NIWA
Taihoro Nukurangi

Guidance for the analysis of temporal trends in environmental data

Prepared for Horizons Regional Council and MBIE Envirolink

April 2021

Climate, Freshwater & Ocean Science

Prepared by:
Ton Snelder (LWP)
Caroline Fraser (LWP)
Scott Larned
Amy Whitehead

For any information regarding this report please contact:

Scott Larned
Chief Scientist – Freshwater and Estuaries
National Institute of Water & Atmospheric Research Ltd (NIWA)
+64-3-343 7834

NIWA CLIENT REPORT No: 2021017WN
Report date: April 2021
NIWA Project: ELF21301

Quality Assurance Statement		
	Reviewed by:	Dr Judi Hewitt Juliet Milne
	Formatting checked by:	Rachel Wright
	Approved for release by:	Steve Wilcox

Contents

Executive summary **7**

1 Introduction **10**

1.1 Scope 11

1.2 Report outline 12

2 Fundamental considerations **14**

2.1 What is trend assessment? 14

2.2 Trend applications and their objectives 14

2.3 Trend assessment methods and model selection 15

3 Acquiring and compiling data **19**

3.1 General data preparation 19

3.2 Formatting censored values 20

3.3 Defining seasons and time periods and ensuring adequately distributed data 21

4 Accounting for confounding factors **24**

4.1 Covariate adjustment 24

4.2 Seasonality assessment 28

5 Trend assessment **31**

5.1 Purpose 31

5.2 Recommended methods 31

5.3 Commentary 35

6 Reporting trend analyses **41**

6.1 Purpose 41

6.2 Recommended methods 41

6.3 Commentary 48

7 Worked trend assessment examples **51**

7.1 Example 1: Flow adjusted, seasonal WQ variable 51

7.2 Example 2: MCI trend 57

7.3 Example 3: Missing data 62

7.4 Example 4: High censoring 67

8 Acknowledgements **73**

9 Glossary of abbreviations and terms 74

10 References..... 76

Appendix A Supplementary data preparation guidance 81

A1 Ensuring consistent data structure..... 81

A2 Correcting data errors..... 85

A3 Ensuring comparable measurement methods 87

Appendix B Comparison of flow adjusted and non-flow adjusted trends 89

Appendix C Comparison of seasonal and non-seasonal trend assessments..... 93

Appendix D Climate influence on water quality trends..... 96

Tables

Table 6-1: Level of confidence categories used to convey confidence in trend direction. 42

Table 6-2: Suggested confidence categories in water quality trend are decreasing (C_d). 43

Table 6-3: Tabulation of trend assessment results for multiple site/variable combinations from a hypothetical lake water quality monitoring programme. 44

Table 7-1: Tabulated results of trend assessment for Example 1. 57

Table 7-2: Tabulated results of trend assessment for Example 2. 61

Table 7-3: Tabulated results of trend assessment for Example 3. 67

Table 7-4: Tabulated results of trend assessment for Example 3 for eight alternative analyses. 67

Table 7-5: Tabulated results of trend assessment for Example 4. 70

Table 7-6: Example 3 – Trend assessment results from TimeTrends exploring the impact of the direction of step change in detection limit, the implementation of the hi-censor filter, and based on the complete, uncensored dataset. 72

Figures

Figure 1-1: High level flow chart showing the main steps in the trend assessment procedure. 13

Figure 2-1: Example of monotonic and non-monotonic trends in nitrate-nitrate nitrogen (NNN) concentrations (observations) in the Taieri River at the National Water Quality Monitoring Network site at Outram between 1998 and 2017. 17

Figure 4-1: Example of the relationship between water quality observations and flow showing alternative fitted models. 27

Figure 4-2: Plot showing the distributions of water quality observations within seasons. 29

Figure 4-3: Plot showing the distributions of water quality observations within seasons. 30

Figure 5-1: Pictogram of the steps taken in the trend direction assessment to calculate bthe Kendall *S* statistic and its *p*-value which is used to characterise bconfidence in trend direction. 32

Figure 5-2: Pictogram of the calculation of the Sen slope, which is used to estimate trend rates. 34
Figure 6-1: Graphical representation of three alternative ways of expressing confidence in trend direction. 42
Figure 6-2: Graphical representation of alternative categorisations of confidence level including trend direction. 43
Figure 6-3: Comparison of SSE and C_d for four variables at 69 NRWQN sites for a 10-year trend period. 44
Figure 6-4: Maps of NRWQN sites summarising 10-year trends in four water quality variables. 45
Figure 6-5: Maps of NRWQN sites summarising 10-year trends in four water quality variables categorised by trend direction and confidence in trend direction. 45
Figure 6-6: Maps of NRWQN sites summarising 10-year trends in four water quality variables, categorised into seven classes based on confidence that the trend direction is decreasing (C_d). 46
Figure 6-7: Examples of box plots to describe distributions of trend direction and trend rate across many sites. 47
Figure 6-8: Example of stacked bar charts of categorical confidence that the trend was decreasing. 48
Figure 6-9: Example showing the presentation of state and trend information on a single plot. 50
Figure 7-1: Example 1 – LWPTrends raw data inspection plots. 52
Figure 7-2: Example 1 – TimeTrends scatter plot of raw data. 52
Figure 7-3: Example 1 – scatterplot plot of concentration (value) and flow produced by LWPTrends. 53
Figure 7-4: Example 1 – box plot of flow adjusted values by season produced by LWPTrends. 54
Figure 7-5: Example 1 – box plot of flow adjusted values by season produced by TimeTrends. 55
Figure 7-6: Example 1 – time-series of monthly flow adjusted observations and fitted non-parametric (Sen) regression line (and 90% confidence intervals) produced by LWPTrends. 56
Figure 7-7: Example 1 – time-series of monthly flow adjusted observations and fitted non-parametric (Sen) regression line produced by TimeTrends. 56
Figure 7-8: Example 2 – LWPTrends raw data inspection plots. 58
Figure 7-9: Example 2 – TimeTrends scatter plot of raw data. 58
Figure 7-10: Example 2 – box plot of raw observations by season produced by LWPTrends. 59
Figure 7-11: Example 2 – box plot of observations by season produced by TimeTrends. 59
Figure 7-12: Example 2 – time-series of observations and fitted non-parametric (Sen) regression line produced by LWPTrends. 60
Figure 7-13: Example 2 – time-series of observations and fitted non-parametric (Sen) regression line produced by TimeTrends using the “Median value per season” option. 61
Figure 7-14: Example 2 – time-series of observations and fitted non-parametric (Sen) regression line produced by TimeTrends using the “All values in season” option. 61

Figure 7-15: Example 3 – LWPTrends raw data inspection plots. 62
Figure 7-16: Example 3 – TimeTrends scatter plot of raw data. 63
Figure 7-17: Example 3 – scatterplot of concentration and flow produced by LWPTrends. 63
Figure 7-18: Example 3 – box plot of flow adjusted values by season produced by TimeTrends (seasons defined as months). 64
Figure 7-19: Example 3 – box plot of flow adjusted values by season produced by TimeTrends (seasons defined as bi-monthly, 'summer' and 'winter'). 65
Figure 7-20: Example 3 – time-series of flow adjusted observations and fitted non-parametric (Sen) regression line produced by LWPTrends (seasons defined as months). 66
Figure 7-21: Example 3 – time-series of flow adjusted observations and fitted non-parametric (Sen) regression line produced by LWPTrends (seasons defined as bi-monthly, 'summer' and 'winter'). 66
Figure 7-22: Example 5 – LWPTrends raw data inspection plots. 68
Figure 7-23: Example 4 – box plot of flow adjusted values by season produced by LWPTrends. 69
Figure 7-24: Example 4 – box plot of flow adjusted values by season produced by TimeTrends. 69
Figure 7-25: Example 4 – time-series of observations and fitted non-parametric (Sen) regression line produced by LWPTrends. 71
Figure 7-26: Example 4 – time-series of observations and fitted non-parametric (Sen) regression line produced by LWPTrends. 71

Executive summary

This document provides practical guidance for the analysis and reporting of temporal trends in environmental data. The emphasis is on freshwater physico-chemical and biological variables that are commonly and routinely measured in New Zealand's rivers (collectively denoted, for simplicity, as "water quality variables"). However, the methods are applicable to other environmental variables and domains (e.g., lakes, groundwaters, estuaries and coastal waters) for which a suitable time-series record exists. The guidance has been prepared for Horizons Regional Council on behalf of the regional sector, via an MBIE Envirolink advice grant.

Trend assessments serve multiple purposes, including informing the public about changes in environmental state, assessing the effectiveness of management actions and policies, evaluating relationships between environmental conditions and the factors that influence them (i.e., driver or explanatory variables), and providing early warning of environmental problems. Trend assessments are used in New Zealand for regional and national environmental reporting, reflecting regional council responsibilities under Section 35 (2) of the Resource Management Act. Recently, the requirement for trend analysis has been made explicit in the National Policy Statement for Freshwater Management. Therefore, the primary purpose of this guidance is to facilitate more consistent and transparent assessment and reporting of trends in freshwater and other environmental data.

Trend assessment is a process of building a statistical model of the behaviour of a variable at a site over a time period of interest based on a series of observations. These guidelines describe methods for detecting and quantifying the two most fundamental aspects of the relationship between the variable and time: the direction (i.e., increasing or decreasing) and the rate of change (e.g., change in concentration per year). In addition, because the observations are subject to random fluctuations and only comprise a sample of the variable's behaviour over the time period, the guidelines also describe methods for quantifying the uncertainties associated with the assessment of trend direction and rate.

Obtaining reliable results from a trend analysis and reporting these results appropriately depends on choices that are associated with the analysis objectives. These guidelines start by describing three types of application of trend analysis: "local", "regional" and "national". Local applications are concerned with maximising information about trends at a single site whereas regional and national applications are concerned about obtaining consistent assessments over many sites. The details of trend analysis and reporting process vary to some extent depending on the type of application.

The guidance describes trend analysis in four main steps:

1. acquiring and compiling data,
2. accounting for confounding factors,
3. assessing trend direction, rate, and confidence in these determinations, and
4. reporting the results.

Each of these steps is addressed in a separate section and each section has three subsections that describe the purpose of the step, the methods, and a commentary. As for all data analysis, there are complications and subjective decisions at every step and there is no single 'right' way to carry out a trend assessment. This means that it is not possible or appropriate to rigidly dictate how each step is to be carried out. Users of these guidelines are therefore encouraged to read them in their entirety

and to fully understand the purpose of each step. The methods subsections then provide recommended approaches in considerable detail. In the commentary subsections, we provide background to the methods, and discuss subjective decisions associated with each step and the implications of those decisions.

The fundamental unit of analysis in trend assessments is a specific water quality (or other) variable at a single site, which we refer to as the site/variable combination. The data acquisition and compilation methods are intended to ensure that the data for each site/variable combination are as complete and correct as possible, organised so that they are easily uploaded to statistical modelling software, and associated with information to assist in reporting (e.g., geographic coordinates so that maps can be produced). Aspects of data acquisition and compilation that are particularly important to the statistical modelling of trends are (1) the correct identification of censored values (observations where the true values were too low or too high to be measured with precision), (2) the definition of seasons and the analysis time period and (3) ensuring adequately distributed data (based on filtering rules). Items 2 and 3 are associated with subjective decisions that need to consider the objectives of the application and the constraints associated with the available data.

Some of the variation in a water quality variable of interest may be associated with factors that confound its relationship with time (i.e., the trend). Increased variation associated with confounding factors reduces the detectability of trends. These guidelines explain how the analysis process deals with two types of confounding factors: covariates and seasonality. A covariate is a variable, other than time, that is related to the observations and whose influence obscures the variable – time relationship of primary interest. Examples of covariates for water quality variables are flows in rivers and tidal fluctuation in estuaries. The guidelines describe how to build a statistical model of the relationship between the covariate and the observations to remove the covariate's influence prior to carrying out the trend analysis.

For many site/variable combinations, season is also a confounding factor and seasonality explains a considerable amount of variation in the observations. In these cases, accounting for systematic seasonal variation increases the statistical power of the trend assessment (i.e., increase the confidence in the estimate of direction and rate of the trend). The guidelines describe how to assess seasonal variation and how to deal with seasonality in trend analyses.

The guidance recommends the use of two non-parametric statistical models for assessing trend direction, rate, and confidence in these determinations: the Mann Kendall correlation assessment and Sen slope regression and their seasonal counterparts. These models have been the basis for trend analyses of freshwater water quality variables and other types of environmental variables in New Zealand and worldwide for approximately 30 years. The guidance explains the methods for using these models and provides a commentary on quantifying confidence and dealing with serial correlation and multiple censoring levels. There are strong practical reasons for using non-parametric models for many applications of trend analysis rather than parametric alternatives; the reasons are explained in the guidelines. In some situations, alternative models are more appropriate; the guidelines do not cover alternative models in detail but provide references to other sources of guidance.

In the past few years multiple changes have been recommended for using the non-parametric statistical models recommended here, and for interpreting outputs from those models. The most important recommendation is that null hypothesis significance testing (NHST) be replaced by a continuous measure of confidence in trend direction. This change is beneficial as it reduces the chance of misinterpreting the results of trend analyses and provides information that is more helpful

for management. The guidelines provide an up-to-date description of recent changes and why they are recommended. We note that the recommended discontinuation of NHST concerns only the interpretation of model outputs (the *p*-value returned by the Mann Kendall correlation assessment) and therefore any trend assessment performed using the recommended method can still be interpreted using the significance testing approach.

The key metrics for reporting the results of trend assessments are: trend direction (positive or negative), confidence about the trend direction (a probability that the evaluated direction is correct), trend rate (the rate of change of the environmental variable per year) and confidence about the trend rate. It is generally necessary to report these metrics for multiple site/variable combinations, and this is always the case for regional and national applications. These guidelines provide methods for effectively reporting the four metrics using tabular, graphical or map format. In addition, graphical methods are suggested for communicating aggregated trend direction and confidence information. Aggregated summaries across sites (e.g., proportion of sites with increasing and decreasing trends, by variable) provide informative overviews of water quality changes across a domain of interest (e.g., region, environmental class, the entire country). The guidance also notes the importance of considering current state alongside trend assessment results.

The trend analysis methods set out in the guidance are used to determine whether an environmental variable of interest has changed over time. These analyses provide no information about the causes of temporal trends. Attribution of trends to causes is not covered by these guidelines. However, trend detection may often trigger the need for management action and this will require investigation of possible causes. Robust attribution of water quality trends to causes is complicated and challenging. One of the complications is the influence of natural climate variation. Appendix D summarises a recent study that quantified the influence of the El Niño Southern Oscillation climate processes (ENSO) on river water quality trends in New Zealand. Results of the study indicated that the ENSO climate signal translates into a predictable effect on water quality trends. This means that climate variation may amplify or counteract the effects of other drivers of water quality trends, such as land use and land management.

In the future, alternatives to the trend assessment methods recommended here will be trialled in New Zealand. The alternative methods are likely to have advantages and disadvantages that need to be evaluated. However, in our view there are several issues of greater immediate importance to improving trend assessment and reporting in New Zealand. First, further consideration should be given to understanding trend rates that are of environmental and management importance to provide context for reporting and for prioritising management actions. Establishing important trend rates would also help with the issues surrounding statistical inference (e.g., this would enable the use of equivalence tests, which pose more realistic hypotheses). Second, more work is required to develop methods for assessing the causes of trends (i.e., attribution). A robust understanding of the causes of water quality trends, both degrading and improving, will enable effective management actions to be prescribed to arrest and reverse degradation and drive recovery.

1 Introduction

Detecting temporal trends in a core set of environmental variables (or 'indicators') is one of the primary aims of many environmental monitoring programmes, particularly long-term state of the environment (SOE) monitoring programmes. In New Zealand – and worldwide – environmental trend analyses are carried out by different agencies (e.g., councils, central government, industries, research agencies) at different spatial (e.g., local, regional, national) and temporal (e.g., annual, decadal) scales. These analyses serve multiple purposes, including informing the public about changes in environmental state, assessing the effectiveness of management actions and policies, evaluating relationships between environmental conditions and the factors that influence them (i.e., driver or explanatory variables), and providing early warning of impending environmental problems (McMellor and Underwood 2014; MFE & StatsNZ 2020; Murphy 2020; Tomperi et al. 2016).

A key use of trend assessment in New Zealand is in regional and national environmental reporting, reflecting regional council responsibilities under Section 35 (2) of the Resource Management Act (RMA), to monitor the state of the environment of their region, and the efficiency and effectiveness of policies and methods in regional policy statements and plans. More recently, the need for trend analyses has been made explicit in national legislation, with Section 3.19 of the National Policy Statement for Freshwater Management (NPS-FM 2020; NZ Government 2020) directing regional councils to assess trends in freshwater attribute states (where attributes comprise the freshwater indicators listed in Appendix 2 of the NPS-FM). Section 3.19 of the NPS-FM requires regional councils to determine appropriate trend periods and sampling frequencies, and to specify the likelihood (i.e., the level of confidence about trend direction) of trends. Section 3.20 of the NPS-FM directs councils to "*take action to halt or reverse degradation*" and requires the action to be "*proportionate to the likelihood and magnitude of the trend, the risk of adverse effects on the environment, and the risk of not achieving target attribute states*".

In a general sense, carrying out a trend assessment means building a statistical model of the changes in an environmental indicator over time. For the last three decades, two types of statistical models[1] have been widely used in trend analyses of freshwater water quality variables and other types of environmental variables, both in New Zealand and worldwide (e.g., Ali et al. 2019; Larned et al. 2016; Sa'adi et al. 2019). However, in the past few years there have been changes to some of the details concerning how these models are implemented and interpreted[2], in New Zealand (McBride, 2019) and elsewhere (e.g., Choquette et al. 2019; Helsel et al. 2020). These changes are generally beneficial as they improve the rigour and scope of trend analyses, but they can also create problems for comparing the results of successive analyses if environmental changes are confounded by methodological changes. In addition, there are numerous steps in the model building process such as data preparation, deciding on temporal resolution (e.g., months, seasons or years within multi-year trend periods), removing or including the influence of covariates, and deciding whether to account for regular seasonal fluctuations (Helsel et al. 2020). These steps in the model building process require the analyst to make decisions. Different decisions between analyses can lead to differences in reported trends, even when these are derived from identical datasets (e.g., Oelsner et al. 2017; Rangeti et al. 2015). The subsequent differences in trend results, and the variation in methods that produce them, can be a source of confusion about and mistrust of environmental reports.

[1] These models are based on the Mann-Kendall test of correlation and Sen slope regression and are discussed in detail in Section 6.

[2] Briefly, these changes involve improvements in the way censored values are handled and the rejection of the use of null hypothesis significance testing in favour of assessing confidence in trend direction. These changes are discussed in detail in Section 5.

In light of the growing requirements and evolving use and application of trend assessment, a selection of the regional sector's Surface Water Integrated Management (SWIM) Special Interest Group supported the preparation of a guidance document to set out the key steps and decisions involved in the trend analysis and reporting. The Horizons Regional Council obtained an MBIE Envirolink large advice grant (HZLC154) for NIWA and LWP to prepare the guidance.

1.1 Scope

This guidance document focuses on temporal trend assessment and reporting methods appropriate for freshwater physico-chemical and biological variables that are commonly and routinely measured in New Zealand's rivers (e.g., visual clarity, dissolved inorganic nitrogen and the Macroinvertebrate Community Index (MCI)). However, the methods are applicable to other environmental variables (e.g., chlorophyll a, Secchi depth) and domains (e.g., lakes, groundwaters, estuaries and coastal waters) for which a suitable time-series record exists. For simplicity, we refer to all variables that are routinely measured in freshwater monitoring programmes as "water quality variables". The methods discussed in this guidance are appropriate for the analysis of variables that are collected in SOE monitoring programmes used for assessing and reporting environmental state and trends. These programmes are characterised by routine but infrequent observation of variables (e.g., monthly, quarterly, or annually). Low frequency observations mean that the statistical assumption of independence of observations is generally not violated. It has become common in recent years for regional councils to undertake near-continuous sampling of some water quality variables (e.g., dissolved oxygen, turbidity). These data exhibit temporal autocorrelation and their analysis must be undertaken using methods that are not covered in this document. An additional characteristic of SOE monitoring programmes is that observations generally occur on a specified date irrespective of the conditions prevailing at the time of measurement. This means that the sample data (i.e., the observations) are representative of the population (i.e., the water quality over the entire monitoring period). Representative samples are a general requirement for estimating statistics that characterise populations, such as mean or median values, or in the case of trend assessment, directions and rates change in a variable through time.

The statistical methods discussed in this document address two primary questions:

1. What was the direction of change in a water quality variable over a time period (i.e., was water quality degrading or improving)?
2. What was the rate of the change over that time period?

In addition, the statistical methods provide an estimate of confidence associated with the answer to each of these two questions. The directions and rates of trends and the confidence in these assessments have been fundamental components of SOE reporting in New Zealand for the past two decades and are now requirements of the NPS-FM 2020. Analyses of water quality time series can also answer more complicated questions such as whether the data exhibit cyclic variation and whether the changes through time are non-linear or stepped. Methods used to detect more complicated types of trends are out of scope for this guidance. However, we provide some discussion of these other methods in the commentary sections of this document, and direct readers to publications with further guidance (e.g., Helsel et al. 2020).

The statistical methods discussed in this document are used to detect and quantify trends but provide no information about their causes. Water quality trends are often attributed to causes in an informal, speculative way (Ryberg et al. 2018). This practice should be avoided, and rigorous, quantitative methods used. However, rigorous trend attribution is a complex topic and advice on

attribution methods is beyond the scope of this guidance. One of the complexities is that trends are invariably influenced by multiple drivers, both natural and anthropogenic. One of the natural drivers of water quality trends that has had limited study in New Zealand is climate variability (Scarsbrook et al. 2003). More recent work indicates that climate variation has considerable influence on the strength and direction of trends in multiple river water quality variables (Snelder et al. submitted). A summary of this study is provided in Appendix D. A pertinent conclusion from the study is that climate variation may amplify or counteract the effects of other drivers of water quality trends, even when those trends are assessed over time windows that are longer than a decade.

The methods in this report have been documented in earlier reports and journal articles (e.g., Gadd et al. 2020; Larned et al. 2016, 2004; McBride 2019; MFE & StatsNZ 2017, 2019). These and other articles and reports represent the ongoing development of trend assessment methods. Although this report sets out the current approaches for analysing and reporting water quality trends, we stress that methods will continue to evolve and there is no single 'right' way to carry out trend assessment. Whatever the approach taken, documenting the steps and any assumptions made is important.

1.2 Report outline

This report is structured to reflect the sequential analysis steps undertaken as part of a water quality trend analysis in New Zealand (Figure 1-1). In Section 2, we discuss some fundamental concepts. This section explains that trend assessment is a field of statistical modelling and that the choice of modelling method, and choices made in the data preparation process, should be made based on the objectives of the analysis. This section also recommends a default method for trend assessment, which is referred to as the "traditional non-parametric method".

In Sections 3-7, we outline the steps in the trend assessment process from obtaining the data to reporting the results. These sections define the purpose of each analysis step, define recommended methods, and provide commentary about the recommendations. Sections 3 to 5 concern aspects of data preparation that are specific to the non-parametric trend assessment method. Section 6 describes the statistical models that are the basis of the traditional non-parametric method. Section 7 describes methods for reporting the results of trend analyses.

In Section 7, worked examples are used to demonstrate trend assessment in several common situations (e.g., missing data, high proportions of censored data). These examples follow the steps shown in Figure 1-1. Supplementary data files are available from the LWP website so that readers can implement the examples in Section 7 using one or both of two free computer programmes for conducting trend assessments, LWPTrends[3] and TimeTrends[4]. The worked examples can be used to explore the consequences of some of the choices that can be made at the various stages of trend assessment.

[3] Available at http://landwaterpeople.co.nz/wp-content/uploads/2018/07/LWPTrends_TrendsDec19.zip
[4] Available at https://www.jowettconsulting.co.nz/home/time-1

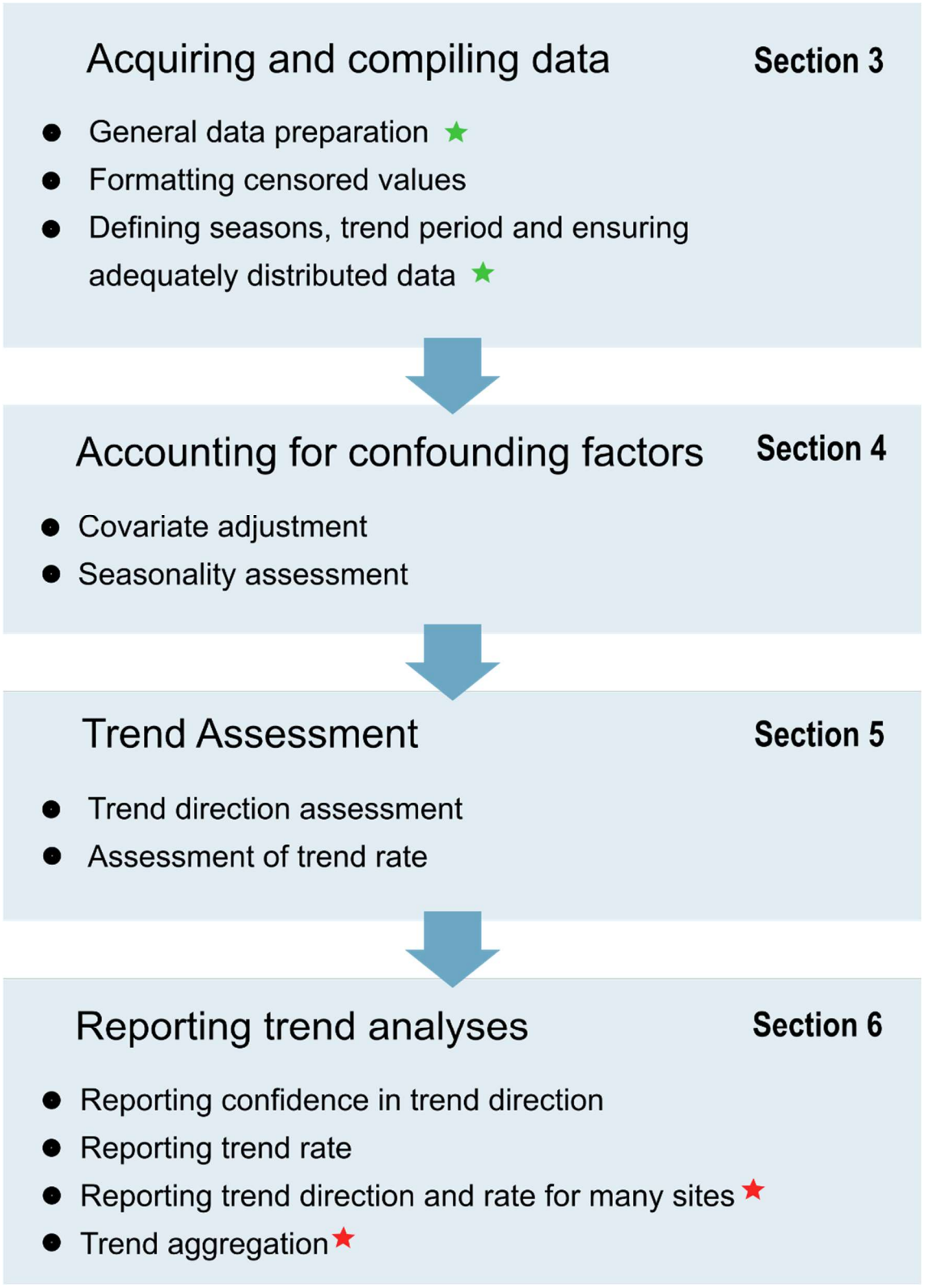

Figure 1-1: High level flow chart showing the main steps in the trend assessment procedure. The guideline section corresponding to each step is indicated. Red stars indicate steps that are relevant to regional and national applications (see Section 2.2). Green stars indicate steps that involve different choices for local applications compared to regional and national applications.

2 Fundamental considerations

2.1 What is trend assessment?

Trend assessment is a process of building a statistical model of the behaviour of some variable over time from a series of observations (Helsel et al. 2020). These guidelines describe methods for detecting and quantifying two aspects of that behaviour, the direction and rate of change of the variable over the time period of interest. The fundamental unit of analysis in trend assessments is a specific water quality variable at a single site, which we refer to as the site/variable combination. The observations pertaining to a site/variable combination comprise a time-series that is a statistical sample of the population (i.e., a sample of the actual conditions over the entire period of interest). Because a sample is only a representation of the population, we can only model the behaviour of the variable over time. Like all statistical models, a trend assessment is a simplification of reality that aims to expose the most important features of the relationship or pattern of interest. The simplest and most salient features of the relationship between a variable and time in trend assessment are the *direction of change* and the *rate of change* in the variable.

Statistical models representing the direction and rate of change of a variable generally consist of one or more of the following: (1) components related to regular cycles such as seasonal or tidal variation; (2) components driven by some exogenous variable (for example, flow as a driver of a contaminant concentration in a river); (3) a long-term trend in the central tendency of the variable; and (4) random variability that is composed of natural variability, measurement error, and possibly serial dependence. The trend (i.e., the pattern of interest) is represented by the long-term trend component of the model (i.e., the direction and rate of change in the central tendency over some multi-year period of interest). If we could measure the variable of interest at a very high sampling rate and with perfect accuracy it is virtually certain that we would conclude that there is a trend (McBride 2019). In reality, monitoring programmes do not have very high sampling rates or perfect measurements, thus statistical models are needed to estimate the direction and rate of change, and to estimate confidence in these estimates.

2.2 Trend applications and their objectives

Trend analyses are undertaken for a variety of purposes and the details of the methods should vary with the objectives. In this document, we compare choices that an analyst would logically make for three types of application trend assessment: a “local application”, a “regional application” and a “national application”. These three applications are not intended to cover all possible objectives of trend assessments, rather they are used to illustrate how the application’s objectives influence choices that are made in some of the analytical steps.

The objective of a local application is to extract as much information as possible about the trend direction and rate of change from the available data for individual site/variable combinations. An example of a local application is associated with implementing the Section 3.19 and Section 3.20 requirements of the NPS-FM in a given catchment where there are highly significant environmental and resource use values at stake. In local applications, choices at various steps in the assessment process are concerned with establishing the trend with minimum uncertainty. We note that a local application might involve assessment of many site/variable combinations. The important point however is that the objective would be to maximise information for each site and variable. The local application may therefore have inconsistencies across different site/variable combinations because maximising the information for each assessment is prioritised over ensuring robust comparisons can be made between sites.

Regional applications, as illustrated in this guidance, entail assessing and reporting trends across many sites and variables in the context of regional SOE monitoring programmes. The objective of a regional application is to allow robust comparison of trends between sites and to provide a synoptic assessment of trends across a region. The objective may also include comparing between sites that are grouped by environmental classes or locations within the region. Therefore, in regional applications, choices at various steps in the assessment process are concerned with ensuring consistency in the individual trend assessments applied to all site/variable combinations, which then enables robust comparisons to be made across all sites.

National applications entail assessing and reporting trends across many sites and variables where the data are derived from multiple regional SOE (or similar) programmes. The objectives of a national application are similar to a regional application and therefore, choices at various steps in the assessment process are concerned with ensuring consistent methods are applied to all site/variable combinations. However, differences between regional SOE monitoring programmes (e.g., differences in sampling frequency, differences in analytical methods) mean that additional steps in the analytical process are required to achieve consistency in the use of data between regions. For example, if there are differences in sampling intervals between regions, achieving consistency may involve removing or coarsening data for some regions.

Differences in the objectives of trend applications lead to differences in all aspects of all analysis steps and therefore the results. This is one reason that, for example, trend assessments for individual sites in local or regional applications may differ from those in a national application.

2.3 Trend assessment methods and model selection

In general, the type of statistical model that is appropriate for assessing trends in water quality variables is a regression model of the water quality observations versus time. There are many types of regression models that can be used to assess different types of trends including sudden, gradual and non-monotonic trends. A classification of these regression models and their strengths and weaknesses is provided by Helsel et al. (2020). An important distinction between statistical models used for trend assessment is whether they are parametric or non-parametric. Parametric models assume that the data conform to specific distributions (e.g., that the data are normally distributed). This means that several conditions of validity must be met for the result of an analysis, particularly the assessment of confidence, to be reliable. Non-parametric models do not rely on distributional assumptions and therefore have fewer conditions of validity.

This guidance provides detailed advice on the use of only one statistical modelling approach for trend assessment, which is referred to hereafter as the “traditional non-parametric method”. This method is the recommended default because:

1. it provides reliable estimates of trend direction and rate over a wide range of data characteristics,
2. it does not require a high level of statistical knowledge to use with confidence, and
3. it will produce assessments that are consistent and comparable for different site/variable combinations.

There are times when alternatives to the traditional non-parametric method may be appropriate, depending on the objective of the trend assessment and the characteristics of the data being analysed. The most familiar alternatives are parametric models. Parametric and non-parametric

models have multiple differences (e.g., data requirements, statistical assumptions); these differences may be strengths or weaknesses depending on the application. Therefore, analysts should be aware of the differences between parametric and non-parametric models so that alternatives to the recommended traditional non-parametric method can be considered should the circumstances suggest this.

To understand the advantages and disadvantages of non-parametric and parametric types of regression that can be used for trend assessment, we first consider some common characteristics of water quality time-series data. These are:

1. non-normal distributions, often with a finite lower bound;
2. presence of outliers;
3. natural cycles such as diurnal, seasonal or tidal;
4. missing values or irregularly spaced observations;
5. the presence of values below analytical detection limits or above reporting limits (i.e., censored values); and
6. associations between the observations and covariates such as river flow.

The most significant feature of non-parametric models is that they are more robust than their parametric equivalents. This means they produce reliable results when data are non-normal, finite lower bounded, and where there are outliers. In contrast, parametric models require a great degree of care to ensure that the model is correctly specified. Parametric models will generally require higher levels of statistical skill by the analyst than the non-parametric default (e.g., correctly transforming the data, dealing with outliers, and applying the correct distribution in the context of a generalised linear model). Although care is still required, the risk of incorrect specification of non-parametric models is lower than for parametric models. This is especially important in regional and national applications where trend analyses are carried out for many site/variable combinations. In these types of applications, non-parametric models can be applied consistently to all site/variable combinations with greater confidence that they will not violate statistical assumptions than parametric alternatives. Another important consideration for regional and national applications is that enough information needs to be provided in reporting the results for the audience to be able to verify that the assumptions were not violated. For parametric models this represents a significant burden because reporting regional and national applications often involves hundreds to thousands of individual models pertaining to site/variable combinations.

One advantage of parametric models over their non-parametric equivalent is that they have greater statistical power. This means the confidence in the results of a trend assessment (direction and rate) will be higher for a parametric regression than the non-parametric equivalent. Helsel et al. (2020) indicate that this advantage is “slight” and depends on the correct specification of the parametric model. They also note that when there is modest departures from normality of the residuals of a parametric model, non-parametric models can be more powerful than parametric regression. When there are many analyses that must be performed (i.e., when the application objective is a regional and national trend assessment), the benefit of increased statistical power that is associated with parametric models is likely to be outweighed by the analytical effort required to undertake detailed case-by-case checking of assumptions.

An important characteristic of the traditional non-parametric method is that it only detects and characterises the monotonic trend through the time period of interest. This means that the model of the behaviour of the water quality variable through the time period is constrained to be either constantly increasing or decreasing. Examples of monotonic and non-monotonic trends are shown in Figure 2-1. In contrast, parametric models can include more flexible representations of the behaviour of a variable through a time period, including non-linear changes and cyclic components. Note that the simplest form of parametric model, an ordinary linear regression, also represents the behaviour of the variable as a monotonic trend. The monotonic feature of the traditional non-parametric method is not a disadvantage in terms of responding to the two questions that are most commonly addressed in trend analyses (i.e., what were the trend direction and rate). However, when the underlying model of the trend is constrained to be monotonic, information about the underlying dynamics may be hidden. A relevant example of the dynamic behaviour of water quality variables that is hidden by monotonic trend assessment is discussed in Appendix D of this guideline. In this case, dynamic behaviour that is linked to climate variation was hidden by individual assessments carried out using the traditional non-parametric method. When these assessments were repeated through time, oscillations in the trend strength and direction were evident.

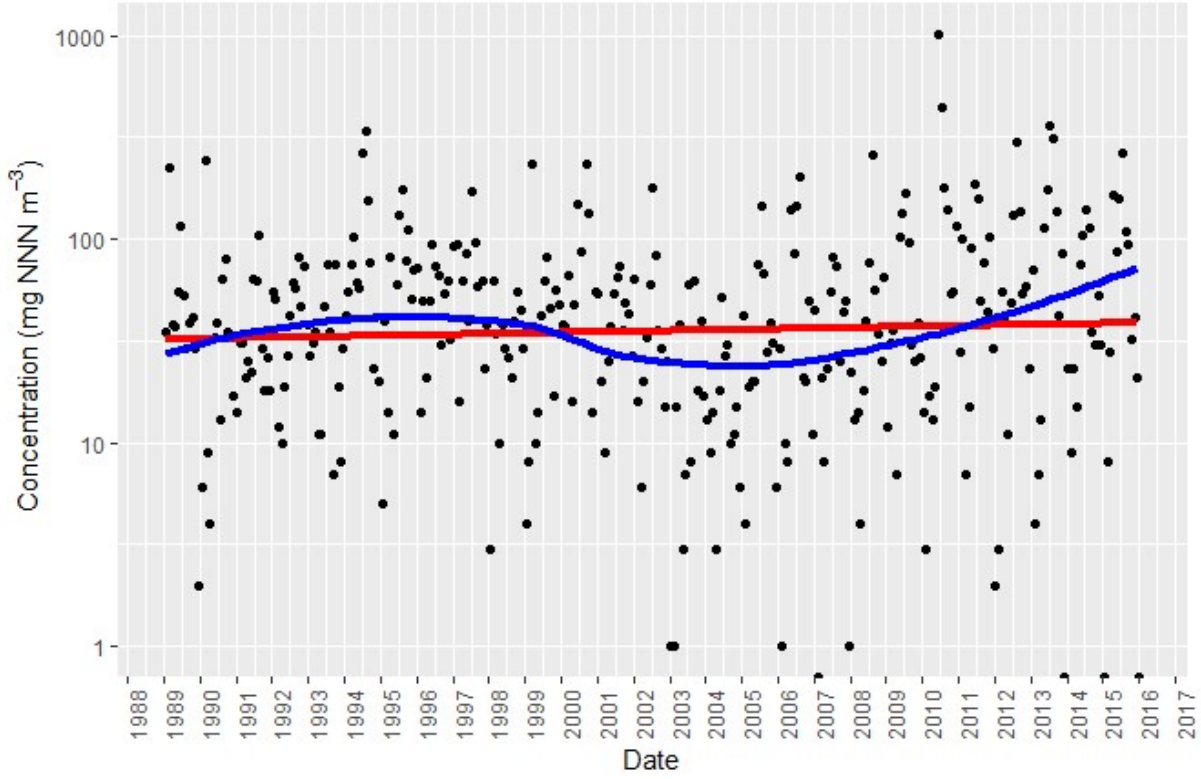

Figure 2-1: Example of monotonic and non-monotonic trends in nitrate-nitrate nitrogen (NNN) concentrations (observations) in the Taieri River at the National Water Quality Monitoring Network site at Outram between 1998 and 2017. The red line represents a model fitted to the observations that is constrained to be a monotonic (increasing) trend. The blue line represents a more flexible model that indicates that the behaviour was not monotonic within the period and suggest NNN concentrations initially increased, then decreased between 1996 and 2005 and increased again after 2005.

Because consistency of trend assessment methods is not an objective of a local application, the use of parametric models to undertake trend assessment is most relevant to this type of application. Local applications involving significant environmental values and resources may drive the need for maximum insight into the form of the water quality change and for maximum confidence in the assessment. The greater statistical power afforded by parametric models may be advantageous even

though ensuring that the model assumptions have not been violated is more onerous than for a non-parametric model. However, there remains a final consideration for the use of parametric models regarding the correct handling of censored values. Censored values are handled robustly by the recommended traditional non-parametric method (see Section 3.2), but censored values complicate the use of parametric models in trend assessments. One option that is often used, but which is inappropriate, is to replace the censored value with some arbitrary value such as zero, the reporting limit, or half the reporting limit. These approaches will give inaccurate assessments (Helsel 2005, 2012). The appropriate parametric approach to modelling trends when there are censored values is the use of a Tobit (or censored regression) model (Cohen 1976; Hald 1949).

The recommended traditional non-parametric method comprises two non-parametric models. The first model is used to assess trend direction based on a non-parametric correlation procedure originally developed by Mann (1945) and further developed by Hirsch et al. (1982). The second model is used to assess trend rate[5] based on a non-parametric regression procedure originally developed by Sen (1968) (the Sen slope[6]) and further developed by Hirsch et al. (1982). These two models have been the basis for numerous trend assessments of water quality monitoring data (i.e., monthly, quarterly, or annual observations) carried out by agencies in New Zealand and internationally. However, there have been recent advances in the interpretation of the outputs from both models (McBride 2019). These advances reflect the need for risk-based decision making in environmental management and are described in detail later in these guidelines.

For readers interested in alternatives to the traditional non-parametric method in this report, we recommend Helsel et al. (2020). We note that for river water quality trend assessment there is increasing use of a model called Weighted Regressions on Time, Discharge, and Season (WRTDS; Choquette et al., 2019; Hirsch et al., 2010, 2015; Woodward and Stenger, 2020). WRTDS is designed to allow for flexibility in representation of the long-term trend, seasonal components, and discharge-related components of the behaviour of the water-quality variables. WRTDS also characterises water quality trend directions and rates with uncertainty estimates. However, WRTDS is data intensive (it requires a minimum of 120 monthly samples and a mean daily flow record; Hirsch et al. 2015).

Practitioners may seek a formulaic approach that specifies a single correct way to undertake a trend assessment in each application and given a specific type of dataset. We do not believe this is possible. Because there is a range of statistical models that can be used to represent the behaviour of a variable over time, and because there are subjective elements in the model building process, there are likely to be several good approaches for a given application.

As a closing note to this section, we recognise that the field of statistical modelling, including trend assessment, is evolving. It's evolution is characterised by the development and provision of alternative methods (e.g., Hirsch et al. 2015) and debates about the advantages and validity of those methods. Common topics of debate include Bayesian versus frequentist methods (Bayarri and Berger 2004; Makowski et al., 2019) and significance tests (Bayarri and Berger 2004; Steidl 2006; Stephens et al. 2007). These debates and the different perspectives held by practitioners means that the methods we recommend in this guidance will not be unanimously accepted. Furthermore, the methods we recommend are not immutable; they will be modified over time and will eventually be obsolete, due to changes in environmental monitoring systems and progress in statistical science.

[5] Trend rates often referred to as magnitudes. We use trend rate because the Sen slope indicates direction and magnitude and its units indicate change per unit time (i.e., a rate).

[6] Note this is also referred to as the Theil–Sen estimator, the Kendall *S* robust line-fit method and the Kendall–Theil robust line.

3 Acquiring and compiling data

Most raw water quality data in New Zealand are held by individual regional and unitary councils and NIWA (via the National River Water Quality Network, NRWQN). When compiling datasets for national or regional applications, where possible, we recommend that data be accessed directly from online data servers (e.g., Hilltop and KiWIS servers) or council websites as the consistency of server formatting can ease the data compilation process. Alternatively, data may be compiled from spreadsheets provided by the monitoring agency. A compiled national water quality dataset is also available from Land Air Water Aotearoa (LAWA), collated from regional council and NIWA data. This dataset has been through a data grooming procedure (e.g., decisions on site inclusion and which water quality variables to include and combine) and is restricted to the previous ten-year period (updated annually). Therefore, the LAWA dataset represents a readily available subset of national water quality data, but may not be suitable for all purposes.

3.1 General data preparation

3.1.1 Purpose

Aggregating data and metadata from multiple sources inevitably results in inconsistencies in data structure, including dissimilar data matrices (i.e., arrangements of data in rows and columns) and multiple forms of variable names, geographic coordinates, date and time formats, units of measurement, and other metadata elements. Consistency in data structure is needed for sorting, searching, manipulating and displaying data, updating and exporting datasets, and carrying out statistical analyses. General principles for data organisation and recommendations for consistent metadata elements have been set out in several recent publications (e.g., Hart et al. 2016; Sprague et al. 2017; Wickham 2014).

Data errors are common in water quality and ecology datasets and can originate at many steps in the sampling, measuring and recording process. Among the most common causes of data errors are faulty or poorly calibrated field and laboratory sensors, sample contamination, calculation errors, taxonomic identification errors and data-entry errors (Davies-Colley et al. 2012, 2019; Rangeti et al. 2015; Rode and Suhr 2007). The resulting data errors include extreme values (for a given variable), negative values, zeros, non-numeric or alpha-numeric entries, and strings of repeated values. Good quality datasets with minimum errors are required to achieve robust trend assessments.

3.1.2 Methods

We provide recommendations for general good practice for preparing water quality time-series data in Appendix A. These recommendations are relevant not only for trend assessments, but other analyses and reporting based on water quality datasets. In particular, we recommend that data should be prepared with:

- consistent data structure;
- corrections to data errors; and
- correct location information, and variable names and units (and any other appropriate metadata).

Analysts should also be cognisant of any changes in sample collection and laboratory procedures, as trend assessments based on time-series with changes between incomparable methods are likely to

be unreliable (for further commentary, see the National Environmental Monitoring Standards (NEMS) for Water Quality)[7].

3.2 Formatting censored values

3.2.1 Purpose

For several water quality variables, true values are occasionally too low or too high to be measured with precision. The "detection limit" is the lowest value that can be reliably measured by an analysis and the "reporting limit" is the greatest value of a variable that can be reliably measured. Water quality datasets from New Zealand rivers and lakes often include phosphorus and ammoniacal nitrogen concentration measurements that are below detection limits (referred to as left-censored), and visual water clarity measurements that are above reporting limits (referred to as right-censored).

Censored values are managed in a special way by the traditional non-parametric trend assessment method (described in Section 5). It is therefore important that censored values are correctly identified in the data. Detection limits or reporting limits that have changed through the trend time period (often due to analytical changes) can induce trends that are associated with the measurement methods rather than actual changes in the variable. Methods are available to account for changes in censoring levels (see worked example in Section 7.4). This is another reason why it is important that censored values are correctly identified within datasets.

3.2.2 Method

Cases where variable values are below a detection limit or above a reporting limit are often indicated by data entries such as "<detection limit", "<DL" and "<0.01", all of which refer to values below laboratory detection limits. Data entries such as "> RL" and ">1500" refer to values above laboratory reporting limits.

Censored values should not be omitted prior to trend assessment, as this results in biased trend estimates. Similarly, censored values should not be replaced with fabricated values (e.g., 0.5×detection limit, 1.1×reporting limit) without retaining the information that the data were censored and the actual detection and reporting limits.

When using trend assessment packages, care needs to be taken to format censored values as required by those packages. For example, the LWP-Trends library and TimeTrends packages recognise censored values when they are formatted as positive numeric values preceded with an inequality symbol (e.g., "<0.025", ">1500"). Neither of these software packages recognise non-numeric entries such as "<DL", ">RL" or "BDL".

3.2.3 Commentary

Values in water quality datasets can be classified as censored when they are not and not classified as censored when they are (actually censored). These misclassification errors can lead to inaccurate trend assessments. In particular, assignment of values as censored when they are not has the potential to cause large inaccuracies. This is because one option in analysis of data with multiple detection limits (i.e., right-censored) is to set all observations whose values are below the highest detection limit to that value (i.e., the highest detection limit; see Section 5.3.3 for details). If a high value in the dataset is misclassified as censored, much of the information about changes in water quality over the time period will be lost when lower values are set to the 'highest' detection level.

[7] http://www.nems.org.nz/documents/

3.3 Defining seasons and time periods and ensuring adequately distributed data

3.3.1 Purpose

The trend assessment methods recommended in this guidance are robust to missing data, irregular sampling intervals and small datasets (see Section 5). However, there are several reasons why it is generally important to define seasons and time periods and to assess whether the observations are adequately distributed over time. First, because variation in many water quality variables is associated with the time of the year or "season", the robustness of trend assessment is likely to be diminished if the observations are biased to certain times of the year. Second, a trend assessment represents a time period; essentially a window of time that has a set starting date and duration. An assessment of the behaviour of a variable over the time period will be hindered if the observations are not reasonably evenly distributed across the time period. For these reasons, important steps in the data compilation process include specifying the seasons, the time period, and ensuring adequately distributed data.

3.3.2 Method

In most regional SOE monitoring programmes, sampling occurs at a set frequency, (e.g., monthly, quarterly). Trend assessment 'seasons' are generally specified to match these sampling frequencies because this maximises the temporal resolution of the available observations. Therefore, in the context of trend assessment, seasons are defined by months or quarters and there is generally one observation for each sample interval (i.e., each season within each year). Sampling frequency for some variables is annual (e.g., freshwater macroinvertebrates). In this case the 'season' is specified by the year.

In practice there are often deviations from the prescribed sampling routine that produce gaps in the record (i.e., missing observations for particular sample intervals). The traditional non-parametric method is statistically robust to such gaps. In addition, assessments can be performed robustly with as few as eight observations (see Section 5.2.1). Trend assessments can therefore be carried out using datasets with high proportions of gaps and small numbers of observations. The quantitative consequence of increasing gaps and reducing observations is to reduce statistical power. Another consequence of increasing the proportion of gaps is a reduction in the representativeness of the observations across the time period. To achieve both adequate statistical power and representativeness, it is desirable that sample size is as large as possible, gaps are limited, and the observations are adequately distributed over the sample intervals within the time period. It is therefore necessary to define the analysis time period so that the adequacy of the data can be evaluated. The time period is defined by nominating the start date and the duration. To avoid bias towards any particular season, we recommend that the time period is specified in complete years and the start date is therefore the beginning of the first year. Note that a "year" can be defined such that it commences in any month (i.e., not just a calendar year). The specification of years starting in months other than January is common in hydrological analyses (and in the annual water quality reports prepared by many councils); these non-calendar years are termed "water years".

There is no specific limitation on the proportion of gaps or requirements for the distribution of observations over the sample intervals. Whether the proportion of gaps is acceptable is a subjective decision that needs to be made by considering the objectives of each trend analysis. The decision is further complicated by the option of coarsening the sample intervals when there are many gaps. Coarsening the seasons (e.g., from months to quarters) can reduce the proportion of sample

intervals that are gaps in the trend assessment period, thereby increasing the representativeness of the observations. Coarsening is usually achieved by taking the median of the original observations within each of the coarsened seasons (e.g., the median of each monthly observation where these exist, within each quarter). The trade-off associated with coarsening the seasons is a reduction in the statistical power (i.e., confidence) and precision of a trend analysis.

The definition of non-calendar ("water") years has at least three benefits. First it can enable an analysis to be more up to date by allowing the time period to end at any date (e.g., the date of the last observation) rather than the end of the last complete calendar year in the record. Second, it can allow more appropriate beginning and ending of sampling intervals. For example, macroinvertebrate monitoring might be carried once every summer, and sampling dates may range from December to February; using a calendar year could inadvertently pool two samples into a single year, leaving no samples in the next year. In this case, a "water-year" would better reflect the intended annual sampling. Third, it can allow seasons (if these are defined by coarsening the sampling interval) to be aligned with the expected seasonal response of the observations.

One deviation from the prescribed sampling routine is the collection of more than one observation in a single sample interval (e.g., two observations within a month). In this situation, taking the median within each sample interval is generally the default approach. As with coarsening the sample interval, the effect of this is a reduction in statistical power and precision. There is an alternative approach that utilises all the observations (and therefore maximises power and precision). The alternative approach assumes that the observations within a sample interval are independent but treats these as occurring at the same time (referred to as ties in time). If the samples are not independent, the median of the observations should be used. The worked example in Section 7.3 demonstrates the implementation of some of these subjective decisions around season definition for a site with an irregular monitoring history.

Another common deviation from the prescribed sampling routine occurs when there is a change in sampling interval within the time series. In this situation, the coarser sampling interval should be used to define seasons. For the part of the record with a higher sampling frequency, the observations in each season should be defined by taking the observation closest to the midpoint of the coarser season. The reason for not using the median value in this case is that it will induce a trend in variance, which will invalidate the null distribution of the test statistic (Helsel et al. 2020).

Choosing an appropriate trend assessment time period is dependent on the objective of the application. Time periods may be selected to coincide with planning cycles (e.g., to assess whether water quality changes between the time that a regional plan took effect, and the next plan change), to assess water quality changes between the commencement of management actions or land-use changes and the present day, or to determine whether water quality changes are correlated with temporal environmental fluctuations. In practice, the selection of an appropriate trend period length is a trade-off between several factors. Shorter periods are appropriate for assessing the impacts of recent changes in land and water management or in environmental conditions. However, for a given sampling frequency, statistical power decreases with reducing time period duration because there are fewer observations. We note also that shorter time periods are more likely to be influenced by climate variation, which can confound signals that may be of primary interest such as those associated with changes in resource use or management (see Appendix D). While the time period for local applications can be tailored to best answer the assessment objectives, (e.g., choosing a time period prior to and after the implementation of a specific management action), for regional and national trend assessments the same trend assessment period(s) should be applied across all site

variable combinations. Choosing an appropriate time period for regional and national applications involves further trade-offs, between maximising the trend period duration (to obtain trend assessments that are associated with larger sample sizes and therefore greater confidence) and maximising the number of sites (to improve spatial coverage and representation) that can be included in the assessment. In the most recent national trend assessments for rivers and lakes (Larned et al. 2018a, b), three time periods were used 10, 20 and 28 years; for each of nine water quality variables, the number of sites that could be included decreased as the time period lengthened.

For a local application, decisions concerning the definition of seasons and tolerance of gaps should be based on maximising the information represented by the observations. However, in regional and national applications it is important that trends are commensurate in terms of their statistical power and representativeness of the time period. In these applications, the usual practice is to define consistent time periods so that all sites are subjected to the same conditions (i.e., equivalent environmental and management conditions). Regional and national applications also require analysts to define the acceptable proportion of gaps and how these are distributed across sample intervals so that the reported trends are assessed from comparable data. The acceptable proportion of gaps and representation of sample intervals by observations within the time period are commonly referred to as site inclusion or filtering rules, or 'completeness criteria' (e.g., Larned et al. 2018a, b).

There are no universally agreed data requirements or filtering rules for regional and national applications. The selection of these rules is complicated by a general trade-off: more restrictive rules increase the robustness of individual trend analyses but generally exclude a larger number of sites, thereby reducing spatial coverage. In most cases, this trade-off is also affected by the trend period. Progressive expansion of freshwater monitoring effort in New Zealand over the last two decades means that data are available for a larger number of sites over short and more recent time periods.

The application of filtering rules for variables that are measured at quarterly intervals or less requires two steps. First, retain sites for which observations are available for at least X% of the years in the time period. Second, retain sites for which observations are available for at least Y% of the sample intervals. For variables that are measured or determined annually such as the Macroinvertebrate Community Index (MCI), the filtering rules are applied by retaining sites for which values are available for at least X% of the years in the trend period. In many recent national and regional trend analyses, where trend periods of 10 and 20 years have been reported the values for X and Y have been either 80% or 90% (e.g., Larned et al. 2018a, b).

In the recent national analyses (Larned et al. 2018a, b), the definition of seasons was flexible in order to maximise the number of sites retained. If the distribution of observations for a given site did not meet a requirement for monthly sampling intervals, a coarsening of the data to quarterly seasons was applied and the distribution of observations was reassessed. It is noted that this decision implies a tolerance of variable levels of statistical power and representativeness across the sites that were represented in the analysis. It is also noted that when a variable is sampled annually, the only filtering rule concerns the distribution of observations across years.

3.3.3 Commentary

Ultimately choices concerning trend time periods and seasons need to consider the objectives of the trend assessment, the constraints associated with the available data and, for regional and national applications, the trade-offs between statistical power and site numbers.

Site inclusion rules used for trend assessments are subjective, and published reports of water quality and ecology trends across multiple sites have employed a wide range of rules (e.g., Mast 2013; Myers and Ludtke 2017; Oelsner et al. 2017; Sprague and Lorenz 2009). We know of no cases where the effects of different inclusion rules have been evaluated, including comparisons of stringent versus lenient rules.

4 Accounting for confounding factors

Some of the variation in a water quality variable of interest may be associated with factors that confound its relationship with time (i.e., the trend). Increased variation associated with confounding factors reduces the detectability of trends. The effect of two types of confounding factors on the variability of water quality observations can be accounted for prior to conducting the trend assessment: covariates and seasonality. In this section, we consider methods for pre-processing time series of water quality observations to ensure that covariates and seasonality are accounted for appropriately.

4.1 Covariate adjustment

4.1.1 Purpose

Water quality trend assessments are used to characterise relationship between water quality variables and time. In this context, a range of different factors can be considered as "covariates". A covariate is a variable that is quantitatively related to the observations, and this relationship confounds or obscures the water quality variable – time relationship that of primary interest. Statistical analysis can be used to remove the influence of the covariate on the observations. For river data, a common covariate is flow and this statistical analysis is called "flow adjustment". The same principle is applied to other environments (e.g., lakes, estuaries and groundwater) and other covariates (e.g., wind, tide and barometric pressure), therefore the more general term is covariate adjustment. This section discusses flow adjustment in detail, but the principles are relevant to any other form of covariate adjustment.

Where water quality observations are made in a river and are associated with solutes or particulate matter (e.g., concentrations, optical measures such as clarity and turbidity) some of the variation is usually associated with the river flow (i.e., discharge) at the time the observation was made. The observed values can vary systematically with flow rate due to flow-dependent physical and biological processes (e.g., dilution, microbial and plant metabolism, sediment and bacterial entrainment from channel deposits; Smith et al. 1996). Different processes may dominate at different sites so that the same water quality variable can exhibit positive or negative relationships with flow. Some water quality variables can be associated with a combination of dilution and wash off with increasing flow. For example, a portion of the *E. coli* load may come from point sources discharges such as sewage treatment plants (dilution effect), but another portion may be derived from surface wash-off. Increasing flow in this situation may result in an initial dilution at the low end of the discharge range, followed by an increase with discharge at higher values of discharge.

Covariate adjustment has two purposes. First, it theoretically increases the statistical power of the trend assessment (i.e., increase the confidence in the estimate of direction and rate of the trend) by removing some of the variability that is associated with the covariate. Second, it removes any component of the trend that can be attributed to a trend in the covariate (e.g., a trend in the flow on sample occasions such as increasing or decreasing flow with time). Whether it is appropriate to

undertake covariate adjustment depends on the objectives the trend assessment; this is discussed in the commentary below.

4.1.2 Method

Covariate adjustment generally involves fitting a model that describes the relationship between the observations and the covariate, and then using the residuals of that model instead of the original observations in the subsequent trend assessment step. In the description of the covariate adjustment method below, we have focused on flow adjustment (i.e., removing the influence of flow at from water quality observations made in a river). However, the principles and the method are the same for any other type of covariate adjustment.

To flow adjust water quality variables, it is necessary to have estimates of flow associated with each water quality observation. We note that in New Zealand, water quality monitoring is often carried out at locations that do not have flow recorders. Ideally, flow is directly measured at the water quality monitoring site, but flow estimates can also be derived from direct measurements made at nearby sites (appropriately calibrated to the observation site), or flow predictions from an accurate hydrological model. In New Zealand rivers, flow is often measured at 15-minute intervals and the time of water quality sample collection is often recorded. Therefore, it may be possible to describe the relationship between the water quality observations and flow using data derived from the high temporal resolution data, or it may be that better relationships can be derived from coarsening the temporal resolution of the flow data (e.g., from 15-minute to daily mean). Choice of temporal resolution is best decided by experimentation with the aim of determining the resolution that yields the best relationship.

One or more regression models are fitted to describe the relationship between the water quality observations and flow. We recommend that censored water quality values are retained for this model fitting step. The approach to this taken by the LWPTrends and TimeTrends computer programmes is to use the raw values (i.e., the numeric component of the censored values) multiplied by a factor. The factor is generally 0.5 for detection limit censoring and 1.1 for reporting limit censoring. The reason for this approach is that the censored values represent the lower and upper range of the observations and therefore help to define the shape of the relationship being modelled. The removal of these values will often greatly diminish the definition of the relationship between the observations and flow. Since flow will only ever explain a proportion of the variation in the observations, inaccuracies associated with replacing censored values are likely to be of lesser importance than the loss of information about the shape of the relationship being modelled. Alternative regression models that are commonly applied in covariate adjustment are: linear regression, log-log regression, locally estimated scatterplot smoothing (LOESS) and generalised additive models (GAM).

The next step is to select the best water quality-flow model from the alternatives. We strongly recommend that the model selection step is not automated and includes inspecting the data and the fitted models. This is because unsupervised fitting of regression models to relationships between water quality observations and flows can result in the selection of unrealistic and therefore inappropriate models (Figure 4-1 and examples in Section 7). We recommend that expert judgement is used to choose the most suitable model based at least three considerations: (1) homoscedasticity (constant variance) of the regression residuals, (2) model goodness of fit measures and (3) plausibility of the shape of the fitted model. We note that model goodness of fit measure alone should not be relied on because they can indicate good model performance but describe unrealistic

relationships. This is particularly likely when more flexible models are used such as LOESS and GAM models and therefore these models should be used with caution.

When the relationship between flow and the water quality variable is poor, it is appropriate to conclude that that there is not a systematic relationship between the observations and flow. In this case, no model is selected, and no flow adjustment is performed. The trend assessment in this case should be performed on the raw data. Choosing not to flow adjust should take into consideration the balance between the potential to reduce variance in the observations, and the risk of selecting an implausible/inappropriate model of the relationship between the observations and flow.

It should not be surprising to conclude that the relationship between flow and the water quality variable is poor for some site variable combinations. In a recent national trend assessment (Larned et al., 2018a), flow adjustment was performed using only log-log models for which the R^2 value for the regression was greater than 20%. Where model R^2 value was less than 20%, no flow adjustment was performed. This choice of model and R^2 value was determined based on a "calibration" to expert judgement. It was concluded that log-log regression models produced realistic relationships for which water quality observations decreased or decreased monotonically with increasing flow. In contrast, the LOESS and GAM models tended to fit unrealistic relationships (Figure 4-1). For ten-year trends in the Larned et al. (2018a) study, only 14% of site/variable combinations with flow observations or estimates were flow adjusted. Some variables were more frequently judged to have relationships with flow including total nitrogen, nitrate-nitrogen and turbidity, while other variables were judged to rarely vary systematically with flow. We note that many of the sites in that study had flow estimates derived from a hydrological model. Uncertainties associated with the flow predictions may have reduced the frequency with which relationships between water quality observations and flow were identified.

Figure 4-1 shows the relationship between water quality observations and flow for a site from the Larned et al. (2018a) study based on four different models. In this example, the log-log regression model has a lower R^2 than the alternative GAM and LOESS models. However, it was concluded by Larned et al. (2018a) that the additional curvature in the alternatives to the log-log regression model could not be reasonably justified by the data and in some cases implied unrealistic relationships (see for example the Figure 4-1 LOESS0.5 model).

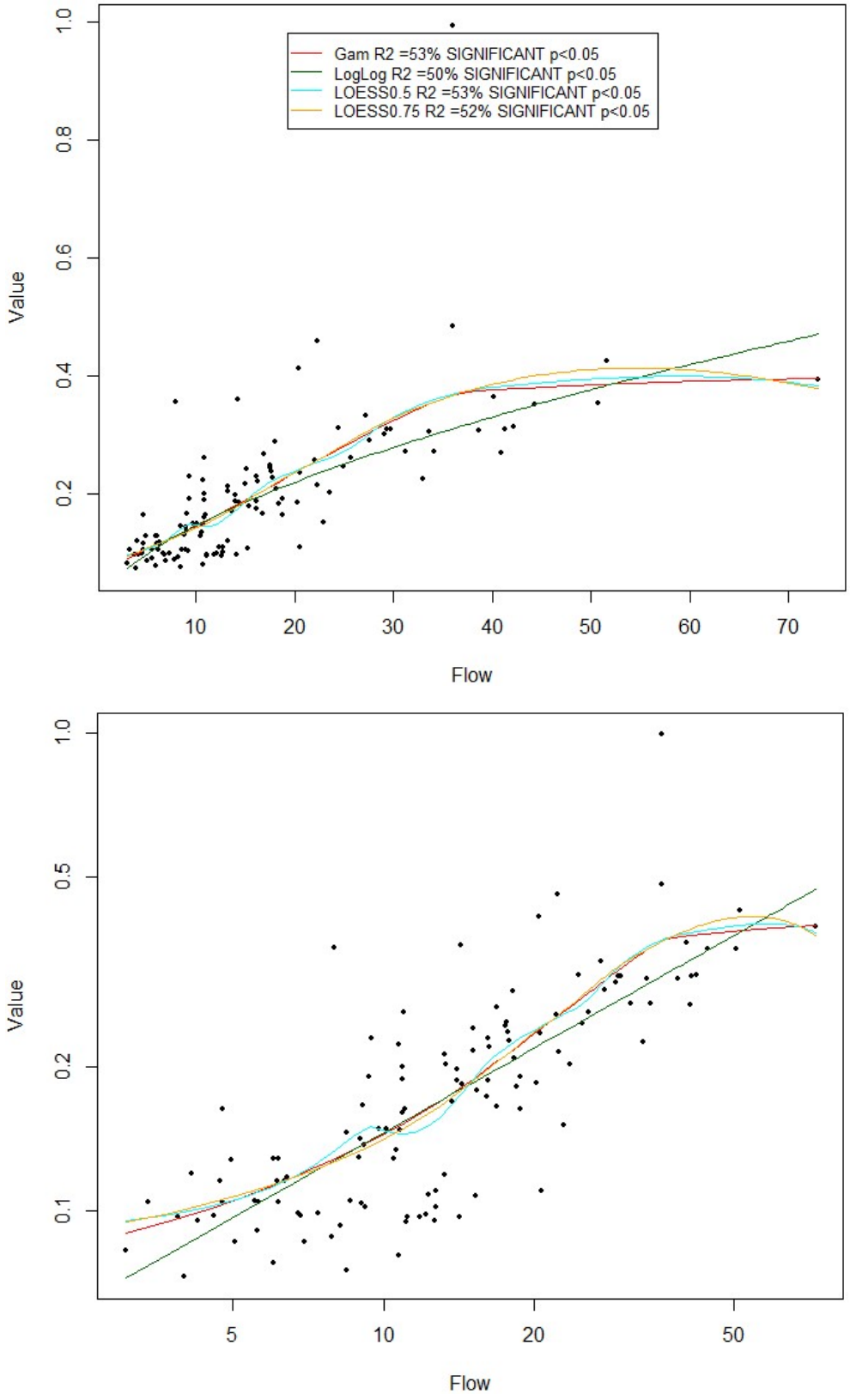

Figure 4-1: Example of the relationship between water quality observations and flow showing alternative fitted models. The top plot shows data plotted on unscaled axes, and the bottom plot shows data plotted on log-transformed axes. The LOESS0.50 model is a LOESS model with a 50% span (more flexible), whereas the LOESS0.75 model is a LOESS model with a 75% span (relatively less flexible).

4.1.3 Commentary

Covariate adjustment involves modelling the relationship between water quality observations and flow, precipitation or other potential confounding variables. This model is necessarily a simplification of reality and introduces subjective decisions about which (if any) model is the most appropriate. It is likely that different analysts will make different judgements concerning covariate adjustment and this will lead to variation in the results of their trend assessments.

Whether covariate adjustment is appropriate depends on the aim of the trend assessment. If the aim is to understand whether a management action has affected water quality over time, then the contribution of a naturally varying covariate to the trend is a confounding factor and covariate adjustment will increase confidence in the trend assessment. In contrast, if the aim of the assessment is to quantify the water quality trend that actually occurred, then covariate adjustment may not be applicable. An example where quantification of the actual (unadjusted) trend might be required would be where a biological change has occurred in a stream and there is interest in whether this was associated with changes in water quality variables. In this case, relationship between the water quality trend and the biological response is of primary interest and it is not important that a component of the water quality trend was due to a covariate.

For regional and national applications, trend assessments pertaining to multiple sites are often aggregated to describe broad-scale patterns in trends (see Section 2.2). When aggregating trends to describe broad-scale patterns, it is important to use consistent methods across sites. Since relatively few river water quality monitoring sites in New Zealand are also used for flow monitoring, requiring flow adjustment can significantly reduce site numbers and the representativeness of data with which to describe aggregate trends. Therefore, a subjective decision needs to be made concerning the trade-off between site numbers and the potential increase in confidence that might be achieved through flow adjustment. In our experience, conclusions drawn from aggregated trend analyses made using flow-adjusted and non-flow adjusted water quality data are similar (e.g., Fraser and Snelder 2018 and see Appendix B).

The covariate adjustment method described above is applied when using the traditional non-parametric method recommended in these guidelines (see Sections 2.3 and 4). Alternative trend assessment procedures can accommodate covariates in a single model. For example, parametric multiple linear regression can be used to fit a model to the relationship between the observed water quality data and time that includes non-periodic covariate values and a periodic representation of season. For further examples and a discussion of alternative trend assessment methodologies and their approaches to including covariates, see Helsel et al. (2020).

4.2 Seasonality assessment

4.2.1 Purpose

For many site/variable combinations, season will explain a considerable amount of variation in the water quality observations. As described in Section 3.3, seasons are defined as any equal subdivision of a year into sample intervals; there can be between 2 and 12 seasons in a year. Note that for annually sampled data (e.g., MCI), the following discussion of seasonality does not apply.

In cases where seasons are a significant source of variability, accounting for systematic seasonal variation should increase the statistical power of the trend assessment (i.e., increase confidence in the estimates of direction and rate of the trend). The purpose of a seasonality assessment is to identify whether seasons explain variation in the observations. If this is shown to be the case, then it

is appropriate to use the seasonal versions of the traditional non-parametric method recommended by these guidelines at the trend assessment step (Section 5).

4.2.2 Method

Because there are often regular seasonal fluctuations in river flow, we recommend that flow adjustment is performed first (as this may account for some or all of the water quality seasonal response), followed by a seasonality assessment. The seasonality of the water quality observations (either raw or flow adjusted) is then evaluated based on user defined seasons, as described in Section 3.3. The recommended method of seasonality assessment is the Kruskall-Wallis multi-sample test for identical populations. This is a non-parametric ANOVA that determines the extent to which season explains variation in the water quality observations. Following Hirsch et al. (1982), we recommend that site/variable combinations are evaluated in terms of seasonal fluctuations using the *p*-value from the Kruskall-Wallis test with α=0.05. For those sites/variable combinations that are determined to fluctuate seasonally, subsequent trend assessments should follow the "seasonal" variants described in Section 5.2. For regional and national applications, a fixed value of α should be applied across all site/variable combinations. For local applications, there is more flexibility in selecting an appropriate value of α for the specific site (see commentary below).

Figure 4-2 and Figure 4-3 show distributions of water quality observations by season. The Kruskall-Wallis tests performed on these data indicate that the site/variable combinations shown in shown in Figure 4-2 and Figure 4-3 are seasonal and non-seasonal, respectively. The plots provide a useful visual corroboration of the seasonality assessment.

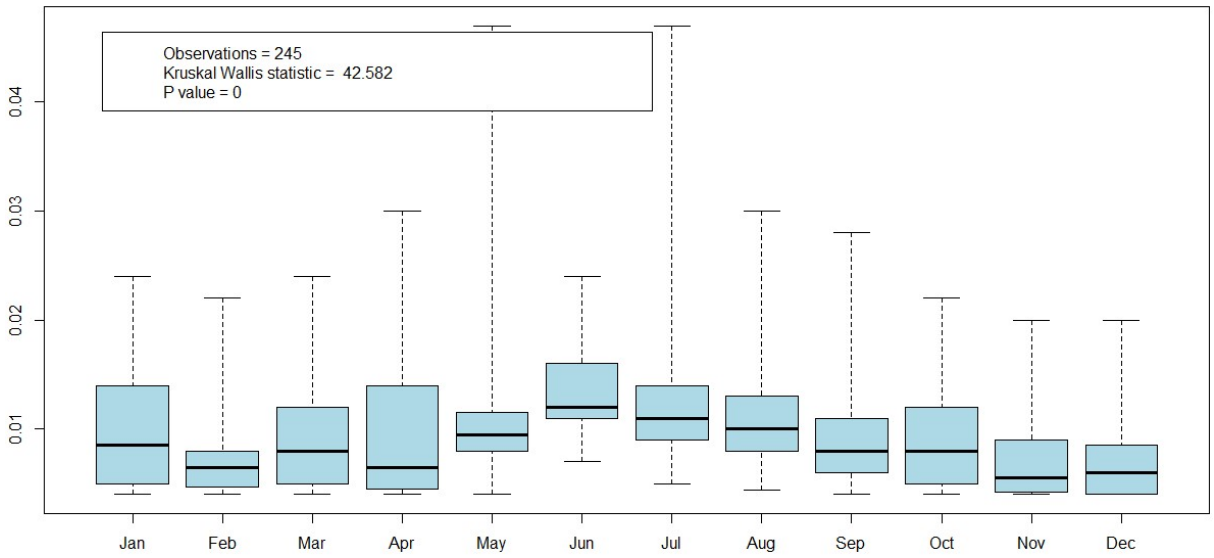

Figure 4-2: Plot showing the distributions of water quality observations within seasons. For this example, the Kruskall-Wallis test was significant at α=0.05.

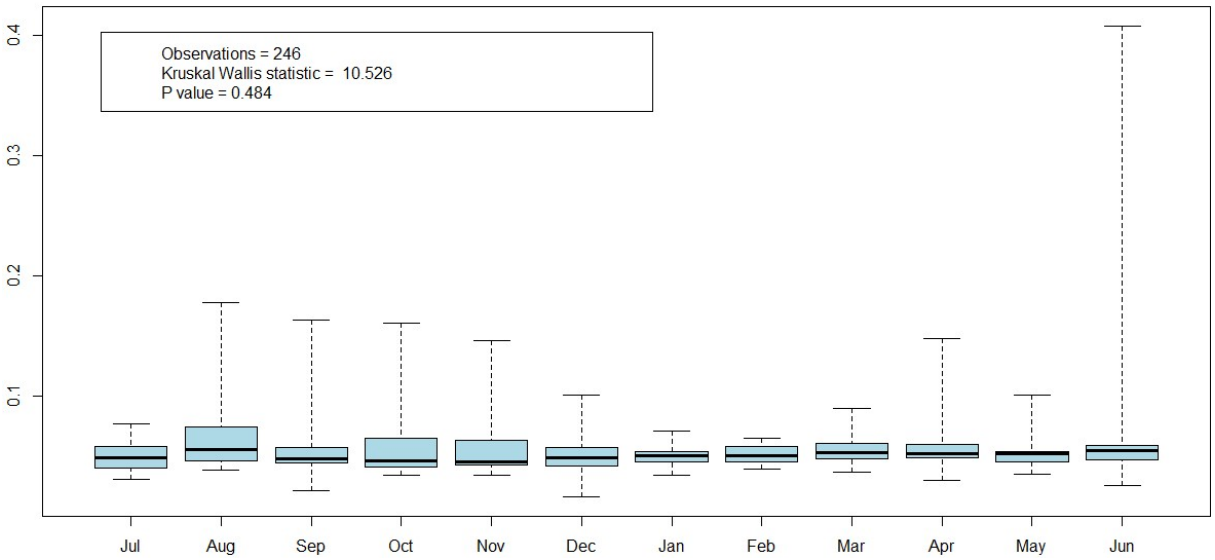

Figure 4-3: Plot showing the distributions of water quality observations within seasons. For this example, the Kruskall-Wallis test was not significant at α=0.05.

4.2.3 Commentary

The choice of α in Kruskall-Wallis test, is subjective and a value of 0.05 is associated with a very high level of certainty the data exhibit a seasonal pattern. However, in our experience there are generally diminishing differences between the seasonal and non-seasonal trend assessments for *p*-values values larger than 0.05 (see Appendix C).

Seasonality differs across sites and variables. In the most recent national trend assessment for rivers (Larned et al. 2018a), 64% of all site/variable/duration combinations (excluding MCI) were found to be seasonal (at α=0.05). The proportion of sites for which observations were determined to be seasonal varied between variables; the variables most frequently found to fluctuate seasonally were total nitrogen and nitrate-nitrogen (90% and 82% of sites, respectively), and ammoniacal nitrogen was the least frequently seasonal variable (38% of sites).

5 Trend assessment

5.1 Purpose

As already established, the primary purpose of trend assessment is to evaluate the direction (i.e., increasing or decreasing) and rate of the change in the central tendency of observed water quality values over a specified time period. Because the observations are subject to random fluctuations and are only a sample of the behaviour of the variable over the period of analysis, there is uncertainty about both the assessed direction and rate of the change. Therefore, as well as determining the trend direction and rate, trend assessments evaluate the uncertainty of these determinations. Trend direction and trend rate are both relevant and important. Focusing solely on trend direction may limit the value of a trend assessment because, while the direction provides information about the existence of a trend, it provides no information about the magnitude of the trend and therefore provides an insufficient basis to assess its importance or relevance.

5.2 Recommended methods

5.2.1 Trend direction assessment

Trend direction and the confidence in trend direction are evaluated using either the Mann Kendall assessment or, where seasonality is identified in the observations (see Section 4.2), the Seasonal Kendall assessment. Although the non-parametric Sen slope regression also provides information about trend direction and its confidence, the Mann Kendall assessment is recommended because it handles censored values more robustly. However, Sen slope regression is the recommended method for assessing the trend rate (see Section 5.2.2).

The Mann Kendall assessment requires no *a priori* assumptions about the distribution of the data but does require that the observations be randomly sampled and independent (no serial correlation) and that there is a sample size of ≥ 8. Both the Mann Kendall and Seasonal Kendall assessments are based on calculating the Kendall *S* statistic, which is explained diagrammatically in Figure 5-1.

The Kendall *S* statistic is calculated by first evaluating the differences between all pairs of water quality observations (Figure 5-1, A and B). Positive differences are termed 'concordant' (i.e., the observations increase with increasing time) and negative differences are termed 'discordant' (i.e., the observations decrease with increasing time). The Kendall *S* statistic is the number of concordant pairs minus the number of discordant pairs (Figure 5-1, C1). The water quality trend direction is indicated by the sign of *S* with a positive or negative sign indicating an increasing or decreasing trend, respectively (Figure 5-1, C2). In the special case that Kendall's *S* is equal to zero, the trend is pronounced "indeterminate" (i.e., the trend direction cannot be determined).

The seasonal version of the Kendall *S* statistic *S* is calculated in two steps. First, for each season, the *S* statistic is calculated in the same manner as shown in Figure 5-1 but for data pertaining to observations in each individual season. Second, *S* is the sum of values over all seasons ($S = \sum_1^n S_i$), where S_i is the number of concordant pairs minus the number of discordant pairs in the i^{th} season and n is the number of seasons. The variance of *S* is calculated for each season and then summed over all seasons.

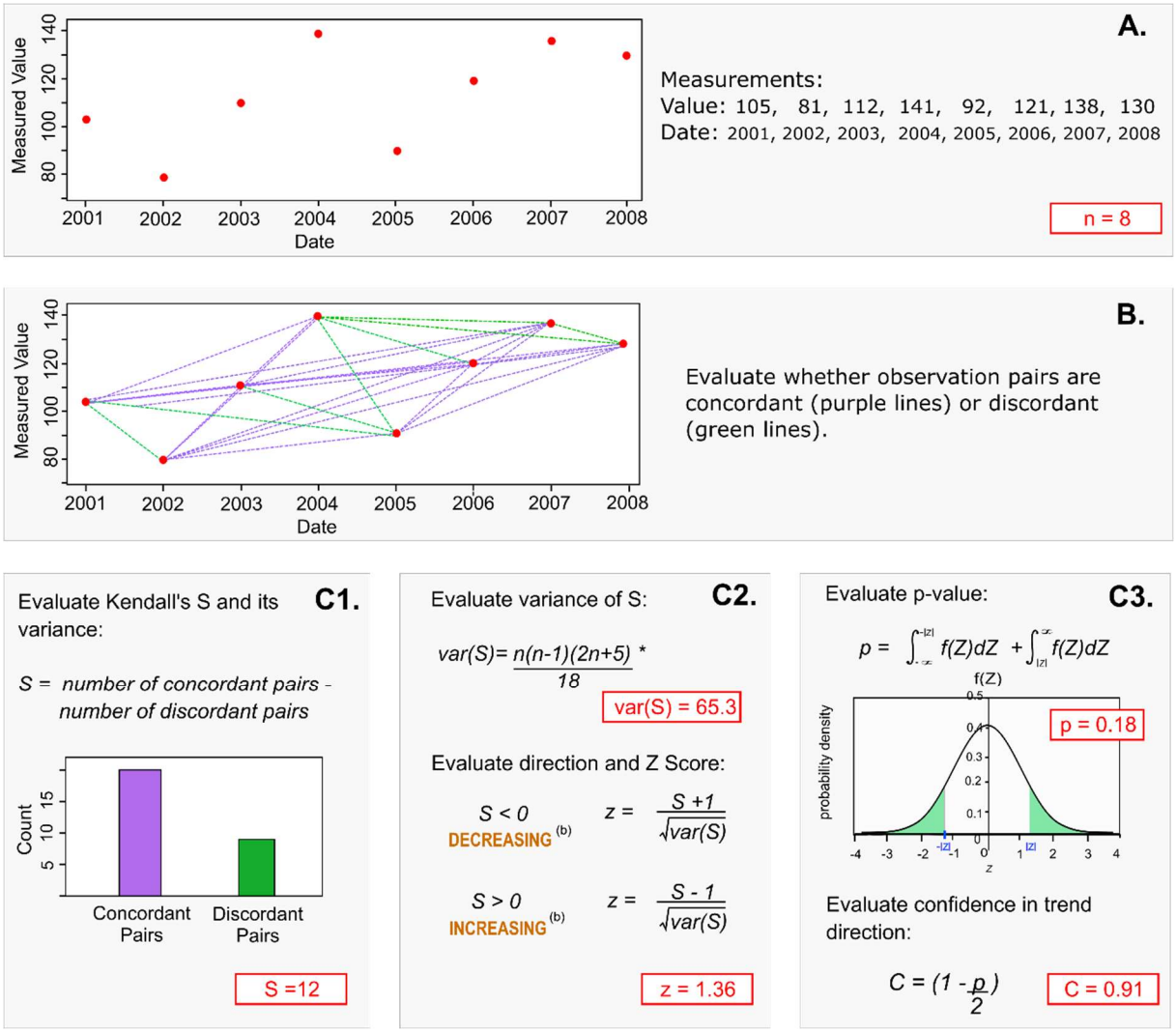

Figure 5-1: Pictogram of the steps taken in the trend direction assessment to calculate the Kendall *S* statistic and its *p*-value which is used to characterise confidence in trend direction. Notes: [a] the calculation of the variance in *S* has some adjustments to account for ties (numerically equal values) and censored values. Details of these adjustments can be found in (Helsel 2005, 2012). [b] There is a third alternative, where *S*=0. In this case the *p*-value is 1 and *C* is 0.5, and the trend direction is classified as "indeterminate". Values of *S* equal to -1 or 1 will also result in a Z value of 0, a *p*-value of 1 and a *C* value of 0.5 and the trend direction is similarly classified as "indeterminate".

The sign (i.e., + or -) of the *S* statistic calculated from the sample represents the best estimate of the population trend direction but is uncertain (i.e., the direction of the population trend cannot be known with certainty). Confidence in the calculated *S* statistic in Mann's (1945) original trend test and subsequent extensions by Hirsch et al. (1982) is based on null hypothesis significance testing (NHST). The null hypothesis is that there is no trend (or the trend is zero). Mann (1945) showed that the *S* statistic was normally distributed, and *S* could be converted to Z-scores based on the formula shown in Panel C3 of Figure 5-1. This model describes the expected range of values of *S* if they were repeatedly calculated from many random samples, each with the same number of observations as the actual water quality data and drawn from a population with no trend (i.e., the null hypothesis was true). The derived distribution allows the evaluation of the probability of observing a value of *S* that is as least as extreme as the observed value, if the null hypothesis was true. That probability is the *p*-value and is shown by the areas of the distribution that are cut off at the calculated value of *S*.

Note that for a two-tailed test, the *p*-value includes the area defined by both tails because the test is concerned with the extremity of the value and does not consider if *S* is positive or negative.

NHST is based on rejection of the null hypothesis when the *p*-value is smaller than an arbitrary value known as the significance level or alpha value (α). Confidence is indicated by two categories, significant and non-significant, which correspond to two tailed *p*-values ≤ $\alpha/2$ or *p*-values > $\alpha/2$, respectively. The *p*-value represents the probability of observing a value of *S* that is at least as extreme as that calculated from the sample if the null hypothesis were true. Recently McBride (2019) highlighted several criticisms of the use of NHST in water quality trend analysis that are discussed briefly below and in more detail in Section 5.3.1. Two of these criticisms are the rationale for our recommendation to use an alternative, continuous measure of confidence, which we call confidence in the trend direction (C). Confidence in the trend direction is calculated as:

$$C = 1 - {}^{p}/_{2}$$

where *p* is the *p*-value calculated for either Kendall *S* or its seasonal variant (Mann 1945; Hirsch et al. 1982).

The value C can be interpreted as the probability that the sign of the calculated value of *S* indicates the direction of the population trend (i.e., that the calculated trend direction is correct). The value C ranges between 0.5, indicating the sign of *S* is equally likely to be in the opposite direction to that indicated by the true trend, to 1, indicating complete confidence that the sign of *S* is the same as the true trend. Further discussion of the derivation of *C*, the benefits of *C* over traditional NHST significance testing, as well as the equivalence of *C* with a Bayesian measure of confidence, is presented in Section 5.3.1.1.

The difference between the recommended confidence in the trend direction compared to NHST is demonstrated in the following example. Consider two trend assessments A and B with positive *S* values and *p*-values of 0.04 for A and 0.14 for B. Under a classic null-hypothesis test with α = 0.05, we would say that for A the null-hypothesis was rejected at the 95% confidence level and for B the null-hypothesis was not rejected at the 95% confidence level. Following the method that we recommend, the conclusion for assessment A is a positive trend with 98% confidence in the direction, and for assessment B, a positive trend with 93% confidence in the direction.

As the size of the sample (i.e., the number of observations) increases, confidence in the trend direction increases. When the sample size is very large, *C* can be high, even if the trend rate is very low. It is important therefore that *C* is interpreted correctly as the confidence in direction and not as the importance of the trend. As stated at the beginning of this section; both trend direction and the trend rate are relevant and important aspects of a trend assessment.

5.2.2 Assessment of trend rate

Trend rate and the confidence in trend rate are evaluated using non-parametric Sen slope regressions of water quality observations against time. The Sen slope estimator (SSE; Hirsch et al. 1982) is the slope parameter of a non-parametric regression. The SSE is calculated as the median of all possible inter-observation slopes (i.e., the difference in the measured observations divided by the time between sample dates; Figure 5-2).

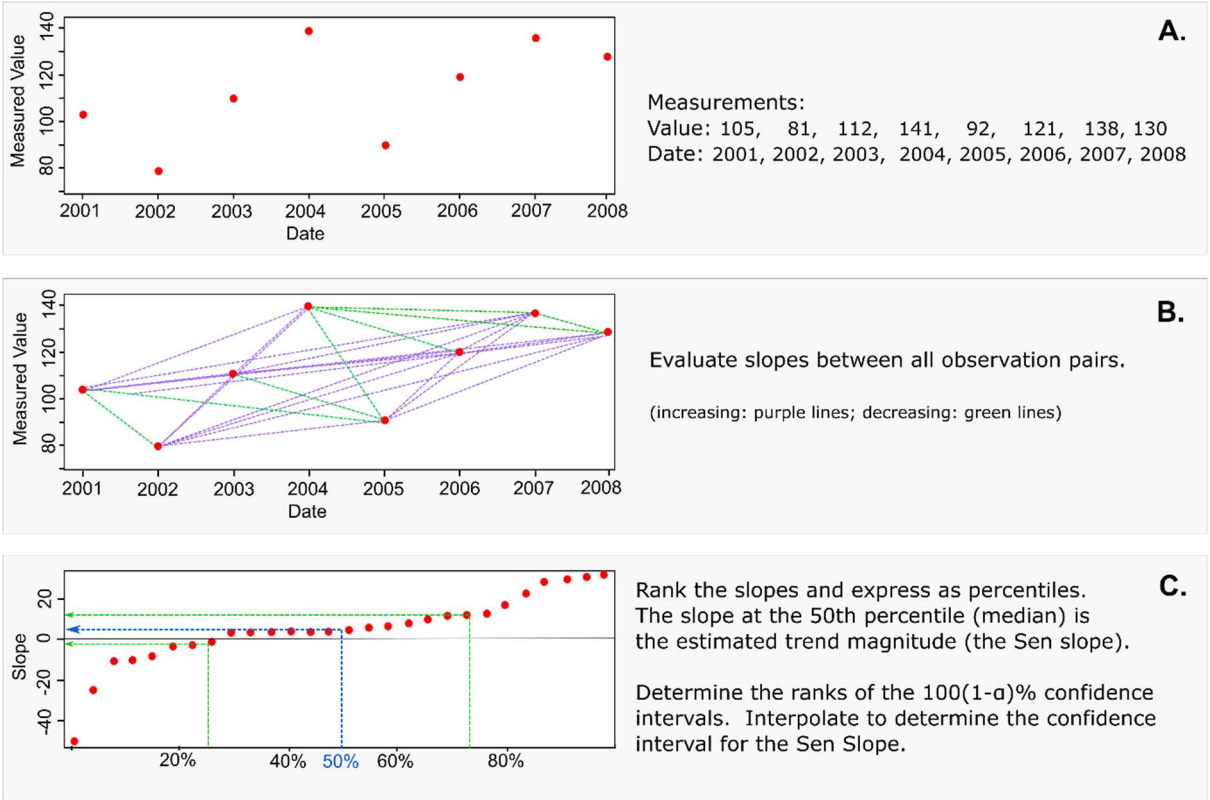

Figure 5-2: Pictogram of the calculation of the Sen slope, which is used to estimate trend rates. The seasonal Sen slope estimator (SSSE) is calculated in two steps. First, for each season, the median of all possible inter-observation slopes is calculated in same manner as shown in Panel B, but for data pertaining to observations in each individual season. Second, SSSE is calculated as the median of the seasonal values.

Uncertainty in the assessed trend rate can be expressed by calculating its confidence interval following a methodology outlined in Helsel and Hirsch (1992). To calculate the 100(1-α)% two-sided symmetrical confidence interval about the fitted slope parameter, the ranks of the upper and lower confidence limits are determined, and the slopes associated with these observations are applied as the confidence intervals.

The inter-observation slope cannot be definitively calculated between any combination of observations in which either one or both observations comprise censored values. Therefore, it is usual to remove the censor sign from the reported laboratory value and use just the 'raw' numeric component (i.e., <1 becomes 1) multiplied by a factor (such as 0.5 for left-censored and 1.1 for right-censored values). This ensures that in the Sen slope calculations, any left-censored observations are always treated as values that are less than their 'raw' values and right censored observations are always treated as values that are greater than their 'raw' values. The inter-observation slopes associated with the censored values are therefore imprecise (because they are calculated from the replacements). However, because the Sen slope is the median of all the inter-observation slopes, the Sen slope is unlikely to be affected by censoring when a small proportion of observations are censored. As the proportion of censored values increase, the probability that the Sen slope is affected by censoring increases.

Helsel (1990) estimated that the impact of censored values on the Sen slope is negligible when fewer than 15% of the values are censored. However, this is a rule of thumb and is not always true. Depending on the arrangement of the data, a small proportion of censored values (e.g., 15% or less) could affect the computation of a Sen slope (Helsel 2012). Alternatively, there may be a larger

proportion of censored values, but if the detection limit is small relative to the median of the non-censored values, the impact of censored values on the Sen slope estimate is likely to be small. In the past, a proportion of censored values greater that 15% was used as a filtering rule (i.e., the trend was not assessed for the corresponding site/variable combination; see Section 3.2.2), but this is no longer recommended. Our current recommendation is to use all available observations, whether censored or not. The LWPTrends package provides an 'analysis note' with all calculated Sen slopes to indicate whether the reported Sen Slope is likely to be affected by censored values. Sen slope calculations that are affected by censored values indicate that the trend rate is smaller than can be detected given the detection limit. The precision of any Sen slope that has been calculated from at least one censored value is equal to the detection limit divided by the length of the time period (years). It is noted that the precision of Sen slopes that are not affected by censored values is equal to the precision with which the observations are reported divided by the length of the time period (years). In practice, these two levels of precision have the same or similar magnitude.

5.3 Commentary

5.3.1 Confidence in trend direction

The approach to trend assessment recommended in this guidance is based on a long-standing traditional non-parametric approach to water quality trend assessment. However, we recommend a departure from the use of NHST as a method for assessing confidence in the assessment of trend direction in favour of the continuous measure of confidence in trend direction (C). While this change is small in terms of implementation (and results), it is a conceptual shift that recognises the widespread criticisms of statistical significance testing (Greenland et al. 2016; Helsel et al. 2020; McBride 2019; McBride et al. 2014).

The continuous measure of confidence in the assessed trend direction addresses two issues that were raised by McBride (2019). The first issue is a non-significant trend (i.e., failure to falsify a null hypothesis that "the trend is zero" for some nominated alpha value). This conclusion is often interpreted as evidence that the null hypothesis is true and therefore that the trend is zero or "stable". This is an incorrect conclusion; a 'large' *p*-value (i.e., $p > 0.05$) indicates only that the data are not unusual if the null hypothesis were true, and none of the other assumptions were violated (Greenland et al. 2016). However, the same data would also not be unusual under many other hypotheses.

The second issue highlighted by McBride (2019) is associated with the arbitrary classification of trends as significant or non-significant based on the significance level. The significance level (α) represents the probability of categorising a trend as non-significant when it is in fact significant (i.e., of rejecting the null hypothesis when in fact it is true). Generally, α is set at a low value (e.g., 0.05) to minimise the risk of incorrectly rejecting the null hypothesis (also known as a Type I error). However, from a management perspective, the acceptable Type 1 error risk should not be defined by an arbitrary statistical rule. This risk should include consideration of the probability of incorrectly assessing the trend direction, its rate and its consequence (such as impacts to environmental values)[8]. Therefore, the acceptable risk of a Type 1 error is a normative decision that should consider the importance of the environmental values that may be under threat from a trend, and the magnitude of the trend rate. A further problem with classifying trends for site/variable combinations as either significant or non-significant is that there is a loss of information about trend direction

[8] This is consistent with Section 3.20 of the NPS-FM 2020 which directs councils to "take action to halt or reverse degradation", that is "proportionate to the likelihood and magnitude of the trend, the risk of adverse effects on the environment".

compared to the alternative continuous measure of confidence in trend direction that we recommend. This is particularly relevant to regional and national applications involving trend assessments performed over many sites. The alternative continuous measure of confidence in trend direction can highlight a collective tendency in trend direction across sites even when numerous site-specific trends are non-significant at a nominated α value.

McBride (2019) proposed a procedure to assess trend direction and confidence based on Sen slope regression. However, as outlined in Section 5.2.1, we recommend that trend direction assessment is undertaken based on the Mann Kendall or Seasonal Kendall statistic, rather than the Sen slope (although we recommend using Sen slope regressions for estimating trend rates). This is because the Mann Kendall and Seasonal Kendall assessments handle censored values more robustly than Sen slope regression. In simple terms, for all Mann Kendall assessments, it is sufficient to know whether there is a positive or negative difference between pairs of observations (see Helsel (2012) for details). In contrast for Sen slope calculations, the absolute difference between all pairs of observations must be known so that that the slopes can be calculated. While the absolute difference between a censored and non-censored value cannot be known, we can know whether that difference is positive or negative. Therefore, the Mann Kendall assessment makes more robust use of the data and this is the reason it is the recommended basis for assessing trend direction. When there are few censored values and ties in the data, confidence in trend direction (C) is numerically equal to the approach of McBride (2019) based on the Sen slope.

One of the criticisms of the use of the historic use of NHST in trend analysis made by McBride (2019) is that in reality, no trend is ever zero (there is always a trend no matter how small) and therefore the null hypothesis is unrealistic[9]. McBride (2019) and others (e.g., Cohen 1994; Jones and Tukey 2000) distinguish *nil* hypotheses as hypotheses that propose no trend or no difference (e.g., between means). The problem with nil hypotheses, and their associated two-sided tests, is that they suggest that that the null hypothesis might be true and therefore encourage the incorrect interpretation of a non-significant result as indicating no trend. McBride's (2019) approach to trend direction assessment avoids the use of *p*-values entirely but has the disadvantage of not robustly handling censored values. Although the approach we recommend uses the two-sided test *p*-value to calculate the confidence in the trend direction (C), it is not based on a hypothesis test. In addition, C quantifies the evidence that the trend is in the assessed direction (i.e., positive or negative) and therefore is consistent with McBride's (2019) principle that there is always a trend. An important point that arises from this is that there is nothing intrinsically bad or wrong about *p*-values, it is their misuse and misinterpretation that is the problem (Makowski et al. 2019).

We also note that C is numerically equivalent to a Bayesian index of effect existence called Probability of Direction, abbreviated as pd (Greenland and Poole 2013; Makowski et al. 2019). The pd index does not require *a prior* distribution, nor does it rely on a null hypothesis, and is mathematically defined as the proportion of the posterior distribution that is of the median's sign (Makowski et al. 2019). The pd index can be interpreted as the probability that a parameter (such as *S*) is positive or negative and therefore has the same interpretation as C. The equivalence of *C* and pd can be explained in three steps. First, consider that the calculated value of *S* (S_{obs}) represents the best estimate of the actual population value (S_{pop} or the 'true trend'). The posterior distribution of S_{obs} is given by a normal distribution, with mean of S_{obs} and variance as calculated in Figure 5-1. Secondly, consider the probability that the calculated sign of S_{obs} is opposite to the population value S_{pop}. This probability is indicated by area under the distribution where *S* has opposite sign to S_{obs} and

[9] We note however that because water quality data is reported to a fixed level of precision, Kendall's *S* values of *S* can occur in practice.

is cut off by S equal to zero. This probability is equal to half the p-value of a two-tailed test because we are now concerned with whether S is positive or negative and not merely the extremity of the value. Finally, confidence is the probability that S_{obs} is the same as S_{pop} and is therefore the complement (i.e., one minus) of half the p-value; this is equivalent to the area under the distribution for which S has the same direction as S_{obs}, which is the same value as pd.

Recent publications in environmental science have used different language to describe the quantity that this guidance refers to as confidence in the trend direction (C). For example, Choquette et al. (2019) and Murphy (2020) refer to C as L_k, "the likelihood the sign of the trend is correct", and Larned et al. (2018a, b) refer to a mathematically related[10] measure as "the probability the trend was increasing". McBride (2019) described a quantity that is closely related[11] to C as the "probability that the slope was truly below (or above) zero". There is likely to be ongoing debate and development of ideas in this area. We recommend the interested reader refers to McBride (2019) for a full explanation of his trend direction assessment procedure and to Greenland et al. (2016) for a comprehensive discussion about the general interpretation (and misinterpretation) of p-values.

5.3.2 Dealing with serial correlation

For trend direction assessments, observations must be independent and not serially correlated. Serial correlation refers to the situation where the values of consecutive samples are correlated after accounting for the long-term trend in central tendency, regular cyclic fluctuations and covariates such as river flow. This situation can arise in water quality monitoring when consecutive samples are collected during discrete events such as floods and droughts, or when consecutive samples are taken from poorly mixed volumes of water. The presence of serial correlation can lead to an overestimation of the confidence in the evaluated trend direction. Hirsch and Slack (1984) proposed a modification to the traditional non-parametric method to account for serial correlation. They recommend using their modified method with datasets of > 10 years length. For shorter time periods, accounting for seasonality (as described in section 4.2) and using the seasonal variant of the Mann Kendall trend direction assessment will generally account for serial correlation. The TimeTrends package offers an option to use the Hirsch and Slack (1984) method, but it is recommended that serially adjusted p-values should not be used when considering whether there is a trend in a time-series of observations but should be used if the fitted model is being used to extrapolate beyond the measured data (i.e., to make predictions about future trends). A rationale for this recommendation is given in McBride (2005). However, as is the case for covariate adjustment (see Section 4.1), whether or not to adjust for serial correlation largely comes down to the purpose of the assessment; in the case that the purpose is purely to evaluate the observed trend over a specific time period, adjustment is not necessary (McBride 2005), whereas if attribution of the long term trends is of interest, then correction for serial correlation is important (see Helsel et al. 2020, section 12.9).

In our view, the risk that serial correlation will lead to severe overestimates of confidence in trend directions is minor for trend assessments based on SOE monitoring data. The strength of serial correlation typically decreases as the intervals between samples increases. Therefore, trend assessments based on data from high-frequency water quality sensors are more likely to be strongly affected by serial correlation than trend assessments based on monthly to annual sampling, as in SOE programmes. For analysts who are interested in estimating the serial correlation parameters of water quality trends, several standard statistical methods are available (e.g., see Darken et al. 2002).

[10] Where P(S>0)=p/2 and P(S<0)=1-p/2 where p was the two-tailed Mann Kendall or Seasonal Kendall p-value.
[11] McBride (2019) calculates this quantity as part of the Sen slope calculations.

5.3.3 Dealing with multiple censor limits

Another important practical consideration in assessing both trend direction and rate is the treatment of multiple censoring levels. This is a relatively common occurrence, particularly for long-term water quality records that span periods during which changes in instruments and analytical procedures have resulted in changes in detection limits. When a time-series includes multiple censoring levels, there is a risk that a trend will be detected that is an artefact of the changes in the censoring level rather than a change in the actual water quality within the time period. In our experience, multiple censoring levels are associated with detection limits (i.e., left censored data) and often occur due to changes in analytical methods for variables that are measured by laboratory analysis.

One approach to handling multiple detection limits is to change all observations that are less than the highest detection limit to this value (i.e., change the value of these observations to the same value as the highest detection limit) across the entire time period. The assessment of trend direction and rate is then performed on these data. The disadvantage of this approach is that there is likely to be a loss of information, because all observed values that are less than the highest detection limit will be treated as numerically equal (i.e., as "tied values" or "ties"). The alternative approach is to carry out the assessment of trend direction and rate with the multiple detection limits. In this case, differences between observations that are the highest detection limit and non-censored values that are less than the highest detection limit are counted as discordant (Figure 5-1). In addition, differences between observations that are the highest detection limit and censored values that are less than the highest detection limit will be counted as discordant. This approach has the advantage of retaining all the information about the variability of the observations but the risk, described above, is that the change in detection limit produces a trend that is unrelated to an actual change in water quality.

We recommend that expert judgement is used to decide on the most appropriate action (i.e., set all observations that are below the highest detection limit to that value or perform the analysis with multiple detection limits). There is no objective rule for choosing between these two options; the decision is a judgement that should be informed by inspection of the time series of observations and by taking into consideration the following questions:

- Can the differences in detection limits be explained (i.e., were there changes in laboratory procedures)? (i.e., is the highest detection limit a data error or representative of an actual change in measurement?)
- How many unique detection limits are there?
- Do the changes to the detection limits occur as step changes or are they randomly distributed through the time period?
- What proportion of the observations are associated with the detection limit change(s), and when in the time period did the changes(s) occur?
- How many tied values would there be in the record, if all observations less than the highest detection limit were set to the largest detection limit?

In cases where the highest detection limit can be attributed to a data error this should be corrected. Where there has been a step change in the detection limit and where a large proportion of observations are censored at the highest detection limit, the imposition of the highest detection limit across the entire time period is the best option. On the other hand, if the highest detection

limit applies to only a few observations within the assessment time period (e.g., as a result of the laboratory periodically reporting a 'sample matrix effect'[12] that prevented a reliable measurement to the specified method detection limit), and particularly if these are randomly distributed through the time period, the imposition of the highest detection limit may represent an unnecessary loss of information. In this case a better option is to retain multiple detection limits when carrying out the assessment of trend direction and rate because this maximises the information with low risk that a small number of censored observations with the high detection limit will unduly influence the trend. An example of dealing with a dataset comprising multiple detection limits is provided in Section 7.4.

We recommend that for regional or national applications, data pertaining to each variable should be screened for step changes in detection limits. Detection limit changes will generally be consistent by region and water quality variable and associated with changes in laboratory procedures. If the change in the detection limit for a variable has potentially induced a trend at any of the sites that are included in a regional or national application, we recommend that all observations that are less than the highest detection limit be changed to this value (i.e., to the same value as the highest detection limit) across all sites. This will ensure consistent assessments across the sites and allow for robust comparison of trends between sites. For local applications, we recommend a site-specific approach to dealing with multiple detection limits by considering the questions listed above for each site/variable combination. Because the primary aim of a local application is to maximise the information about the trend for each individual site/variable combination, the analyst should choose to assess trend direction and rate with the multiple detection limits unless it is judged that the variable detection limits unduly influence the results of the trend assessment. Note that LWPTrends and TimeTrends have built in functionality that sets all observations that are less than the highest detection limit to that value (i.e., the first option described above; see Section 6.3.3 for details).

We discourage use of the raw measurement values from laboratories in place of censored values for trend assessments. While councils may have received such advice and routinely receive both 'raw' and 'official' laboratory measurement values (where the latter presents censored values for any result below the analytical method detection limit), the precision of these raw values is very low relative to their numeric value (i.e., the uncertainty of measurement is very high). We therefore consider that these values are best considered as indistinguishable (i.e., analytically they are "ties") to prevent them having an undue influence on the trend analysis.

On a related note, water quality time-series may encompass periods in which the precision of values changes, or analytical methods change, with no change in censoring. In the first case, time-series may be characterised by numerous tied values early in the data record and fewer tied values later in the record. In the second case, there may be step-changes in a time-series corresponding to the change in methods. These situations need to be addressed on a case-by-case basis and we have no general guidance. In some cases, the variable of concern (or the site/variable combination) has been omitted from a trend assessment. To avoid these omissions, data adjustments could be considered (i.e., ensuring comparable precision over time by rounding high-precision values to match the lowest precision levels in the dataset).

5.3.4 Future advances in trend analysis

In future, new methods of trend analysis such as WRTDS (refer Section 2.3) are likely to be used in New Zealand. While these methods may provide benefits over the methods that are currently in widespread use, they will still require subjective decisions by the data analyst; these decisions are

[12] This refers to the combined effects of various components in the sample, other than the analyte (variable) being analysed. A common way to reduce a matrix effect is to dilute a sample, resulting in the reporting of a higher detection limit.

inherent in trend analysis, as described above. In addition, questions and debates about statistical inferences (e.g., frequentist versus Bayesian approaches to assessing confidence) will apply to these new methods as they do to the methods recommended in these guidelines. However, we believe that critical thinking about two issues is of greater immediate importance than adopting new statistical methods. First, further consideration should be given to how trend analyses are used and what information is needed from them. For example, a greater understanding of the trend rates that are of environmental and management importance is crucial. Assessing the importance of trend rates would provide context for reporting and for prioritising management actions. Establishing important trend rates would also help with the issues surrounding statistical inference. For example, this would enable the use of equivalence tests, which pose realistic hypotheses that are possibly true, namely, that a quantity of interest lies either within or beyond an "interval of indifference" (McBride 2019; 2005). The issue of trend importance when reporting on trend analysis is discussed further in Section 6.3.3.

Second, further consideration should be given to causes of trends (i.e., attribution) and methods for carrying out attribution assessments. A robust understanding of the causes of water quality trends, both degrading and improving, will enable effective management actions to be prescribed to arrest and reverse degradation and drive recovery (Ryberg et al. 2018).

6 Reporting trend analyses

6.1 Purpose

The trend assessment procedures described in Section 5 produce four pieces of information that need to be reported for each site and variable combination: trend direction (the sign of *S*), confidence about the trend direction (*C*), trend rate (SSE or SSSE) and confidence about the trend rate. While this information alone can be informative in the case of a local application, it is often necessary to report trends across multiple sites and multiple water quality variables (e.g., regional and national applications). In these cases, effective metrics and techniques to summarise the results in tabular, graphical or map format are needed. Reporting aggregated summaries across sites (e.g., proportion of site increasing and decreasing, by variable) provide informative overviews of water quality changes over a domain of interest (e.g., the entire country, a region, other geographic or environmental classes).

This section describes options for reporting the results of the four key outputs of trend assessments. We emphasise that it is important that the results are accompanied by thorough documentation of the methods used, including the choices and assumptions that have been made at each step in the assessment process. The methods should be described in sufficient detail that it is possible to identify why an alternative set of analyses arrives at different results. We also note that for contextual purposes it is necessary to consider reporting current water quality state alongside temporal trend information.

6.2 Recommended methods

6.2.1 Reporting trend direction and confidence in trend direction

One approach to communicating trend assessment results is reporting the trend direction and *C* directly (e.g., 91% confidence that the trend was increasing). Alternatively, *C* can be discretised into categories; this can be particularly useful when there is a need to summarise trends across a large number of sites. The simplest discretisation is to divide *C* into two categories based on whether confidence in the estimated direction is greater or less than a nominated level of confidence (e.g., 0.90 indicating 90% confidence in trend direction). This binary classification is analogous to historical reporting of the results of trend analyses that were based on a test of statistical significance. Using this approach, confidence in a trend is categorised as either:

- ‘direction established with confidence’ when *C* is greater than the nominated level of confidence; or
- ‘trend direction not established at the X% confidence level’’ when *C* is less than the nominated level of confidence,

respectively (McBride 2019). An example of more nuanced discretisation is given in Table 6-1, where *C* is divided into four categories, and each is assigned a narrative label to communicate the associated level of confidence. These three alternatives (continuous, binary and multi-category) are demonstrated graphically in Figure 6-1.

Table 6-1: Level of confidence categories used to convey confidence in trend direction. These categories were suggested by Choquette et al. (2019).

Categorical level of confidence the estimated direction is correct	Range of *C* value
Highly likely	0.95–1.0
Very likely	0.90–0.95
Likely	0.67–0.90
Uncertain	0.50–0.67

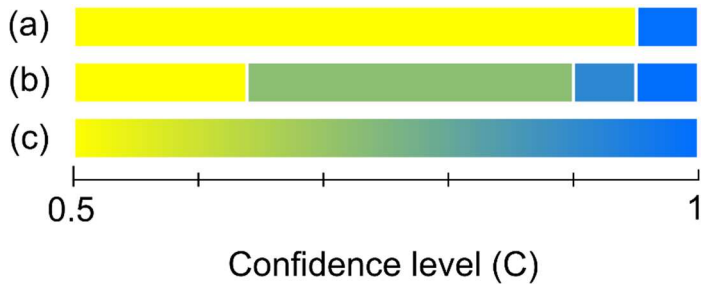

Figure 6-1: Graphical representation of three alternative ways of expressing confidence in trend direction. (a) A binary classification with the category boundary defined by a nominal confidence level of 95%. The blue shaded region indicates that trend direction was established with confidence, yellow indicates that the direction is not established with confidence. (b) Four categories, where the colours indicate confidence categories for the trend direction as described in Table 6-1. (c) Continuous confidence level 'C'.

For reporting purposes, it can be useful to combine the trend direction and the confidence in direction into a single metric. This can be achieved by appending the trend direction to the categorical description of confidence. For example, a simple discretisation is to divide *C* into three categories based on a nominated level of confidence (e.g., 0.95) and increasing and decreasing (i.e., 'highly likely decreasing', 'direction not established with confidence', and 'highly likely increasing'. A larger number of direction and confidence categories can also be defined as shown in Table 6-2.

Alternatively, the confidence in direction can be transformed into a continuous scale of confidence that the trend was decreasing (C_d). For all trends with $S < 0$, $C_d = C$, and for all $S > 0$ a transformation is applied so that $C_d = 1\text{-}C$. C_d ranges from 0 to 1.0 (Figure 6-2). When C_d is very small, a decreasing trend is highly unlikely, which because the outcomes are binary, is the same as an increasing trend is highly likely.

The combined continuous measure of trend direction and confidence (C_d) can be discretised as shown in Table 6-2, but the categories need to cover the range of confidence from 0 to 1.0. The seven categories shown in Table 6-2 are based on modifications of the categories suggested by Choquette et al. (2019). McBride (2019) has suggested nine confidence categories as used by the International Panel for Climate Change (Mastrandrea et al. 2010).

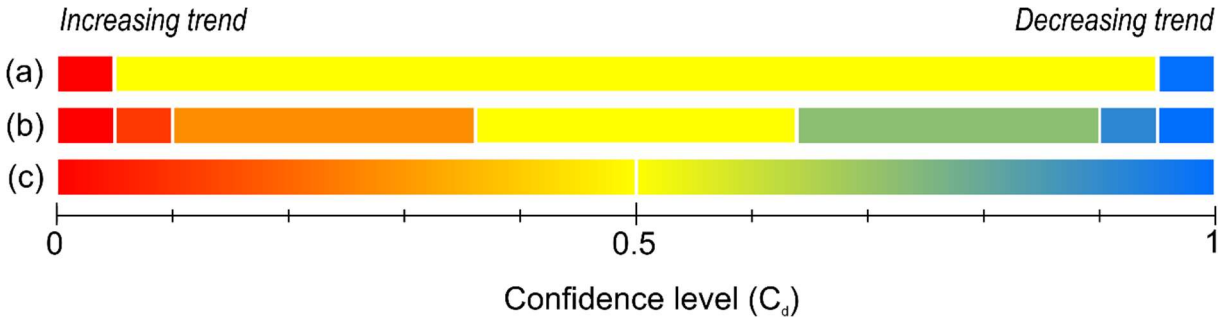

Figure 6-2: Graphical representation of alternative categorisations of confidence level including trend direction. Key: (a) the blue shaded region indicates decreasing trends established at the 95% level of confidence, the red shaded region indicates increasing trends established at the 95% level of confidence, yellow indicates that trend direction could not be established at the 95% confidence level (b) the level of confidence categories as described in Table 6-2 and (c) the continuous confidence level that the trend is decreasing (C_d).

Table 6-2: Suggested confidence categories in water quality trend are decreasing (C_d).

Categorical combined confidence and trend direction	Sign of *S* and value of *C*	Categorical confidence trend was decreasing	Value of C_d
Highly likely decreasing	Negative, 0.95–1.0	Highly likely	0.95–1.0
Very likely decreasing	Negative, 0.90–0.95	Very likely	0.90–0.95
Likely decreasing	Negative, 0.67–0.90	Likely	0.67–0.90
As likely as not	Negative or positive, 0.5–0.67	As likely as not	0.33–0.67
Likely increasing	Positive, 0.67–0.90	Unlikely	0.10–0.33
Very likely increasing	Positive, 0.90–0.95	Very unlikely	0.05–0.10
Highly likely increasing	Positive, 0.95–1.0	Highly unlikely	0–0.05

6.2.2 Reporting trend rate and confidence in trend rate

Trend rate analysis produces two key outputs, the SSE or SSSE value and the associated confidence interval, which are reported in the original units of the variable being assessed per year. The sign of these values indicates the trend direction and should be preserved. When reporting SSE and SSSE values for many variables it is often helpful to standardise the values by dividing by the median value of the observations over the analysis time-period. Trend rates that are made relative to the median in this way are often referred to as the relative SSE or SSSE (RSSE and RSSSE), which are reported as a percentage of the median value per year. A difficulty with standardisation in this way is that sites with observations that are very low values can produce very large values of RSSE and RSSSE. In these circumstances, the relative values are likely to be misleading and should be avoided as should the mixing of results that are reported with and without relativisation.

Note that trend rates across many sites for a single variable can often vary by orders of magnitude. This can make displaying trend rates and their confidence intervals on maps or in graphs challenging.

6.2.3 Reporting trend directions and rates for many sites

Trend directions and rates for regional or national applications that concern many monitoring sites can be reported by tabulation and maps. Tabulation simply displays the relevant analysis results by site and variable (Table 6-3). Exhaustive tabulations for many variables and sites tend to be large and are better suited to appendices or supplementary data.

Table 6-3: Tabulation of trend assessment results for multiple site/variable combinations from a hypothetical lake water quality monitoring programme.

Site	Variable	Median in trend period	Units	N	Trend direction	Confidence that trend is decreasing (%)	SSE (units/yr)	90% confidence limits for SSE	RSSE (%/yr)
Lake Jones	Total N	1.77	mg/L	60	Decreasing	69	0.009	-0.0039 to 0.0085	0.5
Lake Smith	Total N	1.85	mg/L	55	Decreasing	75	0.026	-0.0005 to 0.0017	1.4
Lake Clark	Total N	2.60	mg/L	60	Increasing	22	0.023	0 to 0.0778	0.9
Lake Baker	Total N	0.15	mg/L	60	Decreasing	52	0.001	-0.004 to -0.0009	0.4
Lake Jones	Secchi depth	6.3	m	58	Increasing	31	0.139	-0.062 to 0.3808	2.2
Lake Smith	Secchi depth	7.0	m	55	Increasing	20	0.210	0.0096 to 0.174	3.0
Lake Clark	Secchi depth	4.9	m	58	Decreasing	83	0.044	-0.0024 to 0.0016	0.9
Lake Baker	Secchi depth	8.5	m	58	Increasing	47	0.187	0.02 to 0.15	2.2

Maps are a good graphical method for reporting trend assessment results for many sites. Maps for each variable that show sites colour coded by trend direction and confidence (C_d) or by trend rate and direction convey a great deal of the information obtained from trend assessment. As the distributions of C_d are constrained between 0 and 1, this can help simplify plotting compared to mapping of Sen slope values, particularly across multiple variables (for which the units of the Sen slope can vary). It is noted that C_d and the combination of trend rate and direction are generally highly correlated and therefore maps will show similar patterns. Figure 6-3 shows a comparison of C_d and Sen slope rates from the results of a ten-year period trend assessment (2008-2017) of four water quality variables across 69 sites from NIWA's National River Water Quality Network (NRWQN).

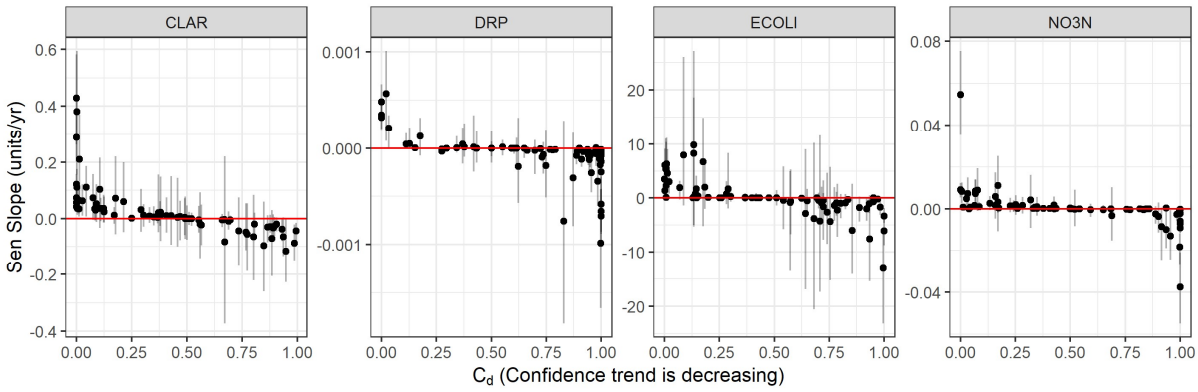

Figure 6-3: Comparison of SSE and C_d for four variables at 69 NRWQN sites for a 10-year trend period. Error bars indicate the 90% confidence interval for the estimated Sen slope.

Maps can also show sites colour coded by discrete confidence categories (i.e., Table 6-1). The most suitable combination of mapping will vary with the application and depends on factors such as the number of sites and their spatial distribution; it is likely that each application will have a mapping solution that optimises communication of the results. Some examples are shown in Figure 6-4, Figure 6-5 and Figure 6-6, based on the categorisations of trend direction and confidence described in Section 6.2.1. The data shown in these plots are the results from a ten-year period trend assessment (2008–2017) for four water quality variables across 69 NRWQN sites.

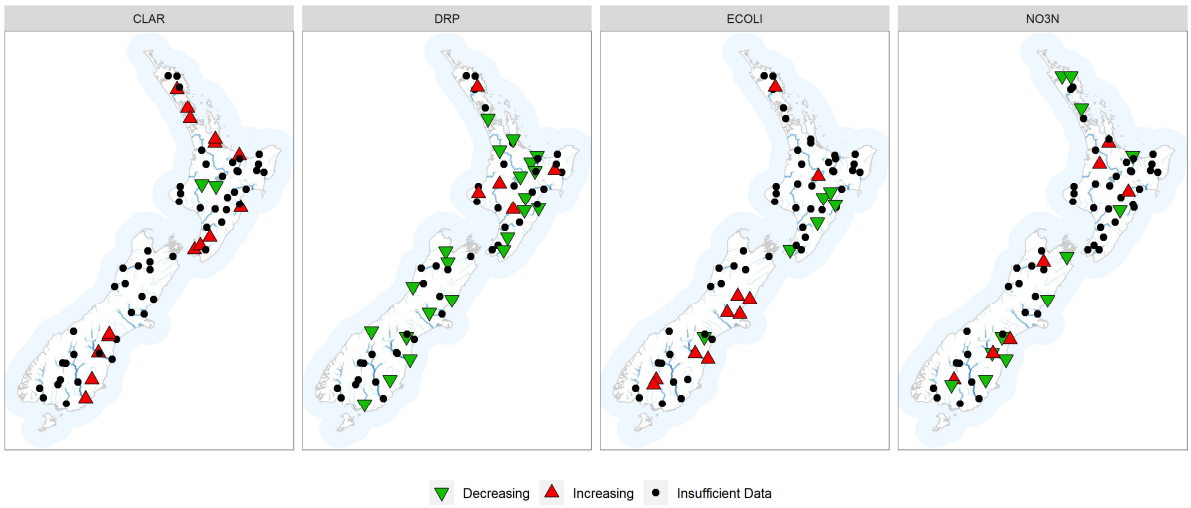

Figure 6-4: Maps of NRWQN sites summarising 10-year trends in four water quality variables. The trends are categorised into three classes: decreasing – 95% confidence that the trend is decreasing; Increasing – 95% confidence that the trend is increasing; and trend direction not established at the 95% confidence level.

Figure 6-5: Maps of NRWQN sites summarising 10-year trends in four water quality variables categorised by trend direction and confidence in trend direction. Shapes indicate trend direction and colours indicate confidence (*C*) in the trend direction.

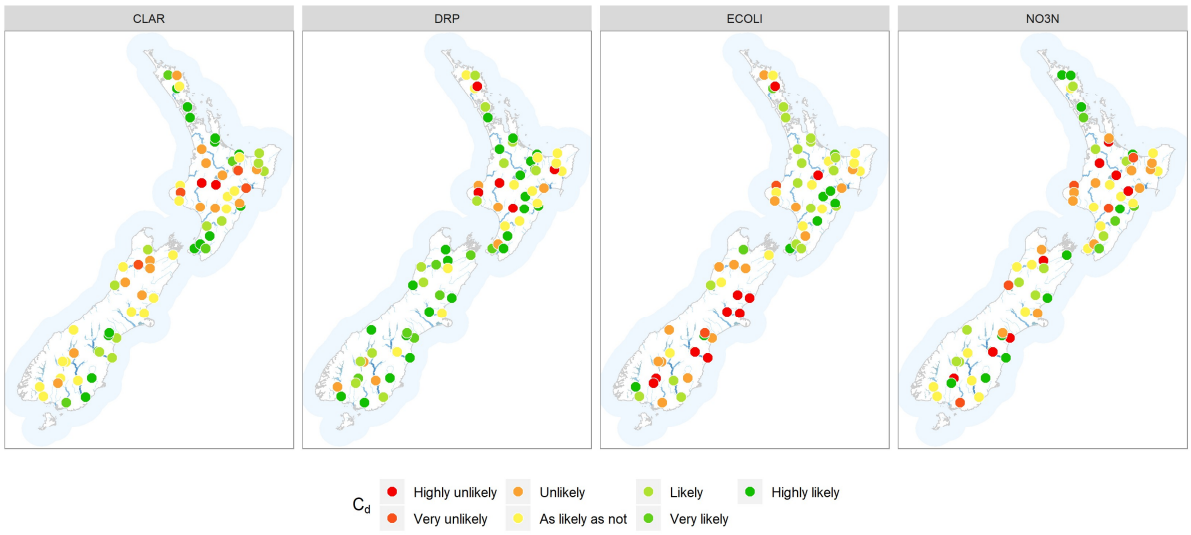

Figure 6-6: Maps of NRWQN sites summarising 10-year trends in four water quality variables, categorised into seven classes based on confidence that the trend direction is decreasing (C_d).

6.2.4 Trend aggregation

Aggregated results of trend analyses performed for many sites (i.e., regional and national applications) reveal general patterns of water quality change over domains of interest such as the country, regions or environmental classes. The simplest types of aggregation are plots showing distributions of site values of either continuous measure of trend direction and confidence (C_d) or the combination of trend rates and direction (Figure 6-7). For example, Larned et al. (2016) used box plots to show the distribution of trend rates and directions for a number of water quality variables for rivers grouped by River Environment Classification (REC) source-of-flow classes.

Tabulation of site trends, categorised by direction and whether the trend was established at a specified level of confidence (generally 95%), and grouped by domains of interest is a simple method of aggregating trends that has been used in the past. However, this approach is not recommended because trends categorised as 'direction not established with confidence' nonetheless contain information about the probable 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). Tabulations of the combination of trend direction by more detailed confidence categories (e.g., Table 6-2) are more informative because they incorporate all information but are not the easiest way to convey the information, especially if multiple variables and domains of interest are involved.

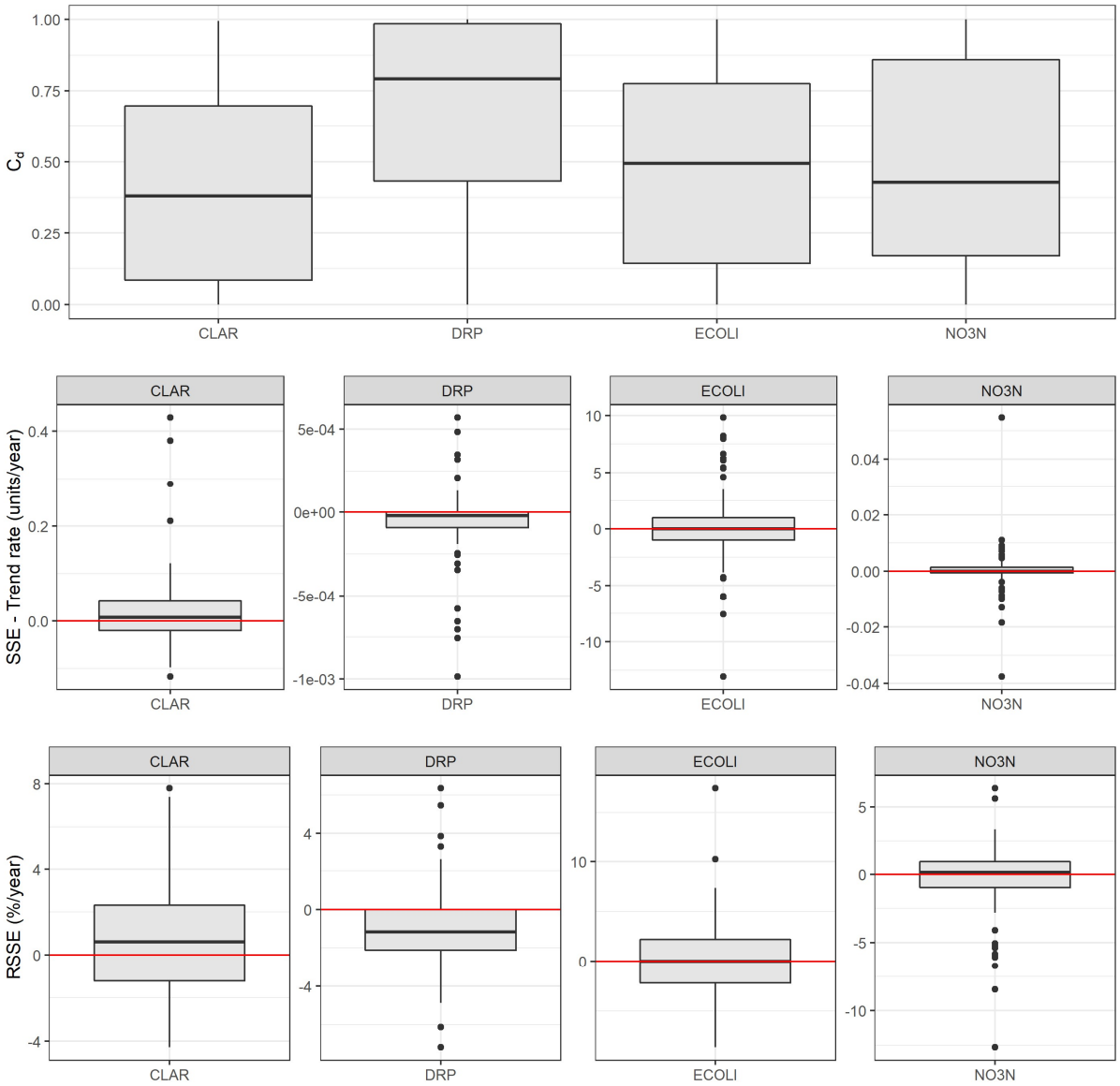

Figure 6-7: Examples of box plots to describe distributions of trend direction and trend rate across many sites. The top, middle and bottom rows show the confidence that the trend direction is decreasing (C_d), the absolute value of Sen slope estimator (SSE; units/year), and the relative Sen slope estimator (RSSE; %/year), respectively.

The graphical method shown in Figure 6-8 is a better way to communicate aggregated trend direction and confidence information (Snelder and Fraser 2018). The method is based on evaluating the number of sites in confidence categories that subdivide the range in C_d such as those Table 6-2. Coarser or finer differentiation could also be used. A choice can be made to categorise confidence the trend was decreasing or to convert this to confidence the trend was improving. Confidence the trend was improving requires taking the complement of C_d for variables for which decreasing values indicate degradation, such as visual clarity and MCI scores. This subjective decision is discussed in Section 6.3.1.

The proportion of sites in each confidence category is calculated for each domain of interest and each water quality variable. The results are then plotted as stacked bar charts where each bar is subdivided and coloured to represent the proportion of sites in each category and in the same order shown in Table 6-2. Different bars or plots can either represent different variables or different domains.

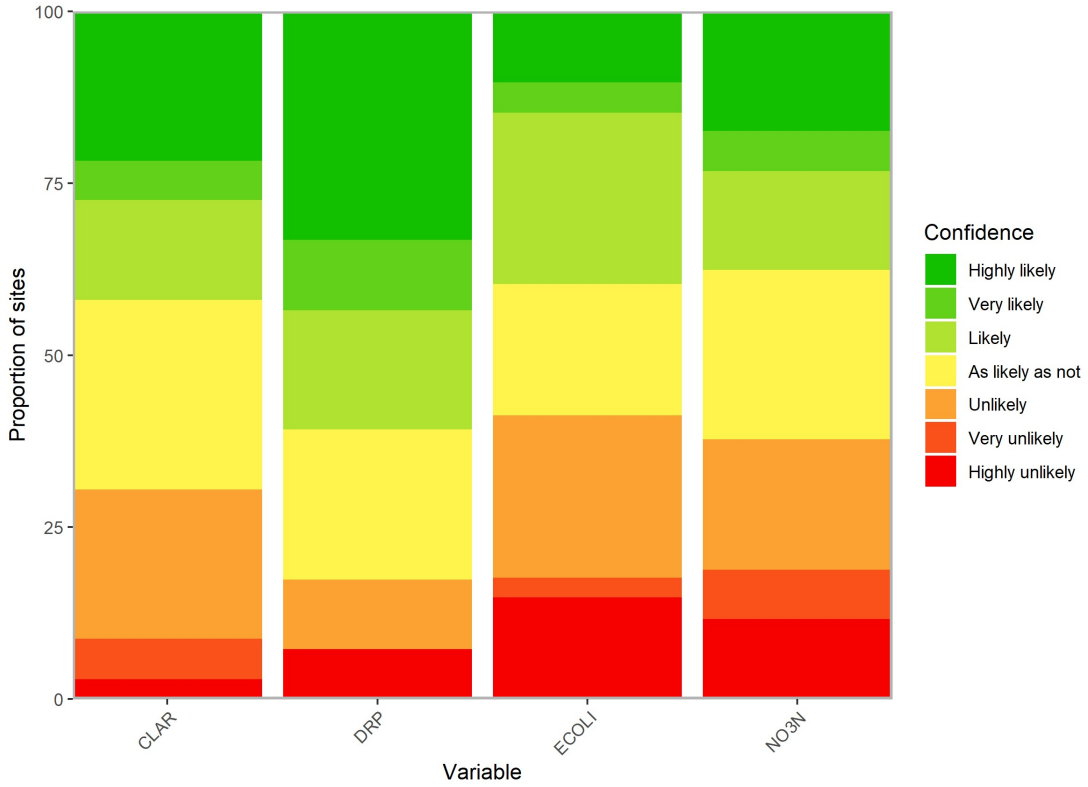

Figure 6-8: Example of stacked bar charts of categorical confidence that the trend was decreasing. Trends are for 69 NRWQN sites, for a 10-year time period, as assessed by Larned et al. (2018a).

6.3 Commentary

6.3.1 Interpretation of trend direction

Judgements about whether trend directions indicate either degradation or improvement is dependent on the variable and are subjective. For example, decreasing trends in nutrient and faecal microbe concentrations generally indicate improving conditions, whereas decreasing trends in other variables, particularly 'ecosystem health' indicators such as MCI scores, indicate degrading conditions. Subjectivity arises because judgements about trend direction implicitly incorporate values. For example, increasing water clarity would generally be interpreted as an improvement. However, increasing clarity downstream of a dam might indicate reduced sediment supply which may lead to undesirable changes in bed substrates (i.e., armouring of the bed). Subjectivity also arises because judgements about trend direction should ideally consider the baseline from which the trend is occurring. An obvious case where the baseline state is important is a trend in pH, for which a decreasing trend could indicate degradation or improvement, depending on the baseline.

Notwithstanding the subjectivity, conversion of trend assessment results from increasing/decreasing into degrading or improving may be useful for reporting purposes (e.g., for the NPS-FM 2020 which refers to "deteriorating" trends). The conversion may reduce the possibility for confusion arising in tabulations, plots and maps that summarise trends across variables that have mixed directions indicating degradation and improvement (such as when trends in contaminant concentrations are reported alongside trends in MCI scores and visual clarity). Although there is a subjective element involved in assigning each increasing/decreasing trend to improving/degrading, this only needs one explanation after which all reporting tables, plots and maps are self-explanatory. By contrast, if

trends are reported as increasing/decreasing, the reader needs to be cognisant of the meaning of increases or decreases for each variable and for each table, plot or map.

6.3.2 Ensuring consistency when comparing trends across many sites

The methods for aggregating trend assessment results across multiple sites (i.e., for regional or national applications) described in Sections 6.2.3 and 6.2.4 assume that the trends that are being simultaneously reported are comparable. Refer to Section 3.3 for a discussion regarding ensuring trends adequately represent the time period and have consistent levels of statistical power.

6.3.3 Interpretation of trend rate and confidence

Historically, trend rates have been classified as environmentally important based on a nominal threshold applied to the RSSE or the RSSSE (e.g., Ballantine et al. 2010; Daughney and Reeves 2009; Scarsbrook 2006) or a rate of change that would cause a state threshold to be crossed within a specified period (e.g., Larned et al. 2015). These definitions of the environmental importance of a trend are arbitrary rules of thumb that may not be appropriate depending on the context. An example of the relevance of context is provided by considering a trend in an influential variable in a stream that is being managed for its significant ecological or cultural value. In this type of environment, any rate of degrading trend, especially where confidence in direction is high, might be judged important. We therefore consider that the trend importance is context specific and is also a subjective judgement for which we have no general guidance.

The same considerations apply to trend confidence. If there is 80% or 90% confidence that there is a degrading trend in an influential variable in a stream that is being managed for its significant ecological or cultural value, then this may be judged as enough evidence to act. The subjective nature of this decision is an important reason that we recommend reporting confidence in trend direction (*C*) rather than the binary classification into 'direction established with confidence' and 'insufficient data', which effectively reduces the information available to the decision maker.

6.3.4 Combining state and trends

The direction, confidence, and rate of trends by themselves are unlikely to provide all the information necessary to identify appropriate management responses (e.g., as required by the NPS-FM). The water quality state (e.g., as described by the median of recent observations) also provides important information when considering actions that should be taken in response to a detected trend. Consider two sites that have equal management significance, and degrading trends of equal rate. If one site was in a pristine state and the other was already severely degraded, the management actions are likely to be different. The implication is that information about trends and state should be presented together and both types of information should be as accessible as possible. We do not offer any prescriptive guidelines about how state and trend information should be combined, but we do recommend that consideration is given to both characteristics. Figure 6-9 shows an example of single plots with state and trend information combined for many sites.

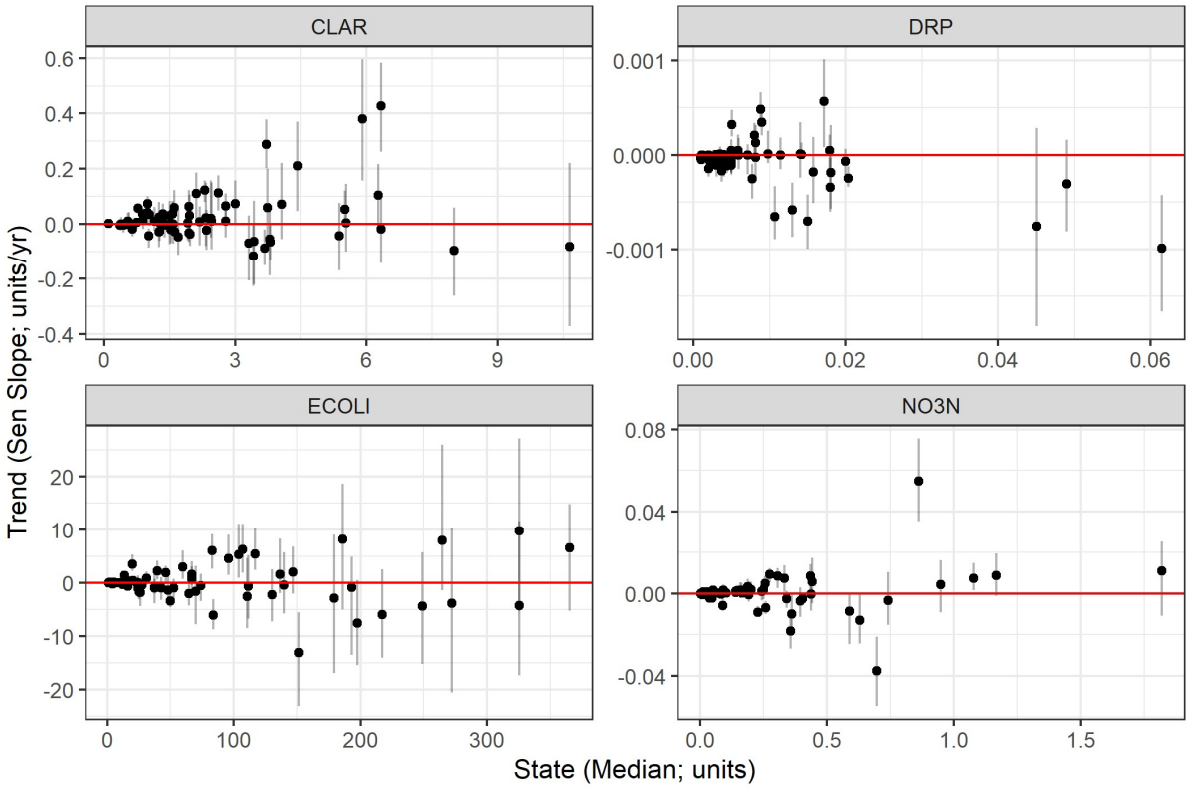

Figure 6-9: Example showing the presentation of state and trend information on a single plot. Here the state is based on the median calculated from the final five years of the trend period to be generally consistent with state assessments (e.g., Larned et al. 2016, 2004). Error bars indicate the 90% confidence interval for the estimated Sen slope. A threshold or guideline could be added as a vertical line to aid interpretation of the median values.

7 Worked trend assessment examples

The four worked examples in this section are designed to demonstrate trend assessments using TimeTrends and LWPTrends, and some of the issues that are frequently encountered when undertaking trend assessments in New Zealand. TimeTrends is a standalone trend assessment software package with a graphical user interface. LWPTrends is a library of functions that can be used with the 'R' statistical computing software (R Core Team 2019).

Other analysis tools can also be used to implement the guidance provided in this document. LWPTrends and TimeTrends are freely available packages that have been developed in New Zealand for water quality data analysis and both have user communities. They are therefore recommended, especially for users with limited experience.

Supplementary files to accompany the worked examples are available from the LWP website[13]. These files include the input data (as csv files) and an R script to reproduce the analyses using LWPTrends (version v2001). The csv data files can also be imported into TimeTrends.. We recommend that a new user consults the help files for the packages for more details.

The analytical choices in the following four examples are representative of local applications (refer Section 2.2). Note that the four assessments are performed over the complete datasets provided, which leads to differences in trend period duration and start dates for each example. The examples explore the implications of some of the subjective choices that must be made in trend assessments . The same subjective choices need to be made in regional and national applications, but there are additional constraints associated with maximising consistency between analyses in these large-scale applications (see Section 2.2).

7.1 Example 1: Flow adjusted, seasonal WQ variable

Example 1 uses a river monitoring site with a long-term record of monthly chemical concentration observations. Simultaneous flow observations are available for the site. The example demonstrates a typical application of flow adjustment, seasonal assessment and trend assessment.

Step 1: Examine raw data

The raw data (i.e., the 'as reported' monthly concentration values) can be examined in LWPTrends using the function "Inspect Data". "Inspect Data" also produces a tabular summary of the data.

The raw data can be examined in TimeTrends by: Analysis>X-Y Plot>Points. The flow data associated with this dataset can also be added to the TimeTrends plot. TimeTrends provides descriptive statistics for the raw data by: Analysis>Descriptive statistics>Overall.

Other packages could also be used to inspect the data (e.g., Excel, Matlab, Minitab, etc).

The data summaries and plots produced by LWPTrends (Figure 7-1) and TimeTrends (Figure 7-2) show:

- there are 25 years of monthly data with no gaps;
- there is limited censoring (2 values below the detection limit for the entire time series); and

[13] http://landwaterpeople.co.nz/wp-content/uploads/2021/01/TrendGuidanceWorkedExamples.zip

- there is a seasonal pattern in observations, with the greatest values over winter. This seasonal pattern is also evident for the flow observations.

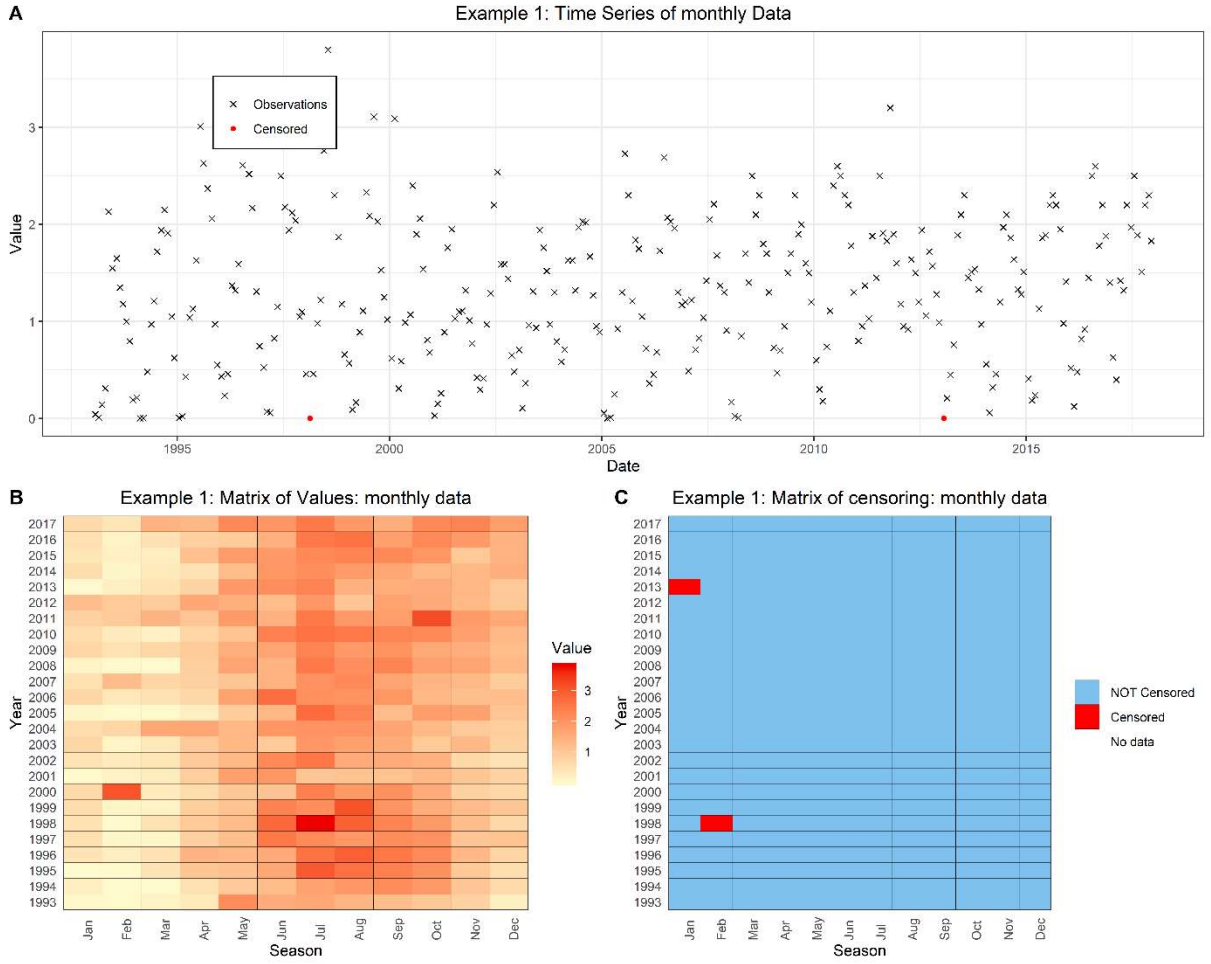

Figure 7-1: Example 1 – LWPTrends raw data inspection plots.

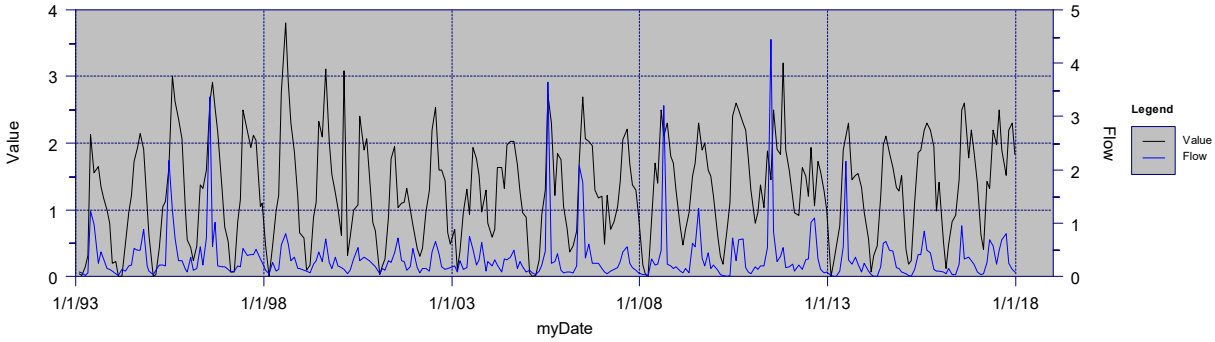

Figure 7-2: Example 1 – TimeTrends scatter plot of raw data.

Step 2: Covariate adjustment and seasonality assessment

Because the dataset includes observations of flow, there is the possibility to flow adjust the raw observations if there is a meaningful relationship between flow and concentration. Flow adjustment is performed as a pre-processing step in LWPTrends, with the function 'Adjust Values'. In TimeTrends, flow adjustment is an integrated component of seasonality testing and the trend evaluation. In TimeTrends, the relationship between the concentration and flow can be visualised using Analysis>X-Y Plots.

In this example, LWPTrends was used to assess four alternative models of the relationship between concentration and flow: Log-Log, GAM, LOESS with a span (i.e., degree of smoothing) values of 0.9 and 0.7. 'Adjust Values' returns a scatter plot of the raw concentration-flow data showing with the fitted models (Figure 7-3).

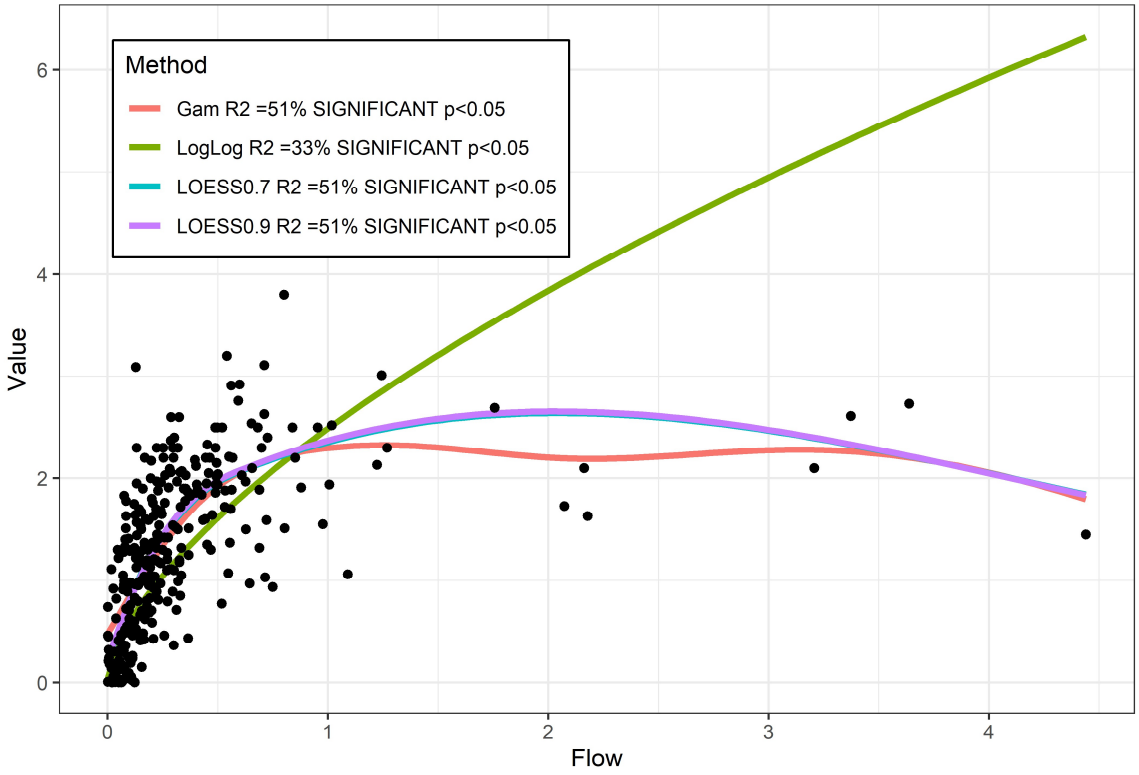

Figure 7-3: Example 1 – scatterplot plot of concentration (value) and flow produced by LWPTrends. The lines represent four alternative models fitted to these data to represent the relationship between concentration and flow.

Based on the scatter plot of concentration versus flow, we selected the LOESS0.9 model. The LOESS 0.7 and LOESS 0.9 produce very similar results, and when there are two models of similar performance, we would recommend selecting the simplest model (in this case LOESS0.9 as it has the widest span). When using LWPTrends, the residuals of all of the possible models (the "flow adjusted observations") are returned from the 'Adjust Values' function and these are added to the data-frame of observations for subsequent analysis.

The next step is to check for seasonality in the flow adjusted observations. For LWPTrends, seasonality is assessed using the function 'SeasonalityTest', and by specifying that the test is performed on the flow adjusted values. In TimeTrends, a seasonality test on flow adjusted data can be performed by Analysis>Seasonality test and selecting flow as a covariate, plus designating a

model type for the relationship. In Example 1, we chose to fit a 'LOWESS' model with 'Lowess % of points fit' set to 90% (note that this is the same smoothing parameter used when the data were flow adjusted above using the LWPTrends 'Adjust Values' function); all other values were left at the TimeTrends defaults. Both packages return box plots showing the distributions of observations within seasons (Figure 7-4, Figure 7-5), as well as details from the Kruskall-Wallis test used to evaluate the significance of any differences in distributions between seasons. Both packages indicate that even after flow adjustment, the data are seasonal (p<0.01).

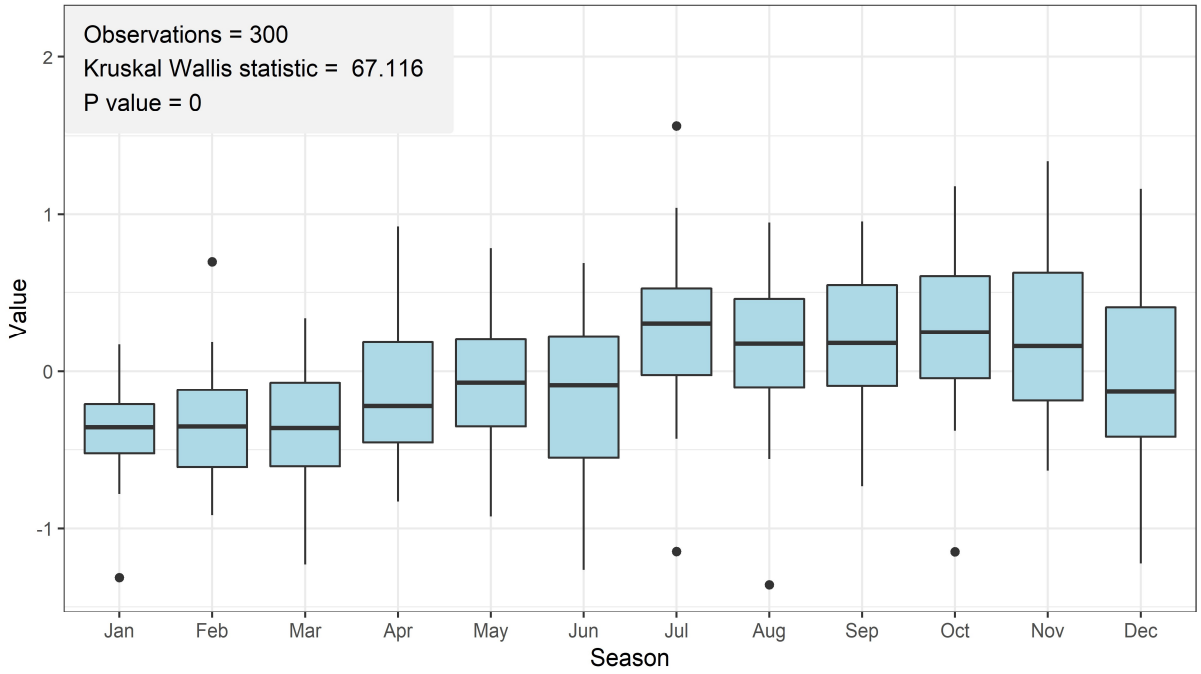

Figure 7-4: Example 1 – box plot of flow adjusted values by season produced by LWPTrends. The black horizontal line in each box indicates the median of the observations, the box indicates the inter-quartile range, the whiskers indicate the 5th and 95th percentiles, and the dots indicate outliers.

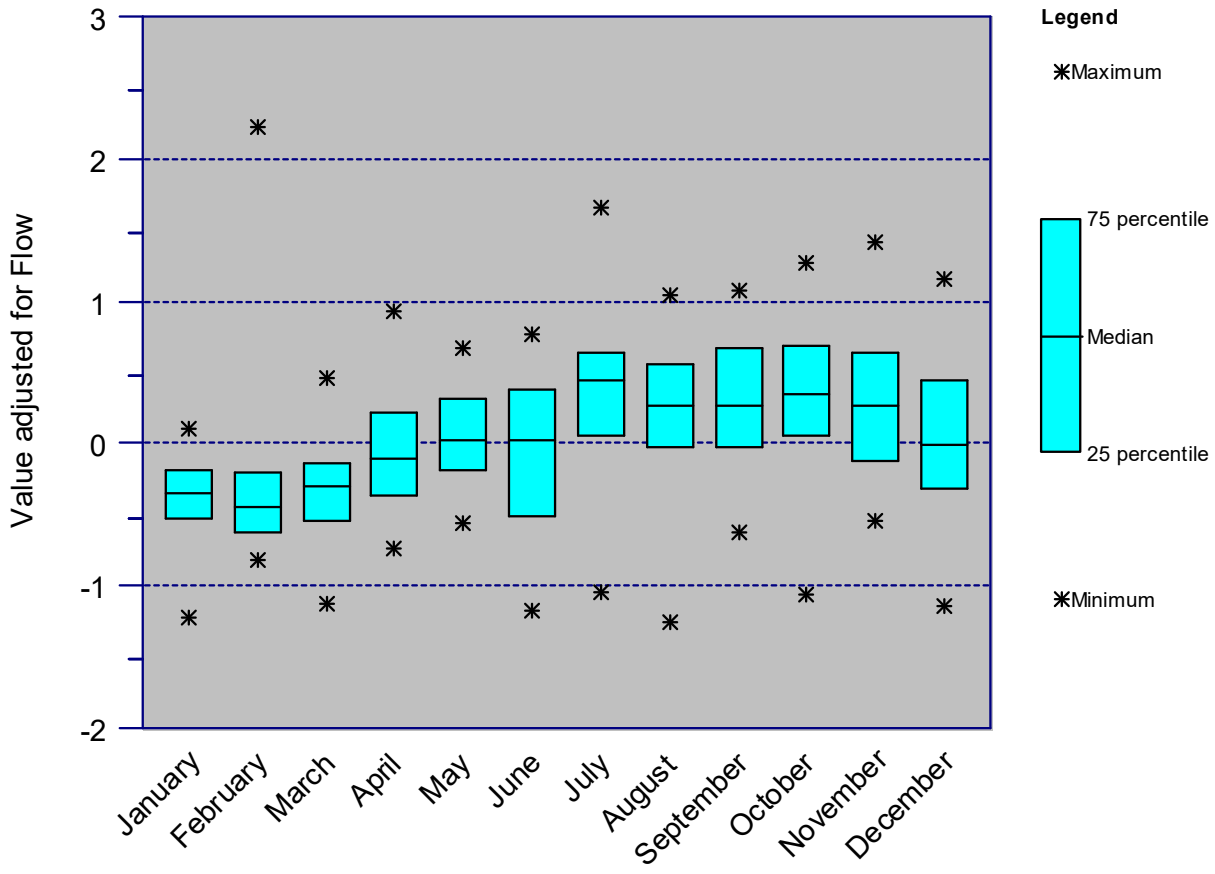

Figure 7-5: Example 1 – box plot of flow adjusted values by season produced by TimeTrends.

Step 3: Perform Trend Assessment

Because the seasonality assessment indicates the observations are seasonal, the trend assessment is conducted using the Seasonal Kendall test and the Seasonal Sen slope. Both LWPTrends and TimeTrends perform these operations as part of one function. For LWPTrends, this analysis is performed using the function "SeasonalTrendAnalysis" (with "ValuesToUse" set to the flow adjusted values), and in TimeTrends using Analysis>Seasonal Kendall Test (with flow selected as a covariate). Both packages output a plot of the observations against time with the evaluated non-parametric (Sen) regression line superimposed (Figure 7-6, Figure 7-7). The LWPTrends package plots the results with the y-axis showing the flow adjusted values, whereas TimeTrends plots the raw values. The LWPTrends plot also includes a summary of the main statistics of the trend assessment. Both packages also output results tables.

TimeTrends does not report confidence in trend direction but this can be calculated as one minus half of the *p*-value. TimeTrends also by default reports raw and flow adjusted trend results when a covariate is selected. To obtain the results of a non-flow adjusted trend using LWPTrends, the user would first need to evaluate whether the raw data were seasonal, and then proceed to use either the "SeasonalTrendAnalysis" or "NonSeasonalTrendAnalysis" functions, depending on the outcome of the seasonality assessment. There are small differences in the flow adjusted trend assessments from the two packages for this example – this is associated with differences in the models used to define the flow-concentration relationships. The resulting small changes in results are seen in the combined trend assessment outputs summarised in Table 7-1). This table includes the results from trend assessments based on the raw (non-flow adjusted) observations, as well as results for a trend assessment on the flow observations.

In this example the non-flow adjusted (raw) Sen slope is smaller than the flow adjusted Sen slope (although the confidence intervals do overlap). The trend assessment of the flow (i.e., river discharge) provides some insight as to why this might be. There was a virtually certain decreasing trend in flow observations over the analysis period (data not shown). The relationship between flow and concentration (Figure 7-3) indicates that lower flows are generally associated with lower concentrations at the monitoring site. Combined with a decreasing flow trend, we might expect this to generate a decreasing concentration trend. In this example, the trend in the raw concentration is increasing, but we see that the flow adjusted trend is larger than the raw trend; the decreasing trend in flow has reduced the upward trend in concentration.

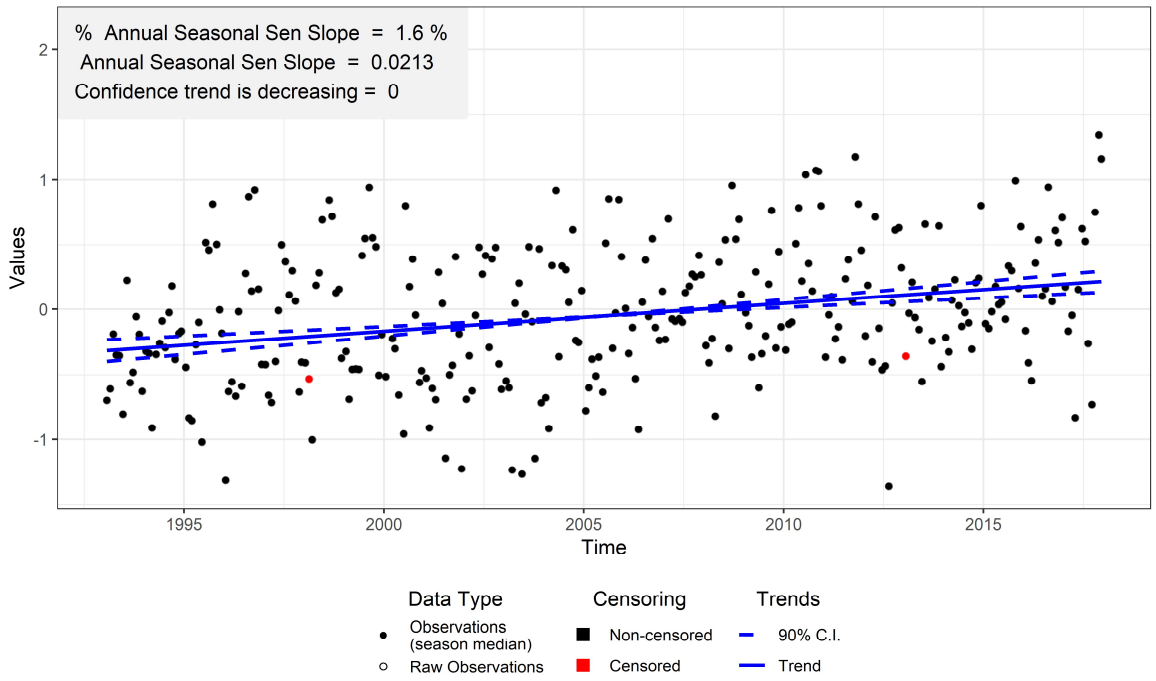

Figure 7-6: Example 1 – time-series of monthly flow adjusted observations and fitted non-parametric (Sen) regression line (and 90% confidence intervals) produced by LWPTrends.

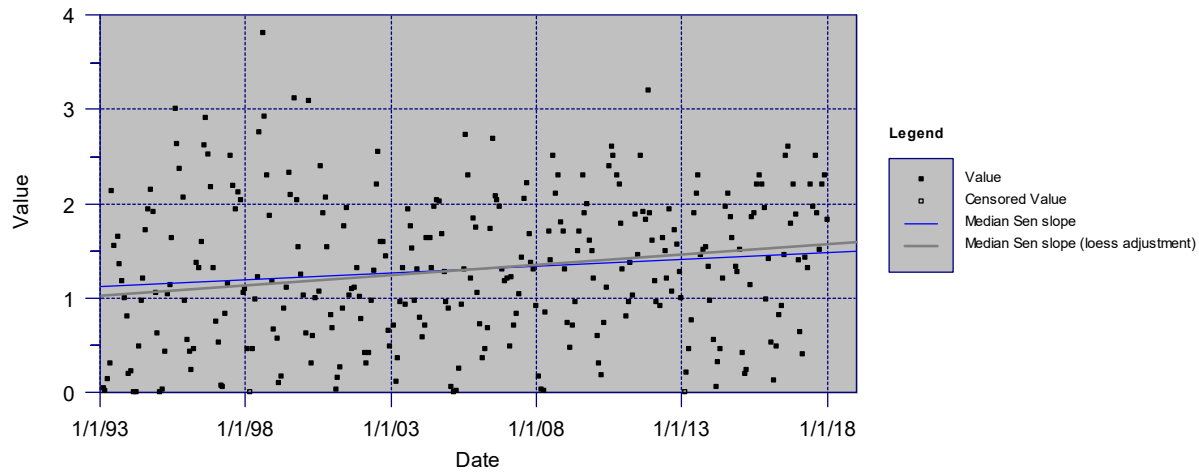

Figure 7-7: Example 1 – time-series of monthly flow adjusted observations and fitted non-parametric (Sen) regression line produced by TimeTrends.

Table 7-1: Tabulated results of trend assessment for Example 1.

	Kendall statistic	Variance	*C* (confidence in trend direction)	Sen slope (annual)	Percent annual change	90% confidence limits for slope	Direction
LWPTrends	793	21979	1.00	0.021	1.639	0.015 to 0.028	Increasing
TimeTrends	840	22000	1.00	0.022	1.657	0.016 to 0.027	Increasing
TimeTrends (raw data)	668	21961	1.00	0.015	1.15	0.010 to 0.021	Increasing
TimeTrends (flow adjusted data)	-341	21999	0.99	-0.002	-0.88	-0.003 to 0	Decreasing

7.2 Example 2: MCI trend

Example 2 uses a river monitoring site with a long-term record of MCI observations. The sampling frequency has changed over time. The objective is to evaluate the long-term trend at the site. This example demonstrates a typical application to data with irregular monitoring frequency.

Step 1: Examine raw data

The raw data (i.e., MCI scores) can be examined using LWPTrends with the function "Inspect Data". "Inspect Data" also produces a tabular summary of the data.

The raw data can be examined in TimeTrends using: Analysis>X-Y Plot>Points. TimeTrends provides descriptive statistics for the raw data using: Analysis>Descriptive statistics>Overall.

The data summaries and plots presented in Figures 7-8 and 7-9 indicate:

- sampling frequency is irregular, but there has been at least one observation in spring and one in summer between September 1995 through to September 2017;
- in the early part of the time series there are observations that appear to be near-duplicates (the values are very similar, and the associated observation dates are no more than one month apart so it may be inappropriate to treat these as independent observations; and
- there appears to be some seasonality, with values in spring (September-November) generally higher than those later in summer (January and February) (Figure 7-8).

The irregular sampling frequency and the presence of dependent observations early in the data record require that some further data grooming is undertaken with careful consideration of the specification of seasons for the analysis prior to the trend evaluation. First, we identified observations that we considered to be dependent. These were observations that were less than two MCI units difference in magnitude and less than 40 days apart. For these pairs, we replaced the two rows of data with a median observation and median date (a groomed dataset with these values is provided with the example files). Second, we considered what appropriate seasons would be. Although trends in MCI are frequently analysed on annual "seasons", this dataset contains more information than a single annual MCI value. However, the measurements are irregular, so defining a monthly season is likely to have many gaps and inconsistent representation of seasons as months. We therefore chose to implement two "seasons" in a year, "spring" (Jul-Dec) and "summer" (Jan-Jun). Note that these decisions would be appropriate in an equivalent regional or national application provided all sites were treated the same way.

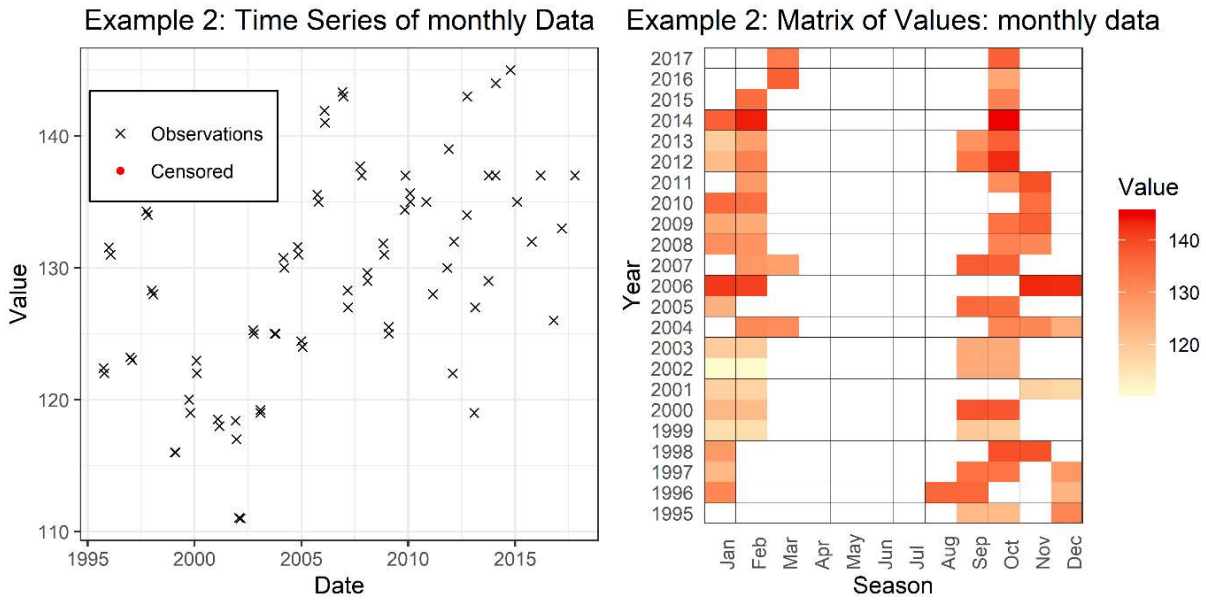

Figure 7-8: Example 2 – LWPTrends raw data inspection plots.

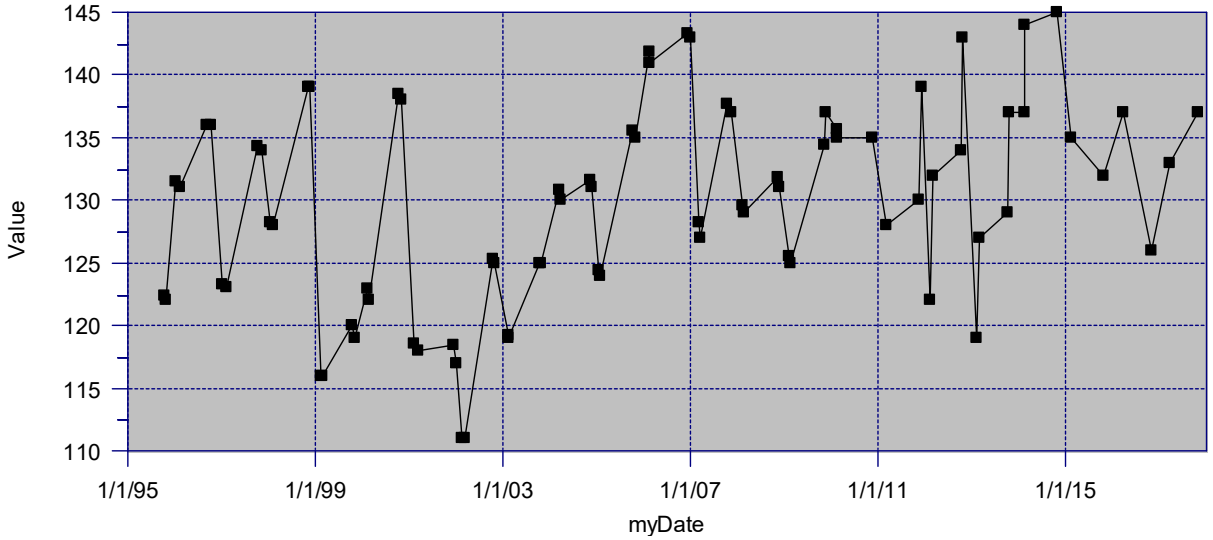

Figure 7-9: Example 2 – TimeTrends scatter plot of raw data.

Step 2: Seasonality assessment

We checked for seasonality in the cleaned observations in LWPTrends using the function 'SeasonalityTest', and in TimeTrends using Analysis>Seasonality test (selecting 2 seasons per year). Both packages return box plots to demonstrate distributions of observations within the specified seasons (Figure 10 and Figure 7-5), as well as the output from the Kruskall-Wallis test. Both packages suggest that the data are seasonal (p=0.015).

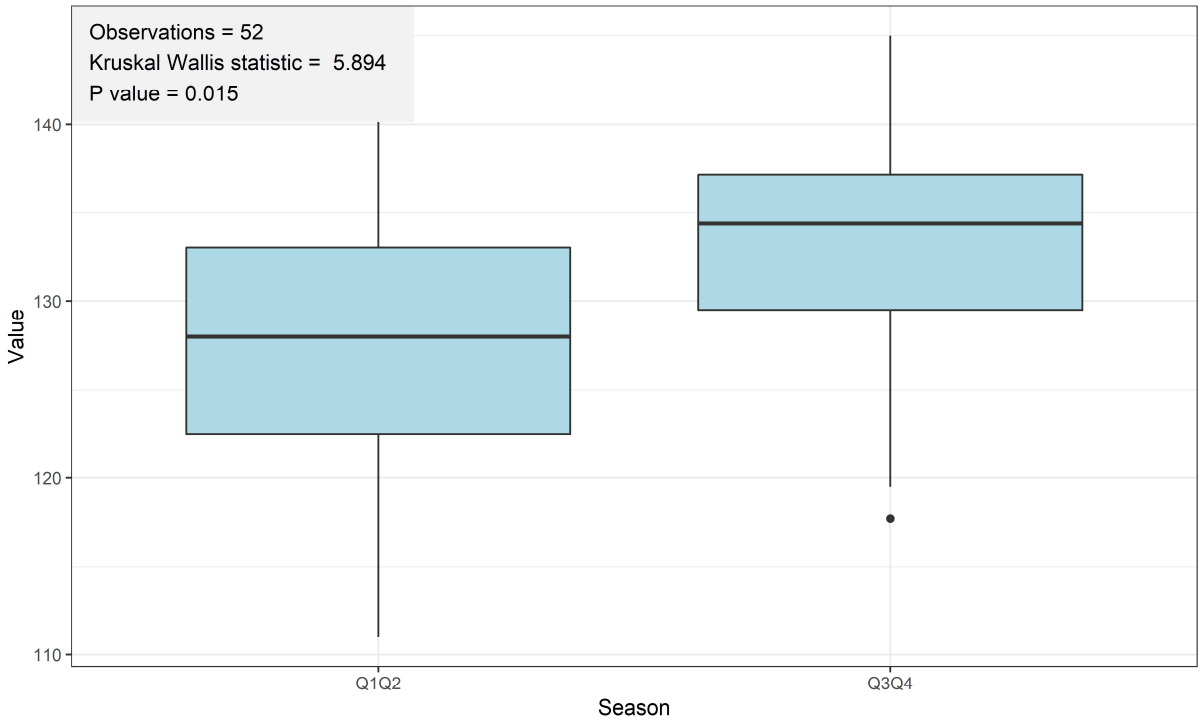

Figure 7-10: Example 2 – box plot of raw observations by season produced by LWPTrends. The black horizontal line in each box indicates the median of the observations, the box indicates the inter-quartile range, the whiskers indicate the 5th and 95th percentiles, and the dots indicate outliers.

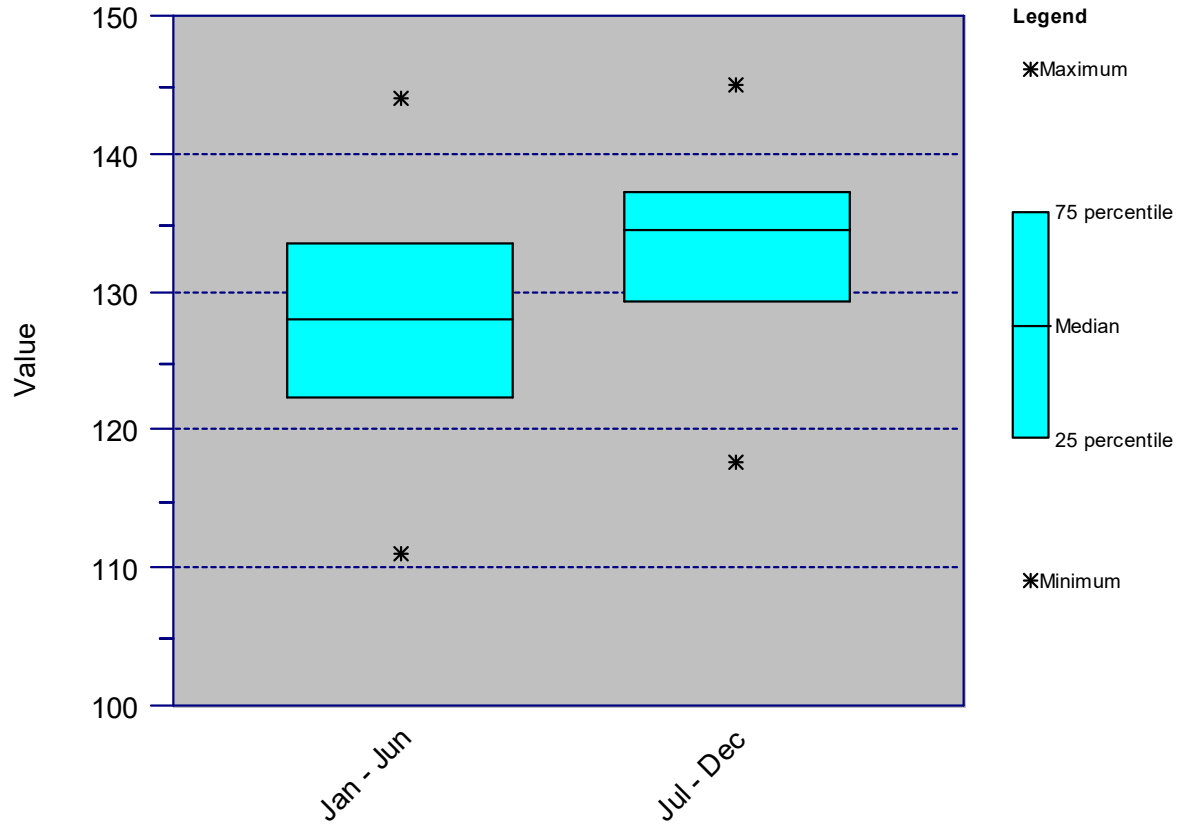

Figure 7-11: Example 2 – box plot of observations by season produced by TimeTrends.

Step 3: Perform Trend Assessment

Because the seasonality assessment indicates that the observations are seasonal, the trend assessment is conducted using the Seasonal Kendall assessment and the Seasonal Sen slope. We followed the procedure outlined in Example 1, with seasons in TimeTrends, (Analysis>Seasonal Kendall Test) set to "2 per year". TimeTrends also offers an additional choice for trend assessments where there are multiple observations within a season. In the box describing "season definition", at the bottom the user may select "Median value per season" (the default, and the only option in LWPTrends), or "All values in season". The second option should only be chosen if the observations within the defined seasons are independent. The "All values in season" option increases the power of the assessment and treats values within the same season as ties in time (i.e., slopes are not calculated between these observations). We performed the trend assessment twice in TimeTrends to demonstrate the impact of this option.

The plot produced by LWPTrends for this example is shown in Figure 7-12. The plots produced by TimeTrends, for the two alternative treatments of multiple observations within a season are shown in Figure 7-13 and Figure 7-14. Table 7-2 presents a summary of the combined trend assessment outputs.

Notable characteristics of these results are:

- Using all observations, rather than using the median within a season, increases the confidence in the trend direction (*C*) and results in differences in the trend rate (i.e., the Sen slope) and the confidence in the rate (i.e., 90% confidence intervals for the Sen slope).
- The exact magnitudes of the Sen slopes evaluated from the LWPTrends and TimeTrends packages differ slightly. This is due to slight differences in the way dates are assigned to the median values when there are multiple observations for a season.

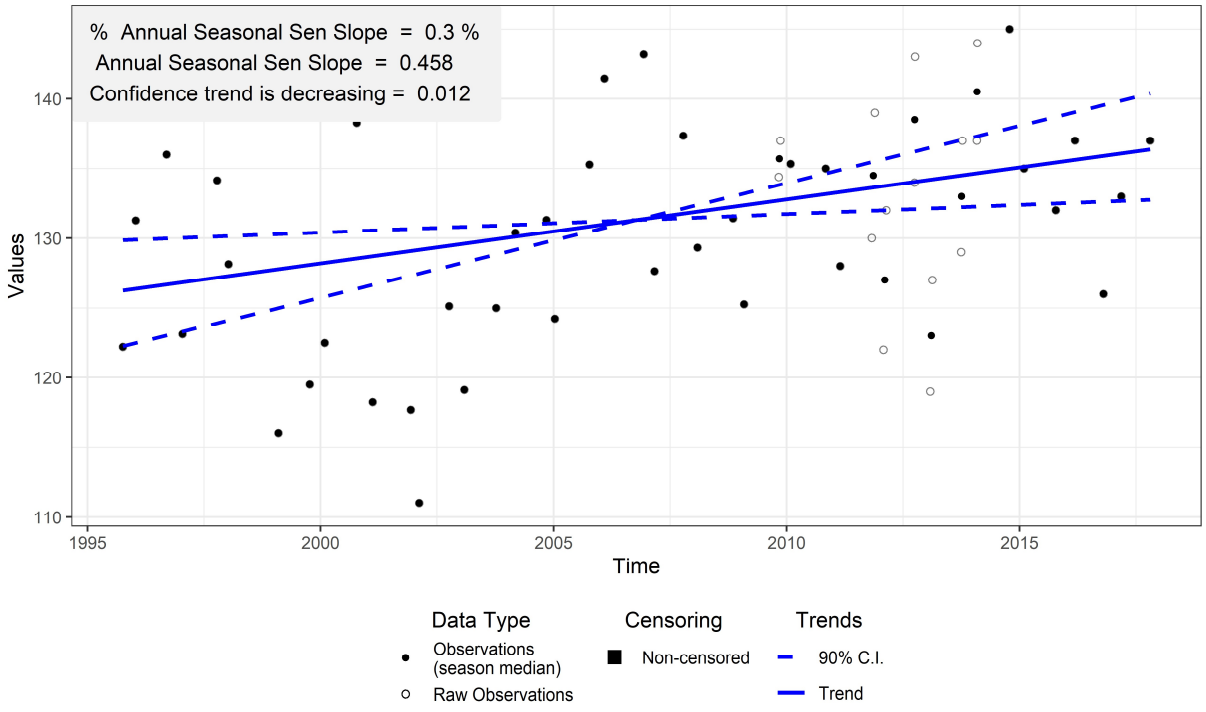

Figure 7-12: Example 2 – time-series of observations and fitted non-parametric (Sen) regression line produced by LWPTrends.

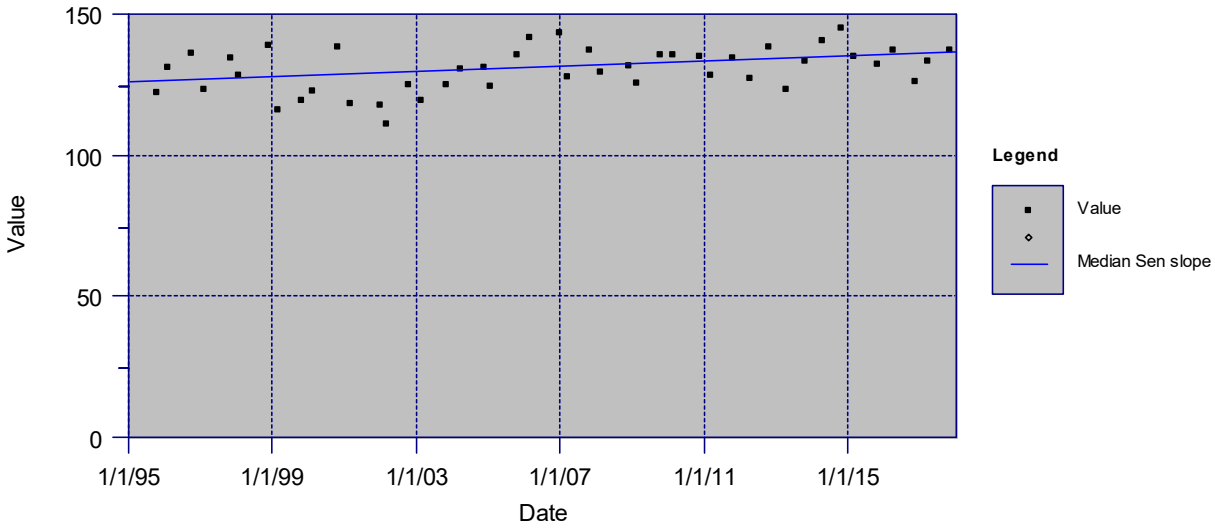

Figure 7-13: Example 2 – time-series of observations and fitted non-parametric (Sen) regression line produced by TimeTrends using the "Median value per season" option.

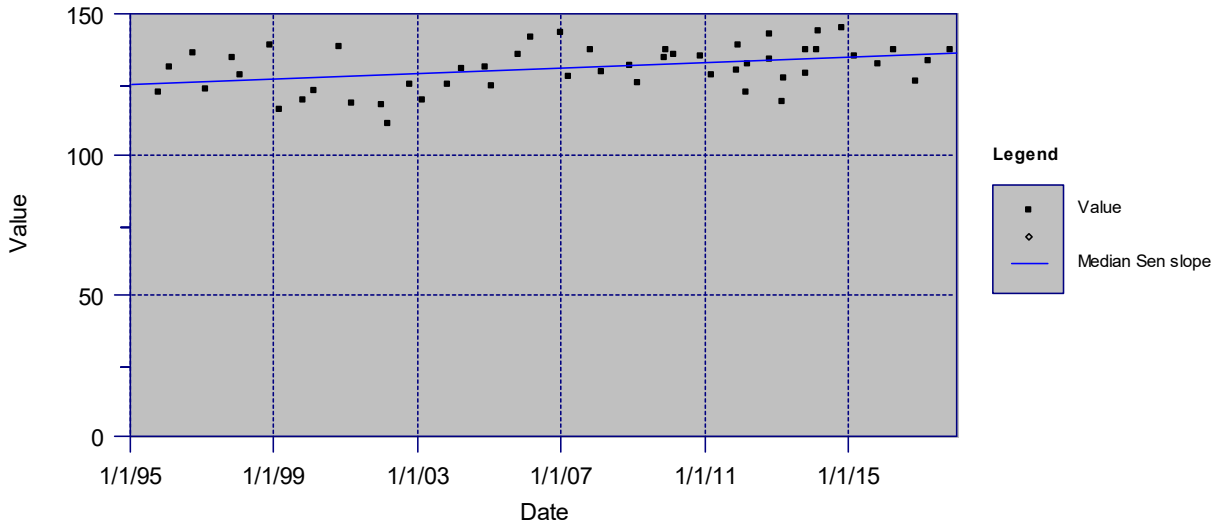

Figure 7-14: Example 2 – time-series of observations and fitted non-parametric (Sen) regression line produced by TimeTrends using the "All values in season" option.

Table 7-2: Tabulated results of trend assessment for Example 2.

	Kendall statistic	*C* (confidence in trend direction)	Sen slope (annual)	Percent annual change	90% confidence limits for slope	Direction
LWPTrends	118	0.988	0.458	0.349	0.132 to 0.823	Increasing
TimeTrends (median for period)	118	0.988	0.462	0.352	0.133 to 0.823	Increasing
TimeTrends (all independent observations)	156	0.992	0.472	0.359	0.179 to 0.785	Increasing

7.3 Example 3: Missing data

Example 3 uses a river monitoring site with a long-term record of observations of a chemical concentration. Simultaneous flow observations are available for the site. The purpose of the example is to demonstrate approaches to dealing with missing data. The dataset was created from a complete time-series of observations (monthly data) with several observations removed to create gaps in the time-series. We use the complete dataset at the end of the example to assess trend rate and direction to compare against the trend assessment for the dataset with gaps.

Step 1: Examine raw data

We generated plots to explore the raw data using both LWPTrends and TimeTrends following the procedure outlined in Example 1. The data summaries and plots, presented in Figure 7-15 and Figure 7-16, show:

- there are 10 years of monthly observations but a considerable proportion of sample intervals are gaps (~25% of all months), with these gaps tending to occur in the summer and winter (the last two years of data have no gaps);
- there is no censoring in the record; and
- it is unclear (from the scatter plot)whether there is a seasonal pattern, or a relationship between flow and concentration.

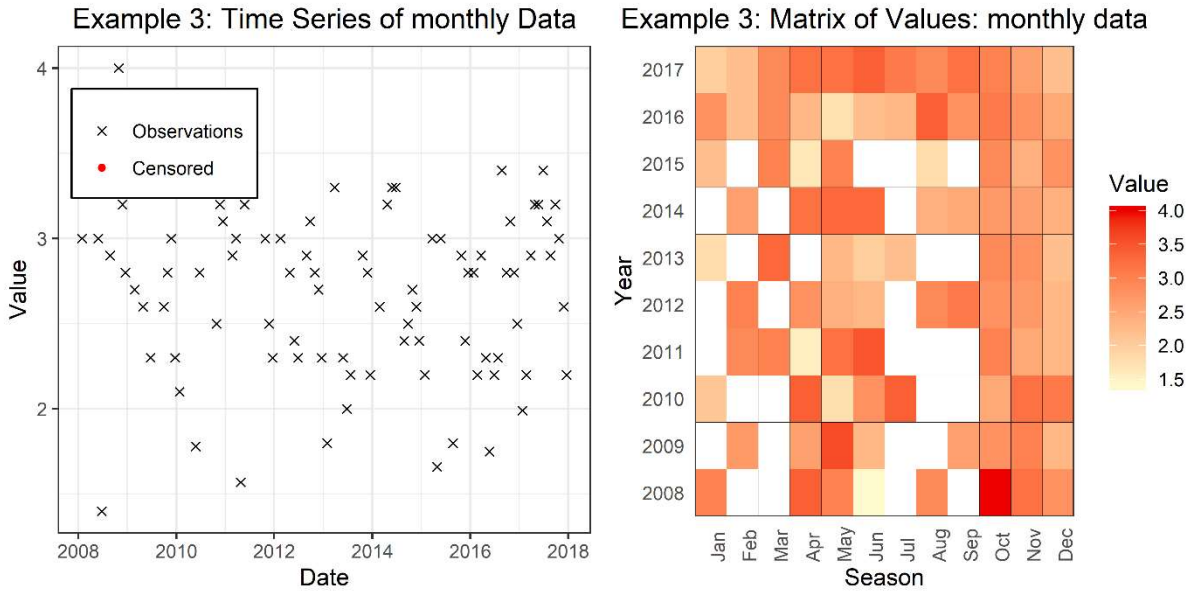

Figure 7-15: Example 3 – LWPTrends raw data inspection plots.

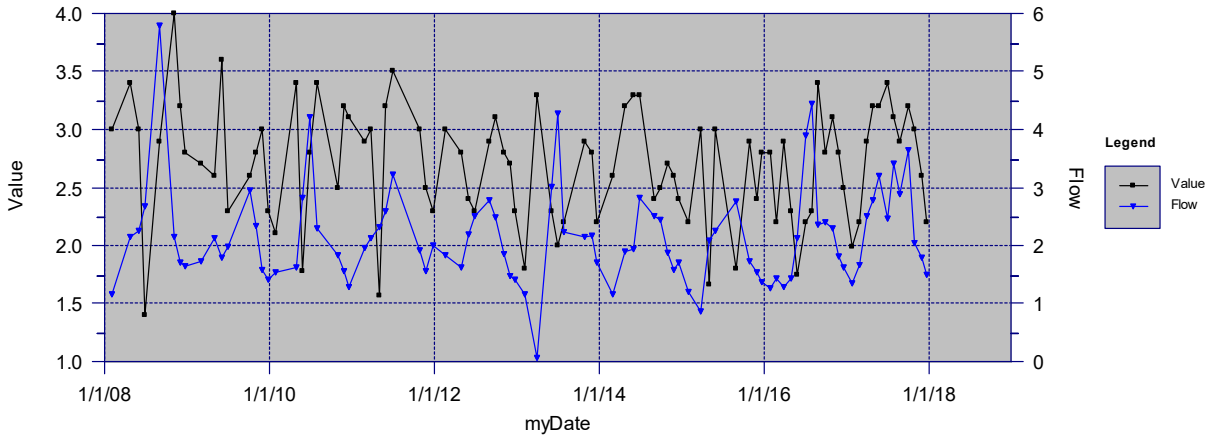

Figure 7-16: Example 3 – TimeTrends scatter plot of raw data.

There is a risk that if seasons are defined by months for the trend assessment, the assessment will be influenced by the large number of gaps, the bias of the gaps to certain times of the year, and the bias of the gaps to the earlier part of the record. An alternative approach would be to select a coarser definition of season so that the seasons are equally represented throughout the time period; although this has the disadvantage of reducing the total number of observations. For Example 3, trend analyses were performed with different definitions of seasons and the results were compared.

Step 2: Covariate adjustment and seasonality assessment

Because the example dataset includes observations of flow, there is the possibility to flow adjust the raw observations if there is a meaningful relationship between flow and concentration. Using LWPTrends we trialled four alternative models to describe the relationship between concentration and flow: Log-Log, GAM, LOESS(0.9 span), LOESS (0.7 span). The LWPTrends function 'Adjust Values' returns a plot of the observed values versus flow along with the fitted models (Figure 7-17).

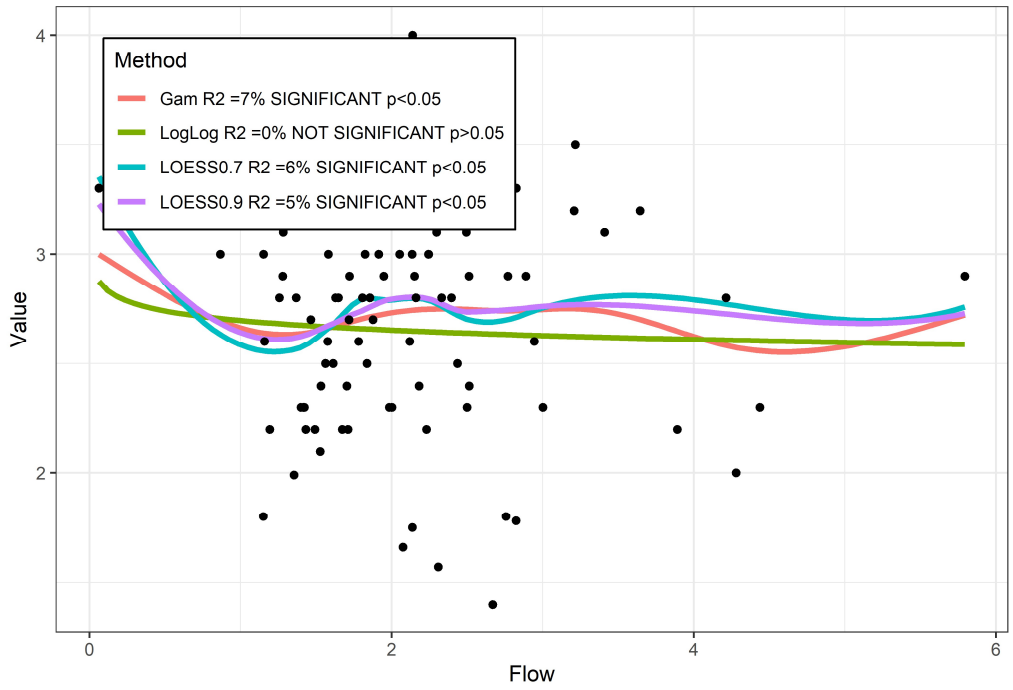

Figure 7-17: Example 3 – scatterplot of concentration and flow produced by LWPTrends. The lines represent four alternative models fitted to these data to represent the relationship between concentration and flow.

Based on inspections of the scatterplot, fitted models (Figure 7-17), and the model diagnostics (R^2 and *p*-values), we concluded that flow adjustment of the data was not justified. Therefore, subsequent steps were conducted with the raw data. We performed a seasonality assessment based on both the monthly and bi-monthly data using both LWPTrends and TimeTrends. Only the seasonality plots from TimeTrends are provided (Figure 7-18, Figure 7-19) because both packages produced identical results. The *p*-values for the Kruskall-Wallis test were 0.312 and 0.153 for the monthly and bi-monthly seasons, respectively. Although these plots do indicate some seasonality, neither alternative met the recommended *p*-value criteria of 0.05.

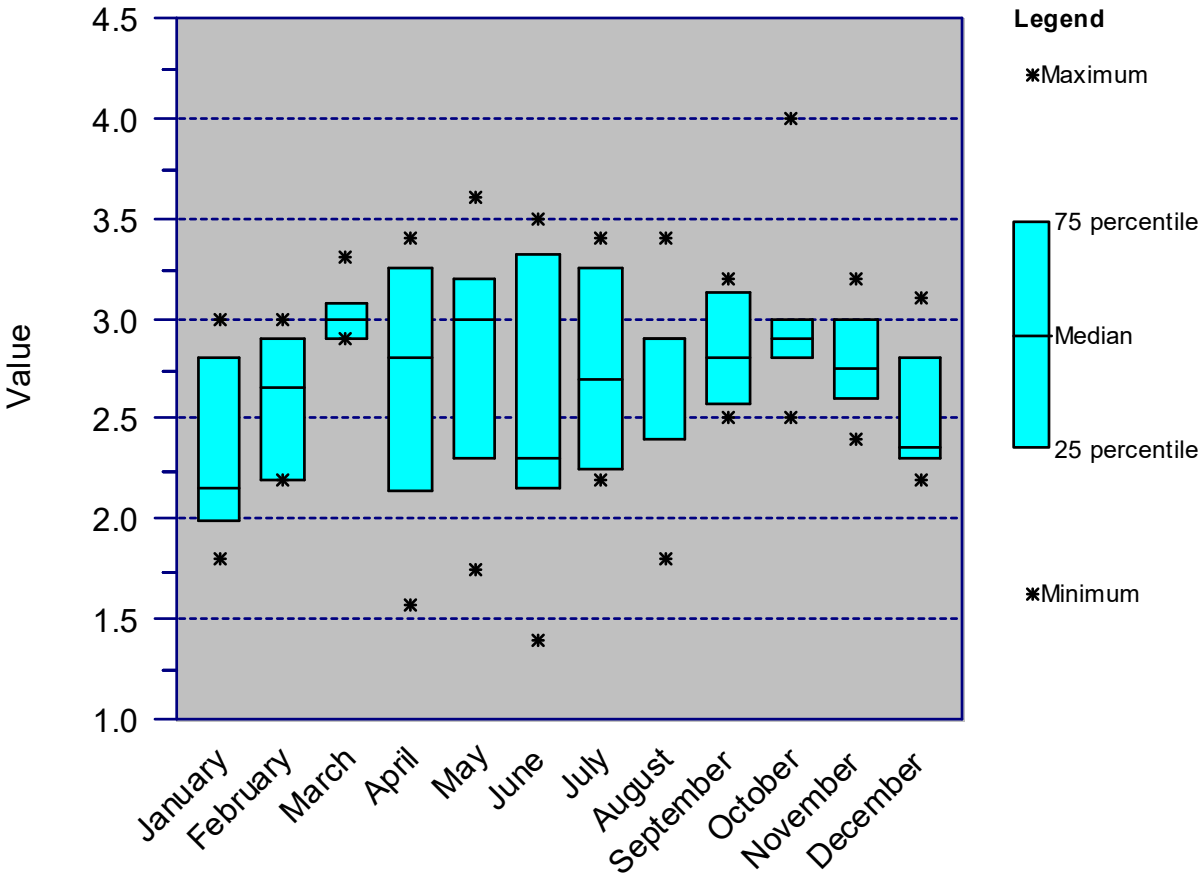

Figure 7-18: Example 3 – box plot of flow adjusted values by season produced by TimeTrends (seasons defined as months).

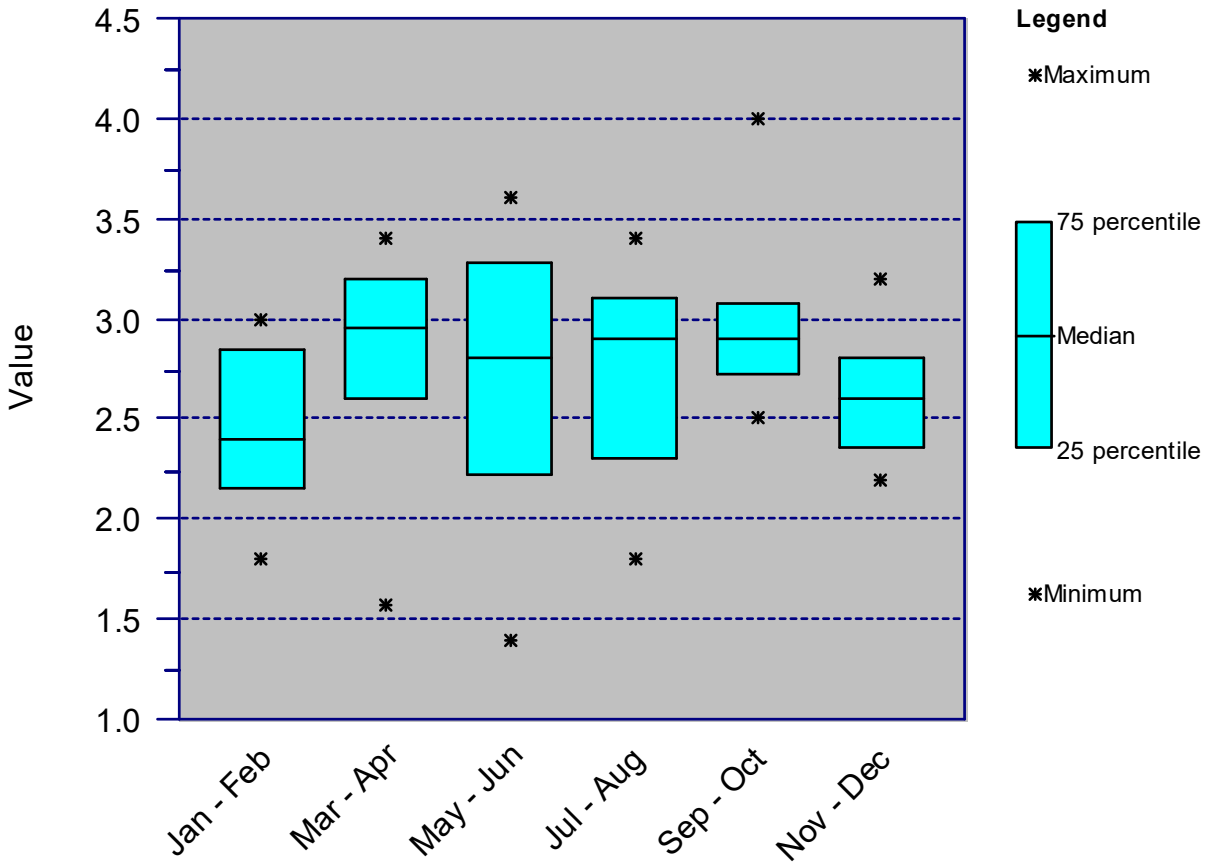

Figure 7-19: Example 3 – box plot of flow adjusted values by season produced by TimeTrends (seasons defined as bi-monthly, 'summer' and 'winter').

Step 3: Perform Trend Assessment

Because the seasonality assessment indicates the observations are not seasonal, the trend assessment is conducted using the Mann Kendall and the Sen slope assessments. For LWPTrends this is performed using the function "NonSeasonalTrendAnalysis", and in TimeTrends using Analysis>Mann-Kendall trend test. We performed the analyses in both packages first with seasons set to months, and then again with bi-monthly seasons. Note that the supplementary files contain the original complete dataset in the ("Example 3 – complete.csv") to allow comparison of the results with the complete dataset. The plots produced by LWPTrends for this example, for the two alternative season designations, are shown in Figure 7-20 and Figure 7-21. The combined trend assessment outputs from both packages are summarised in Table 7-3.

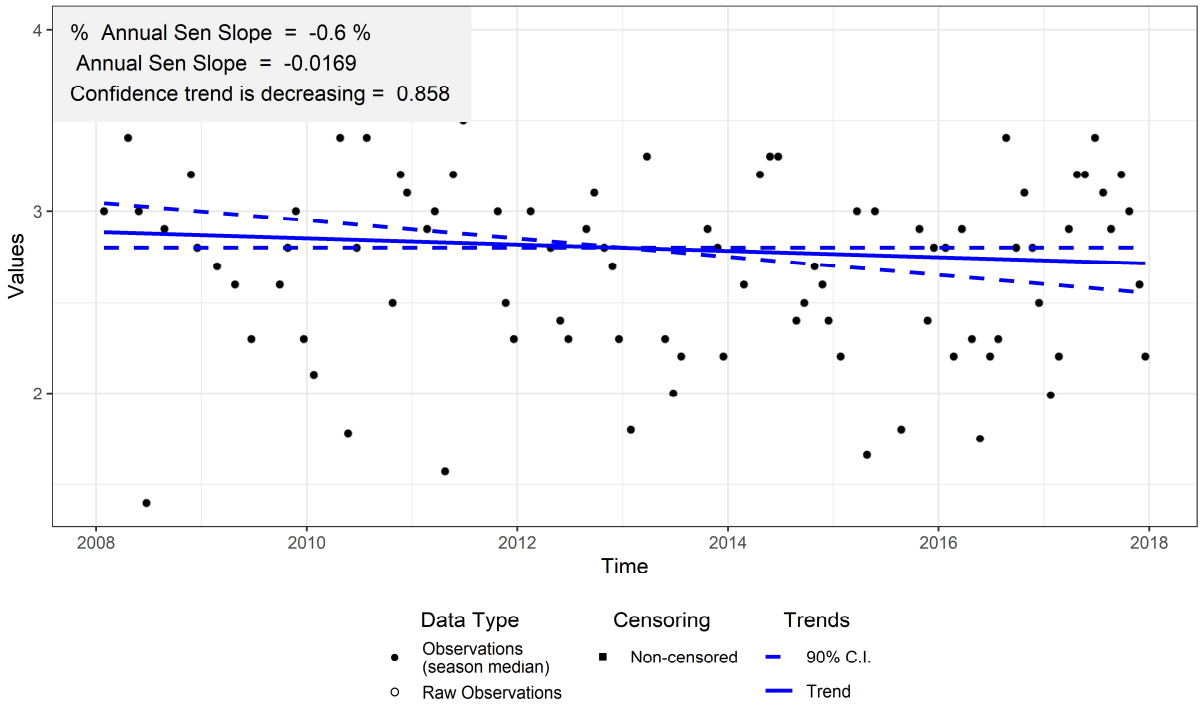

Figure 7-20: Example 3 – time-series of flow adjusted observations and fitted non-parametric (Sen) regression line produced by LWPTrends (seasons defined as months).

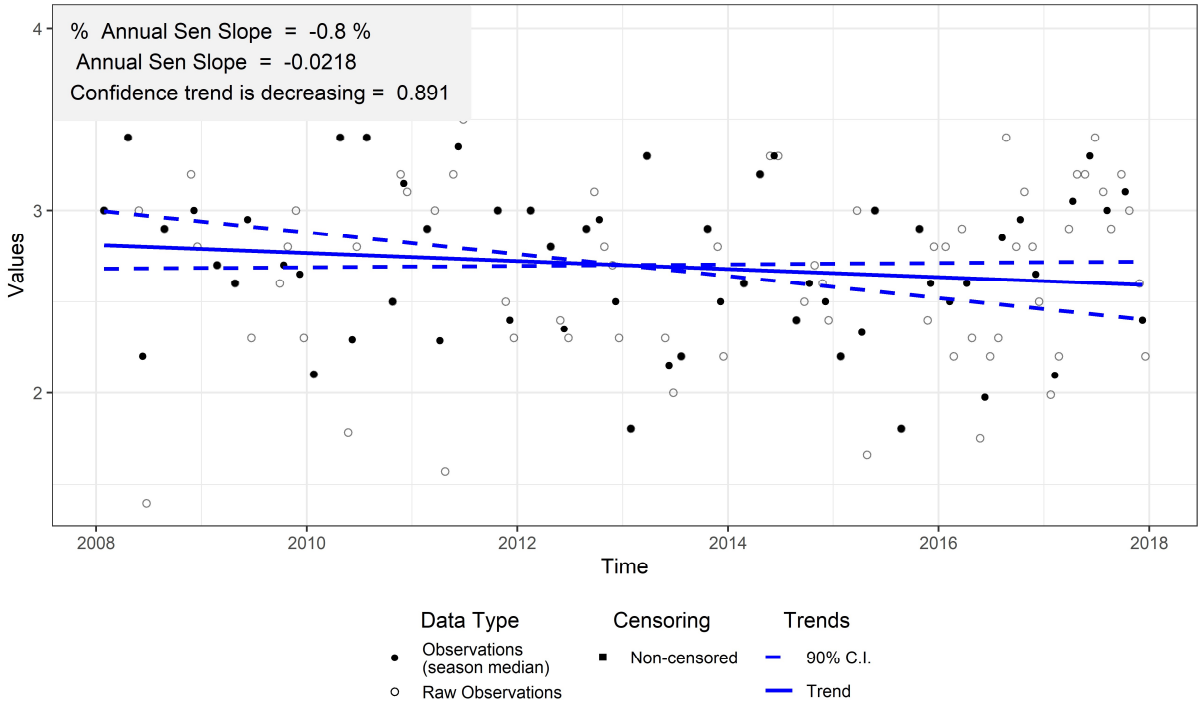

Figure 7-21: Example 3 – time-series of flow adjusted observations and fitted non-parametric (Sen) regression line produced by LWPTrends (seasons defined as bi-monthly, 'summer' and 'winter').

Table 7-3: Tabulated results of trend assessment for Example 3.

	Kendall statistic	C (Confidence in trend direction)	Sen slope (annual)	Percent annual change	90% confidence limits for slope	Direction
LWPTrends (monthly)	-307	0.858	-0.017	-0.603	-0.049 to 0	Decreasing
TimeTrends (monthly)	-307	0.858	-0.017	-0.603	-0.049 to 0	Decreasing
LWPTrends (bi-monthly)	-184	0.891	-0.022	-0.780	-0.060 to 0.004	Decreasing
TimeTrends (bi-monthly)	-180	0.886	-0.022	-0.809	-0.060 to 0.005	Decreasing

There were a number of subjective decisions made in the above trend assessment including:

1. the use of p=0.05 as the cut-off threshold for identifying seasonality;
2. the choice of seasons; and
3. if seasons are less frequent than monthly, whether or not to use the median for the season or to use all datapoints (only an option in TimeTrends).

To explore the implications of these choices, we compared eight alternative analyses performed using TimeTrends, using the seasonal and non-seasonal variants of: (1) the complete dataset with seasons as months; (2) the dataset with gaps and seasons as months; (3) the dataset with gaps, bi-monthly seasons and median values for seasons; and (4) the dataset with gaps, bi-monthly seasons and all data points. The results from these eight alternatives are summarised in Table 7-4.

Table 7-4 indicates there are small differences in the confidence in trend direction and the trend rate (Sen slope) across the range of analysis options. An appropriate conclusion is that it is at least likely (see Table 6-1) that there is a decreasing trend in the concentration at this site.

Table 7-4: Tabulated results of trend assessment for Example 3 for eight alternative analyses. Highlighted rows are based on the original complete dataset. All other rows are based on the data with gaps.

	Season	Kendall statistic	*C* (Confidence in trend direction)	Sen slope (annual)	Percent annual change	90% confidence limits for slope
Non-seasonal assessment	Monthly	-672	0.937	-0.024	-0.856	-0.054 to 0
	Monthly	-307	0.858	-0.017	-0.603	-0.049 to 0
	Bi-monthly [a]	-180	0.886	-0.022	-0.809	-0.060 to 0.005
	Bi-monthly [b]	-298	0.851	-0.016	-0.586	-0.049 to 0
Seasonal Assessment	Monthly	-65	0.953	-0.020	-0.716	-0.050 to 0
	Monthly	-37	0.900	-0.020	-0.714	-0.050 to 0
	Bi-monthly [a]	-37	0.917	-0.020	-0.752	-0.051 to 0
	Bi-monthly [b]	-54	0.843	-0.020	-0.708	-0.050 to 0

Notes: (a) median taken for seasons; (b) all data used within seasons.

7.4 Example 4: High censoring

Example 4 uses a river monitoring site with a long-term record of a chemical concentration. Flow observations are not available for the site. The detection limit changed part way through the record. The purpose of the example is to demonstrate choices around handling changing detection limits in an analysis. The example data file contains three alternative versions of the observation time series:

"Value" – detection limits decrease half-way through the observation period; "Value1" – detection limits increase halfway through the time period; "Value2" – no censored observations. The analysis steps are only shown for the first dataset but results for all three datasets are presented for comparison at the end of the example.

Step 1: Examine raw data

We generated plots to examine the raw data using both LWPTrends and TimeTrends following the procedure outlined in Example 1. The data summaries and plots produced by the LWPTrends package (Figure 7-22) indicate:

- there are 10 years of monthly data with one gap;
- approximately 40% of the observations are censored, with almost 60% of observations censored over the first five years;
- the detection limit has changed halfway through the period from 0.015 to 0.01; and
- there appears to be some seasonality in the observations, with higher concentrations in the summer months and lowest concentrations (and greatest prevalence of censoring) in the spring months.

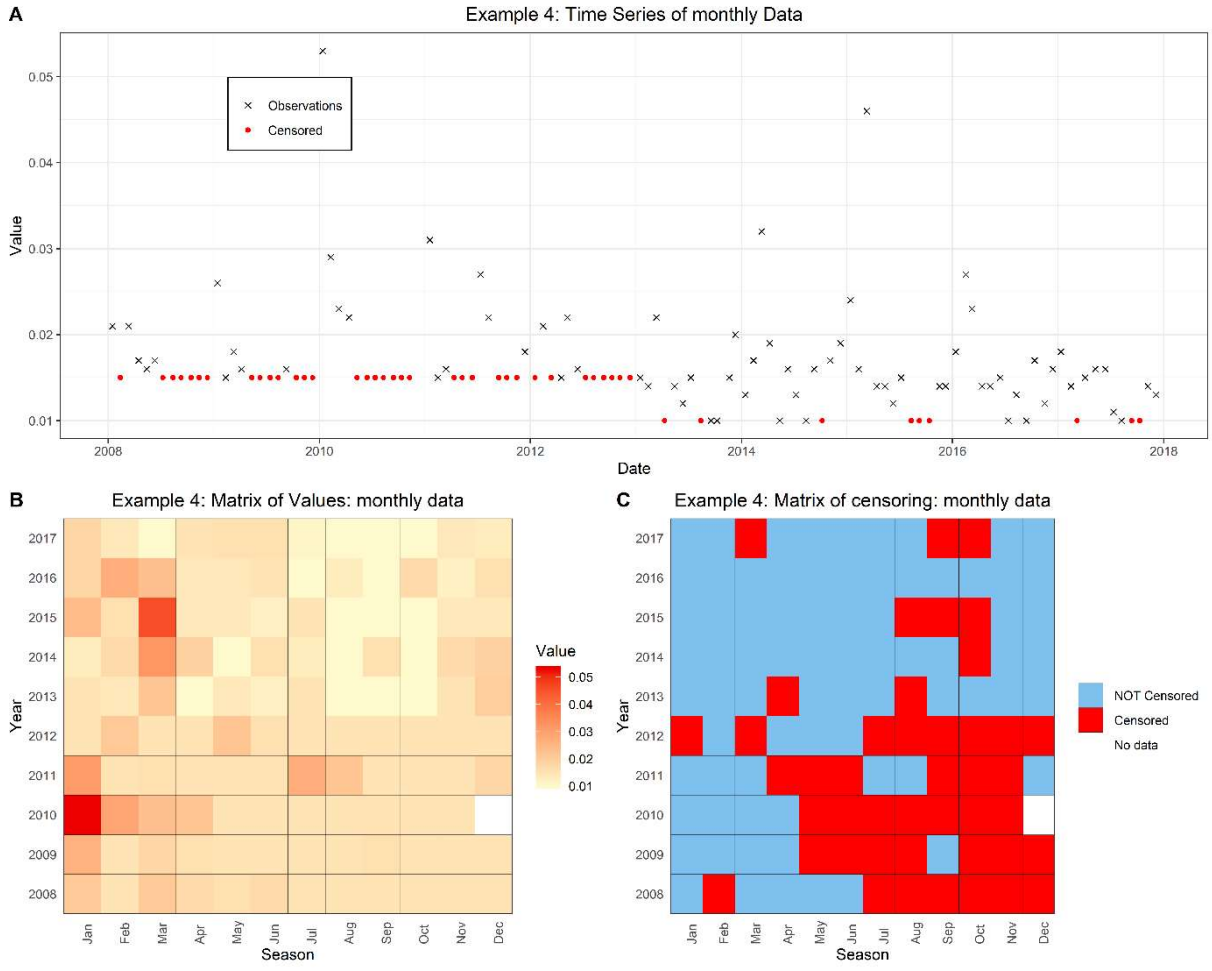

Figure 7-22: Example 5 – LWPTrends raw data inspection plots.

Step 2: Covariate adjustment and seasonality assessment

We checked the seasonality of the data with seasons defined by months using both LWPTrends and TimeTrends as described for Example 1. TimeTrends issued two warnings for this analysis: "Warning - more than 20% of data are censored"; and "Insufficient uncensored data in variable for evaluation of censored values using ROS. Substituted values used instead. If using a seasonal analysis consider using a longer season". If the warnings are suppressed, the analysis continues. Both packages indicate that the data are seasonal (Figure 7-23, Figure 7-24), with the *p*-value from the Kruskal-Wallis test <0.002.

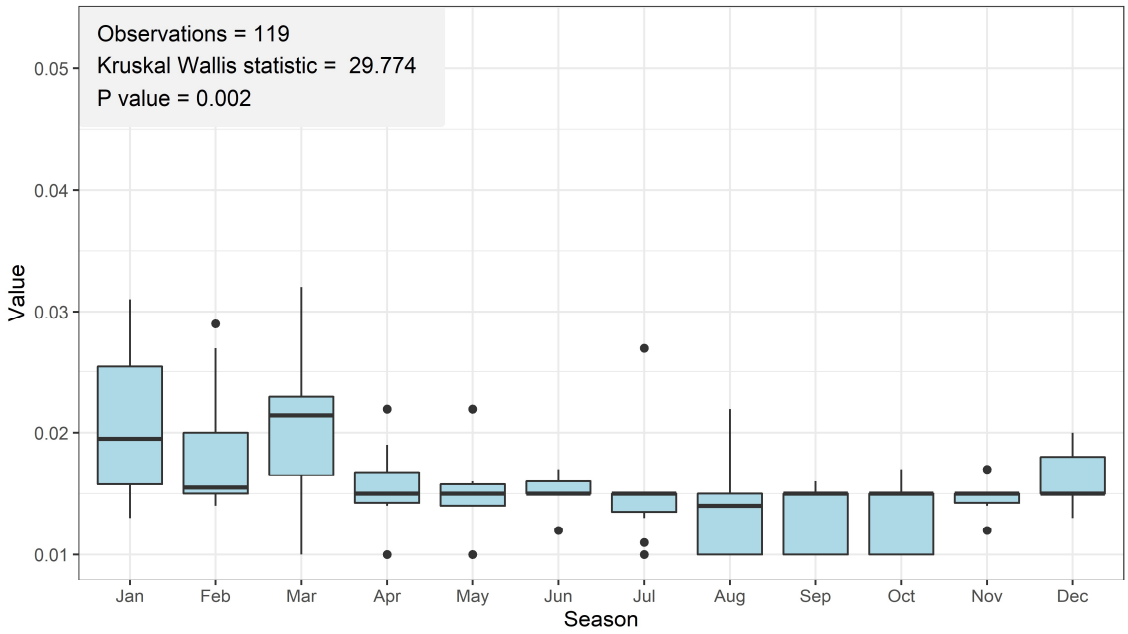

Figure 7-23: Example 4 – box plot of flow adjusted values by season produced by LWPTrends. The black horizontal line in each box indicates the median of the observations, the box indicates the inter-quartile range, the whiskers indicate the 5th and 95th percentiles, and the dots indicate outliers.

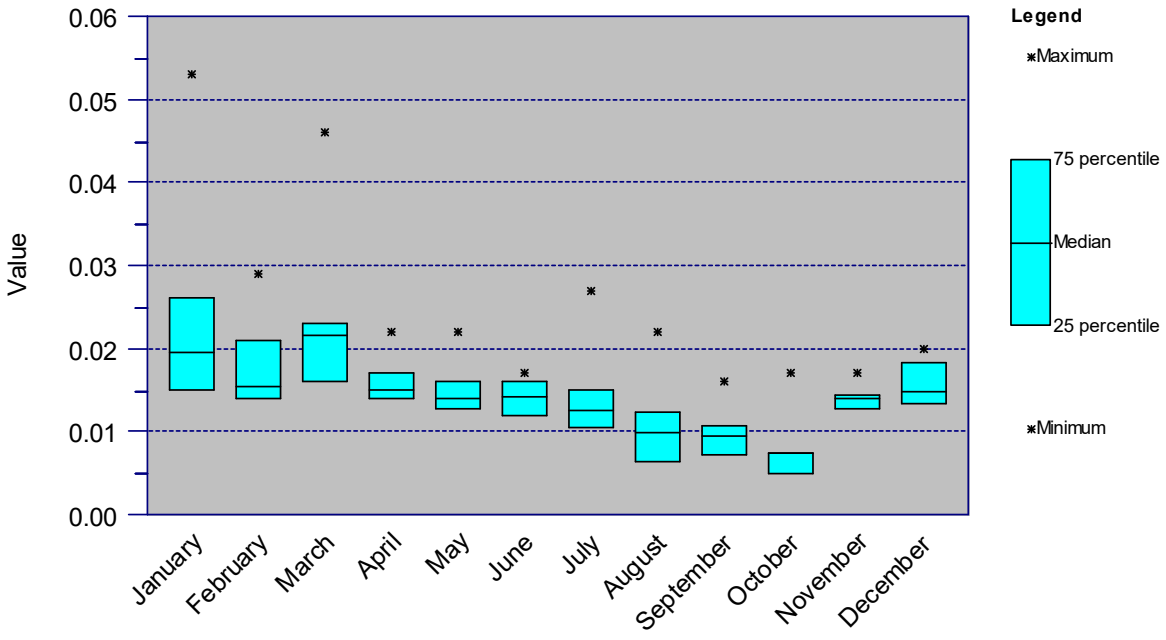

Figure 7-24: Example 4 – box plot of flow adjusted values by season produced by TimeTrends.

Step 3: Perform Trend Assessment

Because the seasonality assessment indicates that the observations are seasonal, the trend assessment is conducted using the Seasonal Kendall test and the Seasonal Sen slope. Using LWPTrends, the assessment is performed using the function "SeasonalTrendAnalysis", and in TimeTrends using Analysis>Seasonal Kendall Test (with the box "Use censored values for slope calculation" checked). The hi censoring filter is implemented in LWPTrends by adding an argument to the "SeasonalTrendAnalysis" function: HiCensor=TRUE, and in TimeTrends by checking the box "Set all values (censored or otherwise) less than the highest detection limit as censored at the highest detection limit". See Section 3-2for an explanation of how the censored values are treated when these options are selected.

Plots produced by LWPTrends showing the raw data with the assessed seasonal Sen slope are shown in Figure 7-25 and Figure 7-26 for the assessments without and with the hi censor filter, respectively. A summary of the trend assessment outputs from both packages is provided in Table 7-5.

The results from the two packages are consistent with each other, although there are some small differences in the calculated variances, confidence levels and SSE confidence intervals. In this example, the implementation of the hi censor filter had very little impact on the assessment results. A noteworthy result of the assessment is that the direction of the trend, as indicated by the *S* statistic, is not consistent with the direction of the trend indicated by the confidence intervals of the Sen slope. This can occur when there are many tied values in the observations; either due to many censored values or the values in the data having low precision relative to the distribution of the observations. This outcome is not a cause for concern and simply indicates that the trend is small compared to the information provided by the data. Assuming that laboratory methods to measure water quality are fit for purpose (i.e., appropriate precision to measure concentrations at the levels of interest), then this issue will only arise at those sites with low concentrations and small trends.

Table 7-5: Tabulated results of trend assessment for Example 4.

	Kendall statistic	*C* (Confidence in trend direction)	Sen slope (annual)	Percent annual change	90% confidence limits for slope	Direction
TimeTrends	-6	0.556	0	0	0 to 0.0004	Decreasing
TimeTrends (HI censor)	-6	0.563	0	0	0 to 0.0004	Decreasing
LWPTrends	-6	0.557	0	0	0 to 0.0004	Decreasing
LWPTrends (Hi censor)	-6	0.568	0	0	0 to 0	Decreasing

The designation of trend direction in Table 7-5 is based on an interpretation of the sign of *S*. TimeTrends returned two alternative narratives to describe "Trend direction and confidence", one associated with the Mann Kendall test ("Trend unlikely") and one associated with the Seasonal Sen slope estimate ("Increasing trend possible"). LWP Trends issued a warning associated with the Seasonal Sen slope estimate: "WARNING: Sen slope influenced by censored values".

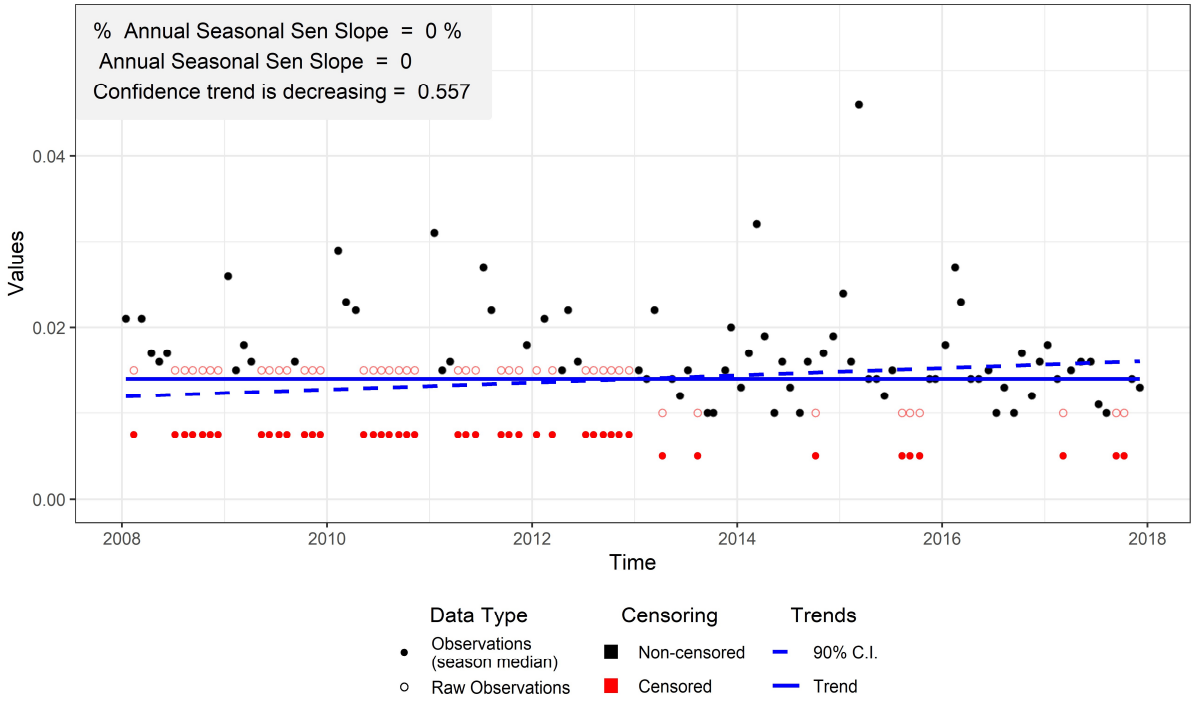

Figure 7-25: Example 4 – time-series of observations and fitted non-parametric (Sen) regression line produced by LWPTrends. Note the change in detection limit in 2013.

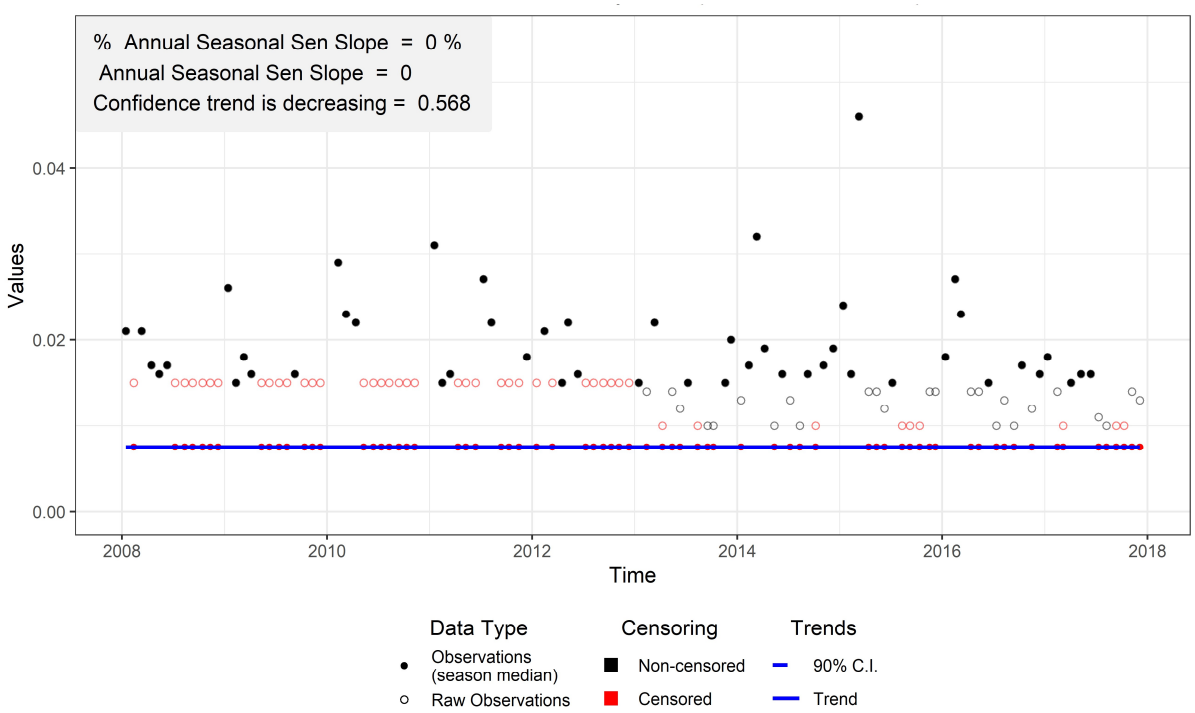

Figure 7-26: Example 4 – time-series of observations and fitted non-parametric (Sen) regression line produced by LWPTrends. Note the hi-censor limit filter has been applied in this case which has reduced all values prior to the change in the detection limit in 2013 to the new detection limit (compare with Figure 7-25).

For comparison, we have also performed trend assessments for Example 4 based on the additional two datasets provided in the example file (one where the change in detection limit goes from low to high, and the other being the original data with no censoring). The results for the trend assessments for all three datasets (as analysed using TimeTrends) are summarised Table 7-6. Note, applying the hi-censor filter for first two datasets produces the same result as they both have the same maximum detection limit.

It is interesting to note that, while the confidence intervals for all Sen slopes are consistent, the directions from the assessments are inconsistent. When there are no censored values (Complete set, "Value2"), the trend is assessed as increasing, whereas the hi-censor filter and the dataset with a step change in detection limit from a higher to lower value ("Value") indicates a decreasing trend, and the dataset with a step change from a low to high detection limit ("Value1") has an indeterminate trend. Censoring also significantly reduces the confidence in the assessed trend direction. Despite these apparent inconsistencies, the conclusions that a pragmatic user would make for all four assessments are: (1) the site has good water quality (evidenced by the frequency that observations are below the detection limit) and (2) the trend is very small, and hence the "direction" is likely to have limited practical consequence.

Table 7-6: Example 3 – Trend assessment results from TimeTrends exploring the impact of the direction of step change in detection limit, the implementation of the hi-censor filter, and based on the complete, uncensored dataset.

Dataset	Kendall statistic	*C* (Confidence in trend direction)	Sen slope (annual)	Percent annual change	90% confidence limits for slope	Direction
High to low DL ("Value")	-6	0.556	0	0	0 to 0.0004	Decreasing
Low to high DL ("Value1")	1	0.5	0	0	0 to 0.0003	Indeterminate
Hi-censor filter	-6	0.563	0	0	0 to 0.0004	Decreasing
Complete set ("Value2")	35	0.815	0.0001	0.43	0 to 0.0004	Increasing

8 Acknowledgements

We thank senior members of the regional sector's Surface Water Integrated Management (SWIM) Special Interest Group for prompting and supporting the preparation of this guidance, and in particular, Abby Matthews (Horizons Regional Council) for seeking the MBIE Envirolink Advice Grant that funded its preparation. Additional in-kind funding was provided by NIWA's Freshwater & Estuaries National Science Centre.

Judi Hewitt and Juliet Milne (NIWA) and Eric Goodwin (Cawthron) provided insightful reviews and discussions that improved the guidance. A number of regional council staff also provided feedback on a draft version of this guidance, in particular Tom Stephens and Coral Grant (Auckland Council), Mike Scarsbrook (Waikato Regional Council), Luke Fullard (Horizons Regional Council), Roger Hodson (Environment Southland) and Andy Hicks (Hawke's Bay Regional Council).

9 Glossary of abbreviations and terms

Term	Description
Censored values	Measurements of water quality variables that are too low or too high to be measured with precision. The "detection limit" is the lowest value that can be reliably measured by an analysis and the "reporting limit" is the greatest value of a variable that can be reliably measured. Values that are reported as either the detection limit or reporting limit are referred to as censored.
Confidence (*C*)	The evaluated trend direction confidence level. *C* can be understood as the probability that the assessed trend direction is the same as the true trend (or that the assessed direction is correct). *C* varies between 0.5 (low confidence) and 1 (high confidence).
Covariate adjustment	A statistical analysis that is applied to the time series of water quality observations to remove the variation that is explained by another observed variable. When the other variable is river flow, this analysis is known as flow adjustment.
Filtering rules	Rules that define the acceptable proportion of gaps and representation of sample intervals by observations within the time period. Also referred to as 'site screening criteria' and 'completeness criteria'.
Local application	An application in which the objective is to extract as much information as possible about the trend direction and rate from the available data for a single (or for each) site/variable combination.
Monotonic	Model of the behaviour of the water quality variable through time that is constrained to be either constantly increasing or decreasing.
National application	An application in which the objective is to assess and report trends across many sites and variables using data obtained from several regional SOE monitoring programmes. The objective is to allow robust comparison of trends between sites and to provide a synoptic assessment of trends across the whole country.
Non-parametric model	A statistical model for which there is no assumptions concerning the distribution of the data.
Parametric model	A statistical model for which there is an assumption that the data have an underlying distribution (e.g., that the data are normally distributed).
Regional application	An application in which the objective is to assess and report trends across many sites and variables using data from a regional SOE monitoring programme (or similar). The objective is to allow robust comparison of trends between sites and to provide a synoptic assessment of trends across a whole region.
Sample interval	A specific time interval in which an observation occurs that is defined by each season within each year.
Season	Water quality observations generally have a set frequency, which is determined by the sampling frequency (e.g., monthly, quarterly). The trend assessment 'season' is generally specified to match this frequency (e.g., seasons are months or quarters). In some circumstances, the temporal resolution of the data is coarsened, for example, monthly data is coarsened to quarterly.
Time period	The interval of time over which a trend is assessed.
Traditional non-parametric method	The approach to water quality trend assessment recommended by these guidelines. The statistical robustness of the traditional non-parametric method make it the "safest" option for the three types of trend applications identified by these guidelines (i.e., local, regional and national applications).
Trend	Behaviour of a variable over time. In this guidance a trend is quantified by the direction and rate of change.

Term	Description
Trend direction	The direction of the trend; either increasing or decreasing.
Trend rate	The assessed rate of change of the trend. The units are those of the water quality variable being analysed per unit time (which is generally expressed as years).

10 References

Ali, R., Kuriqi, A., Abubaker, S., Kisi, O. (2019) Long-Term Trends and Seasonality Detection of the Observed Flow in Yangtze River Using Mann-Kendall and Sen's Innovative Trend Method. *Water*, 11: 1855.

Ballantine, D., Booker, D., Unwin, M., Snelder, T. (2010) Analysis of National River Water Quality Data for the Period 1998–2007. *NIWA Client Report*, *Christchurch*, *New Zealand*.

Bayarri, M.J., Berger, J.O. (2004) The Interplay of Bayesian and Frequentist Analysis. *Statistical Science*: 58–80.

Broman, K.W., Woo, K.H. (2018) Data Organization in Spreadsheets. *The American Statistician,* 72: 2–10.

Campbell, J.L., Rustad, L.E., Porter, J.H., Taylor, J.R., Dereszynski, E.W., Shanley, J.B., Gries, C., Henshaw, D.L., Martin, M.E., Sheldon, W.M. (2013) Quantity Is Nothing without Quality: Automated QA/QC for Streaming Environmental Sensor Data. *BioScience,* 63: 574–585.

Choquette, A.F., Hirsch, R.M., Murphy, J.C., Johnson, L.T., Confesor, R.B. (2019) Tracking Changes in Nutrient Delivery to Western Lake Erie: Approaches to Compensate for Variability and Trends in Streamflow. *Journal of Great Lakes Research*, 45: 21–39.

Cohen, A.C. (1976) Progressively Censored Sampling in the Three Parameter Log-Normal Distribution. *Technometrics,* 18: 99–103.

Cohen, J. (1994) The Earth Is Round (P<. 05). *American Psychologist*, 49: 997.

Darken, P.F., Zipper, C.E., Holtzman, G.I., Smith, E.P. (2002) Serial Correlation in Water Quality Variables: Estimation and Implications for Trend Analysis. *Water Resources Research* 38(7): 22-1.

Daughney, C. Randall, M. (2009) National Groundwater Quality Indicators Update: State and Trends 1995–2008. *GNS Science Consultancy Report 2009/145*, prepared for the Ministry for the Environment.

Davies-Colley, R.J., Hughes, A.O., Verburg, P., Storey R. (2012) Freshwater Monitoring Protocols and Quality Assurance (QA). *NIWA Client Report,* Hamilton, New Zealand.

Davies-Colley, R., McBride, G. (2016) Accounting for Changes in Method in Long-term Nutrient Data: Recommendations Based on Analysis of Paired SoE Data from Wellington Rivers. *NIWA Client Report*, Hamilton, New Zealand.

Davies-Colley, R., Milne, J., Heath, M.W. (2019) Reproducibility of River Water Quality Measurements: Inter-Agency Comparisons for Quality Assurance. *New Zealand Journal of Marine and Freshwater Research*, 53: 437–450.

Ellis, S.E., Leek, J.T. (2018) How to Share Data for Collaboration. *The American Statistician* 72: 53–57.

Fraser, C.E., Snelder, T. (2018) State and Trends of River Water Quality in the Manawatū-Whanganui Region: For All Records up to 30 June 2017. *Landwaterpeople*.

Gadd, J., Snelder, T., Fraser, C., Whitehead, A. (2020) Current State of Water Quality Indicators in Urban Streams in New Zealand. *New Zealand Journal of Marine and Freshwater Research*: 1–18.

Greenland, S., Poole, C. (2013) Living with P Values: Resurrecting a Bayesian Perspective on Frequentisi Statistics. *Epidemiology*: 62–68.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman D.G. (2016) Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations. *European Journal of Epidemiology*, 31: 337–350.

Hald, A. (1949) Maximum Likelihood Estimation of the Parameters of a Normal Distribution Which Is Truncated at a Known Point. *Scandinavian Actuarial Journal,* 1949: 119–134.

Hart, E.M., Barmby, P., LeBauer, D., Michonneau, F., Mount, S., Mulrooney, P., Poisot, T., Woo, K.H., Zimmerman, N.B., Hollister J.W. (2016) T*en Simple Rules for Digital Data Storage*. Public Library of Science San Francisco, CA USA.

Helsel, D.R. (1990) Less Than Obvious-Statistical Treatment of Data Below the Detection Limit. *Environmental Science & Technology,* 24: 1766–1774.

Helsel, D.R. (2005) *Nondetects and Data Analysis. Statistics for Censored Environmental Data*. Wiley-Interscience, New Jersey.

Helsel, D.R. (2012) Statistics for Censored Environmental Data Using Minitab and R. John Wiley & Sons, Inc., Hoboken, New Jersey. http://onlinelibrary.wiley.com/doi/10.1002/9781118162729.ch3/summary. Accessed 19 Aug 2016.

Helsel, D.R., Hirsch, R.M. (1992) *Statistical Methods in Water Resources*. Elsevier, Amsterdam, The Netherlands.

Helsel, D.R., Hirsch, R.M., Ryberg K.R., Archfield, S.A., Gilroy, E.J. (2020) *Statistical Methods in Water Resources.* Report, Reston, VA.

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

Hirsch, R.M., Moyer, D.L., Archfield, S.A. (2010) Weighted Regressions on Time, Discharge, and Season (WRTDS), with an Application to Chesapeake Bay River Inputs 1. *Journal of the American Water Resources Association,* 46: 857–880.

Hirsch, R.M., Slack J.R. (1984) A Nonparametric Trend Test for Seasonal Data with Serial Dependence. *Water Resources Research*, 20: 727–732.

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

Horsburgh, J.S., Reeder, S.L., Jones, A.S., Meline, J. (2015) Open Source Software for Visualization and Quality Control of Continuous Hydrologic and Water Quality Sensor Data. *Environmental Modelling & Software*, 70: 32–44.

Jones, L.V., Tukey, J.W. (2000) A Sensible Formulation of the Significance Test. *Psychological Methods*, 5: 411.

Kilroy, C., Biggs, B.J. (2002) Use of the SHMAK Clarity Tube for Measuring Water Clarity: Comparison with the Black Disk Method. *New Zealand Journal of Marine and Freshwater Research,* 36: 519–527.

Larned, S.T., Scarsbrook, M.R., Snelder, T., Norton, N.J., Biggs, B.J.F. (2004) Water Quality in Low-Elevation Streams and Rivers of New Zealand. *New Zealand Journal of Marine and Freshwater Research,* 38: 347–366.

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

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

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

Larned, S., Whitehead, A., Fraser, C.E., Snelder, T., Yang, J. (2018a) Water Quality State and Trends in New Zealand Rivers. Analyses of National-Scale Data Ending in 2017. *NIWA Client Report*, Christchurch, New Zealand.

Larned, S.T., Snelder, T., Whitehead, A., Fraser, C. (2018b) Water Quality State and Trends in New Zealand Lakes. *NIWA Client Report*, Christchurch, New Zealand.

Makowski, D., Ben-Shachar, M.S., Chen, S.H.A., Lüdecke, D. (2019) Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology,* 10. doi:10.3389/fpsyg.2019.02767.

Mann, H.B. (1945) Nonparametric Tests Against Trend. *Econometrica: Journal of the Econometric Society*: 245–259.

Mast, M.A. (2013) Evaluation of Stream Chemistry Trends in US Geological Survey Reference Watersheds, 1970–2010. *Environmental Monitoring and Assessment* 185: 9343–9359.

Mastrandrea, M.D., Field, C.B., Stocker, T.F., Edenhofer, O., Ebi, K.L., Frame, D.J., Held, H., Kriegler, E., Mach, K.J., Matschoss, P.R. (2010) *Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertaintie*s.

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

McBride, G.B. (2019) Has Water Quality Improved or Been Maintained? A Quantitative Assessment Procedure. *Journal of Environmental Quality.*

McBride, G., Cole, R.G., Westbrooke, I., Jowett, I. (2014) Assessing Environmentally Significant Effects: A Better Strength-of-Evidence than a Single P Value? *Environmental Monitoring and Assessment*, 186: 2729–2740.

McMellor, S., Underwood, G.J.C. (2014) Water Policy Effectiveness: 30 Years of Change in the Hypernutrified Colne Estuary, England. *Marine Pollution Bulletin*, 81: 200–209.

MFE & StatsNZ (2017) *Our Fresh Water 2017*. Ministry for the Environment & Statistics NZ, Wellington, New Zealand.

MFE & StatsNZ (2019) Environment Aotearoa 2019. *Environmental Reporting Series, Ministry for Environment and Statistics New Zealand*, Wellington, New Zealand.

MFE & StatsNZ (2020) Our Freshwater 2020. *New Zealand's Environmental Reporting Series, Ministry for the Environment & Statistics NZ*, Wellington, New Zealand.

Murphy, J.C. (2020) Changing Suspended Sediment in United States Rivers and Streams: Linking Sediment Trends to Changes in Land Use/Cover, Hydrology and Climate. *Hydrology & Earth System Sciences*, 24.

Myers, D.N., Ludtke, A.S. (2017) Progress and Lessons Learned from Water-Quality Monitoring Networks. *Chemistry and Water,* Elsevier: 23–120.

NEMS. (2019) *National Environmental Monitoring Standards for Water Quality, Part 2 of 4: Sampling, Measuring, Processing and Archiving of Discrete River Water Quality Data.* Version 1.0.0, March 2019. Ministry for the Environment, New Zealand.

New Zealand Government. (2020) *National Policy Statement for Freshwater Management 2020*.

Oelsner, G.P., Sprague, L.A., Murphy, J.C., Zuellig, R.E., Johnson, H.M., Ryberg, K.R., Falcone, J.A., Stets, E.G., Vecchia, A.V., Riskin, M.L. (2017) Water-Quality Trends in the Nation's Rivers and Streams, 1972–2012—Data Preparation, Statistical Methods, and Trend Results. *US Geological Survey.*

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

R Core Team (2019) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rangeti, I., Dzwairo, B., Barratt, G.J., Otieno, F.A. (2015) *Validity and Errors in Water Quality Data-a Review. Research and Practices in Water Quality.* Durban University of Technology, Durban, South Africa: 95–112.

Rode, M., Suhr, U. (2007) *Uncertainties in Selected River Water Quality Data.*

Ryberg, K.R., Blomquist, J.D., Sprague, L.A., Sekellick, A.J., Keisman, J. (2018) Modeling Drivers of Phosphorus Loads in Chesapeake Bay Tributaries and Inferences about Long-Term Change. *Science of the Total Environment*, 616: 1423–1430.

Sa'adi, Z., Shahid, S., Ismail, T., Chung, E.S., Wang, X.J. (2019) Trends Analysis of Rainfall and Rainfall Extremes in Sarawak, Malaysia Using Modified Mann–Kendall Test. *Meteorology and Atmospheric Physics*, 131: 263–277.

Salinger, M.J., Mullan, A.B. (1999) New Zealand Climate: Temperature and Precipitation Variations and Their Links with Atmospheric Circulation. *International Journal of Climatology: A Journal of the Royal Meteorological Society,* 19: 1049–1071.

Scarsbrook, M.R. (2006) State and Trends in the National Water Quality Network (1989–2005). *NIWA Client Report*, NIWA, Hamilton, New Zealand.

Scarsbrook, M.R., McBride, C.G., McBride, G.B., Bryers, G.G. (2003) Effects of Climate Variability on Rivers: Consequences for Long Term Water Quality Analysis1. *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., McBride, G.B., Bryers, G.G., Wisse, J., Mink, D.F. (1996) Trends in New Zealand's National River Water Quality Network. *New Zealand Journal of Marine and Freshwater Research*, 30: 485–500.

Snelder, T., Fraser, C. (2018) Aggregating Trend Data for Environmental Reporting. *WP Client Report 2018-01*, LWP Ltd, Christchurch, New Zealand.

Snelder, T., Larned, S., Fraser, C., DeMalmanche, S. (submitted) *Influence of Climate Variation on Physicochemical Trends in New Zealand Rivers.*

Sprague, L., Lorenz, D. (2009) Regional Nutrient Trends in Streams and Rivers of the United States, 1993- 2003. *Environ. Sci. Technol*, 43: 3430–3435.

Sprague, L.A., Oelsner, G.P., Argue, D.M. (2017) Challenges with Secondary Use of Multi-Source Water-Quality Data in the United States. *Water Research*, 110: 252–261.

Steidl, R.J. (2006) Model Selection, Hypothesis Testing, and Risks of Condemning Analytical Tools. *Journal of Wildlife Management*, 70: 1497–1498.

Stephens, P.A., Buskirk, S.W., del Rio, C.M. (2007) Inference in Ecology and Evolution. *Trends in Ecology & Evolution*, 22: 192–197.

Tomperi, J., Juuso, E., Leiviskä, K. (2016) Early Warning of Changing Drinking Water Quality by Trend Analysis. *Journal of Water and Health*, 14: 433–442.

Wickham, H. (2014) Tidy Data. *Journal of Statistical Software*, 59: 1–23.

Woodward, S.J., Stenger, R. (2020) Extension of Bayesian Chemistry-Assisted Hydrograph Separation to Reveal Water Quality Trends (BACH2). *Stochastic Environmental Research and Risk Assessment,* 34: 2053-2069.

Appendix A Supplementary data preparation guidance

Note: This guidance largely relates to the preparation of data from multiple sources (e.g., for inter-regional or national reporting). However, many of the steps are useful for assessing long-term records of data from the same source, particularly where there has been a change in sampling, analytical or data management methodology over time. The National Environmental Monitoring Standard for Water Quality (NEMS 2019)[14] *also provides guidance that can assist with implementing some of the steps described here.*

A1 Ensuring consistent data structure

A1.1 Purpose

Aggregating data and metadata from multiple sources inevitably results in inconsistencies in data structure, including dissimilar data matrices (i.e., arrangements of data in rows and columns) and multiple forms of variable names, geographic coordinates, date and time formats, units of measurement, and other metadata elements. Rearranging data into an internally consistent structure and applying consistent labels, formats and units are the first steps in data processing. Consistency in data structure is needed for sorting, searching, manipulating and displaying data, updating and exporting datasets, and carrying out statistical analyses. General principles for data organisation and recommendations for consistent metadata elements are set out in several recent publications (e.g., Hart et al. 2016; Sprague et al. 2017; Wickham 2014).

A1.2 Method

A. Organise data into a consistent dataset

An example of a tidy and consistent dataset is shown in Figure A-1, where each row represents a single observation and columns store information about the site, the "raw" observed data[15] and the tidied final data after applying rules to ensure consistency of metadata (e.g., variable names, units). Storing data in a "long" data table is preferable over a "wide" data table (i.e., each row represents a single site x date combination with separate columns holding the values for each variable) as it allows additional metadata (e.g., collection and analytical methods, units, censorship status) to be stored with each observation. Using such a format can assist with identifying and correcting data consistency issues, such as inconsistent units between observations of the same variable. In addition, storing the "raw" data alongside the "tidied" data can help identify errors or inconsistencies that may have been inadvertently added during the tidying process. Data should always be stored as numeric values or text and never as colours or other formatting within a spreadsheet as these are easily missed, particularly if the data are to be analysed using a scripting language. Data can be organised into a consistent data table using automated "data tidying" procedures such as the R *tidyverse* package, or manually by copying, pasting and transposing data from different sources into a single data table (Broman and Woo 2018; Ellis and Leek 2018).

[14] There are four parts to the Standard, addressing discrete water quality sampling and measurement (including field and laboratory procedures and instruments) for each of groundwater, rivers, lakes and coastal waters. See: http://www.nems.org.nz

[15] "Raw" as used here applies to the data as received from the data source/supplier. We recognise that these data may already have been modified in some way as part of internal QA/QC procedures.

Site information | Raw observation data | Tidied observation data

	site_id	site_name	x	y	epsg	date	time	raw_variable	raw_value	raw_units	method	datasource	variable	censor	data_flag	multiplier	value	units
1	ARC-00001	Cascades Stream @ Confluence	174.52206	-36.88883	WGS84	1986-07-22	13:15	Ammoniacal nitrogen	0.007	g/m3	filtered	AC Hilltop server	NH4N	=	ok	1	0.007	g/m3
2	ARC-00001	Cascades Stream @ Confluence	174.52206	-36.88883	WGS84	1986-08-19	14:00	Ammoniacal nitrogen	<0.001	g/m3	filtered	AC Hilltop server	NH4N	<	ok	1	0.003	g/m3
3	ARC-00001	Cascades Stream @ Confluence	174.52206	-36.88883	WGS84	1986-09-23	09:02	NH4-N	5	mg/m3	filtered	AC Hilltop server	NH4N	=	ok	0.001	0.005	g/m3
4	ARC-00001	Cascades Stream @ Confluence	174.52206	-36.88883	WGS84	1986-09-23	09:02	NH4-N	5	mg/m3	filtered	AC Hilltop server	NH4N	=	duplicate	0.001	0.005	g/m3
5	ARC-00001	Cascades Stream @ Confluence	174.52206	-36.88883	WGS84	NA	NA	NH4N	7	mg/m3	filtered	AC Hilltop server	NH4N	=	ignore	0.001	0.007	g/m3
6	ARC-00001	Cascades Stream @ Confluence	174.52206	-36.88883	WGS84	1986-12-16	12:47	NH4N	520	mg/m3	filtered	AC Hilltop server	NH4N	=	outlier	0.001	0.52	g/m3

Figure A-1: Example of a "tidy" water quality dataset. Each row represents a single data observation with columns recording the minimal required information for the site, the "raw" observation data as provided by the original source and the tidied observation data after converting to consistent variables names and units. The "censor" column denotes whether the original data were censored (< or >), while the "data_flag" column indicates whether records are duplicated (e.g., records 3 & 4 are identical), have been identified as outliers (e.g., record 6) or should be ignored (e.g., record 5 is missing date information). The "multiplier" column indicates the value used to adjust the original observed value to standardised units (e.g., 0.001 x 7 mg/m^3 = 0.007 g/m^3).

B. Apply consistent metadata elements

Using a consistent format for metadata within a dataset is important for ensuring that data are collated and analysed appropriately, particularly if the data are to be given to someone else for analysis. Development of a "data dictionary" to accompany the tidy dataset that documents the definition of metadata (e.g., quality code definitions) is also recommended. While it is most important that the dataset is internally consistent, we provide some guidelines for suggested metadata formats below.

1. Missing data

   Missing values should be represented by a consistent fixed code (e.g., NA or a hyphen) to help distinguish between those values that are truly missing (e.g., not recorded) and those that are unintentionally missing (Broman and Woo 2018). Numeric values (e.g., 999 or -999) should not be used to denote missing values as they are easy to miss and may inadvertently be included in subsequent analyses as real data. Notes about why data are missing should always be included in a separate comments column and not inserted in place of the data.

2. Site name and IDs

   Ensure that site names are informative and spelt consistently across the dataset. This is particularly important if you are combining data from multiple datasets. Specific issues to watch for include:

   - inconsistent capitalisation;
   - the use of *at* vs *@*;
   - abbreviations of words such as *Bridge* to *Br.* or *Road* to *Rd*; and
   - site IDs (e.g., ARC-00001) which should have a consistent format – check for inconsistencies in underscores, hyphens, spaces and capitalisation.

3. Site coordinates

   Ideally all coordinates representing site locations are recorded in the same coordinate system. Coordinate systems commonly used in New Zealand include latitude and longitude in decimal degrees using the World Geodetic System (WGS84, EPSG: 4326), or northing and easting in NZ Transverse Mercator (NZTM, EPSG: 2193) or NZ Map Grid (NZMG; EPSG: 27200). Coordinates can be converted from one coordinate system to another using tools such as R, ArcGIS or online converters (e.g., https://epsg.io/). It is good practise to include a column that specifies which coordinate system is used, particularly if the original datasets include coordinates from different sources.

4. Date and time format

   Dates and times should be saved in a consistent format, preferably using the ISO 8601 format (e.g., "yyyy-mm-dd" and "hh:mm"). Consistent with National Environmental Monitoring Standards (NEMS) requirements, times should be recorded in 24-hour time (e.g., 18:00 rather than 6:00 pm). Use caution when working with files in Microsoft Excel as date formats may be altered automatically when you open the file and it can be difficult to undo the changes (Broman and Woo 2018).

5. Variable names

   Use a consistent naming convention to describe each measured variable (e.g., NH4N, NO3N, TN – see the NEMS *Water Quality* for recommended naming conventions). Each observation should also be accompanied with a record of the measurement and analytical methods where possible. Units should always be stored in a separate column from the variable name (see below).

6. Measurement units

   Convert multiple measurement units for individual variables to a single, consistent unit for all observations. In some cases, two units are synonyms (e.g., mg/L (or g/m^3) and PPM, µg/L and mg/m^3), and the only requirement is to apply one label consistently. In other cases, shifting to consistent units requires data conversions (e.g., converting from mg/m^3 to mg/L or, in the case of conductivity, from perhaps mS/m to mS/cm or µS/cm). It is good practise to add a *conversion* column to your data that records the multiplier used to convert from one unit to another (e.g., 0.001 to convert from mg/m^3 to g/m^3). This helps to trace any potential errors that may inadvertently be introduced in the data tidying process.

7. Censored data and quality codes

   Information about censoring of numeric data and codes that describe data quality should be stored in separate columns from the measurement values. Use consistent formatting for censorship flags (e.g., < or >) and ensure that the meaning of quality codes is clear.

A1.3 Commentary

Inconsistent or ambiguous variable and site names, data formats, measurement units, and other metadata elements are very common in aggregated water quality and ecology datasets. For the most commonly measured variables in New Zealand freshwater monitoring, multiple names are in widespread use (e.g., dissolved reactive phosphorus, soluble reactive phosphorus, filterable reactive phosphate, orthophosphate and PO4 are all used to refer to the same measurement). Common examples of ambiguous nomenclature include:

1. The use of 'nitrate', 'ammonium', and 'phosphate' without indicating whether the corresponding values refer to ion concentrations (NO_3^-, NH_4^+, PO_4^{3-}) or elemental concentrations (NO_3^--N, NH_4^+-N, PO_4^{3-}-P).

2. The use of "total nitrogen" to refer variously to Kjeldahl nitrogen (organic nitrogen plus ammoniacal nitrogen), total dissolved nitrogen in filtered samples, and total nitrogen in unfiltered samples.

3. The use of "EPT" (invertebrates from the orders Ephemeroptera, Plecoptera, and Trichoptera) without indicating whether the values correspond to number of EPT taxa, percent of EPT taxa or percent of EPT individuals.

When ambiguities cannot be resolved by communicating with the data provider, the data analyst is faced with a choice between guessing the specific meaning of the variable name or omitting the data. Both choices are problematical; the first raises a risk of introducing errors into the data, and the second results in a reduced dataset size, and in some cases, the loss of site × variable combinations.

Some common water quality variables are unitless (e.g., pH, trophic level index, MCI score), but most are reported using measurement units (e.g., concentration in mg/m^3 or mg/L). Missing units are a common problem in water quality datasets. As with ambiguous variable names, missing units must be added if possible or the affected data omitted.

A2 Correcting data errors[16]

A2.1 Purpose

Data errors are common in water quality and ecology datasets and can originate at many steps in the sampling, measuring and recording process. Among the most common causes of data errors are faulty or poorly calibrated field and laboratory sensors, sample contamination, calculation errors and data-entry errors (Davies-Colley et al. 2012, 2019; Rangeti et al. 2015; Rode and Suhr 2007). The resulting data errors include extreme values (for a given variable), negative values, zeros, non-numeric or alphanumeric entries, and strings of repeated values. The NEMS (2019) for Water Quality includes quality assurance (QA) and quality control (QC) guidance and a process for assigning a quality code to individual water quality measurements. With increasing uptake of the NEMS and the passage of time, this should see a reduction in erroneous data in water quality databases and assignment of a 'flag' against data that are potentially erroneous or of lower quality. However, errors will still be present in time-series used for trend assessments (as these time-series often extend back over a decade) and some errors will still occur in the future. Therefore, data analysts must still assess all data for errors and, where possible, correct these errors prior to data analysis. Data error correction occurs in three steps: screening data for potential errors, distinguishing between errors and valid data entries, and correcting the data deemed to actually be erroneous.

There are multiple approaches for screening data for potential errors, including algorithm-based data flagging systems that flag anomalous values, graphical approaches for identifying anomalous values, and statistical outlier tests. Note that these approaches will identify obvious errors, but incorrect entries that do not appear unusual (e.g., fall within the expected range) will not be detected.

After screening data, some assessment is required to determine which of the potential errors are in fact errors (or highly likely to be errors), and which are valid data entries. This step may involve rechecking laboratory and field data sheets, assessing scientific inconsistencies (e.g., DIN concentrations higher than TN concentrations in individual samples, elevated NO3N concentrations in samples from hypoxic or anoxic environments), and investigating possible explanations for anomalous but valid data (e.g., step changes in time series due to land use change). Consideration of NEMS (or other) data quality codes at this step may be helpful if they are present.

Once confirmed, data errors are corrected by reformatting, conversion (e.g., from ion concentrations to elemental concentrations), as well as removing or replacing erroneous data entries. Some water quality data processing routines automate the entire data checking and correction process. However, we recommend manual correction to minimise the risk of removing or altering valid data entries.

A2.2 Method

Where possible, we recommend using an initial automated data processing script to flag potential anomalous data entries, followed by a manual checking process to assess whether flagged records do in fact represent errors. Potential and actual errors should be identified within the tidy dataset in a column that contains a pre-determined set of flags. These allow erroneous records to be removed

[16] The guidance provided here generally assumes that the data lack any formal 'quality stamp', such as a NEMS (2019) quality code.

prior to subsequent analysis, while retaining a record of why they were removed. We suggest that the following flags be used, with a data dictionary created to record their specific definition:

- ok: records that pass the error checking procedure.
- ignore: records that should be removed before the final analysis. Reasons for use may include non-numeric or impossible data, missing metadata, duplicate records or incompatible measurement or analytical methods.
- outlier: numeric measurement values that fall outside the expected range after checking for issues such as incorrect units.
- composite: use if multiple measurements on the same date have been combined (e.g., daily mean) or if data are calculated from multiple measured variables (e.g., MCI or TLI).
- synthetic: use if a measurement value is missing but has been estimated or inferred from a known existing relationship with another variable (e.g., visual clarity based on an established relationship with turbidity or suspended sediment).

Suggested steps for error checking and processing errors in measurement data (also see the NEMS Water Quality)

1. Identify and inspect non-numeric measurement data. These data often occur when censor flags or other metadata (e.g., comments) are included in the same column as measurement data. All metadata should be removed and included in a separate column as a censor flag or comment. If no numeric measurement data are available, flag as *ignore*.
2. Identify impossible (e.g., negative values) and highly improbable data entries (e.g., DIN and DRP concentrations > TN and TP concentrations after taking into account the associated uncertainty of measurement). Where possible, check field and laboratory data sheets to determine whether the values were measurement errors or transcription errors. Where possible, correct errors or flag as *ignore* (if impossible) or *outlier* (if beyond the expected range). Ideally, data corrections should be passed back to the data originators to allow the original data source to be corrected as appropriate.
3. Use graphical methods to help interpret data anomalies. For example, a strongly bimodal distribution of values may indicate that two different measurement units have been used (e.g., nutrient concentrations in mg/m^3 and mg/L, dissolved oxygen concentrations in percent saturation and mg/L). Time-series plots can be used to identify potential step-changes where measurement units and/or detection limits have changed, while quantile plots, histograms and box plots can help identify outliers.
4. Use data summaries (e.g., pivot tables) to identify extreme values and potential detection limits. For example, check for repeated values (>5 % of values within a given variable) that may represent a detection limit and apply the appropriate censor flag if one has not already been assigned (e.g., < or >).

Suggested steps for error checking and processing errors in metadata

1. Flag observations with missing or inconsistent metadata as *ignore* if the metadata are critical to the analysis (e.g., date, spatial coordinates).

2. Flag observations with incompatible measurement or analytical methods as *ignore*.

3. Check for data records that may have been duplicated (e.g., multiple records on the same date). This is particularly important if combining multiple datasets. Where the data are identical, flag duplicates as *duplicate*. Where multiple records for a given site and variable are present on the same date, it may be appropriate to calculate summary statistics (e.g., daily means). If this approach is used, then add a *composite* flag and make sure that the methodology is noted in the comments.

4. Map sites using the supplied coordinates to ensure that site locations match the descriptions. Initial checks include identifying sites with missing coordinates or those outside the geographic bounds of the study region/area (e.g., due to zeros, flipped coordinates or incorrect coordinate systems). If coordinates need to be transformed to a different coordinate system, it is good practise to compare maps of sites using both systems to ensure that errors have not been introduced during the transformation process. Finally, site locations should be checked against site names where possible to ensure that they are mapped in the correctplace.

A2.3 Commentary

Data screening and error correction are needed to ensure that trend assessment results are accurate. However, some data errors are likely to remain undetected using the methods recommended here, because the measurement values appear reasonable. Conversely, some anomalous but valid values may be incorrectly classed as errors and removed. The aim is to maximise the former while minimising the latter. As noted above, a lenient approach increases the number of data errors retained, which can reduce accuracy, and a highly stringent approach can increase the number of site/variable combinations removed, which can reduce the spatial extent of the analysis. To our knowledge, there is no optimal balance. The non-parametric trend assessment methods in Section 6 should ensure that a moderate number of uncorrected errors will have minor effects on the assessment results. Therefore, highly stringent approaches that result in high numbers of site/variable combinations removed) are not recommended.

Automated error identification and correction systems have been developed for processing water quality data, particularly data generated by high-frequency, *in situ* sensors (Campbell et al. 2013; Horsburgh et al. 2015). In these cases, the quantity of data can make manual error identification and correction prohibitively time consuming. However, automated error correction applies to the limited range of errors (e.g., sensor drift, skipped measurements) that can be corrected using simple algorithms. Furthermore, automated procedures lack the intermediate 'confirming and interpreting errors' step set out above. This is an important step that cannot be automated as it involves subjective decisions and in some cases, communication between data analysts and data providers. Assuming that the data used for trend-analyses are primarily from monthly to annual monitoring, we recommend manual error interpretation and correction.

A3 Ensuring comparable measurement methods

A3.1 Purpose

For most commonly measured water quality and ecological variables, two or more measurement methods have been used in New Zealand monitoring programmes (Larned et al. 2016). Alternative methods that produce divergent measurement values are a source of extraneous variation in

statistical analyses. One of the primary aims of the NEMS initiative is to recommend reliable and accurate methods, and compliance with the standard will increase consistency across monitoring programmes. In turn, greater consistency will reduce the effects of methodological variation in data analyses. Unfortunately, the beneficial effects of consistency in methods will be gradual for water quality trend assessments; the NEMS for Water Quality was published in 2019 and the time-series used for trend assessments often extend back over a decade. For this reason, data analysts must address the issue of multiple methods in time-series. This issue can be manifested in several ways: the time-series for a given variable at a single site may be affected by a change in measurement methods part way through the record, the measurement methods for a given variable may differ among sites, or both.

Data analysts must make subjective decisions when dealing with the issue of multiple methods. Retaining all data regardless of the measurement method raises the risk that variation due to variable methods will confound trend detection. Conversely, retaining the data generated from a single method and omitting the data generated from all alternative methods can result in large reductions in site/variable combinations. As a compromise, we recommend pooling data that correspond to 'comparable' measurement methods. Comparable methods produce equivalent values for a given sample. The NEMS standard provides commentary on comparable and non-comparable methods for some water quality variables (e.g., *E. coli,* TN). For other variables, comparability can be assessed using published reports of the comparative accuracy, precision and bias of two or more methods (e.g., Kilroy and Biggs 2002; Davies-Colley and McBride 2016).[17]

A3.2 Method

For each variable to be analysed, determine the largest group of data corresponding to comparable methods. Retain this group and delete the data corresponding to the non-comparable methods. In some cases, only one acceptable method is in use in New Zealand; this method is generally the most widely used. In these cases, retain the data corresponding to the acceptable method and delete the remaining data.

A3.3 Commentary

The problems posed by variation in methods are greatest for trend analyses that incorporate data from many sources, which can include a wide range of measurement methods (e.g., national data compilations from multiple agencies). If non-comparable methods for a given variable are used in each of two or more regions that each encompass numerous sampling sites, then all sites from one or more entire regions will be excluded from the analysis. For example, in previous national-scale analyses, the total nitrogen data were excluded from all sites in several regions that used methods deemed non-comparable (Larned et al. 2018a; Larned and Unwin 2012).

[17] In addition, Davies-Colley et al. (2019) document an approach for assessing the level of agreement between river water quality measurements from two organisations.

Appendix B Comparison of flow adjusted and non-flow adjusted trends

This appendix provides a comparison of flow adjusted versus non-flow adjusted trend assessments. We use "raw" and "flow adjusted" to distinguish analyses performed using the raw (i.e., non-flow adjusted data) and flow adjusted data. The purpose of this comparison is to allow the reader to gauge the impact that flow adjusting might have on the outcomes of trend assessments. To demonstrate these impacts, we used the 10-year trend assessments from Larned et al. (2018a) and extracted only site-variable combinations where Larned et al. (2018a) had chosen to flow adjust the observations (i.e., at sites where there was a plausible relationship between observations and flow and $R^2 \geq 20\%$). This reduced the dataset to from 9,342 to 775 site-variable combinations (in total 8 variables). We present the comparison in terms of the suggested reporting measures outlined in section 6.2:

- Table of trend direction (Table B-1);
- Table of categorical confidence the trend was decreasing (Table B-2);
- Scatter plots of confidence that trend direction is decreasing, C_d, (Figure B-1); and
- Scatter plots of Sen slopes (and uncertainties) (Figure B-2).

For the continuous measures (C_d and Sen slopes), we also provide descriptive statistics (correlation coefficient, root mean square deviation and bias) to describe the relationship between the raw and flow adjusted estimates. For the Sen slopes, which are also provided with 90% confidence intervals, we have also evaluated the percentage of sites where the 90% confidence overlap, and where the flow adjusted Sen slope falls within the 90% confidence interval of the raw Sen slope.

For this dataset we found that:

- Trend direction was consistent between raw and flow adjusted trends for 84% of site/variable combinations. Just 1% of site/variable combinations that had raw confidence levels of "likely" or greater, switched direction to "unlikely" or stronger for the flow adjusted confidence category, while 2% of site/variable combinations that had raw confidence levels of "unlikely" or greater, switched direction to "unlikely" or stronger for the flow adjusted confidence category.
- Categorical levels of confidence that the trend direction was decreasing were consistent between raw and flow adjusted trends for 52% of site/variable combinations. A further 32% of sites only moved up or down by one category.
- There was high correlation between C_d for the raw and flow adjusted trends (0.84 for the entire dataset). The root mean squared deviation (RMSD) represents the mean difference between C_d estimated for the raw and flow adjusted trends (Piñeiro et al. 2008). RMSD varied between variables from 0.13 to 0.25. There was a small, but mostly consistent negative bias (C_d for the flow adjusted trends was on average slightly smaller than for the raw trends) (Table B-3).

- 99% of site/variable combinations had overlapping 90% confidence intervals for their raw and flow adjusted Sen slopes while 89% of site/variable combinations had flow adjusted Sen Slopes that lay within the 90% confidence interval of their raw Sen slopes. This suggests that the differences in Sen slopes are generally small relative to the uncertainty on the Sen slope estimates.

The overall conclusion from this analysis is that flow adjustment can result in appreciable differences in trend assessment results compared to analyses performed on the raw data for individual sites. However, based on this analysis, conclusions drawn from aggregated results of trend analyses performed for many sites are not very sensitive to flow adjustment.

Table B-1: Comparison of trend directions estimated with and without flow adjustment.

		Raw		
		Decreasing	Increasing	Indeterminate
Flow adjusted	Decreasing	377	32	1
	Increasing	78	277	4
	Indeterminate	3	3	0

Table B-2: Comparison of categorical confidence levels that the trend was decreasing (from Table 6-2) estimated with and without flow adjustment.

		Raw						
		Highly unlikely	Very unlikely	Unlikely	As likely as not	Likely	Very likely	Highly likely
Flow adjusted	Highly unlikely	55	15	24	15	1	0	0
	Very unlikely	6	6	24	7	1	0	0
	Unlikely	4	11	41	47	15	2	0
	As likely as not	0	1	15	91	45	5	5
	Likely	0	1	3	28	79	14	19
	Very likely	0	0	0	1	16	14	14
	Highly likely	0	0	3	2	12	13	120

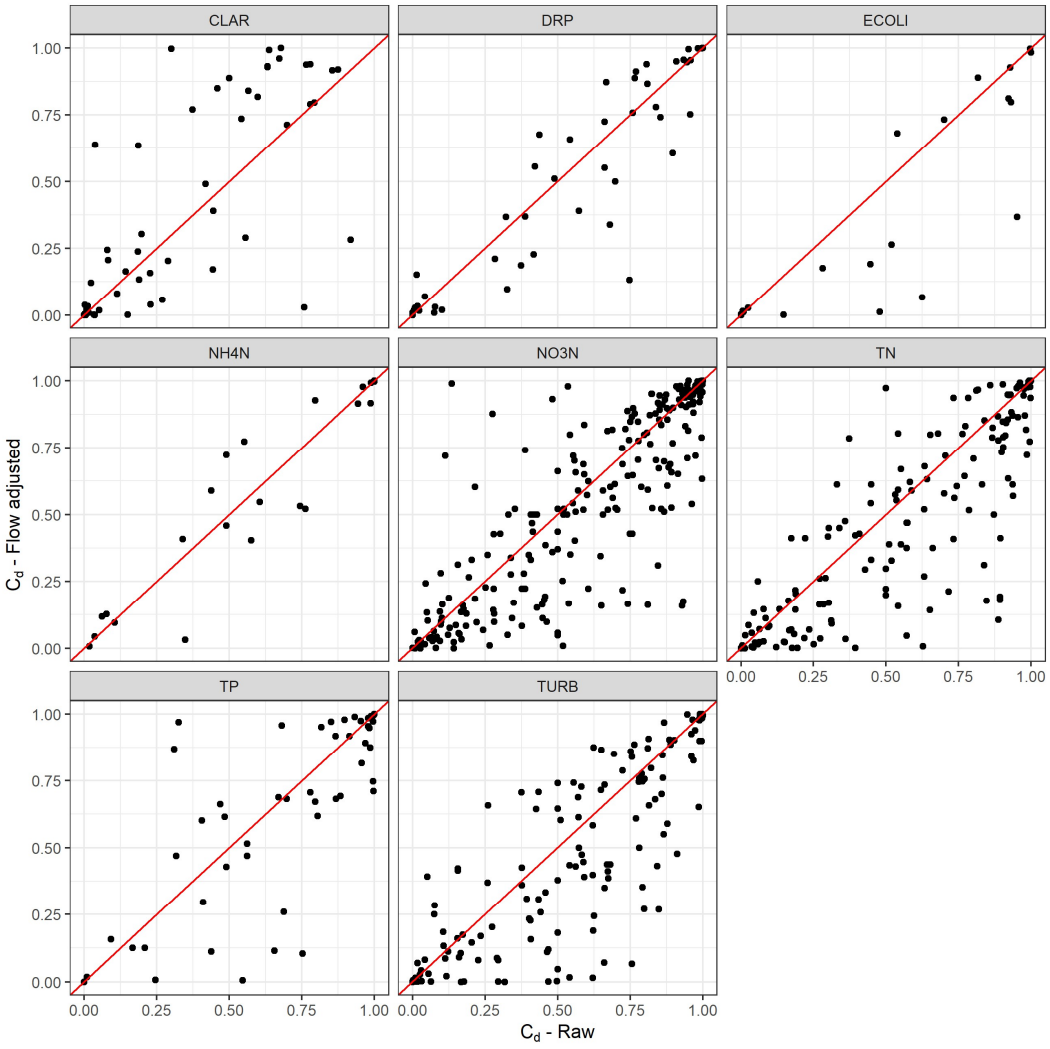

Figure B-1: Scatter plots comparing the confidence that the trend direction was decreasing for raw and flow adjusted trends, by variable.

Table B-3: Comparison of confidence that the trend direction was decreasing (C_d), by variable for raw and flow adjusted trends.

Variable	Correlation coefficient	RMSD	Bias
CLAR	0.77	0.25	0.06
DRP	0.93	0.15	-0.03
ECOLI	0.86	0.23	-0.12
NH4N	0.93	0.13	-0.01
NO3N	0.86	0.19	-0.05
TN	0.85	0.21	-0.08
TP	0.77	0.23	-0.04
TURB	0.82	0.22	-0.08

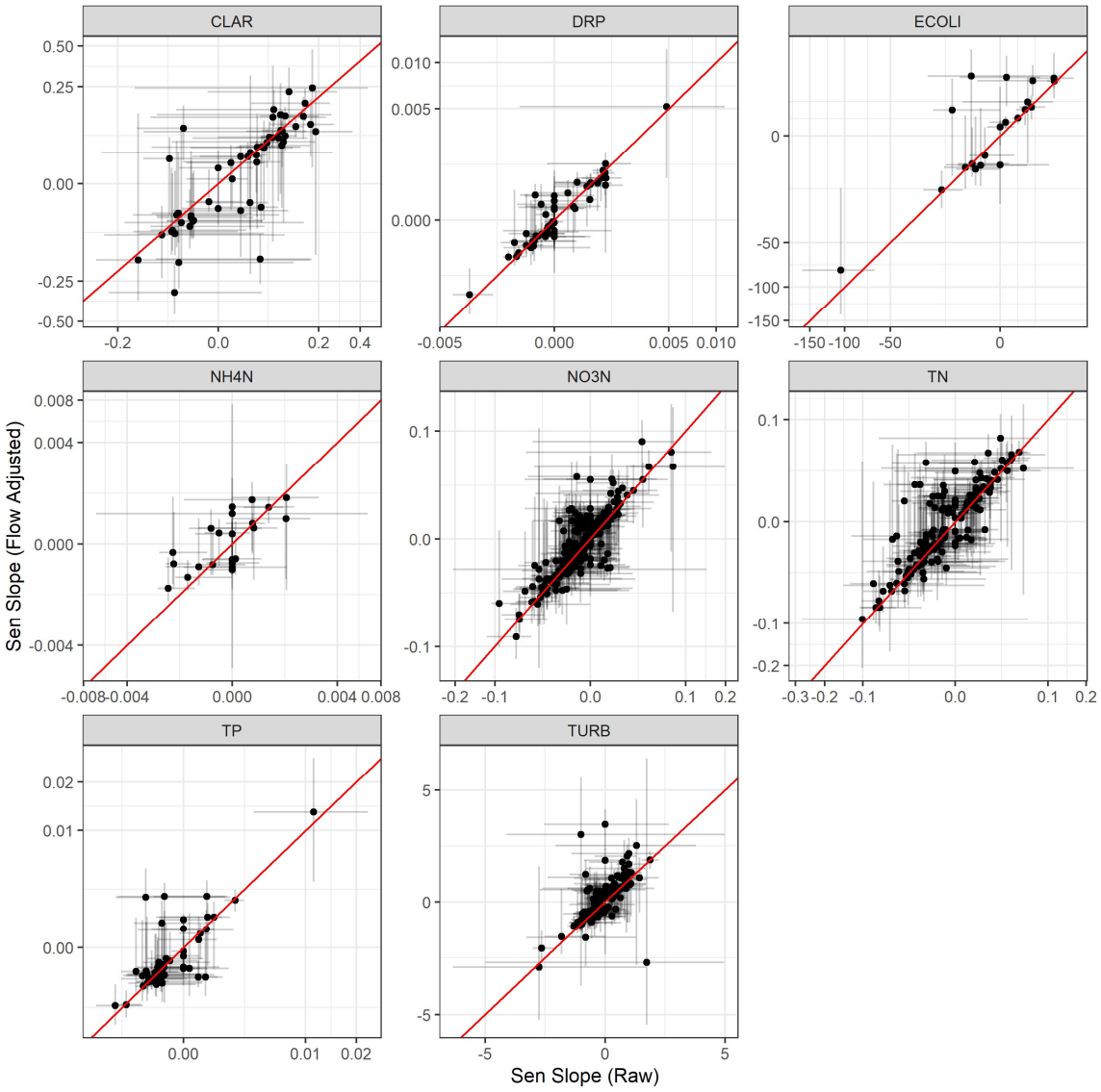

Figure B-2: Scatter plot comparing trend rate (i.e., Sen slopes) for raw and flow adjusted trends, by variable. Grey lines indicate the 90% confidence intervals Note that X and Y scales are non-linear (square root adjusted).

Table 3: Comparison of Sen slopes for raw and flow adjusted trends. CI Overlap: the proportion of pairs of raw trends and flow adjusted trends for the same site for which the 90% CI overlaps. FA SSE in raw CI: the proportion of sites for which the flow adjusted Sen slope falls within the 90% CI of the raw Sen slope.

Variable	Correlation coefficient	RMSD	Bias	CI Overlap	FA SSE in raw CI
CLAR	0.75	6.2E-02	-0.013	100%	91%
DRP	0.98	1.9E-04	0.000028	98%	94%
ECOLI	0.95	8.7E+00	3.9	100%	80%
NH4N	0.69	4.8E-04	0.00016	100%	86%
NO3N	0.85	7.9E-03	0.0018	100%	88%
TN	0.87	1.1E-02	0.0033	99%	92%
TP	0.95	7.0E-04	0.00016	100%	85%
TURB	0.46	3.6E-01	0.069	97%	88%

Appendix C Comparison of seasonal and non-seasonal trend assessments

In this appendix we present a comparison of the effects of using a seasonal versus a non-seasonal trend assessment. We used 10 years of monthly data for 74 National River Quality Network (NRWQN) sites from the study Larned et al. (2018a). The purpose is to demonstrate the implication of the subjective choice of a Kruskall-Wallis test *p*-value of 0.05 as the threshold for assessing seasonality.

For all 74 sites and eight variables, we evaluated the Kruskall-Wallis test *p*-value to quantity the degree of seasonality (with seasons defined as months) for each site/variable combination. We then performed both a seasonal and a non-seasonal trend assessment for each site/variable combination. The confidence that the trend was decreasing (C_d) and the trend rate (i.e., Sen slope, and its uncertainty) for the seasonal and non-seasonal trend assessments are compared.

In general, Figure C-1 and Figure C-2 indicate that seasonal and non-seasonal trend assessments yield similar results. Qualitatively, it appears that the larger differences in the confidence that the trend was decreasing (either positive or negative differences) tend to be more associated with site/variable combinations where the Kruskall-Wallis *p*-value was <0.05 (i.e., they would be identified as "seasonal"). All non-seasonal Sen slope confidence intervals contained the seasonal Sen slope estimate. Therefore, the differences between the estimated Sen slopes were smaller than the uncertainties of these estimates.

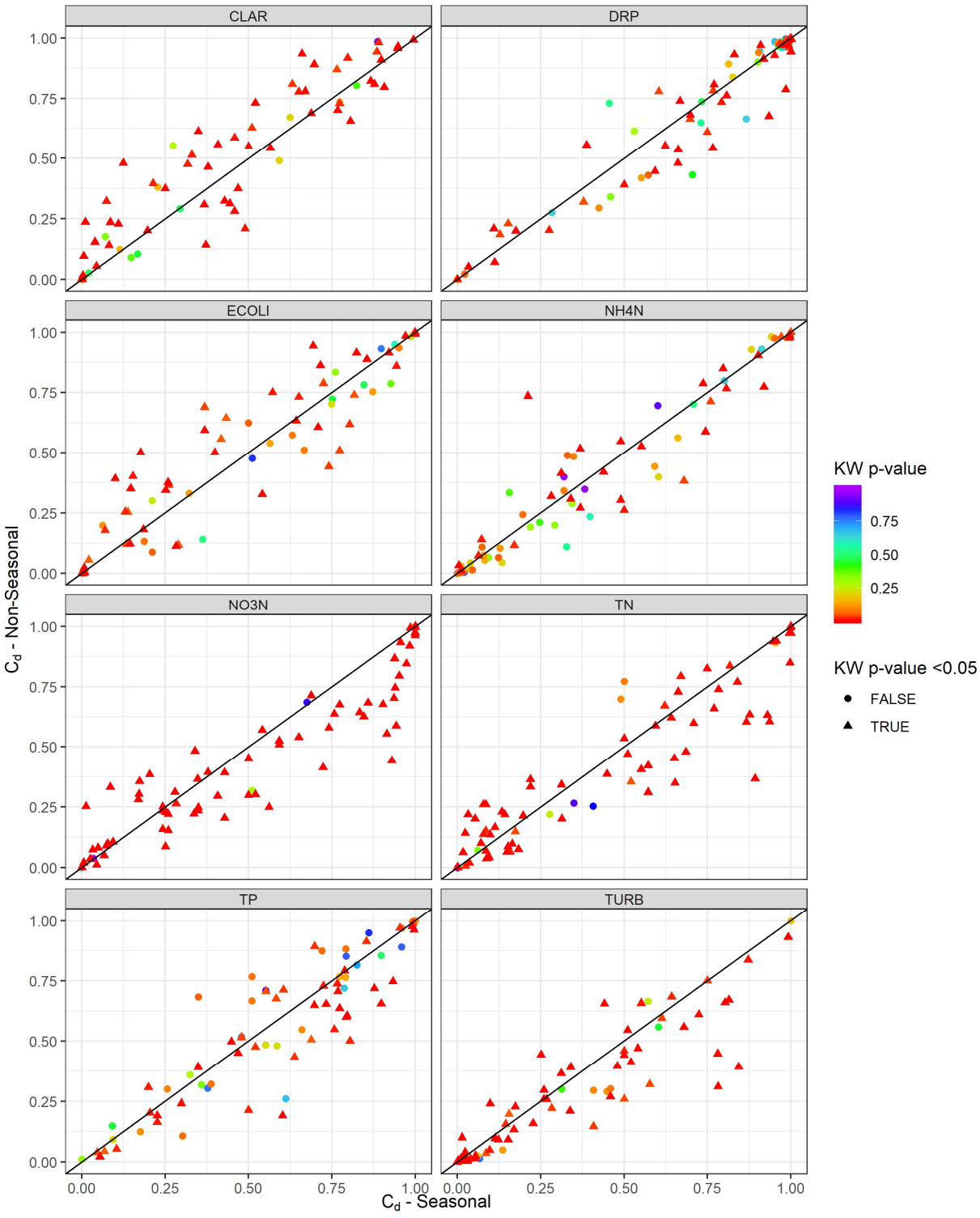

Figure C-1: Comparison of seasonal and non-seasonal estimates of the confidence that the trends were decreasing (C_d).

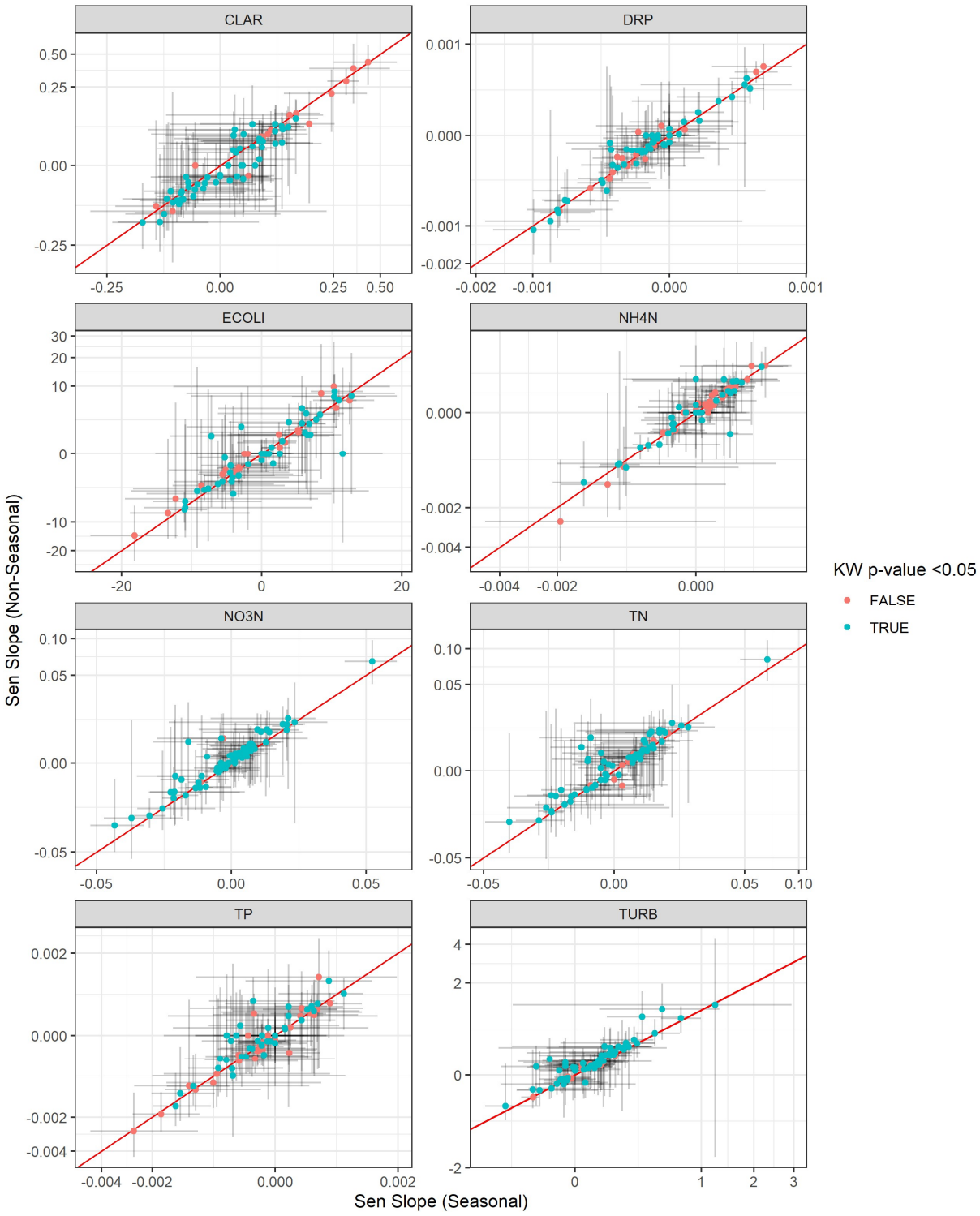

Figure C-2: Comparison of seasonal and non-seasonal estimates of the Sen slope. Grey error bars indicate the 90% confidence intervals for the Sen slope estimates.

Appendix D Climate influence on water quality trends

A recent study has built on the earlier work by Scarsbrook et al. (2003) and quantified the influence of the El Niño Southern Oscillation climate pattern (ENSO) on freshwater water quality trends in New Zealand (Snelder et al., submitted). The study assessed the relationship between the Southern Oscillation Index (SOI), which is an indicator of the ENSO climate pattern, and temporal variability in monthly observations of eight water quality variables, water temperature and flow at 77 National River Water Quality Network (NRWQN) monitoring sites over a 27-year period (1999–2016).

The study comprised three steps. Step one investigated the correlation between the monthly observations of each of the variables at each site and the corresponding monthly values of the SOI. The correlation coefficients for each site and variable combination are referred to as SOIc and ranged between -0.63 and 0.69 with an even split across sites and variables into positive and negative values. This finding is evidence that temporal variation in ENSO strength drives temporal variation the observed variables in New Zealand.

The strength and direction of the correlation between the observed variables and the SOI (i.e., SOIc) were highly variable in space. However, geographic patterns in SOIc for flow (FLOW), turbidity (TURB), electrical conductivity (COND), colour dissolved organic matter (CDOM) and total phosphorus (TP) were consistent with the known effects of ENSO on rainfall and were not explained by categorisation of the NRWQN sites into impacted and baseline (i.e., relatively natural) catchment conditions. The combination of these findings suggests that the responses of these variables to the SOI are linked to the influence of rainfall on their mobilisation and transport rather than variation in land management practices in response to climatic conditions. In contrast, there were no detectable geographic patterns in SOIc for dissolved reactive phosphorus (DRP), ammoniacal nitrogen (NH4N), nitrate and nitrite nitrogen (NNN) and total nitrogen (TN). In addition, between-site differences in SOIc in these variables were not explained by the impacted and baseline categories. These findings indicate that the response of water quality variables to the same climate stimulus is variable across sites even when these sites are in close proximity. This may be because the responses of DRP, NH4N, NNN and TN are more strongly mediated by catchment processes such as mobilisation and biogeochemical transformation within soils and groundwater than the other water quality variables.

Step two of the study investigated variation in the trends assessed for each site and variable for different 'time windows' (i.e., analysis time periods). Rolling windows were defined of 5, 10 and 15-years duration starting in 1990 and incrementing by one year to a final period ending in 2016. This resulted in 23, 18 and 13 time windows of 5, 10 and 15-years duration, respectively. For each site, variable and window, the trend was quantified by Kendall's Tau (τ), which is a standardised version of the Kendall *S* statistic. Values of τ range between -1 and +1; a positive value indicating that the observations increased through time and vice versa. For each variable, time window and duration, the site trend is referred to as τ_w.

For each variable, values of τ_w tended to oscillate between time windows for all three durations (Figure D-1). Within a variable, the magnitude of changes in the τ_w between adjacent time windows decreased with increasing time window duration (Figure D-1). For example, for the 5 and 10-year time window durations, there were frequent changes in the direction of the majority of site trend between time windows that were separated by only one or two years (e.g., from >50% of sites with increasing trends to >50% of sites with decreasing trends). In contrast, changes in direction of the majority of site trends within one or two years were less frequent for the 15-year time window duration.

The findings of step two of the study indicate that results of trend analyses are sensitive to the time window of the analysis and that large changes in trend strength and direction occur between time windows. The oscillation in the direction and strength of trends suggests that cyclic climatic processes are involved.

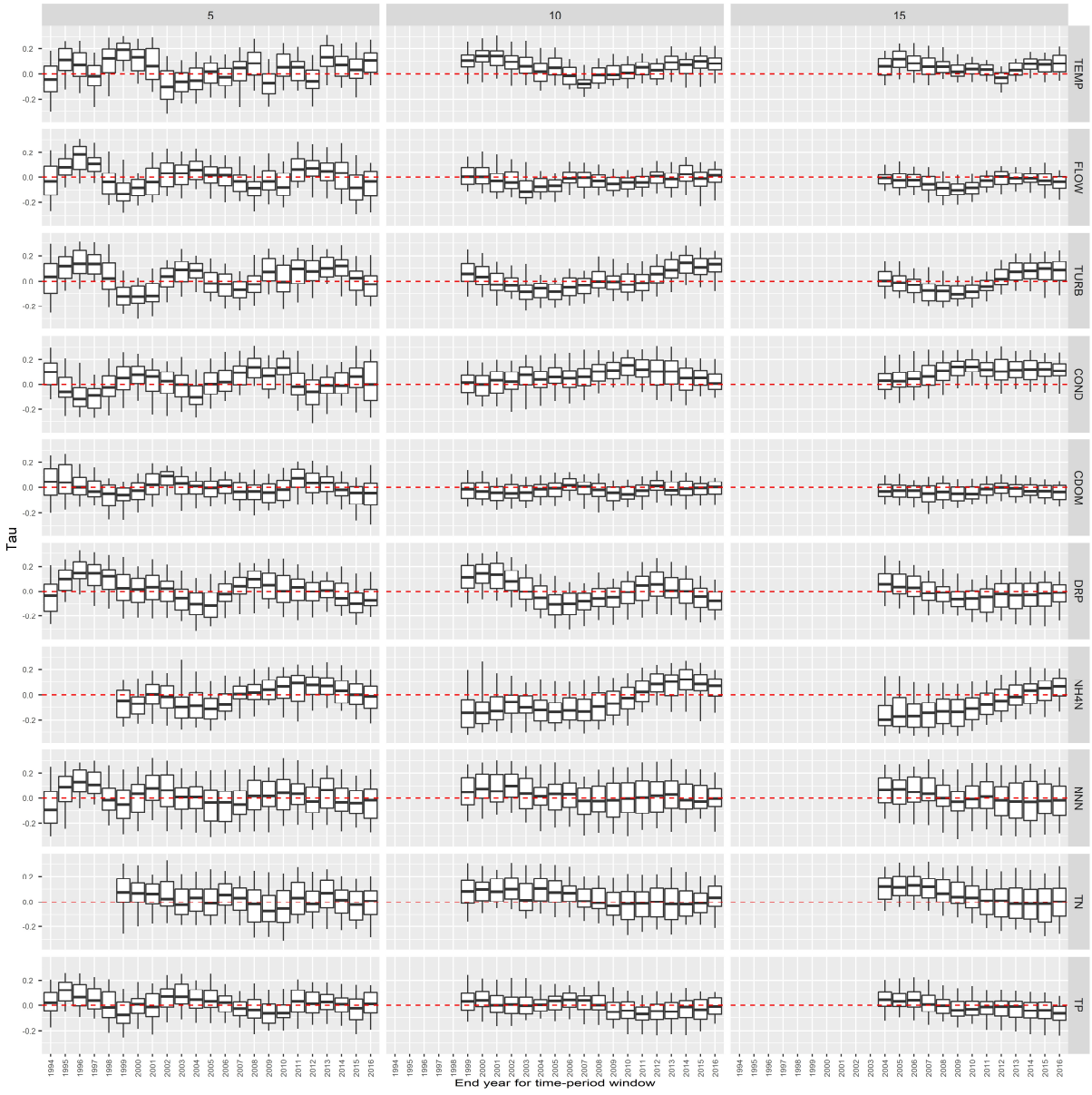

Figure D-1: Distribution of site trends ($\tau_w$statistic) with time window and time window duration. Each panel represents a time window duration (columns) and a water quality variable (rows). Within each panel, the boxes and whiskers represent the distributions of the site trend strength and direction (i.e., the τ_w statistics for the 77 sites) for the time windows ending at the associated end year (horizontal axis). The plot indicates the interquartile range, the central horizontal line in the box represents the median, and the lower and upper ends of the whiskers indicate the 5th and 95th percentile values, respectively. The horizontal red line indicates a τ_w statistic of zero; median values above and below this line indicate the majority of sites (i.e., > 50%) had increasing and decreasing trends, respectively.

Because the ENSO process is quasi-cyclic with irregular phases occurring on average every two to seven years, there are monotonic trends in the SOI for each time window which we refer to here as 'SOI trends' and denoted δSOI_w (Figure D-2). In step three of the study, regression models were used to relate the strength and direction of site trends (τ_w) to the SOI trend.

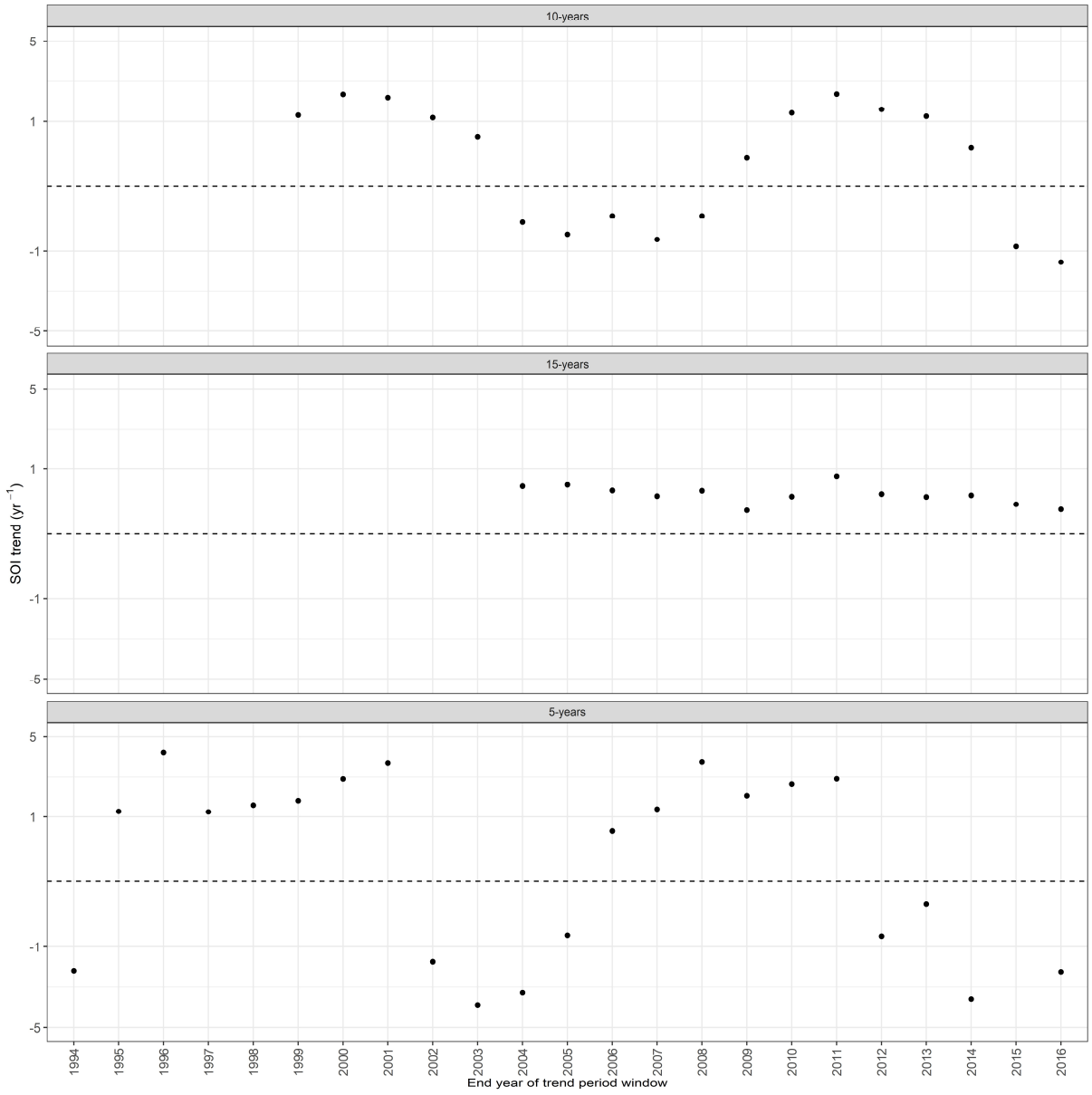

Figure D-2: SOI trends, defined by linear trends in the SOI, for time windows representing time window of three durations of 5, 10 and 15-years. Each point represents δSOI_w for the time window indicated by the end year (x-axis). Note that the scale of the vertical axis is transformed to emphasise values close to zero.

For all variables, variation in site trends between windows were significantly explained by the combination of the SOI trend (δSOI_w) and the correlation between the monthly water quality observations and SOI ($SOIc$). When averaged across all variables, the models explained an average of 11%, 24% and 9% of the variation in site trend strength and direction for trend durations of 5, 10 and 15-years, respectively. The variation in site trends explained by the models were related to the absolute value of $SOIc$. Model R^2 values tended to increase with increasing absolute values of $SOIc$. In addition, the trend direction in any time window was related to the direction of the SOI trend.

When δSOI_w and $SOIc$ were positive, the water quality trend tended to be positive and vice versa and the opposite applied for sites with negative $SOIc$ values.

The findings of step three of the study indicate that the climate signal translates into a predictable variation in water quality trends. During periods when the SOI trend is positive, there is a tendency for increasing water quality trends at sites with positive $SOIc$ values. The opposite applies to sites with negative $SOIc$ values. This pattern is reversed during periods when the SOI trend is negative. Oscillations in aggregate trends (**Error! Reference source not found.**) arise because the ENSO process drives variation in rainfall and temperature over large spatial scales. In turn the variation in rainfall and temperature drives synchronous water quality responses (i.e., changes at the same time across sites). However, the direction and strength of water quality responses varies depending on the specific response to the SOI at each site (i.e., $SOIc$ values).

The SOI is a broad indicator of climate variation that accounts for less than 25% of the year to year variance in seasonal rainfall and temperature at most New Zealand measurement sites (Salinger and Mullan 1999). The SOI is therefore an imprecise representation of the proximate climatic drivers of trends (i.e., rainfall and temperature) at any site and this means it is likely that the study under-estimates the contribution of climate to water quality trends for at least some sites and variables. The characteristics of climate that best explain water quality responses are likely to vary between sites and variables due to the differing and complex mechanisms that mediate those responses.

A conclusion of this study is that effects of climate variation may amplify or counteract the effects of other drivers of water quality trends, even when those trends are assessed over time windows that are longer than climate cycles. In turn, this means that a risk of reporting water quality trends without robust attempts to identify the causes is that it may lead to speculative attribution of the trends to anthropogenic drivers. This may then lead to management actions to mitigate anthropogenic drivers that are ineffective in reversing degrading trends.