



Aggregating Trend Data for Environmental Reporting

August 2018

Prepared By:

Ton Snelder

Caroline Fraser

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: LWP Client Report 2018-01

Prepared for: The Ministry for the Environment

Report Date: August 2018

LWP Project: 2018-01

Quality Assurance Statement


Version	Reviewed By	
Final	Charlotte Jones-Todd, NIWA, Hamilton	

Table of Contents

Executive Summary	v
1 Introduction	8
2 Trend analysis methods	9
2.1 Background to new trend assessment procedure.....	9
2.2 Aggregation of trend analyse from many sites.....	9
2.2.1 <i>Traditional approach</i>	9
2.2.2 <i>Graphical presentation of aggregated trends</i>	10
2.2.3 <i>Evaluation of the proportion of improving trends</i>	11
3 Case study	12
4 Trend analyses	13
4.1 Sampling dates and time-periods for analyses	13
4.2 Statistical analyses.....	14
4.3 Censored values	16
4.4 Seasonality	16
4.5 Flow adjustment.....	17
4.6 Categorisation of trends	17
4.7 Aggregation of trends.....	18
4.7.1 <i>Graphical presentation of aggregated trends</i>	18
4.7.2 <i>Proportion of improving trends</i>	18
4.8 Implementation	20
5 Results	21
5.1 National scale	21
5.2 Source of flow	25
5.3 Land cover	26
5.4 Region	28
6 Conclusions	30
6.1 New trend aggregation methods	30
6.2 Inferences from aggregated trends for 10 -year period ending 2013	31
6.3 Limitations of the PIT statistic.....	32
Acknowledgements.....	33
References.....	34

Figures

Figure 1. Proportion of censored samples in the dataset by variable.	13
Figure 2: Pictogram of the steps taken in trend analysis to calculate the Sen slope and the probability that the true trend was decreasing.	15
Figure 3. Summary plot representing the proportion of sites with improving 10-year time-period trends at each categorical level of confidence.	21
Figure 4. Comparisons of proportions of sites with improving trends over the 10-year time-period based on the PIT statistic and the count-based evaluation.	22
Figure 5. Evaluated PIT statistics for 10-year flow adjusted trends grouped by REC source-of-flow classes.	26
Figure 6. Evaluated PIT statistics for 10-year flow adjusted trends grouped by REC land-cover categories.	27
Figure 7. Evaluated PIT statistic for 10-year flow adjusted trends grouped by region. .	29

Tables

Table 1. Level of confidence categories used to convey the probability that water quality was improving.	11
Table 2. River water quality variables, measurement units and site numbers used in this study.	12
Table 3. Modified river environment classification (REC) source-of-flow classes and land-cover categories used by this study.	19
Table 4. Regional groupings and data collecting agency.	20
Table 5. Estimates of proportions of sites with improving 10-year site-trends (national grouping).	24

Executive Summary

Trend analyses performed on many sites are regularly aggregated by water quality variable and summarised in tabular, graphical or map format as part of environmental reporting. The intention of aggregated site summaries (e.g., proportion improving and degrading, by variable) is to provide an overview of recent water quality changes over a domain of interest (e.g., the entire country, a region, an environment class).

In presenting these types of aggregate summaries, it has been typical to tabulate the number or proportion of sites for which trends are ‘established with confidence’ at a specified level of confidence (generally 95%) and to define the direction of these trends as increasing or decreasing (or improving/degrading). Typically, these tabulations also include the number or proportion of sites for which there are “insufficient data” to determine trend direction with confidence. This categorisation produces two problems. First, the trends for which there are insufficient data can be misinterpreted as “no change” or “stable”. This is an incorrect inference; insufficient data simply indicates a lack of confidence in the analysis at the nominated level. Second, trends with insufficient data to confidently determine direction nonetheless contain information about the likely direction of change that is effectively ignored by these tabulations. An extreme but plausible outcome is a situation in which, over many sites, no trend direction is established with confidence, but all trends are in the same direction at a lower level of confidence. The tabulation would show that all trends have insufficient data, implying that “nothing is known” about the aggregate trend direction. However, it is likely there is a general trend (i.e., the group of sites as a whole exhibit a trend).

Some studies have sought to overcome these problems by ignoring the levels of confidence and considering trends based on the sign of the evaluated trend. This approach is justifiable because over many sites, incorrect trend evaluations will tend to cancel each other out (e.g., as many sites will be misclassified as increasing as sites misclassified as decreasing). Thus, ‘count-based’ assessments of the number of trends in a given direction for a domain of interest simply count the number of individual trends for which the sign of the evaluated trend is in the direction of interest, disregarding the level of confidence in the trend directions. However, such assessments are subject to unquantified uncertainty, because the individual trends are always an uncertain estimate of the true trend.

This study developed an approach to quantifying the uncertainty of an assessment of the proportion of sites for which water quality was improving (or its complement, the proportion of sites that were degrading) based on aggregating site trends. The analysis uses the probability that the true direction of an individual site’s trend indicates improvement, which is evaluated as part of the analysis of each individual site trend. The approach takes trend assessments for multiple sites that represent a domain of interest (e.g., the entire country, a region or a class). The proportion of the individual site trends for which the probability of improvement is greater than degradation is referred to as the proportion of improving trends (PIT). The probabilities of improvement for the individual site trends are used to construct confidence intervals about the estimate of PIT.

The approach was applied to a case study of river water quality trends derived from a national dataset assembled by Larned *et al.* (2015). Site trends were analysed for six water quality variables for the same ten-year time-period assessed by Larned *et al.* (2015): the 10 years ending 2013. The method used to evaluate the individual trends in this study differs in the way censored values are handled to the method used by Larned *et al.* (2015). This change allows a larger number of site trends to be analysed because it was unnecessary to exclude sites with more than 15% of observations being censored values (as was done in Larned *et al.* 2015).

Table S1 shows the results of the analyses. The most important contrast is between the proportion of improving trends (i.e., PIT) and the proportion of the trends whose direction was established with confidence (at the 95% confidence level) that were improving. The second assessment represents conclusions that are likely to be made from a table of results comprising trends that are categorised as improving, degrading and insufficient data. The potential interpretation is to ignore the trends with insufficient data and assess the overall trend considering only the improving trends established with confidence as a proportion of only the trends for which direction was established with confidence. For example, Table S1 indicates that 83%, 90% and 24% of the trends that were established with confidence for NH₄N, TP and MCI were improving. However, the PIT statistics for NH₄N, TP and MCI were 59%, 75% and 40% respectively (Table S1). In addition, the results based on just the trends whose direction was established with confidence were generally outside the 95% confidence intervals of the PIT statistics (Table S1). This indicates that through not including information provided by all the site trends, the traditional approach gives a misguided impression of the proportion of improving sites. The PIT statistic distils the information contained in all the individual trends into a single number (plus its uncertainty), which provides a more robust evaluation of the general (aggregate) trend direction because it uses all the available information. As well as providing a single easily understood statistic (the proportion of improving trends, or its complement), PIT avoids referring to trends with insufficient data and the potential misinterpretation as “no change” or “stable”.

The PIT statistic also has the benefit that it is associated with confidence intervals. For example, Table S1 shows that >50% of sites were improving for CLAR, NH₄N, TP, DRP, and ECOLI. The lower 95% confidence interval was >50% for CLAR, NH₄N, TP and DRP, indicating that there is high confidence that the majority of sites had improving trends for these variables over the 10-year period.

The count-based estimates (i.e., counting all trends based on the sign of the evaluated trend and disregarding the confidence in the trend direction) were always within the 95% confidence interval for PIT (Table S1). This indicates that these count-based assessments are a reasonable approximation of the proportion of improving site trends. However, it should be kept in mind that the count-based assessment is subject to unquantified uncertainty.

PIT statistics for domains of interest (e.g., nationally, regionally or by classes) enables robust identification of spatial patterns in water quality changes that are difficult to perceive by examining the individual site trends. For example, as well as establishing with high confidence (i.e. the 95% confidence interval does not contain 50%) that trends for CLAR, NH₄N, TP and DRP were improving at >50% of sites over the 10-year period, our analyses established with high confidence that NO₃N trends were degrading at >50% of sites over the 10-year period. Furthermore, PIT statistics derived for regional domains indicated that NO₃N was degrading at >50% of sites in six regions for the 10-year period: Waikato, Tasman, Canterbury, West Coast, Otago and Southland. These and other patterns we identified elucidate general water quality changes and provide insights that are important for making robust inferences from trend analyses.

We recommend that the PIT statistic is used in future to represent aggregate measures of water quality change over spatial domains of interest. We also recommend that PIT statistics for a specified spatial domain are presented as distinct from the trend evaluations for individual sites, for which certainty in trend direction (or significance) remains an important piece of information.

Table S1: Estimates of proportions of sites with improving trends (PIT) for the 10-year time-period (national grouping). The PIT statistic and its 95% confidence intervals were derived from the probabilities that the true directions of the individual site trend indicated improvement. The count-based proportion of improving sites was evaluated by counting the number of individual trends for which the sign of the evaluated trend is in the direction of interest, disregarding the level of confidence in the trend directions. The proportions of trends with insufficient data were based on counting the individual trends for which confidence in direction was less than 95%. The proportion improving is based on counting the individual improving trends for which confidence was 95% or more. The proportion of trends with direction established with confidence that were improving is the sum of the number of individual improving trends divided by the total number of trends that were established with confidence (at the 95% level). See Table 2 in the main report for an explanation of the water quality variables.

Water quality variable	Number of sites	PIT (%)	95% confidence interval for PIT (%)	Count-based proportion improving (%)	Proportion with insufficient data (%)	Proportion improving (%)	Proportion of trends with direction established with confidence that were improving (%)
CLAR	393	58	55 - 61	58	58	28	67
NH4N	488	64	60 - 67	63	77	19	83
TN	274	49	46 - 53	49	54	25	54
NO3N	524	43	40 - 46	43	50	18	36
TP	486	79	76 - 81	78	49	46	90
DRP	520	72	69 - 74	72	49	39	76
ECOLI	495	50	46 - 53	49	82	10	56
MCI	462	41	36 - 45	40	79	5	24

1 Introduction

Long term water quality data that are collected at regular intervals (e.g., monthly) at monitoring sites are regularly analysed to assess the direction and magnitude of trends (e.g., Larned *et al.*, 2004, 2016). Trend analyses performed on many sites are regularly aggregated by water quality variable and presented in tabular or graphical form as part of environmental reporting (e.g., Ministry for the Environment, 2015, 2017). The aggregated water quality trends are intended to provide an overview of recent water quality changes over a spatial domain of interest (e.g., the entire country, a region, an environment class). Aggregated trends, for example expressed as proportions of site trends in different trend-direction categories, are intended to represent the recent progress toward or away from environmental objectives for the spatial domain.

Environmental reports tend to tabulate the numbers or proportions of site trends in three categories: increasing, decreasing, and insufficient data to confidently determine direction (“insufficient data”). When tabulating site trends by category, it has been usual to adopt a default alpha value (generally 0.05) to define trends for which direction is established with confidence. This generally means that the insufficient data category can make up a substantial proportion of the sites. This type of tabulation has two important problems. First, the insufficient data category can be misinterpreted as “no change” or “stable”. This is an incorrect inference; the insufficient data outcome simply indicates a lack of confidence in the analysis at the level defined by alpha. The second problem is that trends categorised as insufficient data contain information about the general direction of change that is effectively ignored. For example, a trend’s direction may not be established with confidence at the 95% level but may be established with an 80% level of confidence. An extreme but plausible outcome of these tabulations is a situation in which, over many sites, no trend is established with confidence at the default value of alpha, but all trends are in the same direction at a lower level of confidence. The tabulation would show that all trends are in the insufficient data category, implying that nothing is known about the aggregate trend direction. However, it is likely there is a general trend (i.e., the group of sites as a whole exhibit a trend).

The purpose of this study was to develop new methods for aggregating site trends representing a spatial domain of interest that incorporate all available information and that quantifies the uncertainty of the aggregate statistic. The methods use information produced by a recently adopted modification to trend evaluation that replaces the traditional test of statistical significance with a quantification of the level of certainty in the direction (increasing or decreasing) of the evaluated trend (Larned *et al.*, 2015, 2016; McBride, 2018). The new trend evaluation procedure treats confidence in the trend direction as a probability (i.e., a continuous quantity between zero and one) instead of the traditional binary ‘trend’, ‘no-trend’ interpretation. The new trend aggregation methods reduce the risk of misinterpreting insignificant or insufficient data trends and make maximal use of the available information. This report describes the new aggregation methods and provides a case study of their application to a national dataset of river water data that was first reported by Larned *et al.* (2015).

2 Trend analysis methods

2.1 Background to new trend assessment procedure

Water quality trends are commonly evaluated by fitting a regression to the relationship between the water quality variable (e.g., chemical concentration) and time, using the non-parametric Sen slope estimator (Hirsch *et al.*, 1982; Sen, 1968). The Sen slope estimator is a non-parametric statistic, removing the need to make assumptions about the distribution of the observations. The method is also robust to missing data, which is also a common feature of water quality data (Hirsch *et al.*, 1982). The Sen slope is an estimate of the rate of change in the central tendency of the water quality variable through the time-period.

Evaluations of water quality trends at individual sites are always uncertain. The level of uncertainty depends on the number of observations and the magnitude of the water quality change through the time-period being analysed. Traditionally, a statistical significance test is undertaken that evaluates the uncertainty of the trend by considering if it could have been observed if the true trend were exactly zero (Hirsch *et al.*, 1982). An insignificant test indicates that the observed trend could have been observed by chance (at a defined level of significance---typically denoted by alpha and generally set to 0.05) if the true trend was zero.

Recently, the logic underlying this significance test has been questioned and a new trend assessment procedure has been adopted. The new method posits that there is always a trend, no matter how small, but the ability to confidently infer its direction depends on the power of the statistical analysis. The method evaluates the $100 - \alpha^1$ confidence interval for the estimated trend magnitude (Larned *et al.* 2015, 2016). Briefly, if a symmetric confidence interval around the trend magnitude does not contain zero, then the trend direction (either positive or negative) is “established with confidence”. If it does contain zero, it is concluded that the trend has “insufficient data to confidently determine direction”².

Irrespective of whether the traditional or new confidence-based trend evaluation procedures are used, there are three possible outcomes. Trends are evaluated as increasing, decreasing or insignificant (under the traditional procedure) or insufficient data (under the new procedure). The division of trends into those that are increasing or decreasing and those that are insignificant or insufficient data depends on the value of alpha that is chosen. For a single assessment (i.e., a trend in one water quality variable at one site), the value of alpha should reflect management risks associated with either incorrectly inferring no trend when there is one, or the reverse, (i.e., type 1 and type 2 error rates). However, in practice trends are usually reported by adopting a default alpha value (typically 0.05).

2.2 Aggregation of trend analysis from many sites

2.2.1 Traditional approach

Trend analyses performed on many sites are regularly aggregated by water quality variable and presented in tabular or graphical form in state-of-environment reports as part of environmental reporting (e.g., Ministry for the Environment, 2015, 2017). These tabulations are intended to provide an overview of recent water quality changes over a spatial domain of interest (e.g., the entire country, a region, an environment class).

¹ The symbol α (alpha) represents the tolerance of making an incorrect determination as a probability. So $\alpha = 0.05$ indicates a tolerance of incorrect determinations in 5% of cases.

² It is noted that a $100(1-2\alpha)\%$ two-sided (symmetrical) CI is used in the procedure to define the $100 - \alpha$ level of confidence (see Larned *et al.*, 2016 for details).

It has been common practice when tabulating the numbers or proportions of site trends to present results in three categories: increasing, decreasing, and statistically insignificant (traditional method) or insufficient data (new method). The insignificant or insufficient data category has been defined by adopting a default alpha value (generally 0.05) leading to a substantial proportion of the sites being categorised as being insignificant or having insufficient data and consequently, the two problems outlined in the introduction.

When aggregating trends across many sites, some studies have chosen to accept the trend direction at the face value of the evaluated trend slope (i.e., accept the direction indicated by the estimated Sen slope irrespective of the statistical significance or confidence in the evaluation e.g., Ballantine *et al.*, 2010; Scarsbrook *et al.*, 2003). This approach is justifiable because over many sites, incorrect classifications of direction will cancel each other out (i.e., as many sites will be misclassified as increasing as sites misclassified as decreasing). Thus, 'count-based' assessments of the number of trends in a given direction for a domain of interest are made by simply counting the number of individual trends for which the sign of the evaluated trend is in the direction of interest, disregarding the level of confidence in the trend directions. However, because the evaluated trend at any given site is always an uncertain estimate of the true trend, count based assessments are subject to unquantified uncertainty. For example, if the proportion of improving trends is the statistic being derived, the estimated proportion is uncertain.

2.2.2 Graphical presentation of aggregated trends

The new trend assessment procedure enables the uncertainty associated with individual site trends to be incorporated in any analysis that aggregates trends over many sites. The basis for this is the evaluation of the probability that the true trend (i.e., the trend in the population from which the samples were drawn) was decreasing (hereafter 'probability the trend was decreasing', see details of how this is assessed in S4.2). Note that trend direction is arbitrary and the probability that the true trend was increasing is one minus the probability that it was decreasing. It follows that for any individual site trend, the direction is a Bernoulli distributed variable where the probability of "success" (a decreasing trend) is defined by the evaluated probability. Thus, a trend with an evaluated probability >0.5 indicates success (a decreasing trend) and conversely the probability of "failure" (an increasing trend) is <0.5 .

The probability that the true trend was decreasing facilitates a more nuanced inference rather than the 'yes/no' output corresponding to the chosen acceptable misclassification error rate (McBride, 2018). Confidence categories can be used to express probability that the trend direction is improving (or its complement; degrading). Note that the conversion of the probability that a trend is decreasing to the probability it is improving (and its complement, degrading) depends on whether decreasing values represent improvement or degradation and differs between variables.

The approach to presenting levels of confidence of the Intergovernmental Panel on Climate Change (IPCC; (Stocker *et al.*, 2014) is one way of categorising confidence that trends are improving (Table 1). Note that descriptions of the probabilities of degrading trends are the complement of the categorical levels of confidence in Table 1, i.e. an "exceptionally unlikely" degrading trend is the same as a "virtually certain" improving trend.

Table 1. Level of confidence categories used to convey the probability that water quality was improving. The confidence categories are those used by the Intergovernmental Panel on Climate Change (IPCC; 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

The aggregate proportion of sites in each category shown in Table 1 can be calculated for sites grouped by some spatial domain of interest, and for each variable. The values can then be plotted as colour coded bar charts. These charts provide a graphical representation of the proportions of improving and degrading trends at the levels of confidence indicated by the categories.

2.2.3 Evaluation of the proportion of improving trends

The trends, evaluated at several monitoring sites for a given variable over some domain of interest, can be assumed to represent independent samples of the population of trends, at all sites within that domain. Let the sampled sites within this domain be indexed by s , so that $s \in \{1, \dots, S\}$ and let I be a random Bernoulli distributed variable which takes the value 1 with probability p and the value 0 with probability $q = 1 - p$. Therefore, $I_s = 1$ denotes an improving trend at site $s \in \{1, \dots, S\}$ when the estimated $p_s \geq 0.5$ and a degrading trend as 0 when $p_s < 0.5$. Then, the estimated proportion of sites with improving trends in the domain is:

$$PIT = \sum_{s=1}^{s=S} I_s / S$$

Because the variance of a random Bernoulli distributed variable is $Var(I) = p(1 - p)$, and assuming the site trends are independent, the estimated variance of PIT is:

$$Var(PIT) = \frac{1}{S^2} \sum_{s=1}^{s=S} Var(I_s) = \frac{1}{S^2} \sum_{s=1}^{s=S} p_s(1 - p_s)$$

PIT and its variance represent an estimate of the population proportion of improving trends and the uncertainty of that estimate. It is noted that the proportion of degrading trends is the complement of the result (i.e., $1 - PIT$). The estimated variance of PIT can be used to construct 95% confidence intervals³ around the PIT statistics as follows:

$$CI_{95} = PIT \pm 1.96 \times \sqrt{Var(PIT)}$$

³ Note that ± 1.96 are approximately the 2.5th and 97.5th percentile of a standard normal distribution.

3 Case study

River monitoring data collected by regional councils and NIWA (national river water quality monitoring network; NRWQN) river are periodically acquired and federated into databases for national-scale state-of-environment reports (e.g., Ballantine *et al.*, 2010; Larned and Unwin, 2012). The most recent national scale assessment of river water quality was by Larned *et al.* (2015) who updated databases to the end of 2013 and produced a report on national river and lake state and trends.

This case study used the same river water quality database used by Larned *et al.* (2015). These data comprised 844 sites, representing the 77 NRWQN sites plus 767 regional council state-of-environment river monitoring sites. Several data grooming processes were undertaken to ensure the database was consistent and comparable across sites (see Larned *et al.*, 2015 for details). The final database comprised 653,351 observations of the eight variables shown in Table 2. Each observation was associated with a value, date and an observed or modelled flow at the time of sampling. Each site was associated with meta data, including the geographic location and the unique identification of the segment of the digital river network on which the site was located.

Table 2. River water quality variables, measurement units and site numbers used in this study.

Variable type	Variable	Abbreviation	Units	Number of monitoring sites
Physical	Visual clarity	CLAR	m	454
Chemical	Ammoniacal nitrogen	NH4N	mg/m ³	364
	Nitrate-nitrogen	NO3N	mg/m ³	586
	Total nitrogen (unfiltered)	TN	mg/m ³	354
	Dissolved reactive phosphorus	DRP	mg/m ³	518
	Total phosphorus (unfiltered)	TP	mg/m ³	576
Microbiological	<i>Escherichia coli</i>	ECOLI	cfu/100 mL	485
Biotic Index	Macroinvertebrate Community Index	MCI	unitless	505

Site and variable combinations in the database represented different monitoring period starting and ending dates, numbers of observations and sampling frequencies (see Larned *et al.*, 2015 for details). All variables apart from MCI were associated with censored values and the proportion of censored values was highest for NH4N and DRP (Figure 1).

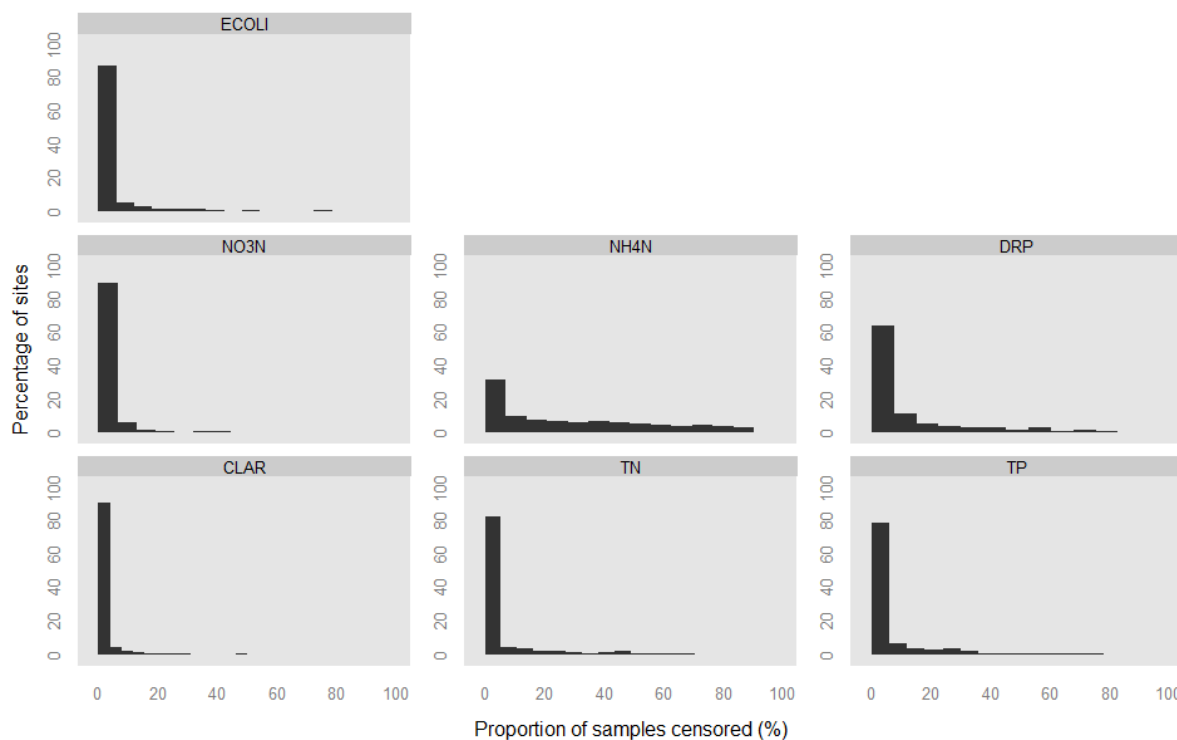


Figure 1. Proportion of censored samples in the dataset by variable.

4 Trend analyses

4.1 Sampling dates and time-periods for analyses

Trend assessments are specific for a given period of analysis. In this study, trends were characterised for the same 10-year time-period assessed by Larned et al. (2015): the 10 years up to the end of 2013.

The dataset had variable starting and ending dates, variable sampling frequencies, and variable numbers of missing values. Filtering rules were used to achieve a reasonable degree of data-representativeness for all site trends that were analysed. We used the same filtering rules as Larned et al. (2015), which restricted site and variable combinations to those for which there were measurements for at least 90% of the years and at least 90% of seasons within the 10-year time-period of analysis. These are more stringent rules than those suggested by Helsel and Hirsch (1992).

We assessed trends for the water quality variables using seasons defined by months preferentially, and quarters when monthly data were not available, provided the filtering rules were met. Because MCI is generally sampled annually, analysis of these trends does not involve seasons. For some sites and variables there were more than one sample within some seasons or years (for MCI). In these cases, we used the median of the values for the season (or year for MCI) to be consistent. We note that when there is more than one sample in a season, the individual within season samples can be used in a trend analysis resulting in increased statistical power and potentially different results. All site by variable combinations that did not comply with these filtering rules were excluded from the analysis.

4.2 Statistical analyses

A simple diagrammatic explanation of the method used for statistical trend analyses is shown in Figure 2. The basis for the method is the Sen slope estimator (SSE), which is the median of all possible inter-observation slopes (i.e., the difference in the measured observations divided by the time between sample dates). Consider 5 years of monthly observations (i.e., $n=60$). There are $(60 \times 59)/2 = 1770$ possible inter-observation slopes. These inter-observation slopes are ranked from the smallest to largest and the Sen slope is the average of two inter-observation slopes with ranks 885 and 886 (i.e., the median of all 1770 inter-observation slopes). The seasonal version of the SSE is used in situations where there are significant differences in water quality measurements between 'seasons'. Seasons are defined primarily by the sampling frequency. In New Zealand, it is common to sample either monthly or quarterly, and in these cases, seasons are defined by months or quarters. The seasonal Sen slope estimator (SSSE) is the median of all inter-observation slopes within each season. Consider monthly data for 5 years of record. All possible inter-observation slopes between data pertaining to January are calculated (10 in number). This is then repeated for all other months giving 120 inter-observation slopes. The SSSE is the average of the two inter-observation slopes with ranks 60 and 61 (i.e., the median of all 120 slopes). The SSE and SSSE values express trends in units of change in the variable per year.

In traditional water quality trend analysis, SSE and SSSE values were accompanied by a statistical test of significance developed by (Hirsch *et al.*, 1982). The statistical test was Kendall's test of rank correlation, which is a nonparametric correlation coefficient measuring the monotonic association between y and x . In water quality trend analysis, y is a sample of water quality measurements and x is the corresponding sample dates. However, the trend direction assessment procedure developed by (Larned *et al.*, 2015) does not use a Kendall test to evaluate the statistical confidence in the trend direction. Rather, confidence intervals (defined based on a nominated alpha value) are interpolated from the ranked inter-observation slopes (McBride, 2018). The confidence intervals can be used to make inferences about trend direction; if a confidence interval around the trend (i.e., the SSE) does not contain zero, then the trend direction (either positive or negative) is "established with confidence" (Larned *et al.*, 2015). If it does contain zero, it is concluded that there is "insufficient data" to determine the trend direction at a given level of confidence.

Confidence intervals are determined by first expressing the ranks of the slopes as quantiles of the standard normal distribution (Z-scores). The probabilities of observing those Z-scores are then calculated using the normal density function (Figure 2). The slopes and associated non-exceedance probabilities can be used to: (1) evaluate the Sen slope, by interpolating the slope at which the non-exceedance percentile = 0.5; (2) evaluate the probability that the true trend is decreasing, by interpolating the non-exceedance probability at which the inter-observation slope is equal to zero (note that direction here is arbitrary and the probability that the true trend was increasing is one minus the probability that it was decreasing); and (3) determine the confidence interval for the Sen slope, by interpolating the slopes at which the non-exceedance percentiles are α and $1 - \alpha$. In this study we have a nominated an alpha value of 0.05, to be consistent with Larned *et al.* (2015).

When the precision of the measured observations is low, there will be many observations with the same value leading to many ties (i.e., inter-observation slopes of exactly zero). This results in a high chance of obtaining an upper or lower confidence interval of exactly zero. The interpretation of confidence in trend direction when a confidence interval bound of exactly zero is equivocal. However, an overarching assumption of the new approach is that there always are differences between observations (leading to the assumption that the trend can never be

zero; McBride, 2018). It follows that an inter-observation slope evaluated as zero, is in fact either an increase or decrease but with a magnitude that cannot be established due to the low precision of the variable being measured. To avoid equivocal assessments of confidence in trend direction, we use the probability that the trend is decreasing. When this probability is interpolated from the slopes and their non-exceedance probabilities, we assume that inter-observation slopes of zero are equally likely to be increasing as decreasing. Therefore, in the case that the probability the inter-observation slope of zero has a non-unique solution, the mean of all the probabilities (associated with an inter-observation slopes of zero) is used. If this probability is <0.05 and >0.95 then we can conclude, with confidence, that the trend is increasing or decreasing respectively.

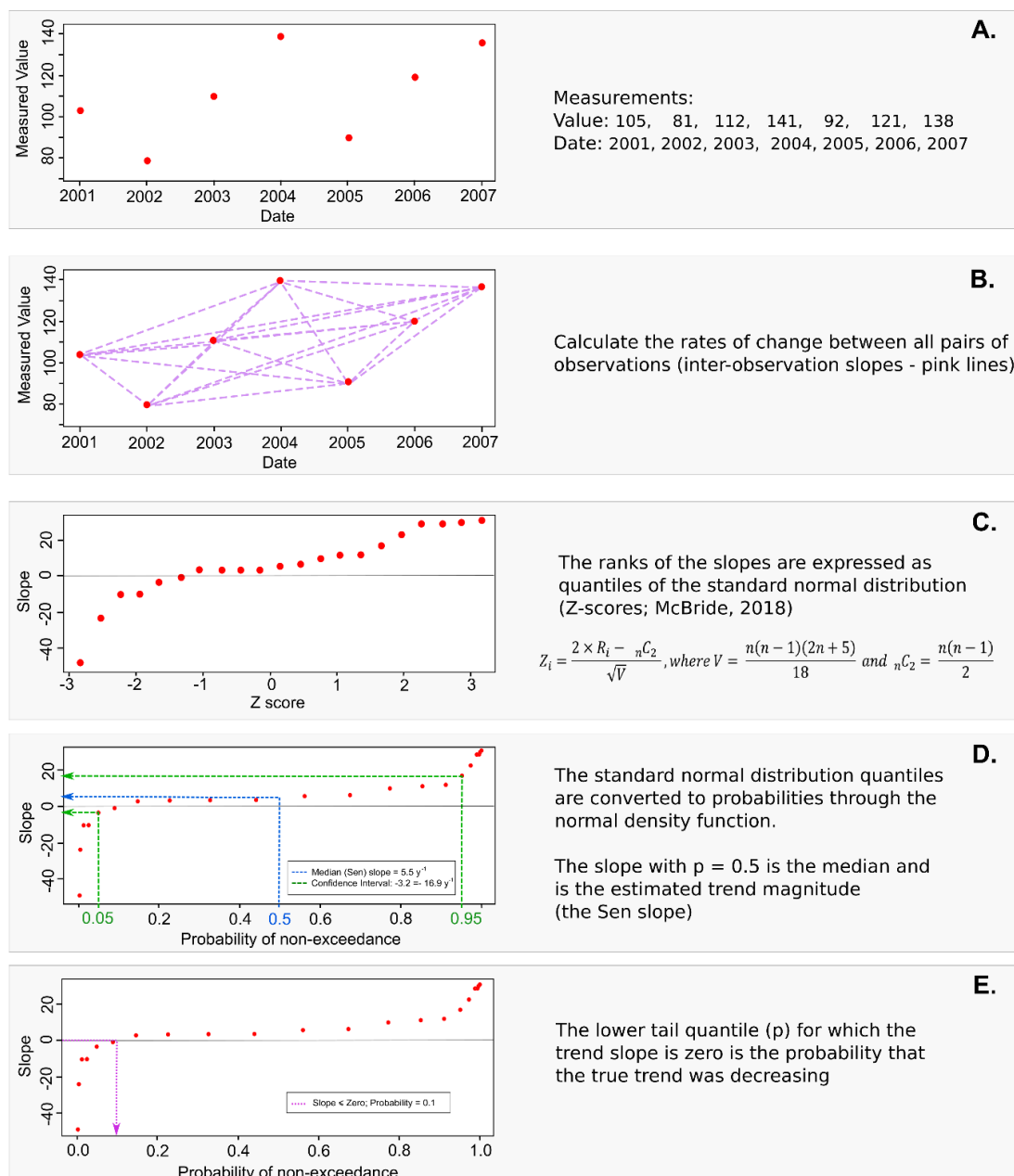


Figure 2: Pictogram of the steps taken in trend analysis to calculate the Sen slope and the probability that the true trend was decreasing.

4.3 Censored values

Censored values are those above or below a detection limit (e.g., >2.5 or <0.001). Values above the detection limit are described as right censored and values below the detection level are described as left censored. 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 $0.5 \times$ detection limit and $1.1 \times$ 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).

The previous method of trend analysis (i.e., 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.

The methods used in this study were based on robust handling of censored values in trend analysis (Helsel, 2012). Key calculations that are affected by censored values are the calculation of Kendall's S and its variance⁴, the estimation of the Sen slope (including the seasonal Sen slope) and the estimation of the confidence intervals for Sen slopes. For left-censored data, increases and decreases in a water quality variable are measured whenever possible. Thus, a change from <1 to 10 is an increase. A change from a <1 to a detected 0.5 is considered a tie, as is a <1 to a <5 , because neither can definitively be called an increase or decrease. Similar logic applies to right censored values. The variance of the S statistic is adjusted for ties (it is reduced) and this influences the computation of confidence intervals.

The slope between any combination of observations in which either one or both are censored cannot be definitively calculated. The slopes associated with censored values are therefore ignored (i.e., removed) and SSE and SSSE are calculated as the median of all real valued slopes between sample dates. The removal of slopes associated with censored values has the effect of decreasing the number of samples used to determine the SSE and SSSE, therefore reducing statistical power and increasing the width of the confidence interval. This means that when there are many censored values, the analysis produces a low degree of confidence in the evaluated trend direction. Where there are fewer than five total and three unique, non-censored observations (but when the other filtering criteria are otherwise met), the method will not analyse the data and these cases are reported as "not analysed" (see Section 3.4).

4.4 Seasonality

When there is seasonal variation in the observations, the seasonal Sen slope estimator (SSSE) should be used (Hirsch *et al.*, 1982). Larned *et al.* (2015) evaluated all trends using the SSSE, however, the seasonal estimator has lower statistical power than the non-seasonal estimator (due to smaller sample sizes). It is therefore advantageous to establish whether the water quality observations are seasonally varying and if this is not the case, to use the more

⁴ Note that although neither the previous or new trend assessment methods use Kendall's rank test of correlation to test the significance of trends, both methods use S and the variance of S to compute the confidence intervals for SSE and SSSE.

powerful SSE to evaluate the trend. The new method of trend analysis commences by testing for the effect of season (i.e., month or quarter) on each site and variable combination using a Kruskal Wallis test. The null hypothesis tested is that observations belonging to all seasons (month or quarters) come from the same population. If there is evidence to reject this ($p \leq 0.05$) a statistically significant effect of season on the value of a variable is inferred, and the SSSE is evaluated, otherwise the non-seasonal SSE is evaluated.

4.5 Flow adjustment

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.*, 2016).

Adjusting the observations to account for the effect of flow (flow adjustment, or any other 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). Flow adjustment 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. Of a total of 785 sites for which we had some water quality data, 547 had no flow information provided. Where flow measurements were available, we used these. Flow measurements were available for all locations with strongly anthropogenically-modified flows, e.g., downstream of hydropower stations, where flows would be otherwise difficult to estimate. Where flow measurements were not available, we used flows estimated using a national hydrological model (TopNet) as described by Larned *et al.* (2015).

In this study we followed the conclusions and recommendations of Snelder (2018) concerning flow adjustment of water quality variables. In particular, we did not rely on the automated flow adjustment procedure used by Larned *et al.* (2015) because unsupervised fitting of regression models to flow versus concentration relationships can result in the selection of unreliable models. We used both generalised additive models (GAM) and locally weighted least squares regression (LOESS) models to fit flow-water quality variable models. We inspected the models and used expert judgement to choose the most suitable model based on the homoscedasticity (constant variance) of the regression residuals and plausibility of the shape of the fitted model. Where there was little difference among models, we used the GAM model to maintain consistency with Larned *et al.* (2015).

4.6 Categorisation of trends

The analyses returned site trend outputs for each site and variable combination and these were classified into four direction categories: improving, degrading, insufficient data and not analysed. An increasing or decreasing trend category was assigned when the 90% confidence interval did not contain zero (i.e. when probability $\geq 95\%$) and the Sen slope was positive or negative, respectively (i.e., the trend direction is established with confidence; Larned *et al.*,

2016). An insufficient data trend category was assigned when the 90% confidence interval contained zero (i.e. when probability $\leq 95\%$; the trend direction was not defined with confidence; Larned *et al.*, 2016). Trends were classified as “not analysed” for two reasons:

- 1) When a large proportion of the values were censored (data has <5 non-censored values and/or <3 unique non-censored values). 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.
- 2) When there is no, or very little, variation in the data because this also results in ties. This can occur because 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 resulting in ties.

4.7 Aggregation of trends

4.7.1 Graphical presentation of aggregated trends

The categorical levels of confidence presented in Table 1 were used to express the likelihood that water quality was improving for each site and variable. Each site trend was assigned a categorical level of confidence that the trend was improving according to its evaluated probability and the categories shown in Table 1. For the chemical and microbiological water quality measures (Table 2), improvement is indicated by decreasing trends (i.e. decreasing concentrations). For MCI and CLAR improvement is indicated by increasing trends.

The aggregate proportion of sites in each category were then calculated for each variable and these values were shown as colour coded bar charts. These charts were produced using all available sites (i.e., national scale aggregation). It is noted that this type of chart can be produced for sites aggregated according to any grouping. Graphical presentations were not produced for other site groupings in this study because we considered that the probabilistic assessments of the proportions of improving trends were a simpler way to represent grouped aggregate trends.

4.7.2 Proportion of improving trends

The proportion of improving trends (PIT) and its uncertainty was evaluated for each water quality variable for site trends grouped in four ways. First, all available sites were grouped at the national scale. Two groupings were based on classes defined by the River Environment Classification (REC; Snelder and Biggs, 2002). The REC is a national classification system of rivers that has been frequently used as a basis for environmental reporting. The REC distinguishes rivers based on the dominant characteristics of their upstream catchments and classes tend to discriminate variation in water quality because this is largely driven by catchment character. The two REC groupings used in this study were:

1. The second (source-of-flow) level of the REC, which distinguishes between catchments based on differences in climate and topography. Source-of-flow classes are denoted by combination of categories that describe the climate and topography of the catchment (Snelder and Biggs, 2002; Table 3). For example, most river segments

on the south-eastern coast of New Zealand are categorized as Cool-Dry climate and Hill topography and, thus, belong to the CD/H class.

2. The REC land-cover category. Land-cover categories are denoted by the land cover type that dominates the catchment (Snelder and Biggs, 2002). For example, much of New Zealand's lowland catchments are dominated by pasture land-cover and mountainous areas are dominated by natural land cover types including native forest, scrub or bare ground.

We used modified source-of-flow classes and land-cover categories defined by Snelder and Biggs (2002), which merged those classes for which there were few monitoring sites into a smaller number of closely related groups (Table 3). The modification reduced the number of groups, which reduced discrimination of environmental variation, but increased the number of sites in each group, thus increasing statistical power.

Table 3. Modified river environment classification (REC) source-of-flow classes and land-cover categories used by this study. The original (Snelder and Biggs, 2002) classes and categories that were merged to form each modified class are shown. See Snelder and Biggs (2002) for a full description of REC classes.

REC class	Class description	Original classes
Source-of-flow level		
CX/H	Cool-extremely-wet hill	CX/H
CX/L	Cool-extremely-wet lowland	CX/L
CW/M	Cool-wet mountain and glacial-mountain, cool-extremely wet mountain and glacial mountains	CW/M, CX/GM, CX/M, CW/GM
CW/H	Cool-wet hill	CW/H
CW/L	Cool-wet and extremely-wet lowland	CW/L, CX/L
CW/Lk	Cool-wet and extremely-wet lake	CW/Lk, CX/Lk
CD/H	Cool-dry hill and mountain	CD/H, CD/M
CD/L	Cool-dry lowland and lake	CD/L, CD/Lk
WX/L	Warm extremely-wet lowland	WX/L
WW/H	Warm wet hill	WW/H,
WW/L	Warm wet lowland	WW/L,
WW/Lk	Warm wet lake	WW/Lk,
WD/L	Warm dry lowland	WD/L
Land-cover category		
EF	Exotic forest	EF
N	Indigenous forest. scrub, bare, wetland, tussock	EF, S, B, W, T
P	Pasture	P
U	Urban	U

The fourth grouping of sites was by region (Table 4). The data available in each region was generally collected by the regional council. The exceptions were sites administered by the

Nelson City Council which were grouped with those of Tasman District Council into the Tasman region and sites administered by Christchurch City Council, which were grouped with those from Environment Canterbury, resulting in 15 regions. In addition, NRWQN sites are administered by NIWA, and were assigned to the region in which the sites were located (Table 4).

Table 4. Regional groupings and data collecting agency.

Group	Region	Data collecting agency
N	Northland	Northland Regional Council and NIWA
A	Auckland	Auckland Council and NIWA
Wai	Waikato	Environment Waikato and NIWA
BOP	Bay of Plenty	Bay of Plenty Regional Council and NIWA
Tar	Taranaki	Taranaki Regional Council and NIWA
G	Gisborne	Gisborne District Council and NIWA
HB	Hawkes Bay	Hawkes Bay Regional Council and NIWA
MW	Manawatu-Wanganui	Horizons Regional Council and NIWA
Wel	Wellington	Greater Wellington Regional Council and NIWA
Tas	Tasman	Tasman District Council, Nelson City Council and NIWA
M	Marlborough	Marlborough District Council and NIWA
C	Canterbury	Environment Canterbury, Christchurch City Council and NIWA
WC	West Coast	West Coast Regional Council and NIWA
O	Otago	Otago Regional Council and NIWA
S	Southland	Environment Southland and NIWA

The PIT statistics and the 95% confidence intervals associated with these estimates were displayed either as tables or plots. The PIT statistics were compared to the proportion of improving trends for which trend direction was established with confidence and with the proportions of improving trends derived from count-based assessments of the trend directions (i.e. by counting all improving trends irrespective of confidence in direction). PIT results for any pair of groups (i.e. domains of interest) that had non-overlapping confidence intervals were conservatively (at $\alpha = 0.05$) interpreted as having statistically significant differences in the proportion of improving trends (Cumming *et al.*, 2007).

4.8 Implementation

All trend analyses presented in this report were undertaken with purpose written functions that implement the new trend assessment method using the R statistical computing environment (<http://www.r-project.org>) that are available here; <http://landwaterpeople.co.nz/pdf-reports/>. The new method of trend analysis has also been implemented in the TimeTrends software (Jowett, 2017), which is commonly used by regional councils in New Zealand and is available here: <http://www.jowettconsulting.co.nz>.

The assignment and plotting of categorical levels of confidence and calculation of the PIT statistics were undertaken using purpose written functions developed using the R statistical computing environment that are available here; <http://landwaterpeople.co.nz/pdf-reports/>.

5 Results

5.1 National scale

Figure 3 shows the proportion of all sites (i.e., nationally), by variable, for which 10-year water quality trends indicated improvement at the nine categorical levels of confidence defined in Table 1. Note that probability of improvement is the complement of probability of degradation, and therefore sites that are classified as “exceptionally unlikely” to be improving, could equally be classified as “virtually certain” to be degrading. The plot indicates that 50% or more of sites were at least likely to be improving (i.e., probability $\geq 67\%$) for CLAR, TP and DRP. The plot also indicates that 50% of sites had NO₃N trends that were at the most unlikely to be improving (i.e., indicating they were at least likely to be degrading). ECOLI and TN had roughly even proportions of sites in the unlikely and likely to be improving categories and MCI had more sites in the unlikely to be improving categories.

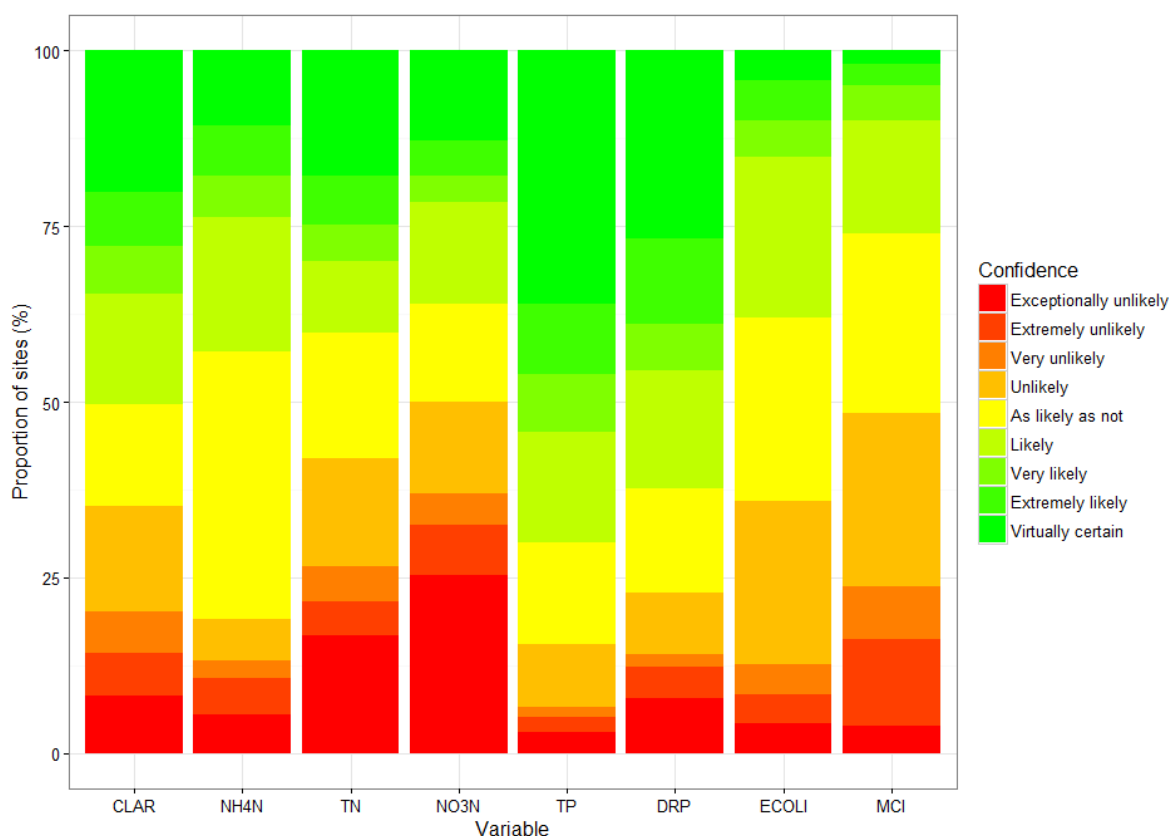


Figure 3. Summary plot representing the proportion of sites with improving 10-year time-period trends at each categorical level of confidence. The plot shows the proportion of sites with improving trends at levels of confidence defined in Table 1.

The PIT statistics produced different results to an assessment of the proportion of improving trends based only on trends that are established with confidence (Table 5). For example, 83%, 90% and 24% of trends established with confidence for NH₄N, TP and MCI were improving, but the PIT statistics indicated 64%, 79% and 41% of sites had improving trends respectively.

(Table 5). The assessments based on the proportion of trends established with confidence that were improving were generally outside the 95% confidence intervals of the PIT statistics (Table 5). This indicates that through not including information provided by all the site trends, the traditional approach gives a misguided impression of the proportion of improving sites.

The PIT statistics produced similar estimates to the count-based evaluation (Figure 4). The 95% confidence interval for PIT always included the one to one line of the plot comparing the two sets of evaluations for all variables (Figure 4, Table 5). This demonstrates that the count-based evaluation of the proportion of improving sites is consistent with the PIT statistics.

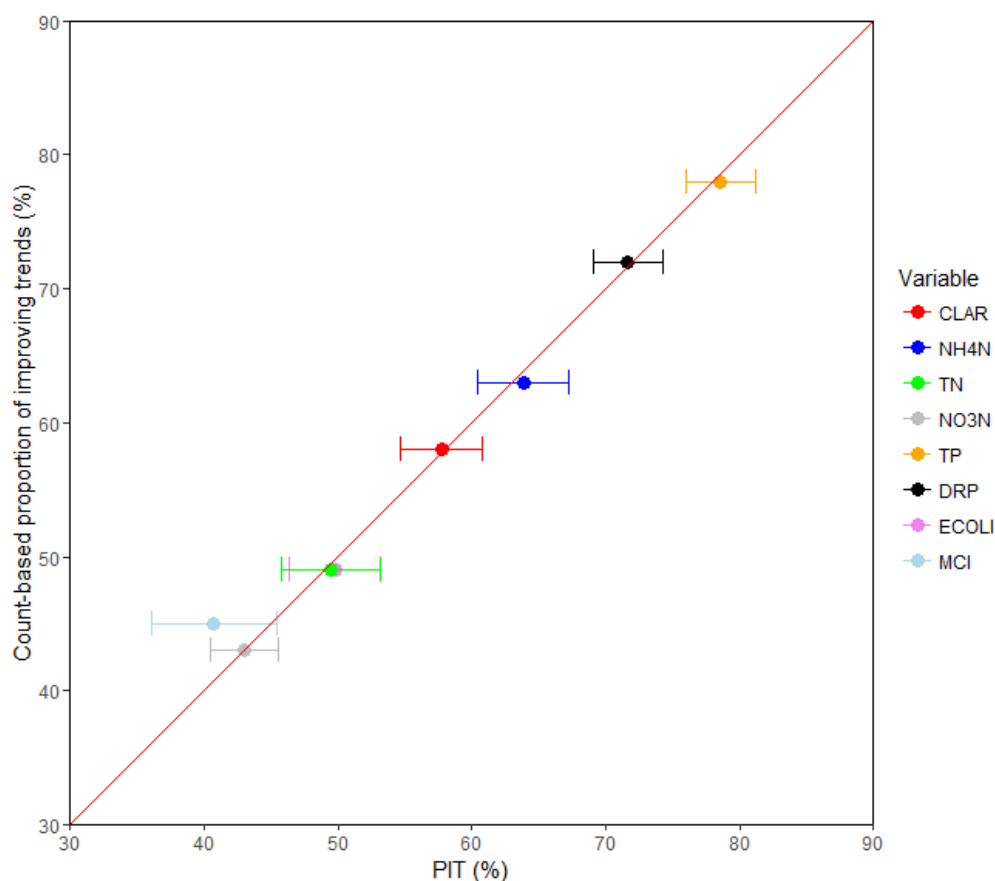


Figure 4. Comparisons of proportions of sites with improving trends over the 10-year time-period based on the PIT statistic and the count-based evaluation. Solid dots are the evaluated proportions of improving sites using both methods. Error bars indicate the 95% confidence interval for PIT.

The PIT statistic was 50% or greater for all variables except TN, NO3N and MCI (Table 5). The variables with the largest proportions of improving sites were TP (79%), DRP (72%), NH4N (64%) and CLAR (58%). NO3N and MCI were improving at only 43% and 41% of sites respectively, or conversely were degrading at 57% and 59% of sites respectively (Table 5). These results are consistent with the proportions of improving site trends at the different levels of confidence shown in Figure 3.

The confidence intervals for PIT were narrow, at between 5% and 9%. As expected, the narrower confidence intervals were associated with the variables for which there were more sites and for which a larger proportion of trends were established with confidence (e.g., TP, and DRP; Table 5, Figure 3).

The trend assessment method used in this study retained a larger number of sites for all variables compared to Larned *et al.* (2015), who filtered (removed) all sites with greater than 15% of censored observations (*Table 5*). The largest differences in the number of analysed sites between this study and Larned *et al.* (2015) were for NH₄N and DRP, which have the largest numbers of censored values (*Figure 1*). The retention of sites with greater than 15% censored samples resulted in a larger proportion of trends categorised as insufficient data by this study compared to Larned *et al.* (2015).

There were differences between the proportion of improving and degrading trends between the two studies (*Table 5*). For example, this study evaluated 19% of sites had improving trends for NH₄N compared to 43% for Larned *et al.* (2015) (*Table 5*). These differences are associated with differences in the numbers of sites included in the two assessments, which is related to the filtering of sites with greater than 15% censored values by Larned *et al.* (2015).

Table 5. Estimates of proportions of sites with improving 10-year site-trends (national grouping). The improving and degrading trends in this study and Larned et al (2015) refer to trend directions that are established at the 95% level of confidence. The proportion of trends established with confidence improving represents the ratio of improving trends to the total number of trends whose directions were established at the 95% level of confidence.

Variable	This study								Larned et al. (2015) results			
	Number of sites	Number of sites not analysed	PIT (%)	95% confidence interval for PIT	Insufficient data (%)	Improving (%)	Degrading (%)	Proportion trend directions established with confidence improving (%)	Number of sites	Insufficient data (%)	Improving (%)	Degrading (%)
CLAR	393	0	58	55 - 61	58	28	14	67	386	48	34	18
NH4N	487	36	64	60 - 67	77	19	11	83	206	31	43	26
TN	273	0	49	46 - 53	54	25	21	54	243	40	32	28
NO3N	523	0	43	40 - 46	50	18	33	36	511	39	24	37
TP	485	0	79	76 - 81	49	46	5	90	421	33	61	6
DRP	519	5	72	69 - 74	49	39	12	76	391	31	51	18
ECOLI	494	0	50	46 - 53	82	10	8	56	396	66	20	14
MCI	249	0	41	36 - 45	78	6	16	24	461	83	13	4

5.2 Source-of-flow

There were differences in the PIT statistics between some pairs of REC source-of-flow classes for all variables (i.e., non-overlapping 95% confidence intervals; Figure 5). Confidence that CLAR was improving at more than 50% of sites exceeded 95% in six REC source-of-flow classes (CX/H, CW/H, CW/L, CD/H, CD/L, WD/L). Confidence that CLAR was degrading at more than 50% of sites exceeded 95% in three REC source-of-flow classes (CW/Lk, WW/H and WW/L). The latter two classes are predominantly located in the upper half of the North Island. Confidence that NH₄N was improving at more than 50% of sites exceeded 95% for five source-of-flow classes (CD/H, CD/L, WW/L, WW/Lk and WD/L). Confidence that TP and DRP were improving at more than 50% of sites exceeded 95% in nine and eight source-of-flow classes. Confidence that TP was improving at more than 80% of sites exceeded 95% in the WW/L and WW/Lk classes. Confidence that TP and DRP were improving at more than 60% of sites exceeded 95% in seven and 5 classes respectively.

NO₃N had the lowest PIT statistics (i.e., the highest proportions of degrading trends), in particular in the CW/H, CW/Lk, CD/H, CD/L and WX/L source-of-flow classes. In these classes, degradation occurred at between 64% to 71% of sites. Patterns in TN were similar to those of NO₃N. Confidence that TN was degrading at more than 50% of sites exceeded 95% in CX/L, CW/H, CD/H, CD/L, WX/L and WW/H classes. There were no REC source-of-flow classes for which there was a 95% level of confidence that ECOLI was degrading and there was only one class (CW/H) for which there was 95% confidence that ECOLI was improving at more than 50% of sites. There were no REC source-of-flow classes for which there was a 95% level of confidence that MCI was improving at more than 50% of sites. In addition, there were three source-of-flow classes for which there was a 95% level of confidence that MCI was degrading at more than 50% of sites (CX/H, CW/M and WD/L, Figure 5).

The uncertainty of the estimated proportion of improving sites in each source-of-flow class was strongly related to the number of sites in the class (Figure 5). Uncertainties were generally larger when there were fewer sites in the class. Where there were few sites in a class but high certainty, the probabilities that individual sites trends were improving were high for most of the sites.

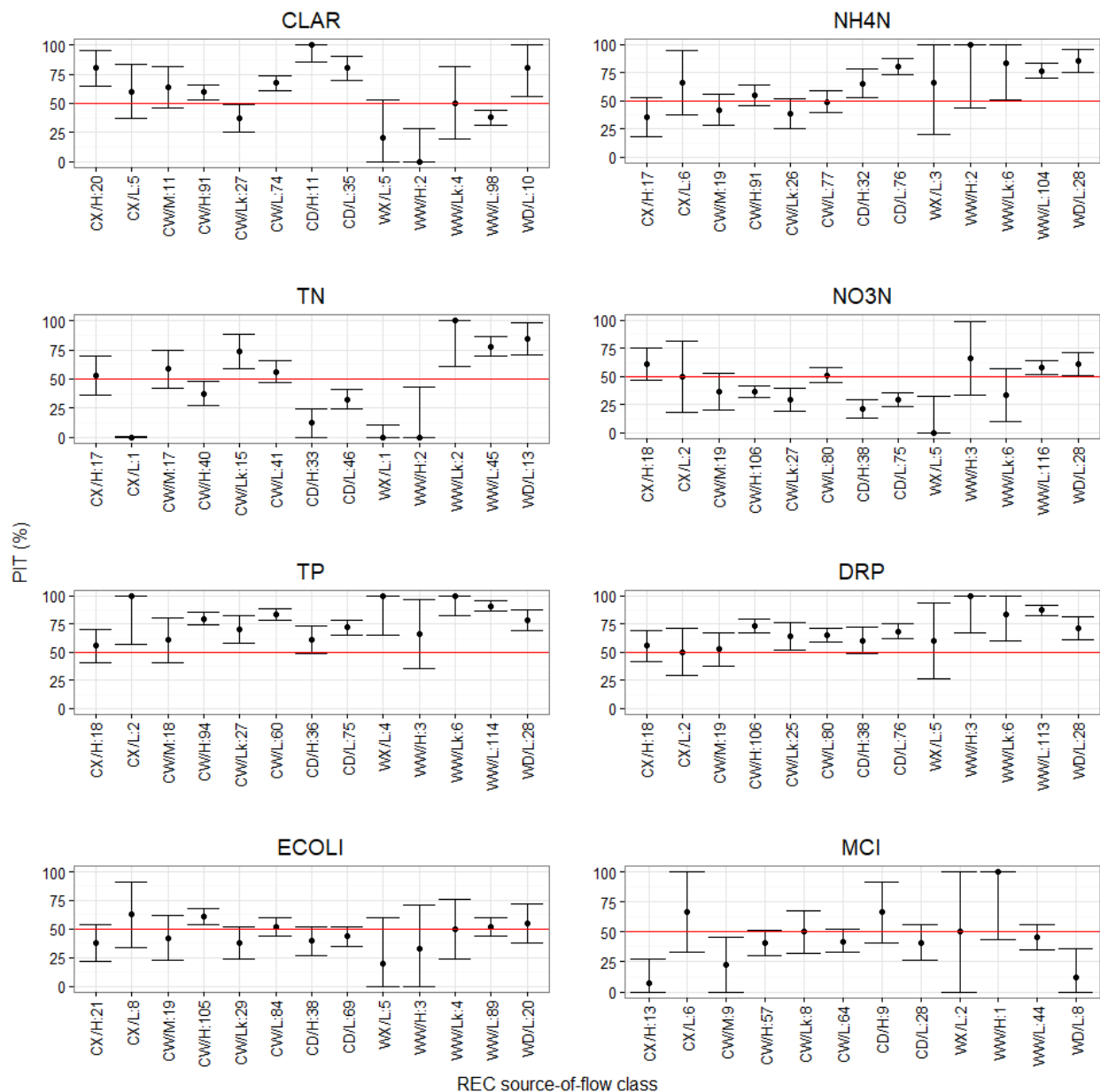


Figure 5. Evaluated PIT statistics for 10-year flow adjusted trends grouped by REC source-of-flow classes. The numbers after the source-of-flow class labels indicate the number of sites used in the evaluation. The error bars indicate the 95% confidence interval for the PIT statistic. The red line indicates 50% of sites with improving trends.

5.3 Land cover

There were differences in the PIT statistics between some pairs of REC land-cover categories for most variables (i.e., non-overlapping 95% confidence intervals; Figure 6). Confidence that CLAR was improving at more than 50% of sites exceeded 95% for three of the four land cover categories (N, P and U; Figure 6). Confidence that NH4N was improving at more than 50% of sites exceeded 95% for two land cover categories (P and U).

Confidence that ECOLI was improving or degrading at more than 50% of sites did not reach 95% for any land-cover category. Confidence that TP and DRP was improving at more than 50% of sites exceeded 95% for four and three land-cover categories respectively and were

improving at more than 76% and 73% of sites respectively in the P (Pasture) category. Confidence that NO₃N was degrading at more than 50% of sites exceeded 95% in the EF (Exotic Forestry) and P (Pasture) land-cover categories. In these categories, there was 95% confidence that degradation occurred at 60% and 56% of sites respectively. Confidence that TN was improving at more than 50% of sites exceeded 95% in only the U land-cover category. Confidence that MCI was degrading at more than 50% of sites exceeded 95% in all land-cover categories except P and there was 95% confidence that it was degrading at more than 58% of sites in the Natural (N) land cover category.

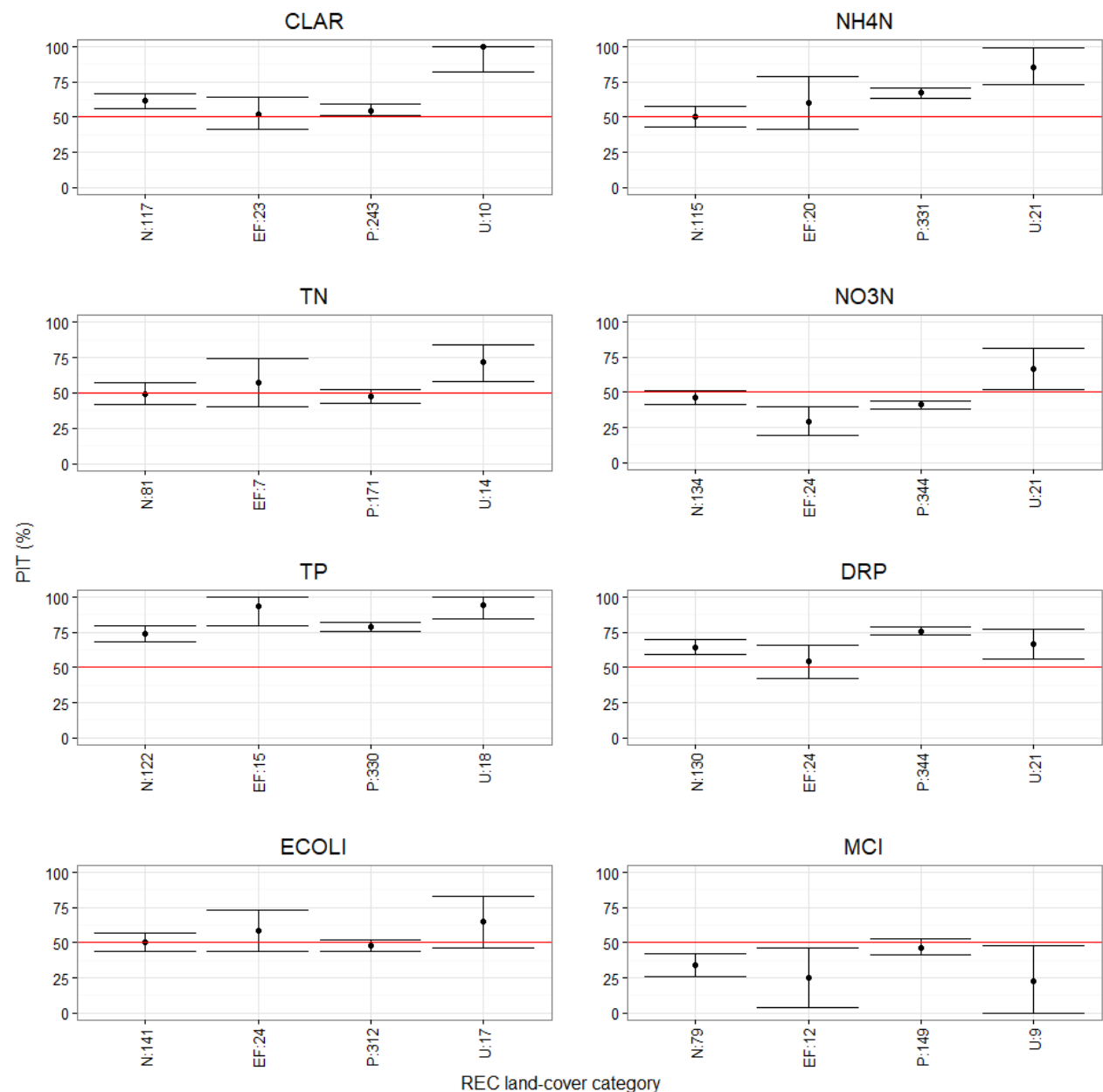


Figure 6. Evaluated PIT statistics for 10-year flow adjusted trends grouped by REC land-cover categories. The numbers after the land-cover category labels indicate the number of sites used in the evaluation. The error bars indicate the 95% confidence interval for the proportion of improving trends. The red line indicates 50% of sites with improving trends.

5.4 Region

There were differences in the PIT statistics between some pairs of regions for all variables (i.e., non-overlapping 95% confidence intervals; Figure 7). Confidence that CLAR was improving at more than 50% of sites exceeded 95% for six regions (Auckland, Bay of Plenty, Wellington, Tasman, West Coast and Southland; Figure 7). Confidence that CLAR was degrading at more than 50% of sites exceeded 95% for the Waikato region (Figure 7). It is noted that the Auckland region had PIT statistic for CLAR of 100% (Figure 7). This arises because both sites representing trends in CLAR in the region had site trends that were virtually certain to be improving.

Confidence that NH₄N was improving at more than 50% of sites exceeded 95% in five regions. Confidence the NH₄N was degrading at more than 50% of sites exceeded 95% in three regions (Marlborough, Taranaki, and Gisborne). Confidence that TP and DRP was improving at more than 50% of sites exceeded 95% for 10 and nine regions respectively. There were no regions for which there was 95% certainty that TP was degrading at more than 50% of sites and only four regions for which there DRP was degrading at 50% or more sites with 95% certainty (Taranaki, West Coast, Tasman and Marlborough).

Confidence that NO₃N was degrading at more than 50% of sites exceeded 95% for six regions (Waikato, Tasman, Canterbury, West Coast, Otago and Southland). Confidence that NO₃N was improving at more than 50% of sites exceeded 95% for five regions (Northland, Auckland, Taranaki, Manawatu-Wanganui and Wellington). Patterns in TN were similar to those of NO₃N.

Confidence that ECOLI was improving or degrading at more than 50% of sites exceeded 95% for in two (Bay of Plenty and Tasman) and four (Taranaki, Gisborne, Hawkes Bay and Otago) regions respectively (Figure 7). Finally, confidence that MCI was improving at more than 50% of sites exceeded 95% in only two regions Northland and Manawatu-Wanganui (Figure 7). Confidence that MCI was degrading at more than 50% of sites exceeded 95% in four regions Auckland, Gisborne, Hawkes Bay and Tasman.

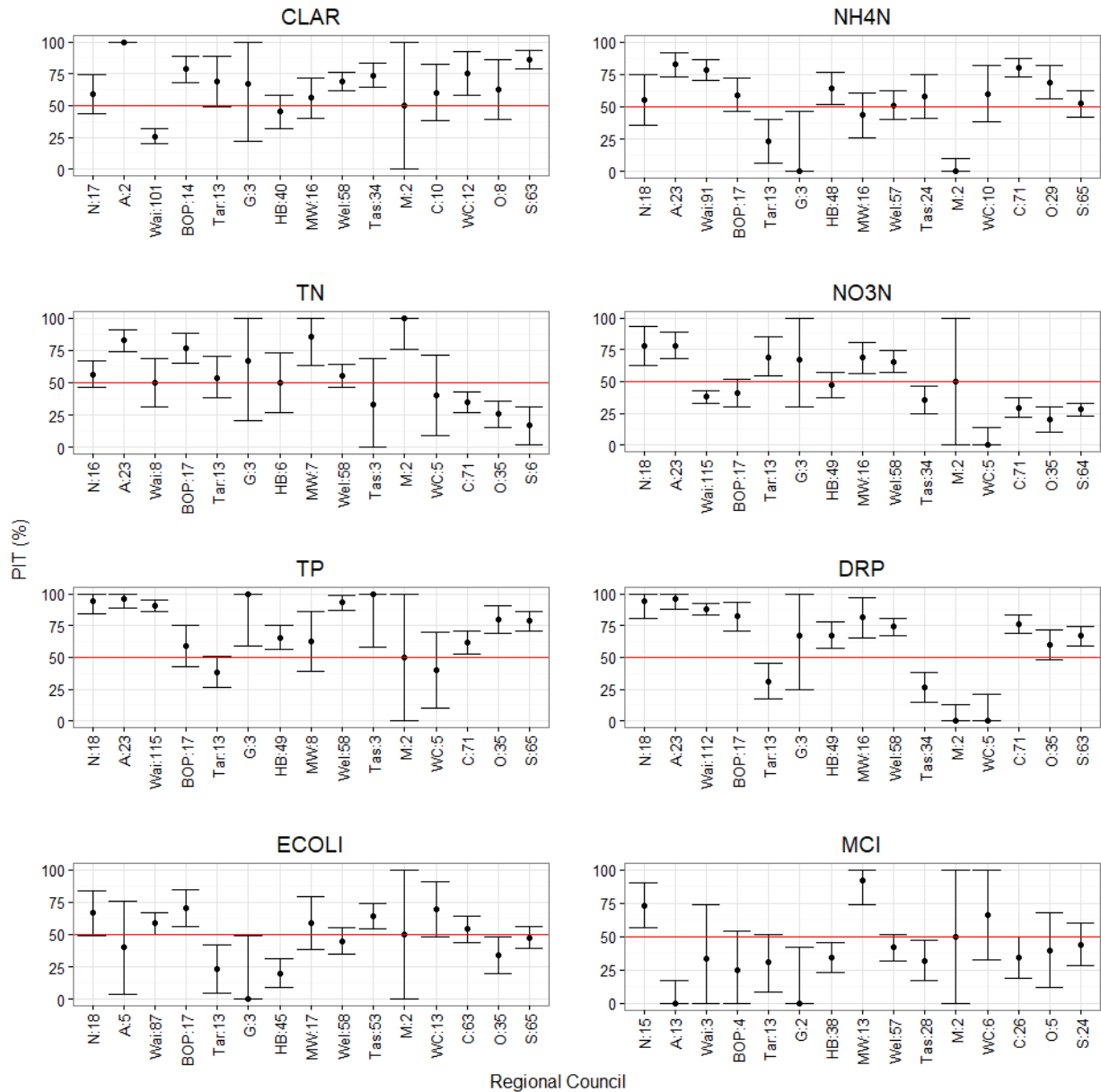


Figure 7. Evaluated PIT statistic for 10-year flow adjusted trends grouped by region. The numbers after the region labels indicate the number of sites used in the evaluation. The error bars indicate the 95% confidence interval for the proportion of improving trends. The red line indicates 50% of sites with improving trends.

6 Conclusions

6.1 New trend aggregation methods

In this study we introduced two new methods for presenting aggregated trends (site trends from many sites that are grouped to represent the ‘overall’ change in water quality that has occurred over some domain of interest). Aggregated trends can be presented as charts that show the proportions of improving sites at different categorical levels of confidence. Alternatively, an assessment of the proportion of improving trends (PIT) can be made that includes the quantification of the uncertainty of this statistic. Both methods treat confidence in trend direction as a probability i.e., a continuous quantity between zero and one, instead of the traditional binary ‘trend’, ‘no-trend’ interpretation. These approaches overcome two problems associated with presenting tables of the proportion or numbers of sites categorised as increasing, decreasing and insufficient data. First, the incorrect inference that trends categorised as insufficient data are “no change” or “stable” is less easily made. Second, information about the direction of change that is associated with the trends categorised as insufficient data is utilised.

Plots representing the proportion of site trends that indicate improvement at each categorical level of confidence demonstrate that there is a continuous, not binary, confidence in trend direction. This is consistent with the philosophy behind the new trend assessment procedure, that there is always a trend but confidence in its direction depends on the available data. The plots of confidence in trend direction provide a visualisation of the confidence in the direction of all the included site-trends and the confidence that the dominant direction of the trends (i.e. the highest proportion) was improving (or its complement; degrading).

Tabulations that include trends categorised as improving, degrading and insufficient data invite the interpretation to ignore the trends categorised as insufficient data and make conclusions about the overall trends based only on the trends whose directions are established with confidence. This approach can potentially lead to incorrect conclusions being drawn as it fails to consider all available information about site trends. To address this problem, we developed the PIT statistic as a more robust alternative for presenting aggregate trend results. The PIT statistic distils the information contained in all the individual trends into a single number (plus its uncertainty). We demonstrated that PIT can potentially be different to the proportion of trends with directions established with confidence that are improving. For example, the PIT statistic indicated 64%, 79% and 41% of sites were improving for NH₄N, TP and MCI respectively (Table 5). However, of trends with directions established with confidence, 83%, 90% and 24% of NH₄N, TP and MCI were improving, and these results were generally outside the 95% confidence intervals of the PIT statistic (Table 5). This indicates that through not including information provided by all the site trends, the traditional approach gives a misguided impression of the proportion of improving sites.

As well as providing a single easily understood statistic (the proportion of improving sites, or its complement), the PIT statistic avoids referring to trends categorised as insufficient data (or the traditional insignificant trends). PIT statistics for domains of interest (e.g., nationally, regionally or by classes) also help to elucidate patterns in water quality changes that are difficult to perceive by examining the individual site trends. We therefore recommend that PIT statistics are used in future to represent aggregate measures of water quality change over a spatial domain of interest. Proportions of improving sites for each spatial domain can be shown as tables or figures, as has been demonstrated in this report. We recommend that the spatial domain(s) is made clear when these types of tables or figures are presented. We also

recommend that PIT statistics are presented as distinct from the trend evaluations for individual sites, for which certainty in trend direction (or significance) remains an important piece of information.

Our study has also shown that count-based estimates are plausible realisations of the PIT statistic (Figure 4) and are therefore a reasonable approximation of the proportion of improving sites. However, it should be kept in mind that count-based estimates are subject to unquantified uncertainty.

6.2 Inferences from aggregated trends for 10 -year period ending 2013

For the 10-year period, sites with improving trends dominate (i.e., are > 50%) nationally for CLAR, NH₄N, TP, DRP and ECOLI. The PIT statistics indicate that the lower 95% confidence interval was >50% for all these variables except ECOLI (Table 5). There is therefore high confidence that the majority of sites had improving CLAR, NH₄N, TP and DRP over the 10-year period. For the 10-year period, sites with degrading trends dominate (i.e., are > 50%) nationally for NO₃N and MCI. The upper 95% confidence interval was <50% for these variables indicating high confidence that degradation occurred at the majority of sites (Table 5).

Some clear patterns at smaller spatial scales are also clarified by the PIT statistics. CLAR was improving at the majority (i.e., >50%) of sites in most environmentally defined river classes and regions. The exceptions to this were REC Source-of-flow classes that are prevalent in the upper North Island (WX/L, WW/H, WW/L and WD/L). Correspondingly of the majority of sites in the Waikato and Hawkes Bay regions had degrading trends for CLAR. In addition, TP and DRP were improving at a majority (i.e., >50%) of sites in the majority of REC source-of-flow classes (Figure 5) and regions (Figure 7).

Larned *et al.* (2016) noted that TP and DRP decreased over the 2004–2013 period at median rates >1.5% yr⁻¹ and that CLAR was improving at a majority of sites. Larned *et al.* (2016) suggested stock exclusion from waterways, improved farm effluent treatment, improved fertiliser management and reductions in phosphorus fertiliser use may be contributing factors. This study shows that TP, DRP and CLAR were improving at a majority of sites belonging to Natural REC Land cover category over the 10-year period (Figure 6). Therefore, changes in clarity and phosphorus are occurring systematically (i.e. >50% of sites) in river classes that have relatively low levels of resource use. It should be noted however, that sites classified as Natural land cover can include up to 15% urban land cover and 25% pastoral land cover. It has been known for some time that both improving and degrading trends in a variety of water quality variables can be associated with climatic variation (Scarsbrook *et al.*, 2003) and that water quality trends occur at minimally impacted sites (e.g., Larned *et al.*, 2004). This means that decision makers need to be cautious about attributing the causes of trends to human activities.

Another pattern that is clarified by the PIT statistics is increasing nitrogen. Over both the 10-year period, NO₃N and TN was degrading at >50% of sites in some REC classes and regions. In particular, the PIT statistic for NO₃N for regions with more than 10 sites indicates that the upper 95% confidence interval was <50% for six regions: Waikato, Tasman, Canterbury, West Coast, Otago and Southland (Figure 7). These results indicate high confidence that the majority of sites had degrading nitrate concentrations over the time-period. It is noteworthy that the 95% confidence interval for the proportion of sites with degrading NO₃N in the Natural REC Land cover category included 50% in the 10-year time-period (Figure 6). This indicates that there were no systemic changes in NO₃N at sites in river classes that have relatively low

levels of resource use (i.e., similar proportions of sites had increasing and decreasing NO₃N). This strengthens the evidence that observed changes in NO₃N in some regions are related to human activities.

Finally, over the 10-year time-period, the PIT statistic indicates that the upper 95% confidence interval for the proportion of sites nationally with improving MCI was 44% (Table 5). This result indicates high confidence that the majority of sites nationally had degrading MCI over the 10-year time-period. The majority of sites also had degrading MCI trends in all land-cover categories (Figure 6). The upper 95% confidence interval for the proportion of sites in the Natural (N) land cover category with improving trend was 45%. This indicates high confidence that degradation occurred at the majority of sites in river classes that have relatively low levels of resource use. This indicates systemic changes may be occurring in MCI and means that decision makers need to be cautious about attributing the causes of MCI trends to human activities. We recommend that further research is carried out on MCI observations to determine if the trends can be explained either by changes in sampling, analysis of samples or calculation of individual MCI scores, or by environmental changes (for example, climatic variation).

6.3 Limitations of the PIT statistic

The PIT statistic is being calculated as though the probability that the trend is decreasing (or improving) were a population parameter associated with each of the individual site trends involved. However, the probability that the trend is decreasing is, itself, an estimate that is uncertain. This means that not all sources of uncertainty in the estimation of PIT are accounted for in our approach. This issue could be addressed by reformulating the trend aggregation problem as a Bayesian model. This was beyond the scope of this study and we consider that it is unlikely to greatly change conclusions when PIT statistics are calculated using reasonable numbers of sites (e.g., >100).

Another advantage of taking a Bayesian approach would be that the uncertainty of PIT could be expressed as a Bayesian credible interval. This would allow the interpretation that the true value of PIT lies within the interval at the level of confidence expressed by the credible interval.

Acknowledgements

We thank continued support from the Ministry for the Environment for development of these ideas and supporting software. Scott Larned, Clive Howard-Williams and Graham McBride provided ideas and statistical advice. We thank Robert Hirsch, Research Hydrologist (U.S. Geological Survey) and Vic Duoba Statistician (Statistics New Zealand) for statistical advice and reviews of early versions of this report. The final version of this report benefitted considerably from formal review by Charlotte Jones-Todd, Statistician (NIWA, Hamilton).

References

- 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/>.
- Cumming, G., F. Fidler, and D.L. Vaux, 2007. Error Bars in Experimental Biology. *Journal of Cell Biology* 177:7–11.
- Helsel, D.R., 2012. Statistics For Censored Environmental Data Using Minitab And R. John Wiley & Sons, Inc., Hoboken, New Jersey.
- Helsel, D.R. and R.M. Hirsch, 1992. Statistical Methods in Water Resources. Elsevier.
- 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., 2017. Time Trends-Trend Analysis and Equivalence Testing for Environmental Data. Jowett Consulting. <http://www.jowettconsulting.co.nz/home/time-1>.
- Larned, S.T., M.R. Scarsbrook, T. Snelder, N.J. Norton, and B.J.F. Biggs, 2004. Water Quality in Low-Elevation Streams and Rivers of New Zealand. *New Zealand Journal of Marine & Freshwater Research* 38:347–366.
- Larned, S., T. Snelder, M. Unwin, and G. McBride, 2016. Water Quality in New Zealand Rivers: Current State and Trends. *New Zealand Journal of Marine and Freshwater Research* 50:389–417.
- Larned, S.T., 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.
- McBride, G.B., 2018. Has Water Quality Improved or Been Maintained? A Quantitative Assessment Procedure. Submitted journal article.
- Ministry for the Environment, 2015. Environment Aotearoa 2015. MFE, Wellington, New Zealand. <http://www.mfe.govt.nz/publications/environmental-reporting/environment-aotearoa-2015>. Accessed 28 Feb 2018.
- Ministry for the Environment, 2017. Our Fresh Water 2017. MFE. <http://www.mfe.govt.nz/publications/environmental-reporting/our-fresh-water-2017>. Accessed 28 Feb 2018.
- Scarsbrook, M.R., C.G. McBride, G.B. McBride, and G.G. Bryers, 2003. Effects of Climate Variability on Rivers: Consequences for Long Term Water Quality Analysis. *JAWRA Journal of the American Water Resources Association* 39:1435–1447.

- Sen, P.K., 1968. Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association* 63:1379–1389.
- 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., 2018. 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. LWP Client Report, LWP Ltd, Christchurch, New Zealand.
- Snelder, T.H. and B.J.F. Biggs, 2002. Multi-Scale River Environment Classification for Water Resources Management. *Journal of the American Water Resources Association* 38:1225–1240.
- Snelder, T., S. Woods, and J. Atalah, 2016. 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.
- 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.