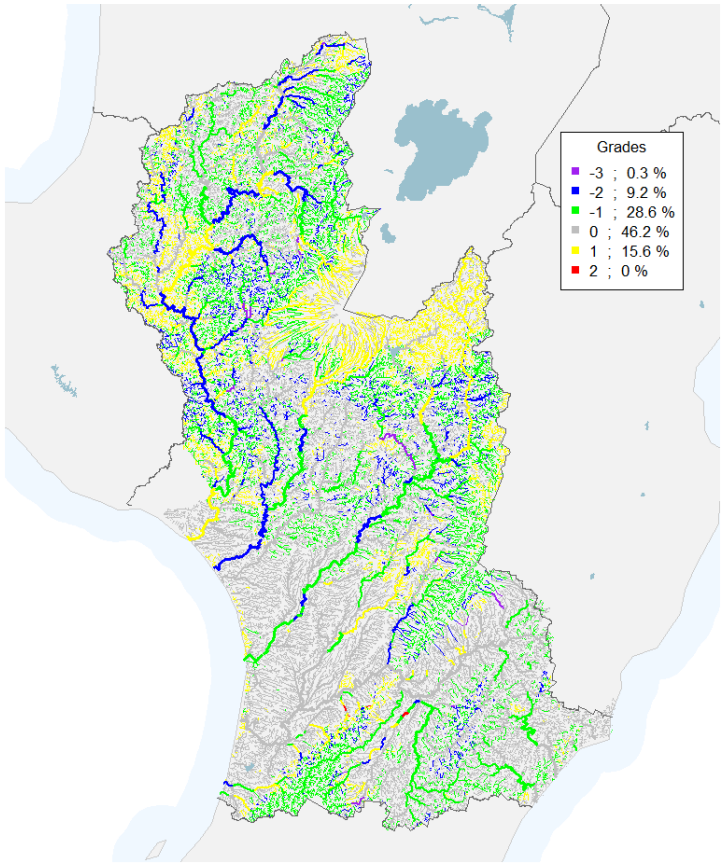

Figure 41. Predicted change in swimming grade for the seven-year time-period. A change in grade of -1 indicates an improvement of one swimming grade, e.g., a change from grade C at the start of the period to grade B at the end.

5.6.3 .Changes in clarity, SSC and turbidity for the seven-year time-period

Input data for step one of the assessment of changes in river swimming grades for the seven-year period are shown in Table 26. The mix of increasing and decreasing sites was reasonably balanced for clarity but SSC and turbidity were dominated by decreasing trends.

Table 26. Sites with increasing and decreasing trends by variable for the seven-year time-period. Trends at all sites were included in this analysis irrespective of confidence in trend direction.

Variable	No sites decreasing	No sites increasing	Median of decreasing trends	Median of increasing trends
Turbidity	55	6	-6.5	2.3
SSC	71	4	-6.2	1.6
Clarity	15	22	-2.5	4.5

The AUC statistics for the random forest models of trend direction for the seven-year dataset indicated good performance (0.6 < AUC < 0.7) for clarity and turbidity and satisfactory performance (AUC < 0.6) for SSC (Table 27). Misclassification rates were low (< 30%) for SSC and turbidity despite the only satisfactory model performance due to the low occurrence of increasing trends (Table 27).

Table 27. Misclassification rates of the RF models predicting trend directions for the six variables included in the seven-year dataset.

Variable	Misclassification rate (%)	AUC
Clarity	34	0.68
SSC	11	0.58
Turbidity	11	0.62

The relationships between the direction of trends in clarity, SSC and turbidity and the model predictors are demonstrated by the PDPs (Figure 42). The partial plots (Figure 42) and the regional predictions (Figure 43) indicate that:

1. The relationships between the response (the probability of trend decreasing) and the predictors were inconsistent across the three variables, indicating differences in the drivers of trends for the different water quality variables. Note that the relationships are generally reversed for clarity for which a decreasing trend indicates degradation.
2. The probability of a site having a decreasing trend increased for SSC and turbidity and decreased for clarity, with catchment area (usArea) and reached a plateau or decreased for catchments greater than ~500 km^2. There was a similar relationship with river mean flow (data not shown).
3. The probability of a site having a decreasing trend for clarity decreased with increasing indigenous forest cover (usIndigForest). Trends in SSC did not have a strong relationship with usIndigForest and trends in turbidity decreased at high values. There were similar relationships with scrub (data not shown).

4. The probability of a site having a decreasing trend in SSC and clarity increased and then reached a plateau as rainfall intensity increased (usRainDays20). Trends in turbidity did not have a strong relationship with usRainDays20.

5. The probability of a site having a decreasing trend for SSC and turbidity increased and reached a plateau (SSC) or decreased (turbidity) with increasing catchment geological phosphorus (usPhos). This indicates that there is an association with trends and geology. Catchments with soft sedimentary geology tend to have high values of usPhos. The complex relationship between the variables and usPhos shown in Figure 42 probably reflects interactions between geology and catchment elevation. Trends in clarity did not have a strong relationship with usPhos.

6. The probability of a site having improving trends for all three variables was maximum at intermediate catchment slopes (usAveSlope), catchment elevations (usCatElev) and catchment temperature (usAvTCold). However, trends in clarity were less responsive to variation in all three predictors than trends in SSC and turbidity.

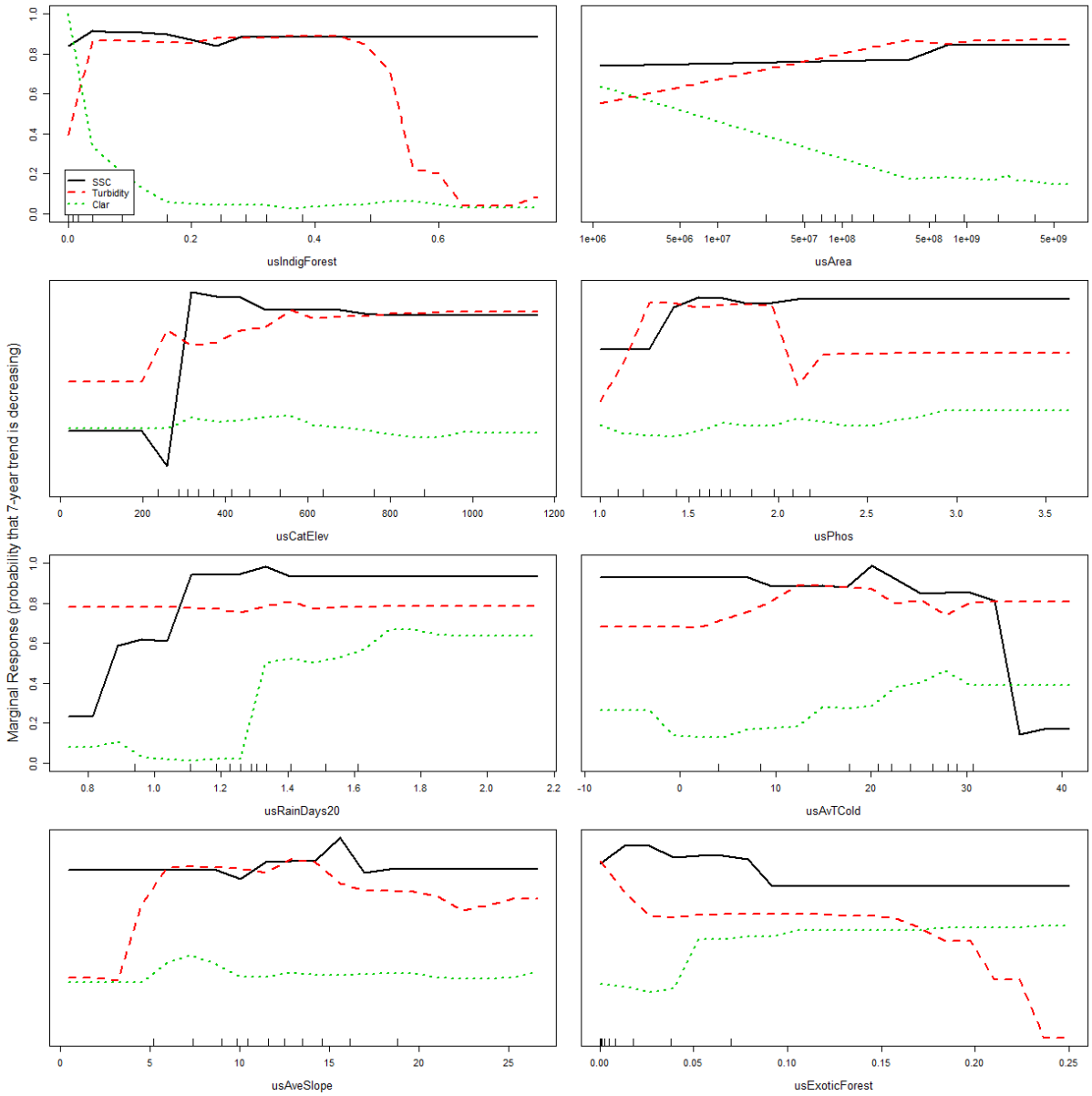

Figure 42. PDPs for the eight most important predictor variables in RF models of the trend direction for clarity, SSC and turbidity included in the seven-year dataset. Each panel corresponds to one predictor. The Y-axis is the standardised value of the marginal change in probability the trend is decreasing for each of the eight modelled variables. Note that a decreasing trend for SSC and turbidity indicates water quality improvement but for clarity indicates degradation. In each case, the original marginal responses over all eight 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. Legend in top left panel applies to all panels. Predictor variables are defined in Table 3.

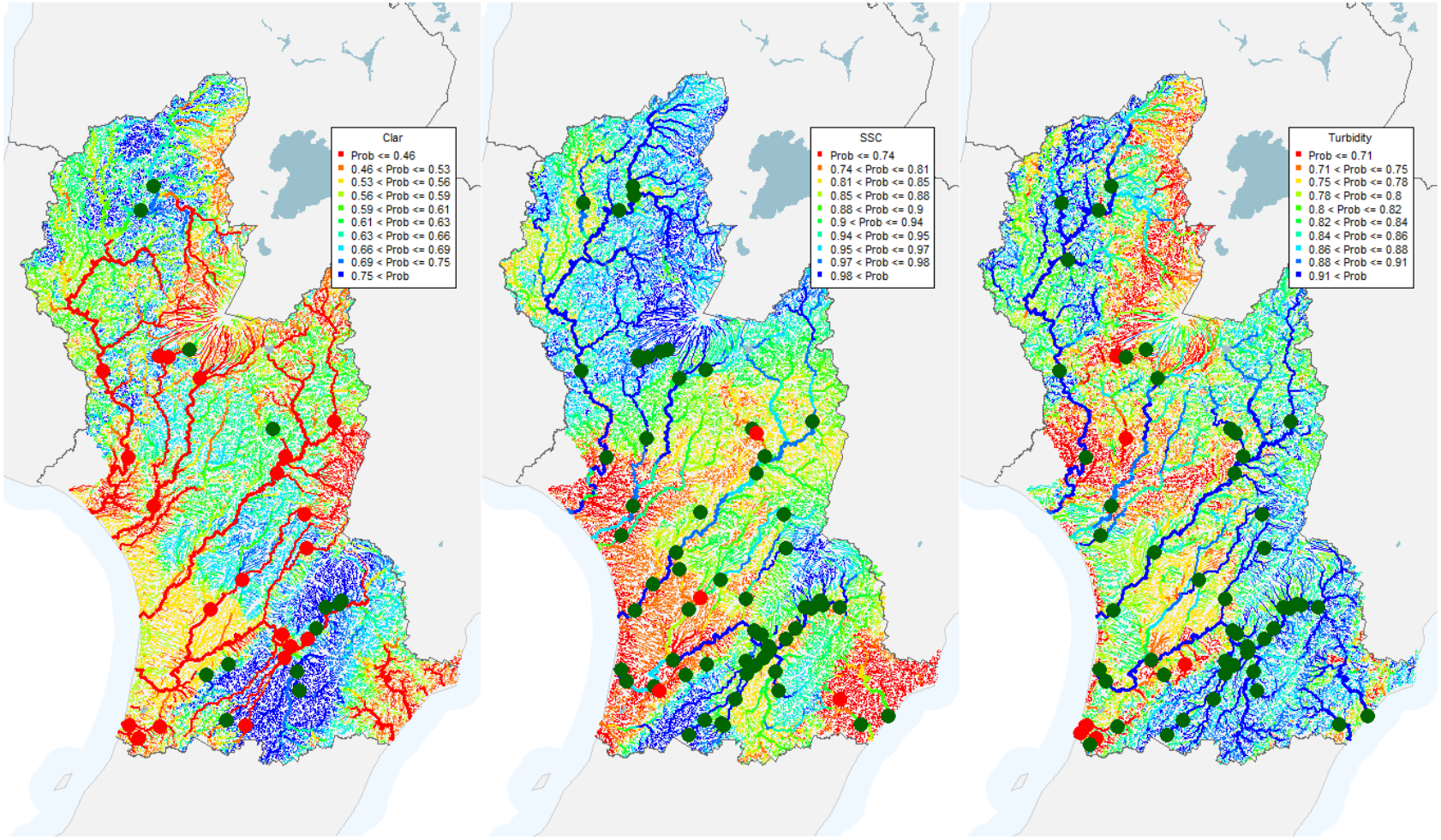

Figure 43. Spatial model predictions made using RF models of trend direction based on the SoE sites represented by visual clarity, SSC and turbidity in the seven-year dataset. The plotted colours represent the probability that the trend is decreasing. Note that for visual clarity a decreasing trend indicates degradation. SoE sites are shown as dots with the colour representing the trend directions, irrespective of confidence (red and green indicate increasing and decreasing trends respectively).

Predictions of trend direction (Figure 43) were combined with the predictions of state (Figure 33) and the median magnitudes of the grouped decreasing and increasing trends (Table 26) to produce predicted changes over the seven-year period. The predictions shown on Figure 44 are consistent with the information provided by the partial plots (Figure 42) indicating, for example, that there was improving trends in many main stem rivers.

The changes in state are quantified in Table 28 and shown in Figure 44. Taken over all segments, clarity decreased, for example, 25% of segments had predicted clarity of <1.7m or more at the start of the seven-year period, which decreased to <1.5m at the end. However, for larger rivers (order 4 or greater) clarity increased over the period. For example, 25% of segments of order 4 or greater were predicted to have a clarity of <1.2m at the start of the seven-year period, which increased to 1.3m at the end. Taken over all segments, and segments larger rivers (order 4 or greater), SSC and turbidity improved. For example, 25% of segments had predicted SSC of <2.6 at the start of the period, which decreased to <1.7 at the end.

Table 28. Predicted state for clarity (m), SSC (g m^{-3}) and turbidity (NTU) for the start and end of the seven-year trend period. The values are the estimated medians that are exceeded by 75%, 50% and 25% of network segments (i.e., 1st, 2nd and 3rd quantiles).

	1st quantile		**2nd quantile**		**3rd quantile**	
	Start	**End**	**Start**	**End**	**Start**	**End**
Clarity All segments	1.7	1.5	2.1	1.9	2.6	2.3
Clarity Segments order 4+	1.2	1.3	1.7	1.6	2.1	2.2
SSC All segments	2.6	1.7	3.6	2.3	5.6	3.6
SSC Segments order 4+	3.1	2.0	5.5	3.5	10.1	6.5
Turbidity All segments	1.9	1.2	3.1	2.0	4.5	2.8
Turbidity Segments order 4+	2.4	1.5	4.3	2.7	6.8	4.3

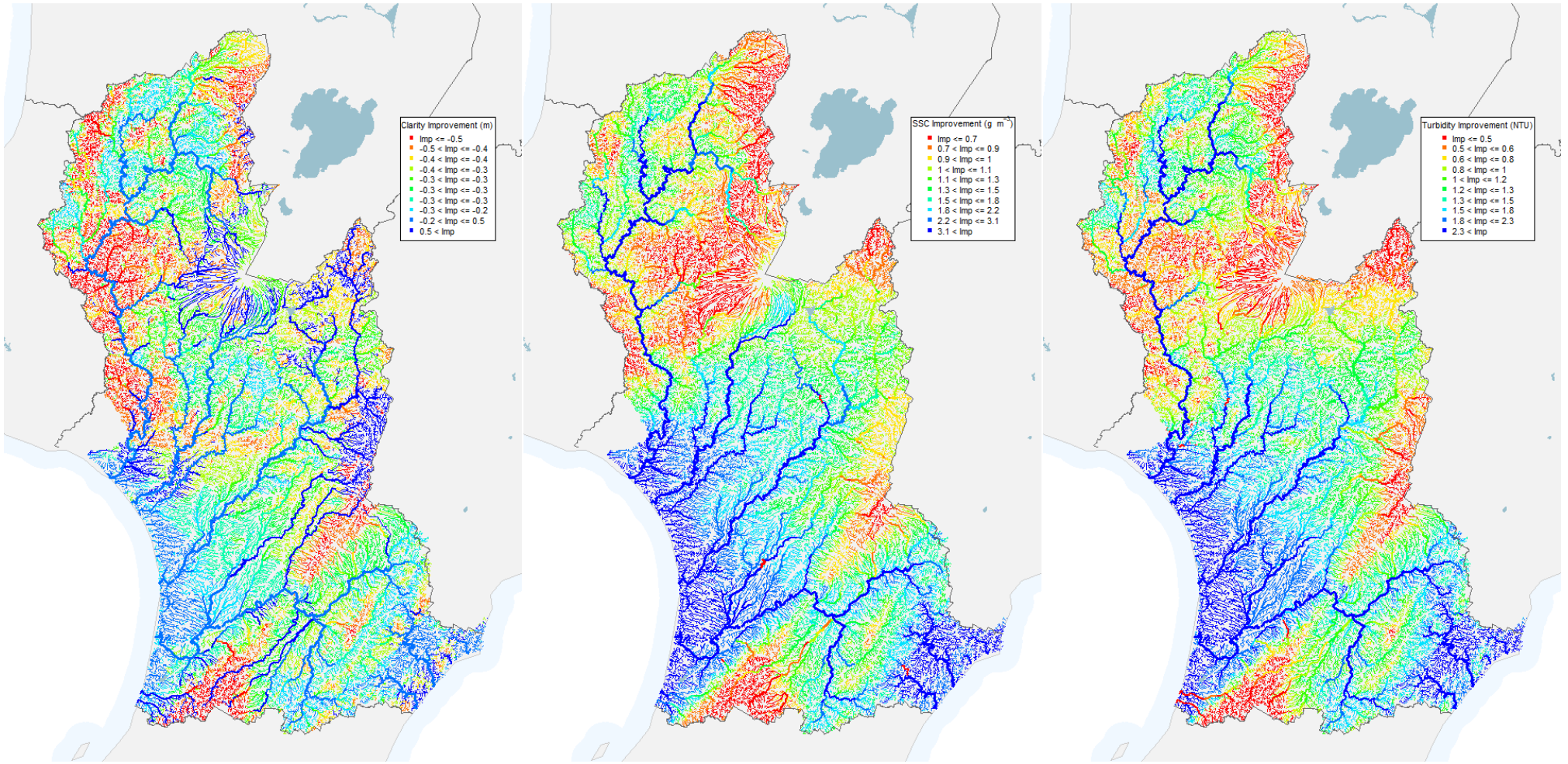

Figure 44. Predicted change in state for clarity, SSC and turbidity for the seven-year time-period. The values are changes in the values in the original units (Table 2) through the time-period. A negative improvement indicates degradation.

5.7 Association between trends and interventions

5.7.1 10-year *E. coli* trends

The directions of the ten-year trends in *E. coli* variables were weakly associated with the predictor variables that represented the interventions and the proportion of the catchment occupied by erosion in 2004 (Figure 45). Decreasing trends for all variables were generally associated with high values of the predictors SLUI and Erosion. Increasing trends for all variables were associated with high values of the predictors Planting and Fencing. The predictors were generally weakly correlated with each other (absolute value of Pearson's correlations coefficient < 0.35) except for erosion and SLUI, which had a Pearson's correlations coefficient of 0.65.

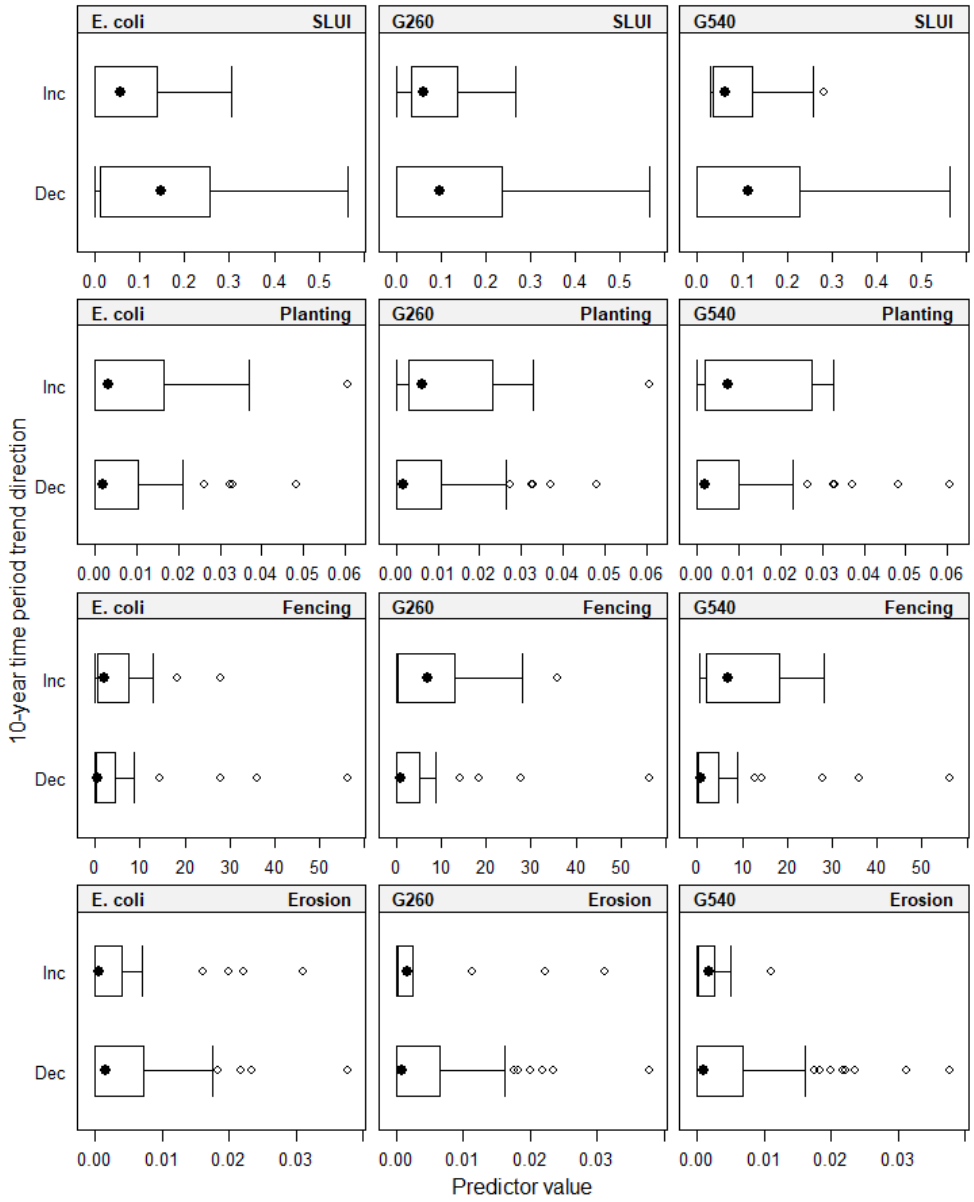

Figure 45. Distribution of the predictor variables representing the interventions and the proportion of the catchment occupied by erosion in 2004 grouped by 10-year time-period trend direction. See *Table 6 for explanation of the predictor variables.*

The misclassification rates and AUC statistics for the reduced RF classification models indicated a statistically significant association between the trend direction and some of the predictors but with poor to satisfactory model performance (Table 29). This is consistent with the weak relationships between trend direction and the individual predictors shown in Figure 45. Lower misclassification rates for G260 and G540 than for *E. coli* partly reflect the low occurrence of increasing trends (Table 29). The proportion of catchments occupied by SLUI farms was included in the *E. coli* model (Table 29). Planting was included in the G260 and G540 models and fencing was included in all models (Table 29).

Table 29. Misclassification rates and AUC statistics for the reduced RF classification models predicting direction of 10-year trends. The models expressed trend direction as a function of the predictor variables representing the interventions and the proportion of the catchment occupied by erosion in 2004.

Variable	Number of sites	Misclassification rate (%)	AUC	Explanatory variables retained
E. coli	69	51	0.54	Fencing + SLUI
G260	62	26	0.58	Planting + Fencing
G540	60	22	0.61	Planting + Fencing

There was a weak association between the magnitude of 10-year trends at SoE sites and the predictor variables (Figure 46). Trend magnitudes decreased with increasing values of SLUI, Erosion and the interaction of Erosion and SLUI. Trend magnitudes increased with increasing values of Fencing and Planting.

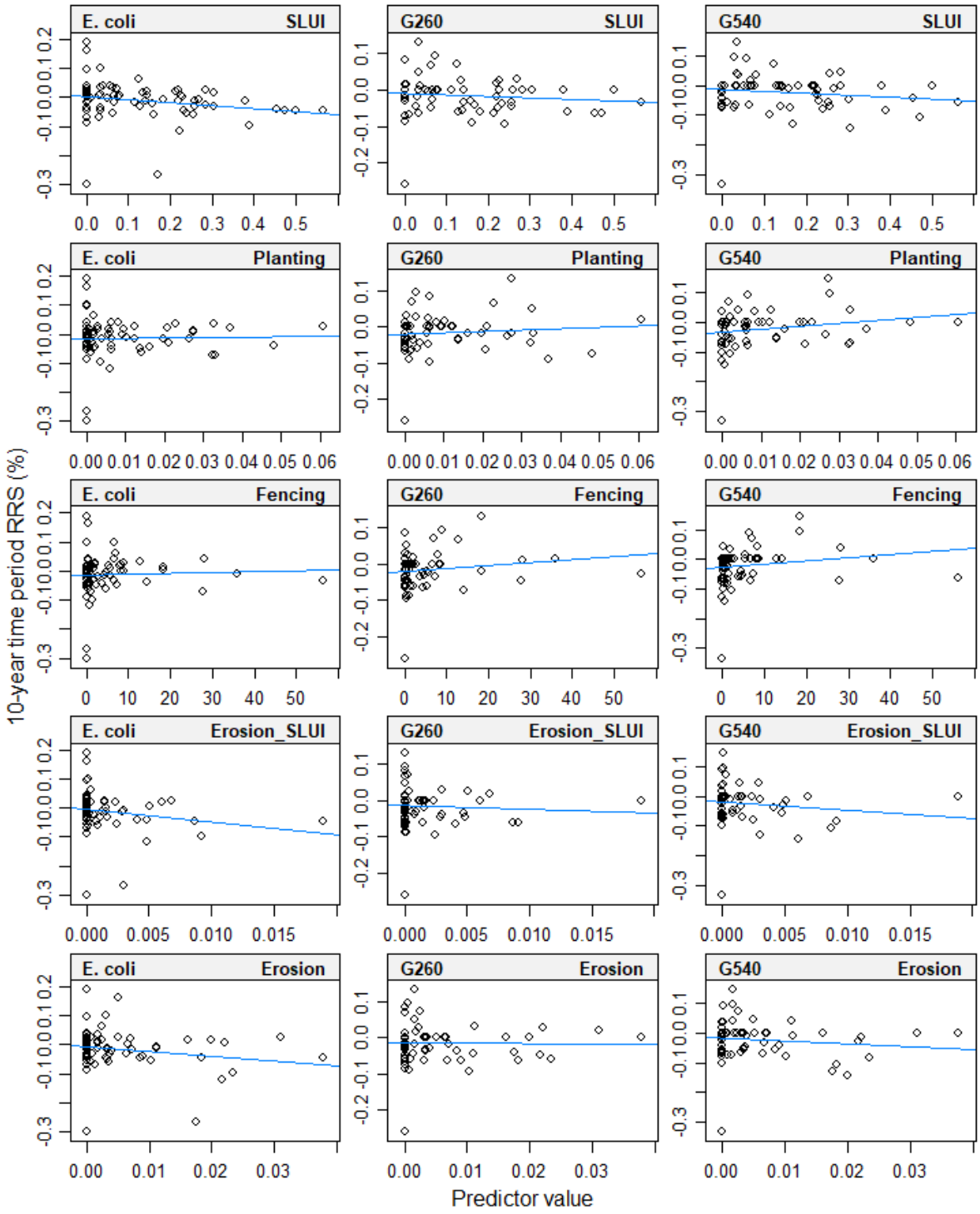

Figure 46. Relationship between 10-year trend magnitudes and predictor variables. The predictor variables represent the interventions and the proportion of the catchment occupied by erosion in 2004. The blue line represents a linear regression. See *Table 6 for explanation of the predictor variables.*

The stepwise linear regression models indicated that the 10-year trend magnitudes for the three *E. coli* statistics were weakly but significantly predicted by combinations of the

predictors (Table 32). Only Erosion and Fencing were included in the *E. coli* and G260 models, respectively, but the G540 model included four predictors.

Table 30. Details of stepwise linear regression models fitted to the magnitudes of trend for each of water quality variables included in the 10-year time-period. Because the models were built by a stepwise process, the retained explanatory variables can be interpreted as significant.

Variable	Number of sites	r^2 (%)	Explanatory variables retained
E. coli	69	4	Erosion
G260	62	6	Fencing
G540	60	17	Erosion + SLUI + Planting + Fencing

5.7.2 Seven-year trends

The directions of the seven-year trends were weakly associated with the predictor variables (Figure 47). Decreasing trends for *E. coli*, G260, G540 and turbidity, were generally associated with high values of SLUI and Erosion. Increasing trends for clarity were associated with high values of SLUI and Erosion. Planting had weak associations with trend direction for all variables but decreasing trends for G260, SSC and turbidity were associated with high values of fencing (Figure 47).

The predictors were generally weakly correlated with each other (absolute value of Pearson's correlations coefficient < 0.4) except for Erosion and SLUI, which had a Pearson's correlation coefficient of 0.62.

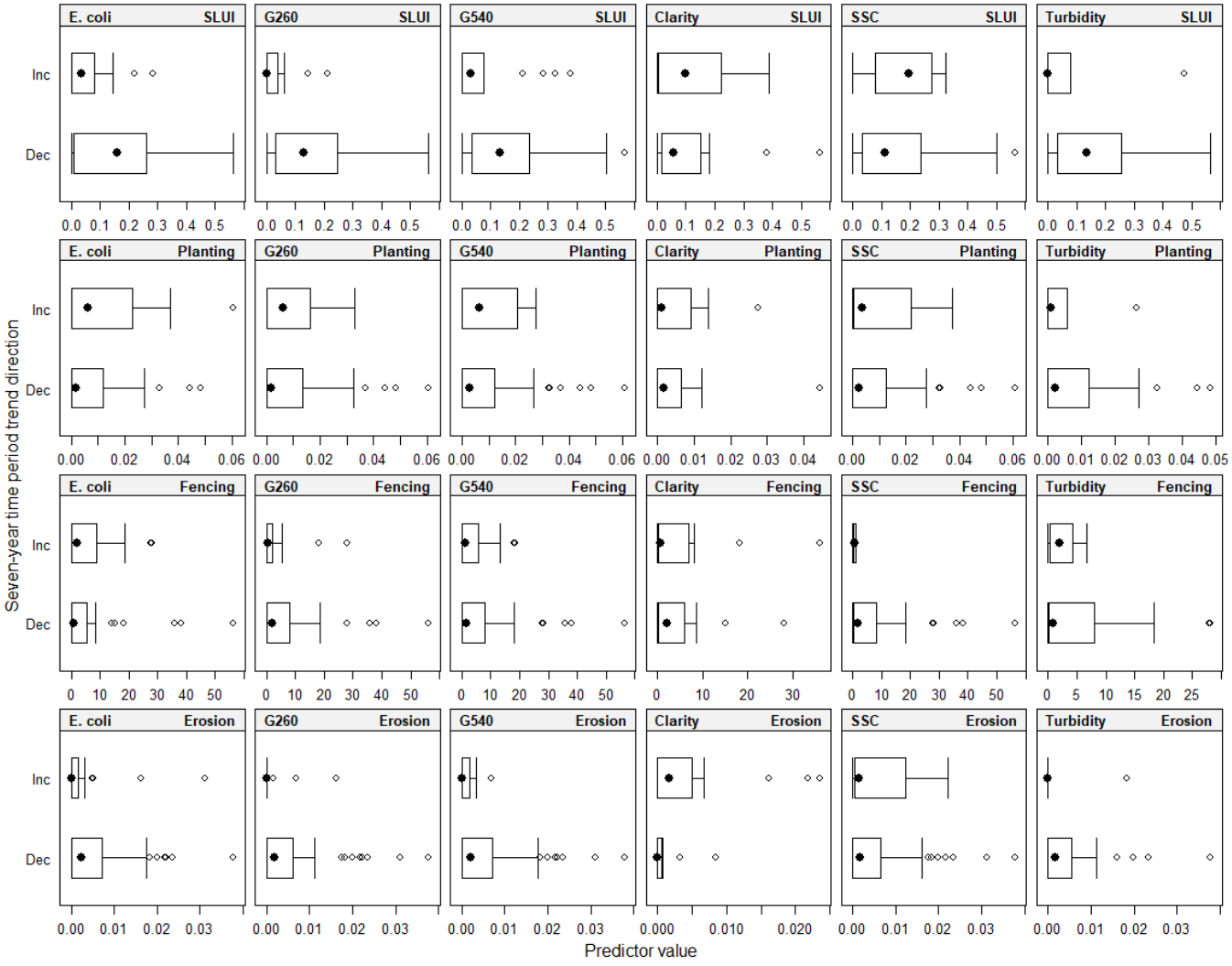

Figure 47. Distribution of the predictor variables representing the interventions and the proportion of the catchment occupied by erosion in 2004 grouped by seven-year time-period trend direction. See *Table 6 for explanation of the predictor variables.*

The misclassification rates and AUC statistics for the reduced RF classification models indicated statistically significant associations between the trend direction and predictors for all models. Model performance varied by variable with the G260 model achieving good performance, *E. coli*, G540 and SSC achieving satisfactory performance and clarity and turbidity being poor (Table 31). Low misclassification rates for SSC, turbidity, G260 and G540 partly reflect the low occurrence of increasing trends (Table 31). The proportion of catchments occupied by SLUI farms was included in the *E. coli*, G260, clarity and SSC models (Table 31). Planting was included in the G540 model and turbidity models, fencing was included in the G260 and turbidity model and erosion was included in *E. coli*, G540 clarity and SSC models (Table 31).

Table 31. Misclassification rates and AUC statistics for the reduced RF classification models predicting direction of seven-year trends. The models expressed trend direction as a function of the predictor variables representing the interventions and the proportion of the catchment occupied by erosion in 2004.

Variable	Number of sites	Misclassification rate (%)	AUC	Explanatory variables retained
E. coli	86	28	0.65	Erosion + SLUI
G260	85	26	0.70	Fencing + SLUI
G540	81	27	0.61	Planting + Erosion
Clarity	37	46	0.57	Erosion + SLUI
SSC	75	8	0.68	Erosion + SLUI
Turbidity	61	13	0.53	Fencing + Planting

There was a weak association between the magnitude of seven-year trends at SoE sites and the predictor variables (Figure 48). Trend magnitudes decreased with increasing values of SLUI, Erosion and the interaction of Erosion and SLUI. Trend magnitudes increased with increasing values of Fencing and Planting.

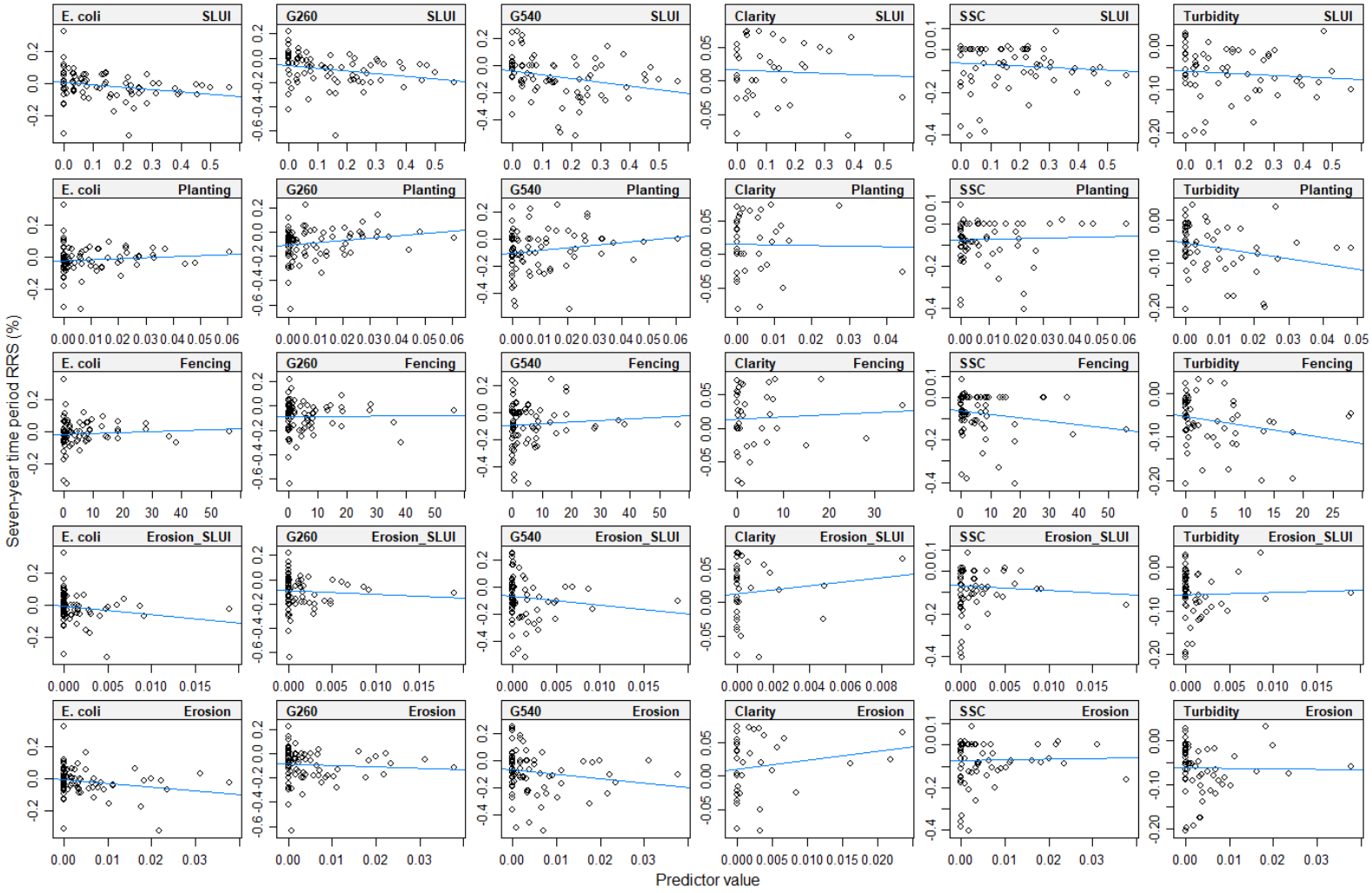

Figure 48. Relationship between seven-year trend magnitudes and predictor variables. The predictor variables represent the interventions and the proportion of the catchment occupied by erosion in 2004. The blue line represents a linear regression. See *Table 6 for explanation of the predictor variables.*

The stepwise linear regression models indicated that seven-year trend magnitudes for all water quality variables, except clarity, were weakly but significantly predicted by combinations of the predictors representing interventions and the proportion of catchment subject to erosion in 2004 (Table 32). The proportion of catchments occupied by SLUI farms was included in all models except clarity (Table 32). Planting was included in the G260 model and Fencing and the interaction of Erosion and SLUI were included the SSC and turbidity models.

Table 32. Details of stepwise linear regression models fitted to the magnitudes of trend for each of water quality variables included in the seven-year time-period. Because the models were built by a stepwise process, the retained explanatory variables can be interpreted as significant. The term Erosion x SLUI indicates the interaction of the erosion and SLUI farm variables.

Variable	Number of sites	r^2 (%)	Explanatory variables retained
E. coli	86	7	SLUI
G260	85	11	SLUI + Planting
G540	81	12	SLUI
Clar	37	0	
SSC	75	10	Erosion + SLUI + Fencing + Erosion x SLUI
Turbidity	61	14	Erosion + SLUI + Fencing + Erosion x SLUI

5.7.3 Relationship between trends at discharge and impact sites

There were 19 pairs of discharge-impact sites for which 7-year trends had been evaluated (see Section 5.4). Concordant trends at the discharge-impact sites were in both the decreasing and increasing direction (Table 33). Of the 19 pairs of sites, 10 had concordant trends for clarity, which was not statistically significant (Table 33). There were 11 and 13 concordant pairs of trends for *E. coli* and SSS respectively, which was highly significant. The paired *E. coli* and SSS trends were predominantly decreasing.

Table 33. Concordance between paired discharge-impact site trends. The p-value indicates the significance of the number of concordant pairs (binomial test, H_0 = paired sites have 50% probability of being concordant).

Variable	Number of paired sites	Number of concordant increases	Number of concordant decreases	Number concordant	*p*-value	Overall direction
Clarity	19	8	2	10	1	Not Significant
E. coli	19	6	11	17	0.001	Decreasing
SSC	19	4	13	17	0.001	Decreasing

5.7.4 Trends in climate and flows

There was only one climate station for which the 10-year annual rainfall trend direction was determined with confidence and this was increasing (Table 34). For the seven-year period two climate stations had certain decreasing trends and one had a certain increasing trend (Table

34). For both the 10-year and seven-year time periods, the only trends in flow that were determined with confidence were decreasing trends (Table 34).

For flows there was a dominance of decreasing trends when certainty in trend direction was disregarded (negative overall median value; Table 34) and sites for which it was at least as likely as not that the true trend was decreasing (Figure 49). Binomial tests indicated a significant (i.e., $p < 0.05$) regional decreasing trend for flows for the seven-year period and a close to significant regional decreasing trend the 10-year period (Table 34). Median RSS values for flow trends for the seven and 10-year periods were -3% and -0.5% respectively (Table 34). By contrast, for rainfall there was a dominance of increasing trends when certainty in trend direction was disregarded (positive overall median value; Table 34) and sites for which it was at least as likely as not that the true trend was increasing (Figure 49). The overall median RSS values for annual rainfall for the seven and 10-year periods were 0.5% and 0.6% respectively (Table 34). Neither time-period was associated with a significant regional trend in annual rainfall.

Table 34. Results of trend analyses on annual rainfall and mean annual flows at climate and flow recording stations.The five left-most columns report the number individual stations and the numbers of trends determined with confidence by direction. The three right-most columns report the overall trends including the results of the binomial tests.

Station type	Time period	Number of stations	Number of certain decreasing	Number of certain increasing	Total decreasing	Overall median RSS	Binomial test *p*-value
Rain	10	13	0	1	3	0.60	0.092
Rain	7	13	2	1	5	0.49	0.581
Flow	10	14	1	0	11	-0.49	0.057
Flow	7	14	4	0	12	-3.09	0.013

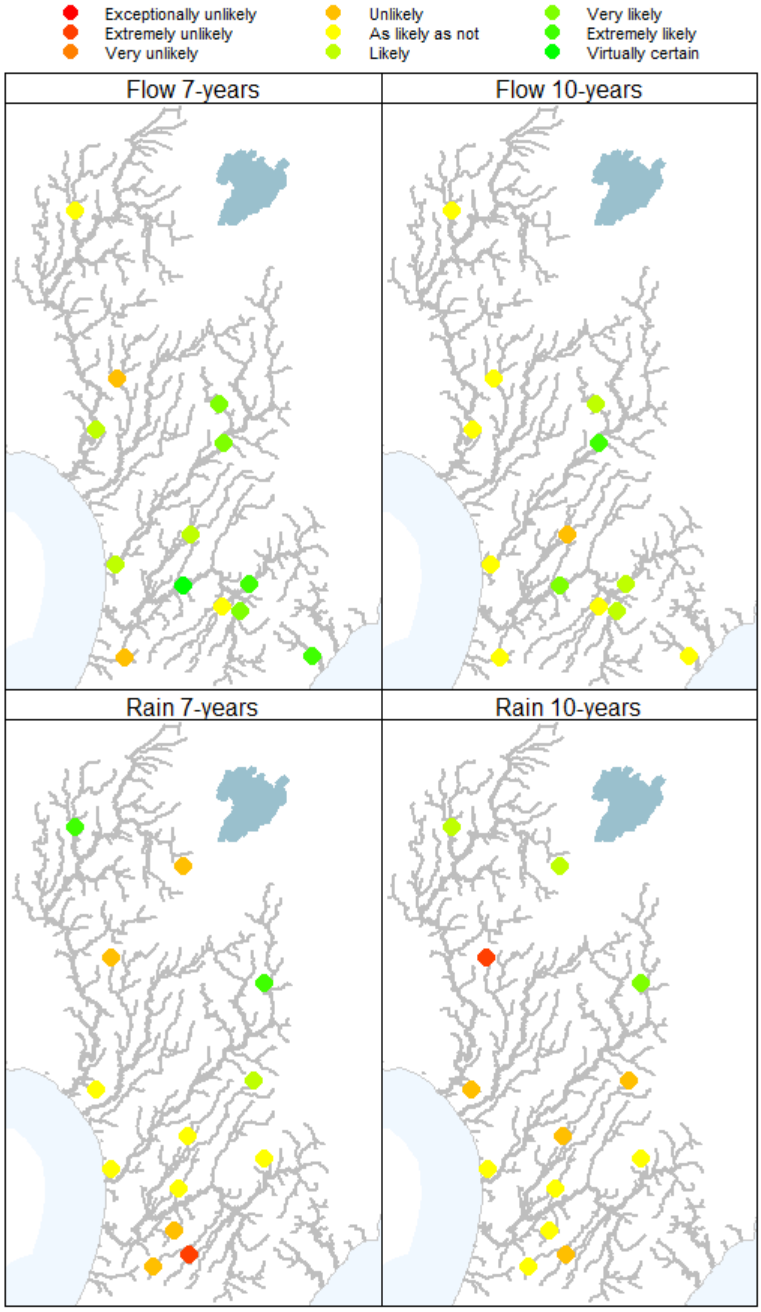

Figure 49. Map of climate and rainfall stations classified by the level of confidence that trends in annual rainfall and mean annual flow were decreasing for the seven and 10-year time periods. See Table 3 for details of the confidence categories.

6 Discussion

This study had objectives that are specific to changes in water quality in the Manawatū-Whanganui Region, but with conclusions and recommendations that are of national interest. The study developed methods for spatially interpolating water quality trends from monitoring sites to a whole region and to quantify water quality improvement. In addition, the study demonstrates methods for evaluating associations between water quality interventions. These methods provide evidence and a quantification of the efficacy of catchment to regional scale management interventions aimed at improving water quality. The regional scale of this study is in contrast to most studies of the efficacy of interventions to improve water quality, which tend to be at the scale of individual mitigation measures, to small sub-catchments (Wilcock *et al.*, 2013).

During this study, some additional detailed research was conducted on questions concerning flow adjustment of water quality data as part of trend analysis and spatial modelling of the *E. coli* statistics G260 and G540. The findings of these additional research items will be of interest in future water quality trend analysis and spatial modelling of river swimming grades.

6.1 Assessment of swimming grades in the region

The regional swimming maps (Figure 28, Figure 31) and quantification of the length of swimmable rivers that were both made using HRC's recent water quality data were broadly consistent with the national swimming maps (Table 15). The data used to produce the national swimming maps differed from that used to produce the regional maps in this study, both in terms of the number of sites representing the region and the treatment of that data. The national swimming maps were generated from a dataset describing *E. coli* measurements at 753 sites throughout New Zealand that comprised at least 30 samples over a period that extended from 1990 at some sites to the end of 2013. There were 82 sites representing the Manawatū-Whanganui Region in the national dataset. By contrast, this study was regional in extent and predictions were generated using two datasets of 69 and 87 sites that pertained to the 10-year and seven-year period ending 2016.

The proportion of swimmable river segments across the region assessed from the seven-year, 10-year and national datasets ranged from 38% to 46% for segments of order four and greater and from 36% to 38% for all segments (Table 15, Table 19). For individual swimming grades, the ranges of estimates of the proportion of river segments were larger (e.g., segments categorised good ranged from 8% to 31% for assessments based on the seven-year, 10-year and national datasets; Table 15, Table 19). Part of the variation in the assessments of swimming grades is associated with instability of the swimming grade assessments through time at individual sites. For example, only 71% of the 69 sites that were in common to the 10-year and seven-year time period had the same grade for both time periods (Table 8). Instability in the grade occurs due to the imprecision of the calculated *E. coli* statistics and is an unavoidable consequence of estimating a population statistic (e.g. median, G260) from a limited number of samples (see McBride (2016) for more details). It is noted that grade instability would likely have been greater if the 95th percentile statistic had been included in grade determinations, due to its high imprecision (Stats NZ, 2017).

Variation between the different swimming maps also arises because the underlying models are sensitive to the mix of sites used in the modelling fitting process. When there are few sites with low values of the G260 and G540 statistics, the model is likely to under predict the occurrence of segments with excellent and good swimming grades. This arises because the 'A' grade is defined by exceedance values less than 0.2 and 0.05 for G260 and G540

respectively. Regional datasets are likely to have few sites with values of G260 and G540 that are at or below this threshold because most monitoring networks under-represent sites with good water quality (e.g., reference sites, (Larned and Unwin, 2012)). It is therefore recommended that modelling that aims to produce regional swimming maps should consider using data from sites in adjacent regions and possibly national models.

In addition to differences in the extent and size of datasets used, some of the modelling details employed by this study differed compared to the modelling underlying the national swimming maps. This study found that logit transformation of the response variable is desirable for modelling the G260 and G540 statistics. The logit transformation has the effect of spreading low values of G260 and G540, which improves the model's fit to this part of the response gradient. It is therefore recommended that G260 and G540 are logit transformed in future modelling of river swimming grades (for more details see Appendix A).

Previous studies have noted that site-scale uncertainty associated with national scale spatial models of *E. coli* statistics is high (Larned et al. 2016; Snelder *et al.*, 2016a). This study found that national models of *E. coli* statistics had better performance than the regional models (Table 13, Table 14). In addition, this study showed that model predictions are sensitive to the input data. Therefore, national and regional swimming grade maps produced using statistical modelling should be regarded as indicative. The greatest confidence should be put in grades derived for individual SoE sites based on monitoring data, but it should be kept in mind that these grades may be sensitive to the assessment time-period.

Swimming grades for the 10-year time-period assessed from year-round data were compared with grades assessed from the same time-period but restricted to data pertaining to the bathing season. These comparisons indicated that a larger proportion of sites were swimmable (grade fair or better) during the bathing season than year-round (59% versus 55%). However, small headwater rivers (order 1, 2 & 3) tended to have lower swimming grades (i.e., less suitable) in the bathing season compared to their year-round grades. By contrast, large rivers (order 4 and greater) tended to have better swimming grades in the bathing season compared to their year-round grades. The reasons for these differences were not investigated. Because of the generally poorer swimming grades for small rivers during the bathing season, swimming maps that represent the bathing season indicate that only 17% of all rivers (by length) are swimmable. The same maps however, indicate that 36% of large rivers are swimmable during the bathing season, which is consistent with the year-round regional swimming map.

6.2 Water quality trends

Most trends at SoE sites were uncertain for the 10 and seven-year time periods. However, this is based on a misclassification (of trend direction) error risk of 5%. When the traditional confidence level of 95% was relaxed there was a clear pattern of improving trends (Figure 17, Figure 20). For example, the proportion of sites with 10-year trends that were at least as likely as not to be improving were 65%, 81% and 80% for *E. coli*, G260, and G540 respectively (Figure 17). For the seven-year time-period, the proportion of sites with trends that were at least as likely as not to be improving were 72%, 91%, 81%, 78%, 99% and 95% for *E. coli*, G260, G540, clarity, SSC and turbidity respectively (Figure 20).

The trend direction classification pertains to an individual site and the error risk describes the degree of confidence in trend direction that its data provides. The misclassification error risk for individual sites can be disregarded when considering water quality trends globally (i.e., for all sites across the region). The logic for this is that over many sites, incorrect classifications of direction will cancel each other (i.e., as many sites will be misclassified as increasing as

sites misclassified as decreasing). Therefore, the general regional change in the water quality variables is summarised by the distributions of RSS values (Figure 18 and Figure 21). These data indicate that the majority (often > 75%) of sites have improving trends. Thus, the trend analyses of SoE sites (i.e., in rivers not subject to specific point source impacts) provide strong evidence of general regional improvement in the four water quality measures over the past decade.

Trend analyses of discharge and impact sites resulted in similar patterns to those of SoE sites. The largest trend category for all variables was 'uncertain' for both types of sites (Table 11, Table 12). However, when more lenient levels of confidence were accepted there was a clear pattern of improving trends for both discharge (Figure 24) and impact (Figure 26) sites. For example, trends at discharge sites were at least as likely as not to be improving at 74% and 59% of sites for *E. coli* and SSC respectively (Figure 24). Trends at impact sites were at least as likely as not to be improving at 71%, 87% and 100% of sites for *E. coli,* SSC and turbidity respectively (Figure 26). Thus, the trend analyses of discharge and impact sites provide strong evidence of regional improvement in water quality specific to point sources over the past decade.

6.3 Predicted regional improvement in swimming grades and sediment related water quality variables

The length of order 4+ rivers that are swimmable was estimated to have increased by 8% and 7% for the 10 and seven-year time-periods respectively (Table 22 and Table 25). These estimates however are based on a combination of analyses (trend analyses and two spatial models) – all of which are associated with uncertainties. In particular, the random forest models of trend direction had only poor to satisfactory performance (Table 24). The combined uncertainty of the two models was not able to be quantified. The estimates of increase in swimmable rivers should therefore be regarded as indicative.

Although the estimated increase in the length of swimmable rivers is uncertain, several lines of evidence support the conclusion that the length of swimmable rivers has increased over the past decade. First, most sites have shown improving trends in the three relevant *E. coli* statistics. In particular, 75% of SoE sites had decreasing trends in the proportion of samples exceeding 260 and 540 *E. coli* $100mL^{-1}$ (i.e., G260 and G540; Figure 18 and Figure 21). Second, there was a dominance of improving trends in water quality variables that indicate sediment contamination (clarity, turbidity and SSC; Figure 21). Because transport pathways for sediment and *E. coli* are similar (i.e., overland runoff; McDowell *et al.*, 2008), the improvements in sediment contamination support the observation that *E. coli* concentrations have reduced and therefore swimming grades have improved. Third, the trend direction classification models were based on relationships that were consistent with known management actions. For example, probability of improving trends was associated with catchments with soft sedimentary geology of intermediate area, slope and elevation (Figure 34 and Figure 38). In addition, the probability of improving trends was negatively associated with the proportion of catchment occupied by indigenous forest and scrub (Figure 34 and Figure 38). The combination of these relationships describes erosion prone hill country areas that have been targeted by the SLUI project and which have been associated with the most significant water quality interventions in the region. Fourth, analysis of trends in point source discharges and associated downstream impact sites indicated that 17 of 19 paired discharge-impact sites had concordant decreasing reducing *E. coli* trends (Table 33). Because point source discharges are located on large main-stem rivers, the improvements in point source discharges will have contributed to an increase in the swimmable length of these rivers.

The study has also shown that the regional river water quality has improved with respect to the sediment related water quality variables. The median SSC values exceeded by 25% of segments were estimated to have reduced by 1.3 g m^{-3} and median turbidity values exceeded by 25% of segments were estimated to have reduced by 1.1 NTU over the seven-year trend period (Table 28). These reductions are considerable when compared to the mean of SoE site median values, which for SSC and turbidity were 10 g m^{-3} (range 1.0 – 72 g m^{-3}) and 6.4 NTU (range 0.5 – 38.5 NTU) respectively. The study estimated that median clarity had increased (i.e., improved) in segments of order four or greater but had decreased in segments of smaller rivers (Table 28). The median clarity values exceeded by 25% of segments of order four or greater were estimated to have increased by 0.3 m. Again, this increase is considerable when compared to the mean of SoE site median values, which was 2.1 m (range 0.2 – 5.6 m).

Regional improvement in the sediment related water quality variables (clarity, SSC and turbidity) was estimated based on a combination of analyses (trend analyses and two spatial models) – all of which are associated with uncertainties (e.g., Table 27). The estimated regional improvement in the sediment related water quality variables should therefore be regarded as indicative. However, for reasons set out above for the *E. coli* trends, several lines of evidence support the conclusion that regional river clarity, SSC and turbidity have improved over the past decade.

6.4 Robustness of regional estimates of water quality improvement

The ability to track progress toward environmental objectives and report on the effectiveness of policies and interventions is an important part of resource management. A major reason for long term water quality monitoring is to contribute to the evaluation of policies and interventions. The national targets for swimming grades in fresh waters (Ministry for the Environment, 2017a) is an example of a management initiative that will require future monitoring and evaluation of management effectiveness. This study has provided an approach to those evaluations. Spatial modelling is fundamental to the approach because regional and national water quality monitoring networks are not representative of general conditions. Previous studies have shown the national river water quality monitoring network is biased to sites in more impacted environments (Larned and Unwin, 2012). Comprehensive assessments and quantification of change in state over a time-period must therefore be based on modelling to produce estimates that reflect the spatial distribution of actual conditions rather than conditions as reflected by the monitoring network (Snelder *et al.*, 2017). The combination of monitoring data and statistical spatial modelling undertaken by this study is an example of the way changes in state over a time period in relation to specific interventions can be assessed. Other types of modelling, such as more mechanistic approaches (e.g., CLUES, (Elliott *et al.*, 2016), could also potentially be used.

The robustness of modelled change in state over a time-period is influenced by the available data. In this study, a crude approach to predicting the direction of trends in all network segments was taken based on temporally static catchment characteristics such as landcover, elevation, slope and geology (e.g., Figure 34). These models only had poor to satisfactory performance, which is probably because these predictors are only associated with the locations at which drivers of water quality were changing through the time-period. For example, the RF classification models indicated that the probability of improving trends was associated with hill country farming areas, particularly in soft sedimentary geology (e.g., Figure 34 and Figure 35). These areas are associated with SLUI farms (Figure 8), which were subject to interventions through the time-period.

Spatial models that included predictors that described actual changes in the drivers of water quality over time would likely be more accurate and informative than the models defined by this study. Water quality changes are generally driven by a combination of management interventions aimed at improvement and resource use intensification that may drive degradation (Wilcock *et al.*, 2013). Therefore, both types of information would be required to improve predictions of change in state over the whole region. To make best use of monitoring data and to maximise model performance, information on interventions and changes in resource use are required at a spatial resolution that is consistent with the spatial grain of the response variables – which is determined by the water quality monitoring network. This means that optimal spatial data describing interventions and resource use would be consistent with the area of the smallest catchments represented by the water quality monitoring network, which is in the order of 1 km^2. However, to date it has not been possible to obtain spatial data describing land use and management changes through time at this level of spatial resolution due to the confidentiality of this type of information. If the ability to track progress toward environmental objectives and report on the effectiveness of policies and interventions is to be improved, access to land use and management data is required. More consideration of ways that this type of data could be provided such that confidentiality was maintained but with sufficient resolution to provide accurate and useful assessments is recommended.

6.5 Association between trends and interventions

This study found statistically significant associations between interventions and water quality improvements. The analysis of associations was possible because HRC had maintained records of the actions that included the geographic location. This highlights the value of not only water quality monitoring, but also monitoring and recording management actions.

These associations are correlative and do not prove that the interventions caused the water quality improvements. In addition, if information describing resource use intensification within the catchments of the water quality monitoring sites were available, different conclusions about the associations between interventions and water quality improvements might be reached.

Trend direction at SoE sites was significantly associated with a combination of the proportion of upstream catchment occupied by SLUI farms, erosion and planting and the proportion of river segment length with new fencing (Table 29, Table 30, Table 31 and Table 32). The interventions were generally positively associated with water quality improvements (i.e., increasing areas of interventions were positively associated with higher probability of improving trends (Figure 45 and Figure 47) or increasing trend magnitude (Figure 46 and Figure 48). There were positive relationships between trends in *E. coli*, G260 and G540 and fencing and planting (i.e., trend magnitude increased with increasing values of the predictors) for both the 10 (Figure 45 and Figure 47) and seven-year (Figure 46 and Figure 48) time periods. This result is counter to expectations and may arise because the interventions are targeted at catchments that have been identified as subject to degrading trends or in a degraded state. In addition, the catchments in which these interventions occurred may also be subject to land use intensification that increased faecal contamination. It is noted the fencing and planting were negatively associated with trends in SSC and turbidity for the seven-year time-period (Figure 47 and Figure 48).

In addition to interventions aimed at reducing non-point sources of *E. coli* and sediment, there has been improvement in point source discharges throughout the region (Table 33). The concordance in trend direction between paired discharge-impact sites is evidence that changes of concentrations in discharges influence downstream receiving environment concentrations for *E. coli* and SSC (Table 33). In addition, changes in concentration in

discharges (and associated downstream receiving environments) were predominantly decreases in the seven-year period. This is evidence that improvements to point sources have contributed to the general regional improvement (at SoE sites) that has occurred over the past decade.

The study also established regional trends in rainfall and flow through the past decade. There was weak evidence for regional increases in rainfall over the 10-year period ended to 2016 but not for the seven-year period (Table 34). There were statistically significant regional decreasing trends in flows for both time-periods (Table 34). These climate induced trends may be at least partly involved in the water quality improvements. Conversely however, reduced flows over the period may be masking the true extent of the improvements because dilution has reduced.

The study provides several lines of evidence of associations between interventions and water quality improvements. This evidence is based on correlations and may be confounded by unmeasured variables such as land use intensification and/or climatic variation. However, the water quality improvements are consistent with mechanistic understanding of the effect of mitigations on the production of *E. coli* (Elliott and Whitehead, 2016, Semadeni-Davies and S. Elliot, 2016) and sediment (Manderson *et al*., 2015) from catchments. It is therefore not possible to conclude with certainty that the water quality interventions have caused the observed water quality improvements in the Region, but it seems likely they have at least contributed.

A mixture of interventions has been deployed by HRC (e.g., farm plans with various individual mitigations, point source upgrades and individual stream fencing and planting) and data describing exactly what occurred where was not available. Because of this lack of mitigation specific data and the potential effects of unmeasured variables, this study (or type of study) cannot be used to quantify the effectiveness of individual mitigations. Specific controlled studies are required for this type of evaluation (e.g., Monaghan *et al.*, 2008; Wilcock *et al.*, 2006). Nevertheless, the results provide encouraging signs that the improvements from local scale interventions are collectively contributing to regional scale water quality improvement.

6.6 Flow adjusting as part of trend assessment

This study considered the issue of flow adjusting water quality data as part of trend assessment in some detail (see Appendix B). There are good reasons to flow adjust. Adjusting data to account for flow (or any covariate) decreases variation and increases statistical power (i.e., increases the likelihood of detecting a trend with certainty, (Helsel and Hirsch, 1992). In addition, flow adjustment can improve trend detection if there has been a bias in the flow on sample occasion (i.e., increasing or decreasing flow on sample occasion with time). However, decisions concerning the appropriateness of water quality variable - flow models that underlie flow adjustment are subjective and site specific. This means that inspection of the data for all trend analysis is required and that trends based on automatic (i.e., non-supervised) flow adjustment should not be relied on.

Based on the examination of a subset of sites with adequate flow data, it was concluded that the findings of this study would not be significantly different if flow adjusted trends had been used. It is not known whether this finding can be extended to other studies and it is recommended similar analyses are undertaken for any study of water quality trends to investigate the importance of flow adjustment.

Ideally there would be a more objective basis for choosing to flow adjust (and for choosing the appropriate model for doing so). There have also been recent developments of techniques for trend analysis that incorporate flow in a more flexible and robust manner than the traditional methods (e.g., Hirsch *et al.*, 2015). Given the importance of trend analysis, it is recommended that flow adjusting and trend assessment in general are further investigated.

Acknowledgements

Evan Harrison of MFE is thanked for assistance with swimming grades and *E. coli* statistics. Thanks to Jon Roygard, Abby Matthews and Staci Boyte of HRC for provision of data, site information review and advice of various types during the study. Caroline Fraser of LWP Ltd provided technical reviews and editorial comments. Formal reviews that improved the original report were provided by John Quinn of NIWA and Vic Duoba and Vince Galvin or StatsNZ. Thanks are also due to Ian Jowett, who gave freely of his time for investigations associated with flow adjustment. Sheree De Malmanche and Jon Roygard were the originators of this project and I am grateful for the resources that they provided.

References

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

Ballantine, D., 2012. Water Quality Trend Analysis for the Land and Water New Zealand Website (LAWNZ): Advice on Trend Analysis. Horizons Regional Council Report, Horizons Regional Council, Palmerston North, NZ.

Ballantine, D., D. Booker, M. Unwin, and T. Snelder, 2010. Analysis of National River Water Quality Data for the Period 1998–2007. Christchurch. http://www.mfe.govt.nz/publications/water/.

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

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

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

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

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

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

Elliott, S. and A. Whitehead, 2016. Effect of E. Coli Mitigation on the Proportion of Time Primary Contact Minimum Acceptable State Concentrations Are Exceeded: Technical Note. NIWA Client Report, Hamilton, New Zealand.

Friedman, J.H., 1991. Multivariate Adaptive Regression Splines. The Annals of Statistics:1–67.

Hanley, J.A. and B.J. McNeil, 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology 143:29–36.

Helsel, D.R., 2012. Reporting Limits. Statistics for Censored Environmental Data Using Minitab® and R, Second Edition:22–36.

Helsel, D.R. and R.M. Hirsch, 1992. Statistical Methods in Water Resources. Elsevier.

Hirsch, R.M., S.A. Archfield, and L.A. De Cicco, 2015. A Bootstrap Method for Estimating Uncertainty of Water Quality Trends. Environmental Modelling & Software 73:148–166.

Hirsch, R.M., J.R. Slack, and R.A. Smith, 1982. Techniques of Trend Analysis for Monthly Water Quality Data. Water Resources Research 18:107–121.

Jowett, I.G., 2017. Time Trends. Hamilton, New Zealand.

Larned, S., T. Snelder, and M. Unwin, 2016. Water Quality in New Zealand Rivers; Modelled Water Quality State. NIWA CLIENT REPORT, NIWA, Christchurch, New Zealand.

Larned, S.T., T. Snelder, M. Unwin, and G.B. McBride, 2016a. Water Quality in New Zealand Rivers: Current State and Trends. New Zealand Journal of Marine and Freshwater Research In Press. doi:http://dx.doi.org/10.1080/00288330.2016.1150309.

Larned, S.T., T.H. Snelder, M. Unwin, and G.B. McBride, 2016b. Water Quality in New Zealand Rivers: Current State and Trends. NZJMFR.

Larned, S.T., T.H. Snelder, M. Unwin, G.B. McBride, P. Verburg, and H.K. McMillan, 2015. Analysis of Water Quality in New Zealand Lakes and Rivers. NIWA Client Report, NIWA, Christchurch, New Zealand.

Larned, S.T. and M. Unwin, 2012. Representativeness and Statistical Power of the New Zealand River Monitoring Network. NIWA Client Report, NIWA, Christchurch, New Zealand.

Manderson, A., J.R. Dymond, and A.-G. Ausseil, 2015. Climate Change Impacts on Water Quality Outcomes from the Sustainable Land Use Initiative (SLUI). Landcare Research Client report, Landcare Research Ltd, Palmerston North, NZ.

McBride, G.B., 2005. Using Statistical Methods for Water Quality Management: Issues, Problems and Solutions. John Wiley & Sons.

McBride, G., 2016. National Objectives Framework; Statistical Considerations for Design and Assessment. NIWA Client Report, NIWA, Hamilton, New Zealand.

McDowell, R.W., D.J. Houlbrook, R.W. Muirhead, K. Muller, M. Shepard, and S.P. Cuttle, 2008. Grazed Pastures and Surface Water Quality. Nova Science Publishers Inc, p. 238.

McDowell, R.W., P. Moreau, J. Salmon-Monviola, P. Durand, P. Leterme, and P. Merot, 2014. Contrasting the Spatial Management of Nitrogen and Phosphorus for Improved Water Quality: Modelling Studies in New Zealand and France. European Journal of Agronomy 57:52–61.

MFE, 1994. Water Quality Guidelines No.2 - Guidelines for the Management of Water Colour and Clarity. Wellington.

Ministry for the Environment, 2015. Research Needed to Develop Sediment Attributes for the National Objectives Framework. MFE, Wellington, New Zealand.

Ministry for the Environment, 2017a. National Policy Statement for Freshwater Management 2014 (Amended 2017).

Ministry for the Environment, 2017b. Swimming Categories for E. Coli in the Clean Water Package: A Summary of the Categories and Their Relationship to Human Health Risk from Swimming. MFE publication, MFE, Wellington, New Zealand.

Monaghan, R.M., C.A. De Klein, and R.W. Muirhead, 2008. Prioritisation of Farm Scale Remediation Efforts for Reducing Losses of Nutrients and Faecal Indicator

Organisms to Waterways: A Case Study of New Zealand Dairy Farming. Journal of Environmental Management 87:609–622.

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

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

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

Schertz, T.L., R.B. Alexander, and D.J. Ohe, 1991. The Computer Program Estimate Trend (ESTREND), a System for the Detection of Trends in Water-Quality Data. US Department of the Interior, US Geological Survey. http://pubs.water.usgs.gov/wri914040/pdf/wri91-4040.aug.pdf.

Schierlitz, C. and J.R. Dymond, 2006. Erosion/Sedimentation in the Manawatu Catchment Associated with Scenarios of Whole Farm Plans. Landcare Research Contract Report, Landcare Research Ltd, Palmerston North, New Zealand.

Semadeni-Davies, A. and S. Elliot, 2016. Modelling the Effect of Stock Exclusion on E. Coli in Rivers and Streams: National Application. NIWA Client Report prepared for the Ministry of Primary Industries, Hamilton, New Zealand.

Smith, D.G., G.B. McBride, G.G. Bryers, J. Wisse, and D.F. Mink, 1996. Trends in New Zealand's National River Water Quality Network. New Zealand Journal of Marine and Freshwater Research 30:485–500.

Snelder, T.H., S.T. Larned, and R.W. McDowell, 2017. Anthropogenic Increases of Catchment Nitrogen and Phosphorus Loads in New Zealand. New Zealand Journal of Marine and Freshwater Research:1–26.

Snelder, T., S. Woods, and J. Atalah, 2016a. Strategic Assessment of New Zealand's Freshwaters for Recreational Use: A Human Health Perspective Escherichia Coli in Rivers and Planktonic Cyanobacteria in Lakes. LWP Client Report, LWP Ltd, Christchurch, New Zealand.

Snelder, T.H., R.M. McDowall, and C. Fraser, 2016b. Estimation of Catchment Nutrient Loads in New Zealand Using Monthly Water Quality Monitoring Data. Journal of the American Water Resources Association. doi:10.1111/1752-1688.12492.

Snelder, T.H., R. Woods, and B.J.F. Biggs, 2005. Improved Eco-Hydrological Classification of Rivers. River Research and Applications 21:609–628.

Stats NZ, 2017. Technical Note on the Initial Assessment of Modelled E.coli Data. Technical Report, Ministry for the Environment & Stats NZ.

Stocker, T., D.Q. Qin, and G.-K. Plattner (Editors)., 2014. Climate Change 2013: The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University

Press.
http://www.climatechange2013.org/images/report/WG1AR5_Frontmatter_FINAL.pdf.

Svetnik, V., A. Liaw, C. Tong, and T. Wang, 2004. Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. F. Roli, J. Kittler, and T. Windeatt (Editors). MCS LNCS 3077??? Springer-Verlag, Berlin Heidelberg, pp. 334–343.

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

Wilcock, R.J., R.M. Monaghan, J.M. Quinn, A.M. Campbell, B.S. Thorrold, M.J. Duncan, A.W. McGowan, and K. Betteridge, 2006. Land-Use Impacts and Water Quality Targets in the Intensive Dairying Catchment of the Toenepi Stream, New Zealand. New Zealand Journal of Marine and Freshwater Research 40:123–140.

Wilcock, R.J., R.M. Monaghan, J.M. Quinn, M.S. Srinivasan, D.J. Houlbrooke, M.J. Duncan, A.E. Wright-Stow, and M.R. Scarsbrook, 2013. Trends in Water Quality of Five Dairy Farming Streams in Response to Adoption of Best Practice and Benefits of Long-Term Monitoring at the Catchment Scale. Marine and Freshwater Research 64:401–412.

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

Appendix A Investigation of alternative transformations and methods for modelling water quality state

A1 Considerations

It was noted by Snelder *et al.* (2016a) that the distributions of the *E. coli* statistics G260 and G260 had values between zero and one (because these are proportions), which could be made more symmetric with a logit transformation:

$$\text{logit} = \log[x/(1 - x)]$$

where x are values in the range 0 to 1 and the results are values between $-\infty$ and $+\infty$. Values in logit space are converted back to the range 0 to 1 by the inverse logit function:

$$\text{inverse logit} = \exp[x]/[1+\exp(x)]$$

Snelder *et al.* (2016a) found that when modelling G260 and G540 a logit transformation of the response variables (i.e., G260 and G540) did not improve model performance and left these variables untransformed when fitting the spatial models. There is however a consideration in addition to model performance that is associated with transformation of the modelled response variable. Because RF models are a partitioning method, the predictions are never outside the range of the response and in fact are always slightly less than the observed range. The truncation of the observed range can be exacerbated by variables that have uniform distributions such as those of G260 and PropGT540. The logit transformation has the effect of stretching out the range of the uniformly distributed G260 and G540 values and reduces the extent to which the predictions truncate the observed range. This reduction in the truncation of the range of model predictions is important in the situation that low (or high) values of 0-1 distributed observations are rare and where low values have special importance. In the national models of Snelder *et al.* (2016a), the predicted range of G260 and G540 values was not severely truncated because there were reasonable numbers of sites with low values of G260 and G540. However, in this study there were few SoE sites with low values of G260 and G540 (i.e., sites with very good water quality) and models fitted to untransformed response data resulted in some unrealistic predictions. In addition, low values of G260 and G540 have special significance as these determine excellent swimming grades (Table 1). For example, if the model is not able to predict values of 0.05 or less for G540 or less than 0.2 for G260 then no locations will be predicted to have swimming grades in the excellent category, even though a small number of SoE sites may in fact have this swimming grade.

A2 Test Methodology

This study investigated two possible approaches to reducing the problem of truncation so that predictions of G260 and G540 are realistic. First, the G260 and G540 values were logit transformed prior to fitting RF models and bias in the back-transformed model predictions were investigated. Second, the use of an alternative type of statistical model called a multivariate adaptive regression spline (MARS) was investigated.

The effect of transformation and model type on predictions of the *E. coli* statistics was examined by numerical experimentation. RF models were first fitted to the *E. coli* statistics without any transformations and compared the range in the fitted and observed values. Models were then fit using $\log_{10}$ transformed median values and logit transformed the G260 and G540 values. The fitted values of both sets of models were then compared with the observed values.

The performance of these models was evaluated as was the performance of the models of the median *E. coli* with and without corrections for re-transformation bias.

MARS models are a type of regression model that can be seen as an extension of linear models (Friedman, 1991). Like RF, MARS has the advantage over traditional statistical models of automatically modelling nonlinearities and interactions between variables. Unlike RF, MARS models are based on fitting piece-wise linear models and predictions are therefore not subject to truncation of the ranges of the observed response values; in fact, MARS models will extrapolate predictions outside the range of the observations like a simple linear regression model. The use of MARS models to make predictions of G260 and G540 was investigated. Of interest was whether MARS-based predictions of low values of G260 and G540 were more realistic than RF models and whether MARS model performance was comparable with RF models. The median values were $\log_{10}$ transformed and the G260 and G540 values were logit transformed before fitting the MARS models. The predictions of the model of the median values were corrected for re-transformation bias when examining the model performance.

The performance of RF and MARS models were tested objectively using leave-one-out cross validation. In this process both models were fitted to the available data for all but one site. The fitted models were then used to make a prediction for the “hold-out” site. This process was repeated for all sites and the independent predictions for each site were then compared with the observed values for both models. The performance of both models was quantified using the performance measures described in Table 5.

A3 Results of tests

The effect of transformation and model type on predictions of the *E. coli* statistics was examined by numerical experimentation using the 10-year time-period dataset. RF models fitted to the untransformed *E. coli* statistics significantly over-estimated the minimum values of all three statistics (Table 35). The fitted minimum values for G260 and G540 were 0.06 and 0.04 respectively whereas the minimum observed values for both statistics were 0.01 (Table 35). The predictions produced low proportions of the network (by segment length) with low values of each statistic (Table 36, Figure 50). This resulted in predictions of very low proportions of segments in the excellent and good swimming grades (Figure 50). This result is unrealistic when compared to the observed grades at the SoE sites (Table 36).

RF models fitted to $\log_{10}$ transformed median values and logit transformed G260 and G540 values produced minimum fitted values that were closer to the observations than their untransformed counterparts (Table 35). Although minimum predicted G260 and G540 values were higher than the minimum observed values, predictions of segments in the excellent and good swimming grades (Figure 28, Figure 50) were more realistic, particularly when compared to the observed grades at the SoE sites (Table 36).

MARS models fitted to logit transformed the G260 and G540 values produced minimum fitted values that were the same as the observations (Table 35). On the other hand, the MARS models produced fitted values for G260 and G540 that were higher than the observed values and fitted values for the Median that were lower than observed. The MARS model predictions produced low proportions of network segments (by length) in the fair and intermittent swimming grades and high proportions in the poor grade (Figure 51), particularly when compared to the proportion SoE sites with these (Table 36).

Table 35. Minimum and maximum observed and fitted values for different transformations and model types. The models were fitted to the three E. coli statistics represented in the 10-year time-period dataset.

Variable	Model	Transformation	Minimum observation	Minimum prediction	Maximum observation	Maximum prediction
E. coli	RF	None	0.5	37	768	565
G260	RF	None	0.001	0.057	0.85	0.78
G540	RF	None	0.001	0.035	0.63	0.50
E. coli	RF	Log10	0.5	3.2	768	553
G260	RF	Logit	0.001	0.011	0.85	0.78
G540	RF	Logit	0.001	0.004	0.63	0.50
E. coli	MARS	Log10	0.5	0.7	768	413
G260	MARS	Logit	0.001	0.001	0.85	0.91
G540	MARS	Logit	0.001	0.001	0.63	0.7

Table 36. Proportion of segments predicted to be in the five swimming grades by the different combinations of model and transformation. The first line shows the proportions of the 69 SoE sites used as fitting data for comparison.

Predictions	Excellent	Good	Fair	Intermittent	Poor
SoE sites	12	12	16	38	23
RF Untransformed All segments	0	2	3	19	76
RF Untransformed Order 4+	1	9	16	38	35
RF Transformed All segments	20	10	6	20	44
RF Transformed Order 4+	13	10	13	34	29
MARS All segments	23	3	4	10	61
MARS Order 4+	13	6	8	25	47

Models and transformation choices produced differing model performance (Table 37). Transformation of the response variables did not markedly affect the performance of the RF models. All RF models had values of NSE, RSR and PBIAS close to, or better than, satisfactory. The RF model of median values had a large bias when the predictions were not corrected for back-transformation bias compared to the corrected predictions (Table 37). Bias increased when the G260 and G540 were logit transformed. It is not known if there is an analytical solution for correcting for this bias (in the same way that the smearing coefficient is used to correct for bias resulting from log and other transformations). However, the analysis suggests that the G260 and G540 values had small bias (small positive Bias and PBIAS values Table 37). In comparison to the RF models, the MARS models performed poorly. The NSE and RSR values were all well below the satisfactory level and the predictions of median values by MARS were strongly biased.

The RF models with transformation of the response variables were judged to be the most realistic and accurate of the tested choices. All further modelling used RF and $\log_{10}$ transformation of median values and logit transformation of G260 and G540. Predictions of the median have been corrected for back-transformation bias but predictions of G260 and G540 have simply been back-transformed with no correction for bias.

Table 37. Performance of the spatial models of E. coli statistics.

Statistic	Model	Transform	NSE	Bias	PBIAS	RSR	RMSD
Median	RF	None	0.46	-5.53	-3	0.73	130
G260	RF	None	0.49	0.00	-2	0.72	0.11
G540	RF	None	0.52	-0.01	-2	0.69	0.15
Median	RF	Log10 Uncorrected	0.43	32.7	16	0.76	134
G260	RF	Logit Uncorrected	0.50	0.02	10	0.71	0.11
G540	RF	Logit Uncorrected	0.57	0.01	3	0.66	0.14
Median	RF	Log10 Corrected	0.45	7.8	4	0.74	131
Median	MARS	Log10 Corrected	-9.04	-119.1	-60	3.17	563
G260	MARS	Logit Uncorrected	0.06	0.00	1	0.97	0.21
G540	MARS	Logit Uncorrected	0.19	0.00	0	0.90	0.14

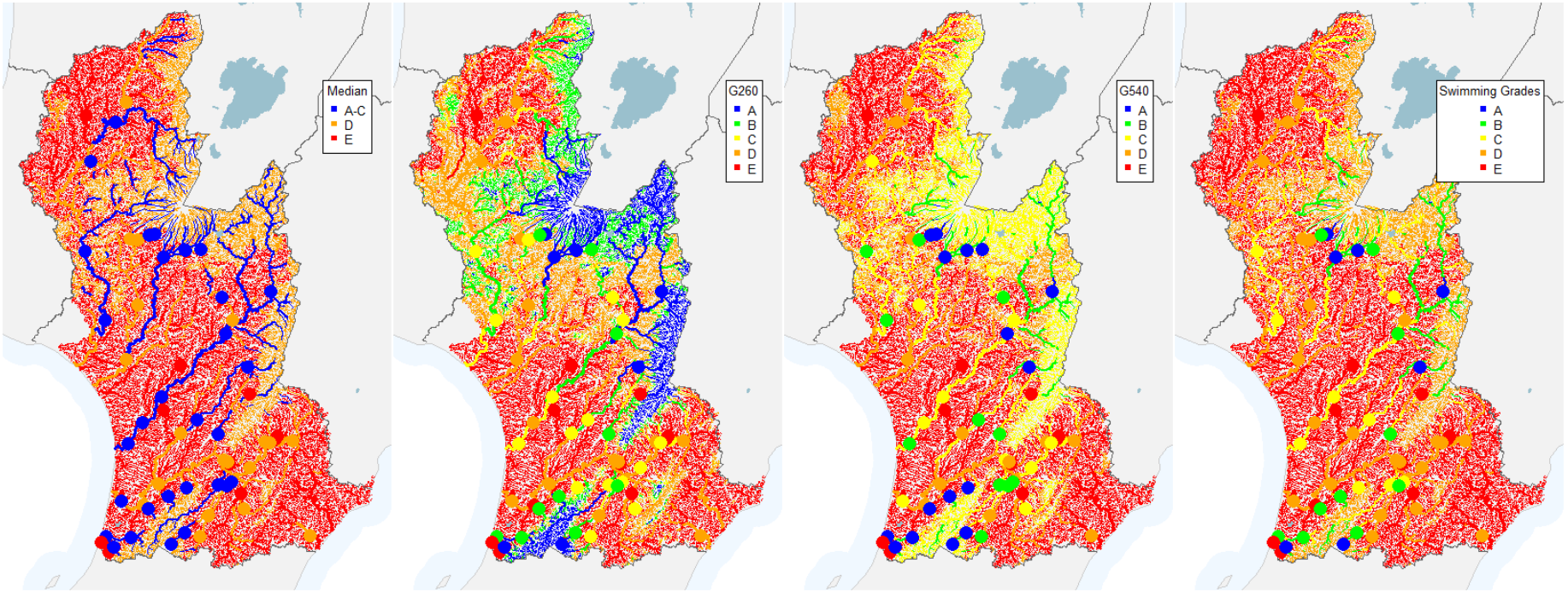

Figure 50. Spatial model predictions made using RF models and untransformed response variables for the 69 SoE sites represented in the 10-year dataset. The right-hand map represents the predicted swimming grade derived from analysis of the predicted values of the three statistics to the left. SoE sites are shown as dots with the colour representing the observed grade for the site (i.e., not the grade predicted by the model).

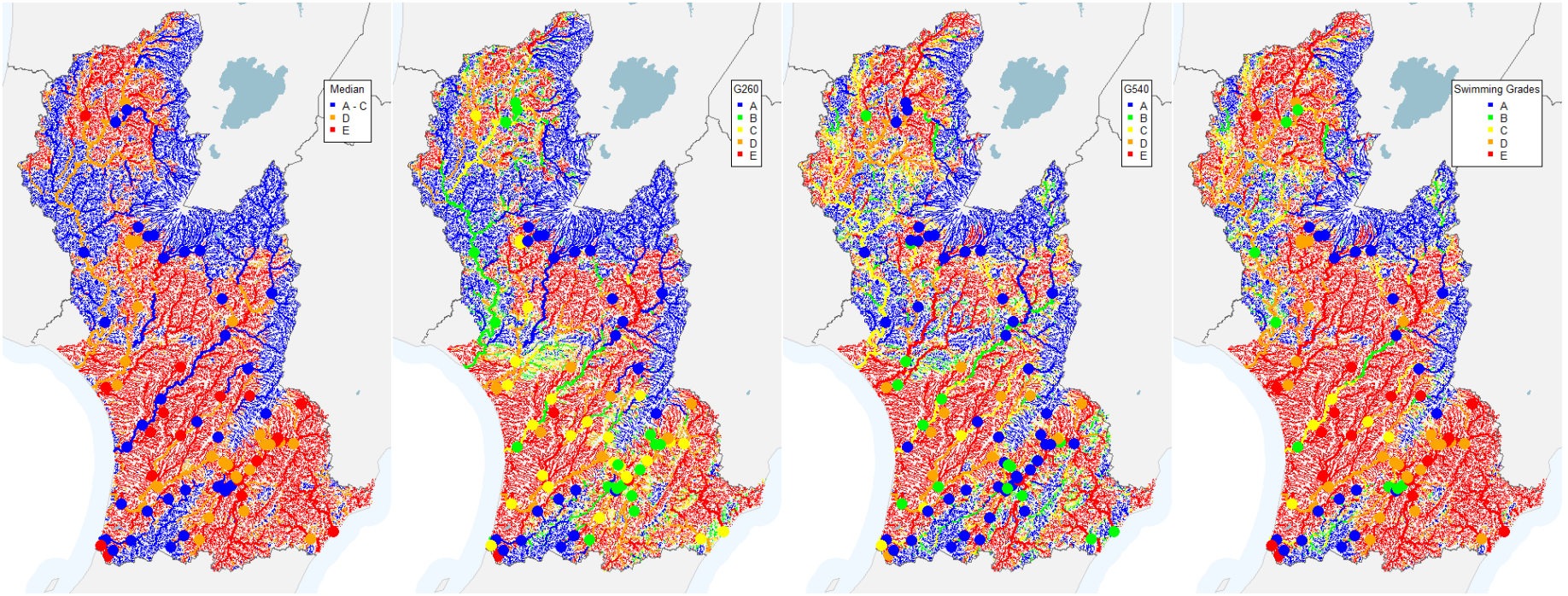

Figure 51. Spatial model predictions made using MARS models and transformed response variables for the 69 SoE sites represented in the 10-year dataset. A log_{10} transformation was applied to the median and logit transformations were applied to G260 and G540 prior to fitting the models. The right-hand map represents the predicted swimming grade derived from analysis of the predicted values of the three statistics to the left.

A4 Conclusions and recommendations for spatial modelling of state

The RF models had better performance than the alternative MARS models. Logit transformation of the G260 and G540 statistics did not improve the performance of the RF models but did improve their ability to discriminate small values of the statistics. Because small values of G260 and G540 are important (i.e., they define the Excellent and Good swimming grades), logit transformation of these variables is recommended. Back-transformation of predictions of a model fitted to logit transformed responses using the inverse logit function does not appear to produce appreciable bias. It is noted that naïve back-transformation of predictions of a model fitted to a log transformed response does produce bias. While both the logit and log10 transformation are non-linear, the logit transformation is not asymmetric, and this may be the reason that back-transformation does not induce a bias.

Appendix B Considerations regarding flow adjustment in trend analysis

B1 Considerations

Flow rate at the time that a river water quality measurement is made can affect the observed values because many water quality variables are subject to either dilution (decreasing concentration with increasing flow) or wash-off (increasing concentration with increasing flow) (Smith *et al.*, 1996). Different mechanisms may dominate at different sites so that the same water quality variable (e.g., *E. coli*) can exhibit positive or negative relationships with flow (Snelder *et al.*, 2016b).

Removing the effect of flow (or any covariate) decreases variation and increases statistical power (i.e., increases the likelihood of detecting a trend with certainty, (Helsel and Hirsch, 1992). In addition, a trend in the water quality variable may arise because there is a relationship between time and flow on sample occasion (i.e., increasing or decreasing flow on sample occasion with time). Removing the effect of flow may change the direction and/or magnitude of the trend and may make an uncertain (i.e., insignificant) trend direction certain.

Flow adjustment uses regression analysis to fit a line or curve to data to represent the relationship between the water quality variable and flow. The differences between the individual water quality measurements and the line or curve are the regression residuals, which represent the variation in the water quality variable that is not explained by, or independent of, flow. Flow adjusted values are derived as outlined by Smith *et al.* (1996):

Flow adjusted value = regression model residuals + median value

Various types of regression models are used to fit a line or curve to the water quality variable and flow data. Traditionally log-log relationships have been used but more flexible relationships have been used since the introduction of locally weighted least squares regression (Schertz *et al.*, 1991). For example, Larned *et al.* (2015) used a generalised additive model (GAM) and Ballantine *et al.* (2010) used locally weighted least squares regression (LOESS).

The problem with flow adjustment is that the adjusted values are sensitive to the underlying model of the water quality variable versus flow relationship. Because the model determines the regression residuals, large differences in trends can arise between raw and adjusted values and between values adjusted using different models. This problem is likely to be encountered when the data are obtained from monthly state of environment monitoring because they tend to be dominated by samples taken at low to median flows, and high flows are poorly represented. This can result in fitted lines or curves that are a poor fit to some of the data. It is therefore difficult to know whether confidence should be placed in trends based on the raw or flow adjusted data or which model is the most reliable basis for flow adjusting.

Advice on assessing the robustness of flow adjustment generally starts by considering if the shape of the fitted relationship is consistent with expectations. For example, typical relationships are monotonic, i.e., increase or decrease as flow increases (Smith *et al.*, 1996). Relationships may be well described by log-log models, but relationships can be curvilinear in log-log space and the rate of change in concentration with flow can plateau or decrease at high flow (Snelder *et al.*, 2016b). For this reason, flexible regression methods such as LOESS are promoted, particularly when large numbers of analyses are being carried out by automated methods (Helsel and Hirsch, 1992; Schertz *et al.*, 1991).

Schertz *et al.* (1991) advise inspection of the residual plots of regression models to check for normality and homoscedasticity (constant variance). However, it is not clear how to determine the extent to which deviations from these regression assumptions can be tolerated. Schertz *et al.* (1991) further advise that flow adjustment only be carried out if the model is significant. However, they acknowledge that removal of small amounts of flow related variability in the water quality variable can improve the detection of significant trends (i.e., establishing trend direction with certainty) and suggest relaxing alpha values to 0.10 or greater.

A more fundamental issue with use of water quality variable - flow models for flow adjustment is the assumption that the relationship applies over the full flow range and for the full period of record. Both assumptions are probably violated for at least some sites and variables. For example, for sediment the relationship varies with flow because the processes that determine sediment concentrations at high flow (i.e., wash-off, bank and bed erosion) are different from those that apply at low flows (i.e., resuspension of bed sediment). The relationship may change with time because sources in the catchment changes (erosion sources healing or being created).

There is therefore considerable subjectivity associated with flow adjusting water quality data that is probably inescapable. In addition, automation of flow adjustment in large analyses by selecting a single method may result in unrealistic flow adjustment for some sites and variables.

B2 Comparison of raw and flow adjusted trends in this study

B2.1 10 year dataset

Of the 69 SoE sites included in the 10-year time-period dataset, 18 had flow data for at least 80% of sample occasions. Regression models based on $\log_{10}$ of the variable versus $\log_{10}$ flow and a LOESS were reasonably consistent with each other for some sites but exhibited considerable departures from each other at other sites[8] (Figure 52).

[8] Plots for the complete set of sites are provided in supplementary file 10-year C-Q plots.pdf

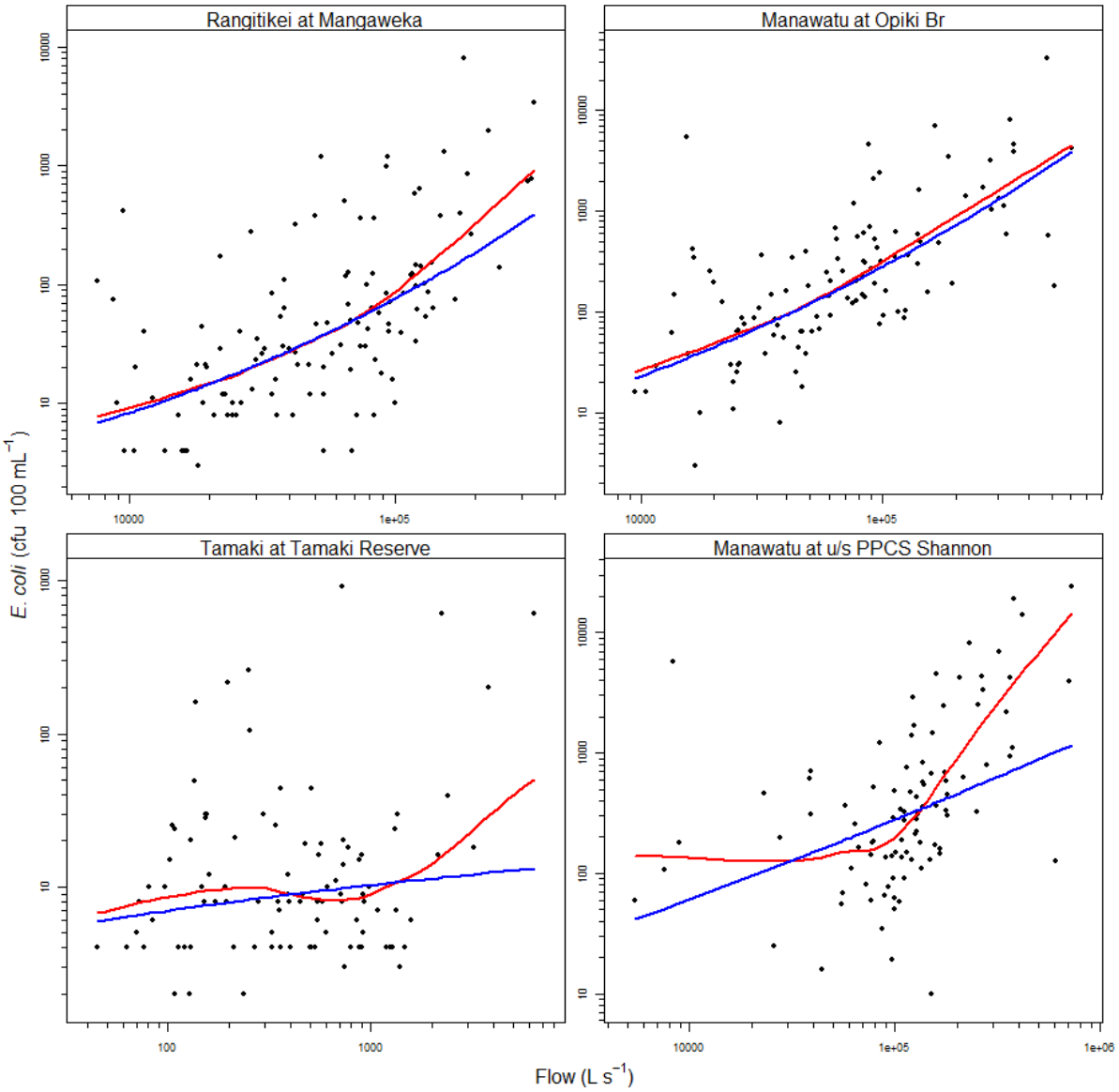

Figure 52. Example of E. coli concentrations versus flow for four sites in the 10-year time-period dataset. The blue line represents a regression fitted to log_{10} of the variable versus log_{10} flow and the red line represents a LOESS model. The flexibility of the LOESS models was determined by the span parameter which had the same value (0.8) for all four sites.

For trends calculated using raw data, 14 sites were categorised as uncertain and one and three sites were categorised as decreasing and increasing respectively. The trends calculated with flow adjusted data (based on a LOESS model), detected an extra increasing trend (Figure 53). The trends categorised as decreasing and increasing agreed for raw and flow adjusted data. One site had an uncertain raw trend and increasing flow adjusted trend.

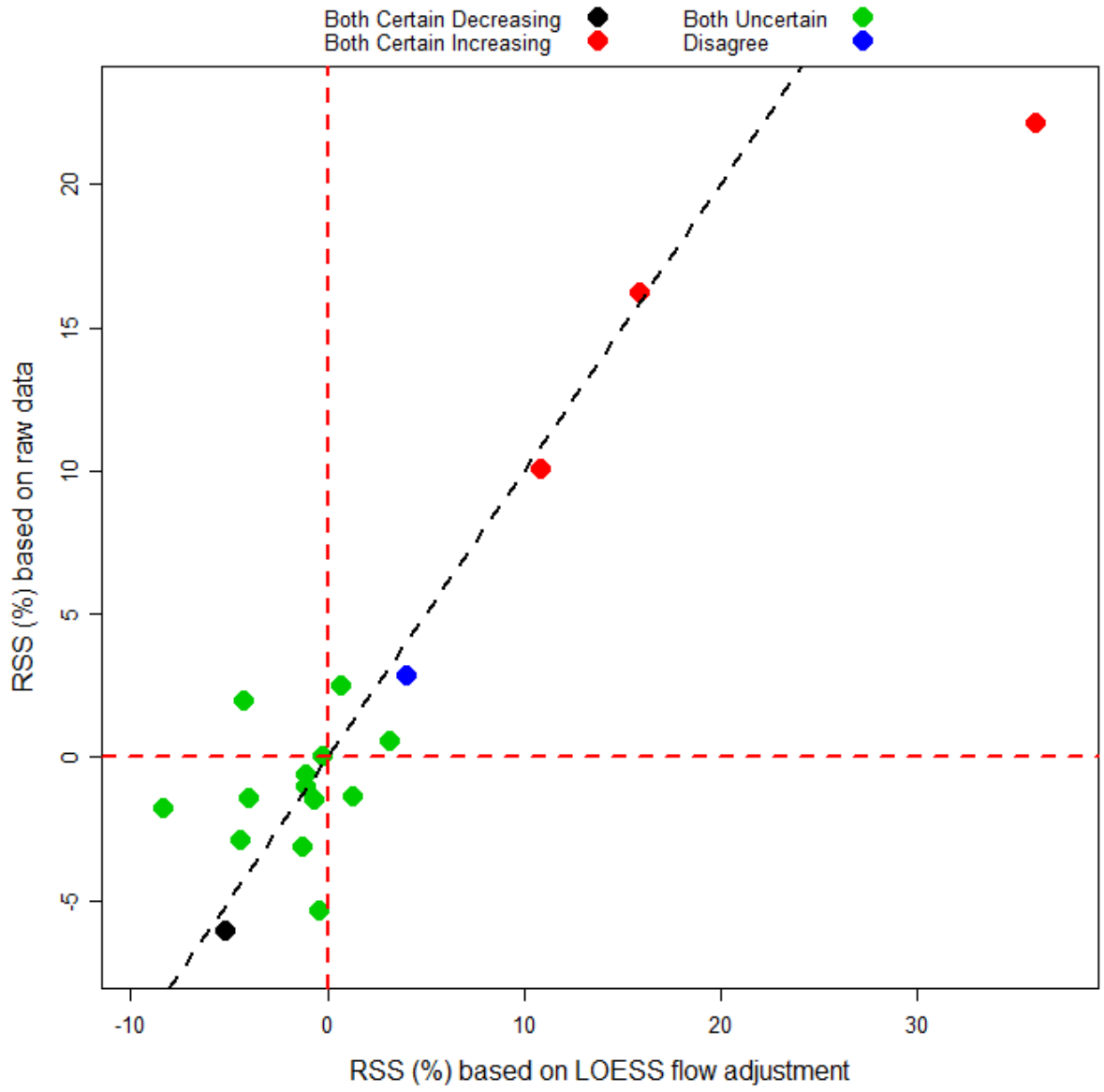

Figure 53. Relationship between magnitudes of trends evaluated using raw and flow adjusted (LOESS) data for median E. coli at 18 sites with flow data in the 10-year time-period dataset. The points are coloured to indicate whether trend categories agreed. Points located in the lower left and upper right quadrants (defined by the horizontal and vertical red dashed lines) indicate agreement in trend direction. Points located on the upper left and lower right quadrants indicate disagreement in the trend direction. The black dashed line is the one-to-one line and represents agreement in magnitude. Trends included on these plots complied with the inclusion rules but their directions were not necessarily established with confidence.

The trend categorisations were generally similar for two sets of trends that were flow adjusted based on log-log and LOESS models (Figure 54). The trends categorised as decreasing and increasing were the same for both sets of flow adjusted data. One site had an uncertain trend based on flow adjustment by log-log model and increasing trend based on the LOESS model (Figure 54).

The log-log flow adjustment produced two trends with large differences in magnitudes compared to the LOESS flow adjustment (Figure 54). These sites were the Manawatū at u/s PPCS Shannon site that had RSS values for log-log adjustment of approximately -28% compared to approximately the -4% for LOESS adjustment and the Tamaki at Tamaki Reserve (RSS values of -8% and 8% for LOESS and Log-log flow adjustments respectively). Inspection of the concentration-flow relationship for these sites (Figure 52) provides some insight into why the flow adjusted trends at these sites differed. The log-log model for the Manawatū at u/s PPCS Shannon site had large residual values for the high concentration samples (associated with high flow) because the fitted model was dominated by samples in the mid-concentration range. By contrast, the residuals for the LOESS fitted model were more homoscedastic because the more flexible model could represent the non-linear relationship. However, the LOESS fitted model exhibited an abrupt change in slope at median flow which may be considered unrealistic and which was influenced by a lack of data representing low flows (Figure 52). The log-log model for the Tamaki at Tamaki Reserve site also had large residual values for the high concentration samples (associated with high flow) because the fitted model was dominated by samples in the mid-concentration range. The residuals for the LOESS fitted model were more homoscedastic because the more flexible model could represent the non-linear relationship. However, the LOESS fitted model exhibited an unrealistic relationship, which was influenced by the distribution of the data (Figure 52). It is important to note that the smoothing parameter for all four LOESS models shown on Figure 52 were the same (span of 0.8). This indicates that decisions concerning the appropriateness of water quality variable - flow models that underlie flow adjustment are subjective and site specific.

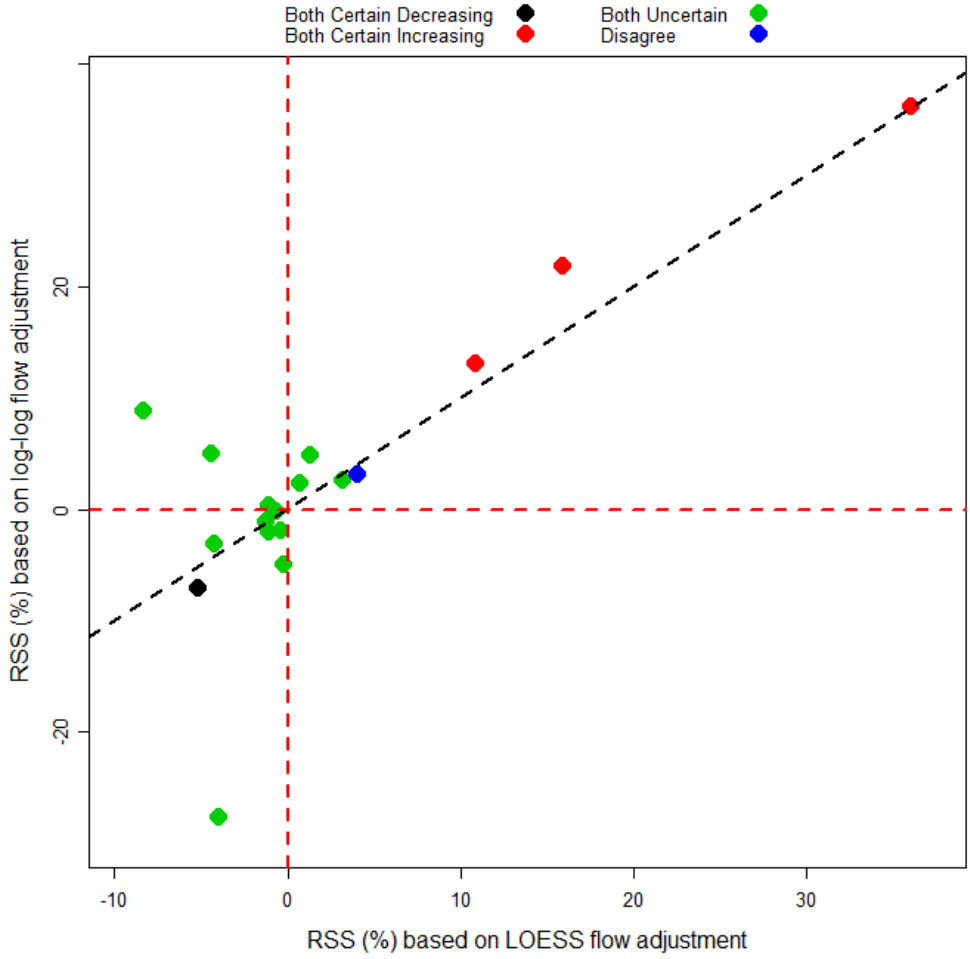

Figure 54. Relationship between magnitudes of flow adjusted trends evaluated using two models of the concentration-flow relationship (log-log and LOESS) data for median E. coli at 18 sites with flow data in the 10-year time-period dataset. The points are coloured to indicate whether trend categories agreed. Points located in the lower left and upper right quadrants (defined by the horizontal and vertical red dashed lines) indicate agreement in trend direction. Points located on the upper left and lower right quadrants indicate disagreement in the trend direction. The black dashed line is the one-to-one line and represents agreement in magnitude. Trends included on these plots complied with the inclusion rules, but their directions were not necessarily established with confidence.

The distributions of trend magnitudes (irrespective of confidence in direction) was reasonably similar for the raw and flow adjusted trends (Figure 55). The distribution of trend magnitudes for flow adjusted trends produced using the LOESS model were more similar to the raw trend magnitude distribution than the trends produced using the log-log model. The global (median over all sites) magnitude of the increasing and decreasing trends were similar for the raw and LOESS adjusted trends (-1.5% and -1.2% for decreasing and 2.8% and 4% for increasing). The global magnitude of the increasing and decreasing trends for log-log adjusted were somewhat larger (-2.5% and 4.9%).

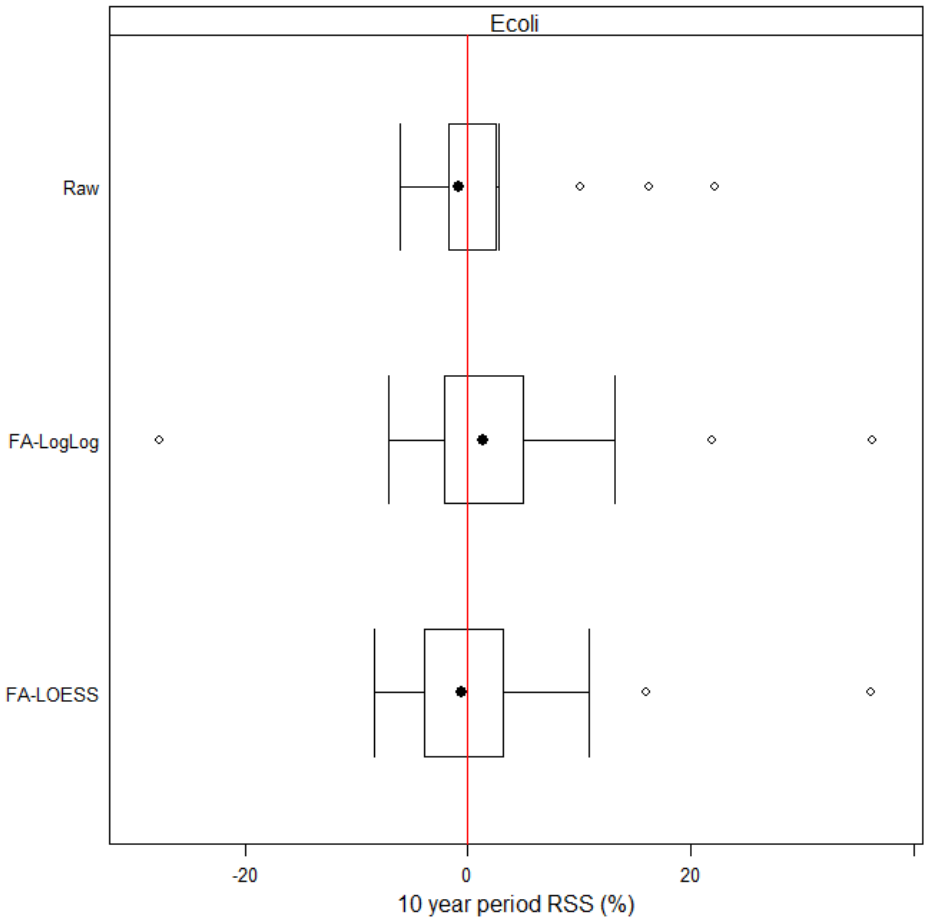

Figure 55. Distribution of trend magnitudes (RSS values) for the median E. coli statistic at 18 SoE sites with flow data in the 10-year time-period dataset.Trends are based on analyses performed using raw and flow adjusted data (indicated by the FA-LogLog and FA-LOESS groups on the y-axis). All sites complied the inclusion rules but their directions were not necessarily established with confidence.

B2.2 Seven-year dataset

Of the 88 SoE sites included in the seven-year time-period dataset, 28 had flow data for at least 80% of sample occasions. Regression models based on $\log_{10}$ of the variable versus $\log_{10}$ flow and a LOESS were reasonably consistent with each other for some sites but exhibited considerable departures from each other at other sites (data not shown[9]).

Trends calculated using raw data, were categorised as uncertain at 17, 10, 20 and 14 sites for *E. coli*, clarity, SSC and turbidity respectively. The trends categorised as certain decreasing and certain increasing agreed for raw and flow adjusted data (i.e., both certain increasing and decreasing Figure 56). Trends calculated with flow adjusted data (based on a LOESS model) included two additional decreasing trends and one additional increasing for *E. coli*; one decreasing and one increasing trends for SSC and three additional decreasing and one increasing trends for turbidity (Figure 56).

[9] Plots provided in supplementary file 7-year C-Q plots.pdf

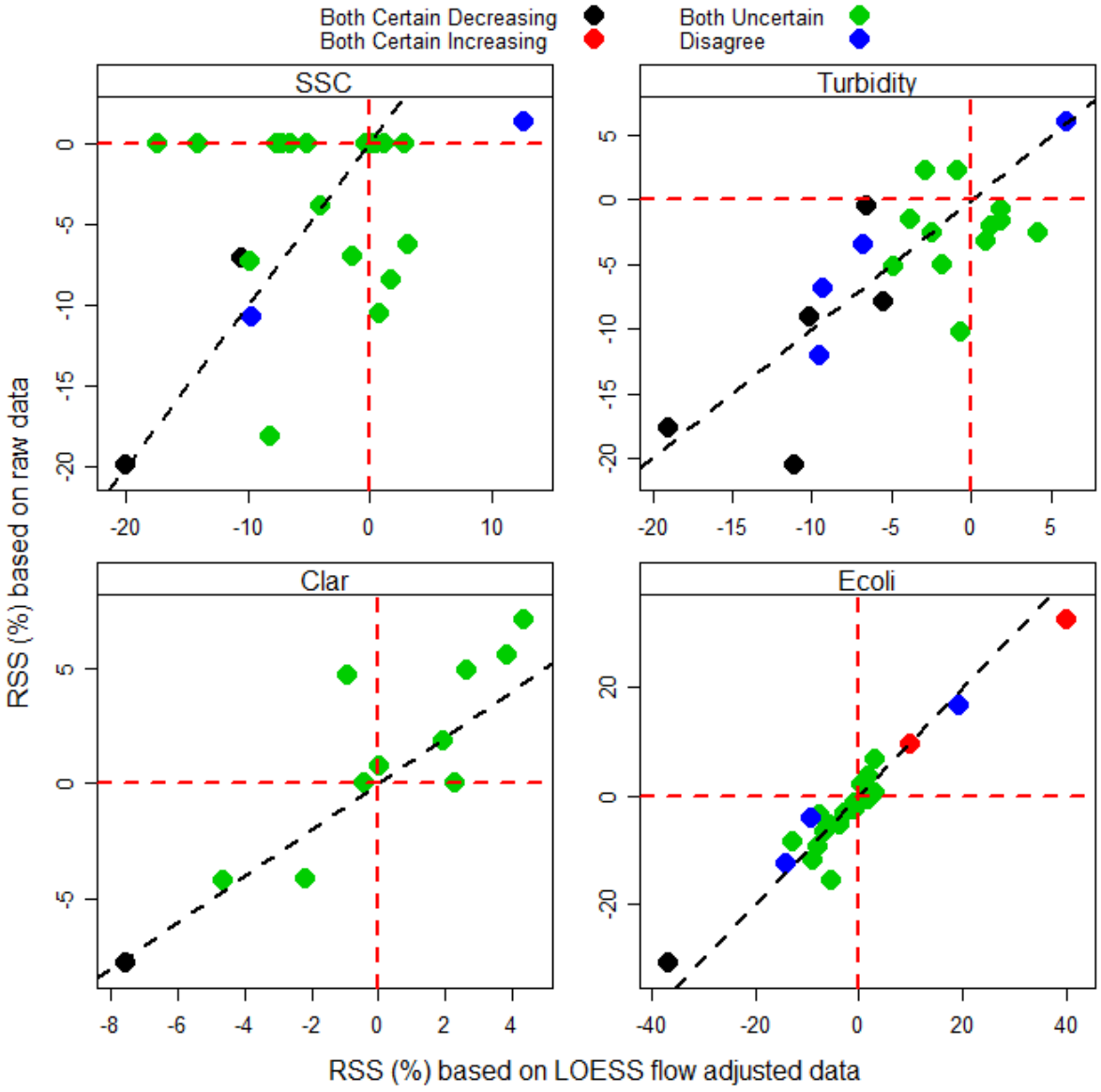

Figure 56. Relationship between magnitudes of trends evaluated using raw and flow adjusted (LOESS) data for the four water quality variables at 28 sites with flow data in the seven-year time-period dataset. The points are coloured to indicate whether trend categories agreed. Points located in the lower left and upper right quadrants (defined by the horizontal and vertical red dashed lines) indicate agreement in trend direction. Points located on the upper left and lower right quadrants indicate disagreement in the trend direction. The black dashed line is the one-to-one line and represents agreement in magnitude. Trends included on these plots complied with the inclusion rules, but their directions were not necessarily established with confidence.

The directions and magnitudes were generally similar for the two sets of trends that were flow adjusted based on log-log and LOESS models (Figure 57). The trends categorised as certain decreasing and certain increasing agreed for both sets of flow adjusted data (i.e., disagreements shown on Figure 57 occurred when one trend was categorised as uncertain

and the other was increasing or decreasing).

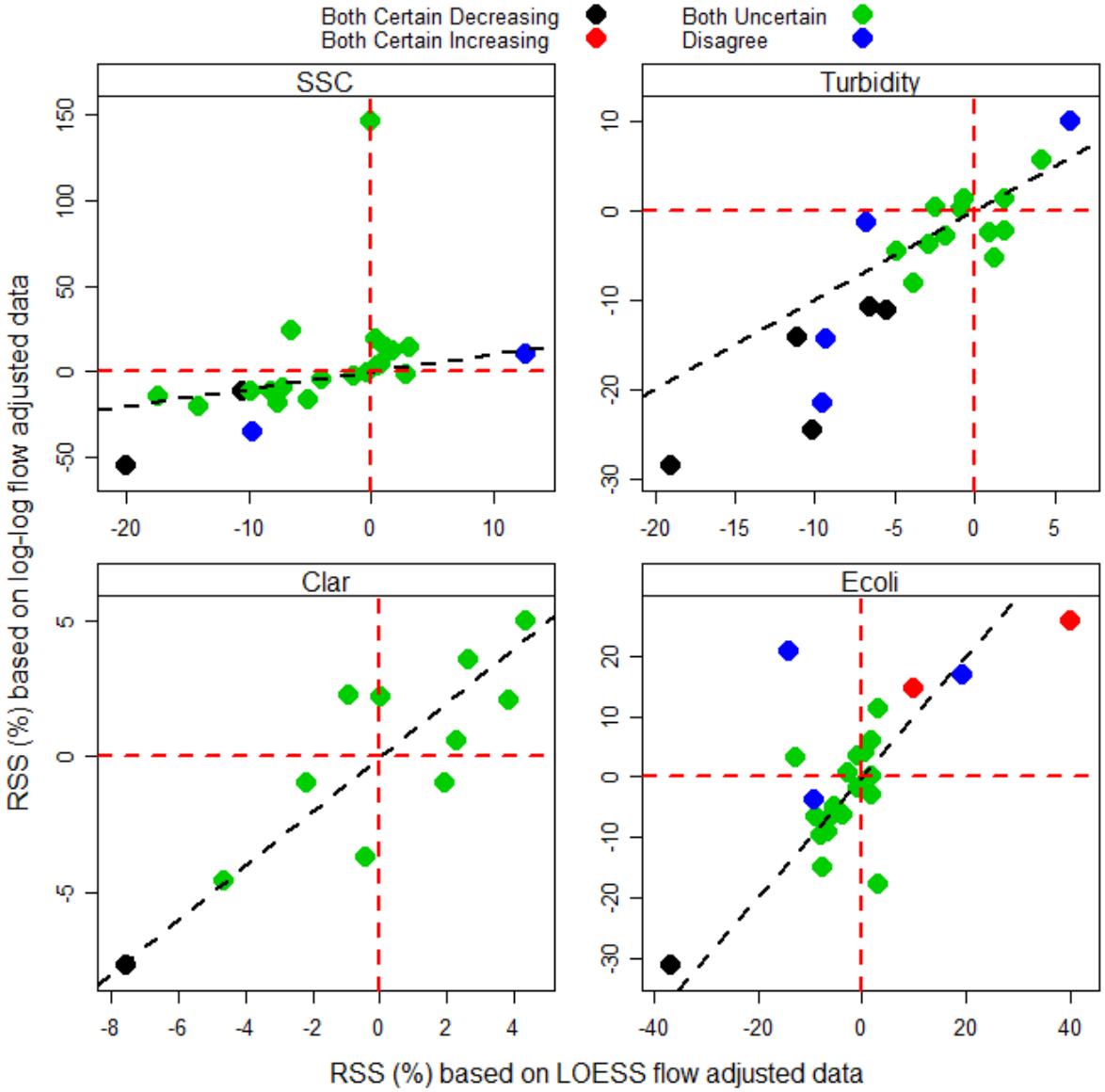

Figure 57. Relationship between magnitudes of flow adjusted trends evaluated using two models of the concentration-flow relationship (log-log and LOESS) data for median E. coli at 18 sites with flow data in the seven-year time-period dataset. The points are coloured to indicate whether trend categories agreed. Points located in the lower left and upper right quadrants (defined by the horizontal and vertical red dashed lines) indicate agreement in trend direction. Points located on the upper left and lower right quadrants indicate disagreement in the trend direction. The black dashed line is the one-to-one line and represents agreement in magnitude. Trends included on these plots complied with the inclusion rules, but their directions were not necessarily established with confidence.

The distributions of trend magnitudes (irrespective of confidence in direction) were reasonably similar for the raw and flow adjusted trends (Figure 58). The distribution of trend magnitudes for flow adjusted trends produced using the LOESS model were more like the raw trend magnitude distribution than the trends produced using the log-log model. For *E. coli*, the global (median over all sites) magnitude of the increasing and decreasing trends differed for the raw

and LOESS adjusted trends (-4.2% and -6.5% for decreasing and 4.1% and 2% for increasing). The global magnitude of the increasing and decreasing trends for log-log adjusted were somewhat larger (-7.9% and 4.7%).

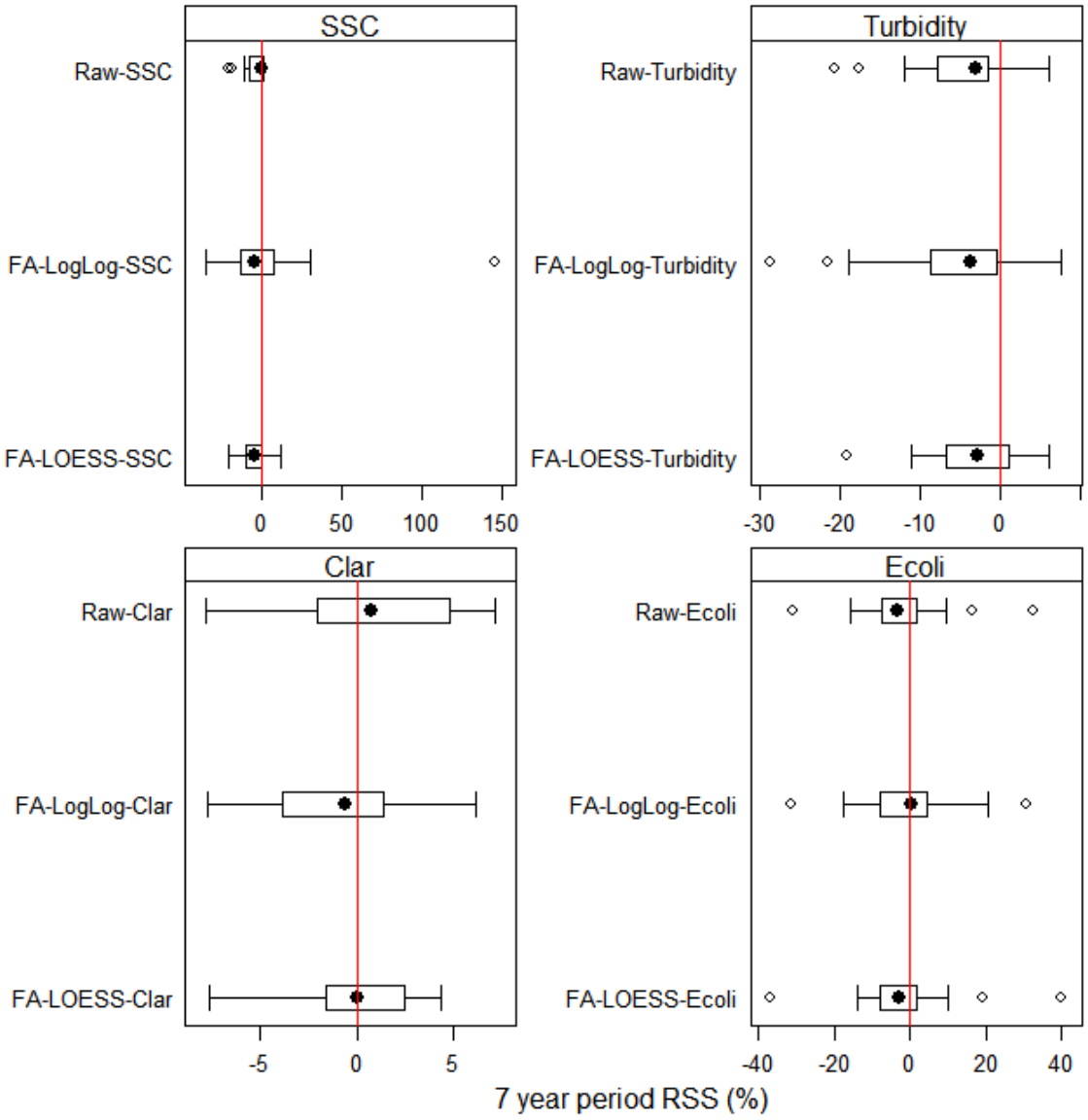

Figure 58. Distribution of trend magnitudes (RSS values) for the four water quality variables at 28 SoE sites with flow data in the seven-year time-period dataset.Trends are based on analyses performed using raw and flow adjusted data(indicated by the FA-LogLog and FA-LOESS groups on the y-axis). All sites complied the inclusion rules but their directions were not necessarily established with confidence.

B3 Conclusions and recommendations

Based on the examination of a subset of sites with adequate flow data, it is concluded that this study's findings would not be significantly different if flow adjusted trends had been used. The case study above indicates that decisions concerning the appropriateness of water quality variable - flow models that underlie flow adjustment are subjective and site specific. Because water quality variable versus flow relationships vary across sites and variables, automation of

flow adjustment in large analyses can result selection of unreliable models. Water quality variable - flow models that are fitted using flexible regression methods, such as LOESS, are more likely to achieve significance and satisfy the assumptions that the residuals are normally distributed and homoscedastic. However, the degree of flexibility (e.g., as defined by the smoothing parameter in a LOESS model) is subjective. Increasing the flexibility, and therefore improving the fit, should not result in obtaining a model whose shape is inconsistent with the mechanisms underlying the relationship. Choosing the most appropriate flow adjustment therefore requires expert judgement and is subjective.