NIWA
Taihoro Nukurangi

Assessing and accounting for the influence of changes in laboratory measurement methods on the interpretation of long-term time-series data

Prepared for HBRC and MBIE Envirolink

December 2024

Prepared by:
David Wood

For any information regarding this report please contact:

David Wood
Water Quality Scientist
Freshwater Modelling
+64 3 343 8050
david.wood@niwa.co.nz

National Institute of Water & Atmospheric Research Ltd
PO Box 8602
Riccarton
Christchurch 8440

Phone +64 3 348 8987

NIWA CLIENT REPORT No: 2024055CH
Report date: December 2024
NIWA Project: ELF23502

Revision	Description	Date
Version 1.0	Final version sent to client	27 March 2024
Version 1.1	Final version after response to client review and feedback	20 December 2024

Quality Assurance Statement		
	Reviewed by:	Neale Hudson
	Formatting checked by:	Rachel Wright
Phillip Jellyman	Approved for release by:	Phil Jellyman

Whilst NIWA has used all reasonable endeavours to ensure that the information contained in this document is accurate, NIWA does not give any express or implied warranty as to the completeness of the information contained herein, or that it will be suitable for any purpose(s) other than those specifically contemplated during the Project or agreed by NIWA and the Client.

Contents

Tables

Figures

Executive summary

Long-term monitoring programmes, such as those used in State of the Environment (SoE) reporting, occasionally require changes in laboratory measurement methods. An unintended consequence of method changes can be a "step" change in time series data, which can complicate the interpretation of state and trends. The question is how to deal with step changes, as there is no well-established practice for water quality data. Hawke's Bay Regional Council, the lead local government agency for this work, commissioned NIWA through the Envirolink scheme to develop guidance for the regional sector to manage changes in laboratory methods. The scope of this guidance was to evaluate and, where possible, account for the impact of changes in laboratory measurement methods on water quality parameters such as nutrients, turbidity and microbiological variables on river, lake and coastal water quality time-series data.

The broad approach recommended in this guidance involves taking parallel measurements using the old and new laboratory measurement methods. These measurements are then used to estimate the systematic difference between the results of the two methods, known as bias. Then, these estimates of bias can be used to make corrections. The approach is not well established and is best considered an "emerging practice." It presents a step forward in managing changes in laboratory methods, but it has limitations. The overall approach to evaluate and account for changes in laboratory methods is

1. Plan the collection of pairwise data using the old and new laboratory measurement methods in parallel; a minimum of 12 water quality measurements using both methods should be taken if sampling is done monthly.
2. Take parallel measurements using old and new laboratory methods and record results.
3. Evaluate if the measurements are representative. The measurements need to be representative if we are to apply the result of the subsequent level of agreement analysis to account for changes in laboratory methods.
4. Assess the level of agreement (LoA) between the old and new methods and record the results. Assessing LoA involves estimating the systematic difference (bias) and the random differences between the two laboratory measurement methods. Ideally, the results of the LoA should be compared with results from other LoA assessments.
5. Using information from step 4, adjust data collected using the old laboratory method before assessing state and/or trend.
6. Assesses state and/or trend according to usual protocols.
7. An optional step is to carry out a sensitivity analysis to investigate how sensitive or not the resulting state and trend are to bias.

When following this guidance, each step should be documented to ensure that the decisions made are transparent. A brief description of the broad steps and limitations is given, and the document's body provides more details.

Data collection – parallel measurements using old and new methods

Steps one and two draw on the National Environmental Monitoring Standards (NEMS) for discrete groundwater, river, lake and coastal water quality. The NEMS recognise the issue of bias associated with method changes and recommend "…measurements using both the old and new methods are carried out for a period of time (at least 12 months where sampling occurs monthly) to provide sufficient data to enable a conversion factor to be derived to 'align' the old and the new data." In this guidance, measurements using both old and new methods simultaneously are referred to as parallel measurements. The NEMS provides advice on how to gather data to estimate a conversion factor but does not provide any practical advice on how to estimate the conversion factor. This guidance, steps three to five, provides practical advice as to how to do that.

Several factors can influence bias, not just the methods. These factors include how the laboratory applies the measurement methods, the level of the water quality parameters in the water, interference effects (ANZG 2018), and factors associated with water sample collection. As a result, bias may differ between monitoring sites.

Analysing paired data, estimating bias and making corrections

Steps three to five provide practical advice about "aligning" the data from the old and new methods.

The level of bias is estimated in step four and based on well-established approaches to assess the level of agreement (LoA) between the two measurement methods, such as Bland and Altman analysis and Deming regression . LoA involves estimating the systematic and random differences between two measurement methods. In step five, we can correct for the systematic error, also known as the bias. However, we cannot correct the random differences between methods.

Step five involves using the estimate of bias to align/adjust the old measurement with the new measurement methods. In this process, we also need to consider differences in precision between measurement methods, generally by rounding to align high-precision values with lower-precision values.

The results of LoA analysis can be used to adjust measurements when parallel measurements have not been taken. So, we assume the results of the parallel measurements are transferable to times when parallel measurements were not made, for example, when you are just using the new method. There is no way of knowing if the results are transferable, but ensuring the results are representative in step three provides us with some confidence that the results are.

Using the adjusted data as the inputs to trend and/or state analysis

Step six is about using the adjusted data as the inputs to trend and/or state analysis. It is recommended, as routine, that data are adjusted for bias before assessment of trend and/or state. The data can then be analysed using protocols such as *Guidance for the analysis of temporal trends in environmental data*.

Routine adjustment of data overcomes the issue of deciding if bias may or may not make a difference to a trend and/or state analysis. Sometimes, a small bias can change the assessment; at other times, a large bias can make no difference. So, we do not recommend ignoring bias if it is less than some arbitrary cut off value.

Step seven is optional and involves assessing how sensitive the trend and/or state analysis is to bias.

Limitations

The above recommendations may not be applicable in all situations. Practical limitations may prevent the collection of parallel measurements using old and new methods. Even if it is possible to take parallel measurements, estimating bias from the resulting paired data may not be possible.

Focusing on situations where paired data from parallel measurements are unavailable, the only option would be to use assessments of bias from other locations. This approach may not account for all the factors associated with measurement method change bias, including how laboratories apply the methods, the level of the analytes in the water, interference effects, and factors associated with water sample collection.

The recommended approaches for estimating the level of agreement (LoA) between the two measurement methods assumes that the measurements are on a continuous scale and can take any value within the range of interest. The approaches described are not designed to be used with censored or discrete data.

Censored data is when the measured value is between zero and the laboratory reporting limit or above a maximum reporting value. If the measured values are highly censored, it may not be possible to estimate bias, and there may be little value in taking parallel measurements.

Though the approaches are not designed to be used with discrete data such as microbiological data, and estimating the level of agreement between discrete measurements is a research topic, the pragmatic thing to do is assume it can be handled just like continuous measurements.

1 Introduction

Regional Councils and other agencies monitor water quality for several reasons, including determining the current state and estimating trends. The data used to make these determinations are derived from a chain of activities, including physical grab sampling, storage of samples prior to laboratory analysis, and analysis in the laboratory, etc. Sometimes, it may be necessary to change one or more of these activities.

Guidance is available to assist councils in analysing temporal trends in environmental data (Snelder et al. 2021). However, the current guidance does not consider or address how occasional and inevitable methodological changes should be addressed. These changes include laboratory analytical method changes, changes to sampling protocols or changes to field measurements. From here on, for simplicity, we consider all changes as “method changes”.

An unintended consequence of switching from one method to another can be a "step change" in the reported values of a measured water quality variable. So when it comes to interpreting the resulting trend assessment, when there has been a change in method, the question is, are the results telling us about changes in the environment that are of interest or a change in the monitoring programme? A similar question may arise when interpreting the differences between estimated states before and after a method change. This guidance explores the issues and options associated with managing the sampling changes and recommends how to handle these changes. It assumes that the laboratory method will change and sets out how to evaluate and account for any resulting differences rather than addressing whether the method should be changed.

While there are challenges to interpreting data in the presence of method changes, method changes have benefits. Benefits include lower detection limits, improved measurement precision, and improved comparability of measurements between sites (NEMS 2019).

Hawke's Bay Regional Council commissioned NIWA under the Envirolink scheme to develop guidance for evaluating and, where possible, accounting for the impact of changes in laboratory measurement methods on water quality time-series data. Regional Councils are required to monitor and assess trends in water quality (Ministry for the Environment 2023).

1.1 Background

Recognition of the regional sector's need for consistent monitoring of the quality of fresh and coastal waters through time is evidenced by the establishment of The Environmental Managing and Reporting[1] (EMaR), closely related National Environmental Monitoring Standards[2] (NEMS), and Land Air Water Aotearoa[3] (LAWA) initiatives. More recently, the requirement for robust state and trend assessments in environmental management has been enshrined in the National Policy Statement for Freshwater Management (NPS-FM) 2020, which indicates that councils are required to take action where they detect *"a trend indicating a deterioration"* in any freshwater attribute state. This has heightened the need to ensure that observed changes in attribute state or temporal trend in data are accurate, not just an artefact of a change in method. Further, where a step change has occurred, guidance is needed to support the evaluation of the significance of this step change on both attribute state and trend reporting, including how (or if) the change can be accounted for in the assessment

[1] https://environment.govt.nz/facts-and-science/environmental-reporting/improving-environmental-reporting-data/

[2] http://www.nems.org.nz/

[3] Land, Air, Water Aotearoa (LAWA) - The homepage

and interpretation of state and trend (e.g., through the determination of an adjustment or correction factor).

The NEMS for sampling, measuring, processing and archiving of discrete water quality data (hereafter the “Standard”) was introduced in 2019 (2021 for coastal waters) to improve national consistency in State of the Environment (SoE) water quality monitoring of rivers, lakes and groundwater. However:

- implementing the NEMS has required many councils to change existing laboratory measurement methods for some variables – these changes in methodology could introduce 'step' changes in their long-term time-series data, and
- as noted in the NEMS, changes in laboratory measurement methods are inevitable as analytical instruments change and new methods emerge over time.

Consistent with NEMS advice, where existing laboratory measurement methods differ from those specified in the Standard, multiple councils (e.g., Horizons, Environment Southland, Auckland, Greater Wellington, Hawke's Bay, Gisborne) have completed a period of parallel measurements taking measurements of water samples from routine monitoring sites using existing (old) and NEMS (new) methods. The period of parallel measurement is intended to:

- provide sufficient paired data to determine if a step change is present - such step changes have been reported for some nutrient species and turbidity in New Zealand and elsewhere (Davies-Colley and McBride 2016; Oelsner et al. 2017; Lindenmayer et al. 2022), and if so
- determine if a relationship exists between the two sets of measurements that may be used to enable the dataset to be 'adjusted' or 'corrected' to obtain a consistent time-series record for assessment that includes both sets of observations.

The NEMS did not establish a process for evaluating paired datasets, and although the *Trend Assessment and Reporting Guidance* (Snelder et al. 2021) developed for the regional sector under Large Advice Grant HZLC154 acknowledged the potential existence of step changes arising from changes in measurement methods, addressing this issue was out of scope. Consequently, no nationally specified or recognised procedures exist to interpret the results of paired measurements, derive conversion factors, or even decide how or when two paired datasets can be deemed equivalent. Establishing “equivalency” must consider the variable in question and measurement accuracy, precision and uncertainty. The absence of a nationally consistent process has potentially left councils with 'broken' time-series records arising from changes in laboratory providers or laboratory methods (including by picking up former National River Water Quality Network monitoring sites from NIWA). The existence of 'broken' time-series records will hinder Councils’ ability to confidently identify temporal changes in water quality.

This report addresses a gap in the advice on implementing the NEMS. Specifically, it builds on Horizons Regional Council's recent *Paired Laboratory Nutrient Analysis* (Hunter et al. 2022*)* guided by NIWA under a small advice grant (HZLC157) and earlier work by NIWA with Greater Wellington Regional Council (e.g., Davies-Colley et al. 2019), and makes use of other relevant paired data sets (e.g., turbidity data sets held by Horizons, Environment Canterbury (ECan) and Environment Southland). The work by Horizons employed multiple statistical approaches to assess paired data sets for several nutrient species (including automation of the assessment process through the development of an *R* script); further work was needed to confirm the most robust and transparent approach(es) and to document the steps followed. This report draws on experience and data from Horizons and other councils to develop guidance.

Completion of this Guidance is consistent with the priorities of several SIG (Special Interest Groups) research strategies and one of several key principles on which EMaR is based: *Standardised protocols and methods, and robust quality assurance.*

1.2 Scope of the guidance

This guidance intends to assist councils with accounting for the impact of changes in laboratory measurement methods on river, lake and coastal water quality time-series data (e.g., State of the Environment (SoE) data). The approach is expected to apply to other changes, such as a change of laboratory or, in some instances, a change of sampling protocols and changes in field measurements. The guidance covers:

- a summary from a brief search of the literature on approaches to evaluating differences in paired measurements
- parallel testing protocols and paired data
- statistical procedures to assess the impacts of changes in laboratory methods, including the treatment of censored values
- variable-by-variable evaluation of paired nutrient data
- worked examples illustrating the paired data assessment and interpretation steps, and
- consider what can be done in the absence of paired data.

The intention was to evaluate a range of soluble inorganic and total nutrients and other parameters such as chlorophyll *a*, turbidity, *E. coli*, and enterococci. The specific variables covered would depend on the provision of council data. The guidance development process would include an on-line workshop with laboratory chemists, water quality scientists, data analysts and practitioners.

Councils provided data for several water quality variables that have been subject to changes in analytical methods, including nitrate, nitrite, Dissolved inorganic nitrogen (DIN), Total Kjeldahl nitrogen (TKN), Total Nitrogen, Total Oxidised Nitrogen (TON), Total Phosphorus and Turbidity. The overall guidance is anticipated to be generalisable and applicable to many water quality variables. However, specific challenges were encountered with *E. coli* and enterococci, which are measured on discrete scales[4] and where laboratory method changes often involve changes in enumeration systems, were out of scope and would not be covered.

The guidance development approach included a workshop with water quality scientists, data analysts and practitioners. Limited discussions with laboratory chemists and microbiologists took place outside of the workshop. The scope of the work included limited literature reviews and data analysis.

Underlying the approach are some assumptions. Firstly, monitoring is carried out in accordance with the appropriate National Environmental Monitoring Standards (NEMS), which lay out the procedures for the sampling, measuring and processing of water quality data. These procedures are documented

[4] Discrete data take distinct values, such as how many pupils are in a class. Continuous data is not constrained to take specific values (at least in theory). For example, the height of a pupil is measured on a continuous scale.

in a Field and Office Manual or equivalent (NEMS 2019b). Following this assumption, any changes to laboratory methods will be made to a NEMS-recognised test method, so it is already established that the replacement method is fit for monitoring water quality. The evaluation of new or non-NEMS methods falls outside the scope of this guidance and may require a more thorough investigation of the comparability of the new with the existing methods. A second assumption is that this guidance deals with discrete sampling associated with SoE reporting as opposed to high-frequency monitoring. Another critical assumption is that time-series analysis aligns with the g*uidance for the analysis of temporal trends in environmental data* prepared by Snelder et al. (2021).

It is recognised that there are instances when it is not possible to follow NEMS best practice guidance, such as when addressing retrospective method changes in the absence of paired data. Limited discussion and advice is given for dealing with changes without paired data.

While advice on the decision as to whether to change the laboratory method or not is beyond this report's scope, the recommendations given in this guidance are expected to help those making these decisions. However, this guidance focuses on how to account for the changes for time series analysis once the decision to change has been made.

The recommendations only apply to the transition to a NEMS-approved method. High-frequency data may require the approach to be modified, and the approval of new laboratory methods (other than NEMS-approved methods) is out of this project's scope.

1.3 Report layout

The report has five main sections:

- method section (Chapter 2) outlines the overall approach, which includes a literature review, workshop and data analysis
- issues section (Chapter 3) examines the challenges related to assessing and accounting for the impact of changes in laboratory measurement methods on the interpretation of long-term time-series data, along with potential solutions
- observed level of agreement section (Chapter 4) presents an analysis of levels of agreement between two laboratory methods from New Zealand pairwise methods studies
- final two sections (Chapters 5 and 6) provide recommendations on a general approach with worked examples to assist those accounting for and assessing laboratory changes in time-series data.

2 Methods

This work draws on multiple lines of evidence to identify issues and options to form recommendations. The evidence and opinions came from four sources: a literature review, an analysis of council data, a stakeholder workshop and feedback on the draft report.

A brief narrative literature review[5] identified themes associated with water quality trend analysis in the presence of laboratory method changes. The review included published academic papers and grey literature such as NIWA and Regional Council reports.

The council data was analysed to explore the similarities and differences between pairs of laboratory methods, test ideas from the literature review, and determine the level of agreement between methods. Three councils - Horizons, ECan and Environment Southland - provided paired water quality datasets. These datasets include a range of nutrients plus turbidity, collected using NEMS and other methods.

Potential issues and options were identified based on the information gathered and lessons learned during the data analysis. The issues raised were discussed with a small group of water scientists and analysts. Points raised in these discussions and feedback on the draft report were used to inform this guidance.

[5] In contrast to a systematic review, which is highly structured and typically more thorough, a narrative or conventional literature review offers a subjective summary of a particular area.

3 Issues

This section examines the challenges related to assessing and accounting for the impact of changes in laboratory measurement methods on the interpretation of long-term time-series data, along with potential solutions. Laboratory method changes are inevitable and, to a certain extent, unavoidable due to changes in technology, equipment becoming obsolete, requirements for standardisation and the evolution of monitoring practices. This section includes findings from the literature and issues raised during a workshop and leads on to the next section, which analyses regional council paired data.

3.1 Findings from literature

A key purpose of water quality monitoring is to assess state and trend. From time to time, it may become necessary to change water quality monitoring procedures, such as laboratory methods. There are benefits to moving to new methods, including national consistency and better precision, particularly for low-nutrient environments where levels may be near the detection limits of the current methods (Davies-Colley and McBride 2016). However, changes can complicate the interpretation of estimated state and trends and must be considered (Newell et al. 1993; Smith and McCann 2000; Coats et al. 2016; Domagalski et al. 2021). Failure to consider these changes can raise concerns that the estimated trends reflect changes to the monitoring methodology rather than the environment (Newell and Morrison 1993; Meals et al. 2011) and can undermine confidence in the reliability of estimated trends.

Failure to take into account changes has real-world implications. A long-term reported decline in silica concentrations in Lake Michigan now appears to be due to a change in analytical methodologies (Shapiro and Swain 1983). The fallout of the trend assessment led to legal action (Lindenmayer et al. 2022). However, method changes do not automatically invalidate the results of trend analysis. In many cases, it can be established that the changes are less than minor and have no bearing on the interpretation of trends (Robson and Neal 1997; Oelsner et al. 2017).

Several themes were identified in the literature, including:

- bias - systematic differences between methods
- differences in limits of detection between methods
- differences in precision between methods
- approaches to estimate the level of agreement between methods
- factors influencing levels of agreement
- accounting for differences between methods and making adjustments for use in trend analysis.

Artificially created trends in time series can occur due to bias or differences in detection limits. Addressing these issues requires estimating the level of agreement between methods. Factors such as confounding variables can also affect the level of agreement, and each of these themes is explored in more detail in the following sub-sections.

3.1.1 Bias

In the context of this guidance, bias is the systematic difference between two methods. At times, bias can result in a noticeable step change in time series data. However, Davies-Colley and McBride (2016) pointed out that method-associated step changes in water quality data are not always obvious, and they illustrated the issue using two graphs (**Error! Reference source not found.**). The left-hand graph shows an apparent upward trend in a water quality variable over five years. However, not all is as it appears, as reporting changed to include data derived from a new method at 30 months. Knowing that fact could raise suspicions, but it is not definite proof that the observed trend results from method changes. The right-hand graph, which plots data derived from the original and new methods together, clearly demonstrates that changes in laboratory methods can influence the observed trend. Though there is an upward trend in the data derived from the original method, switching to the new method changes the magnitude of the trend.

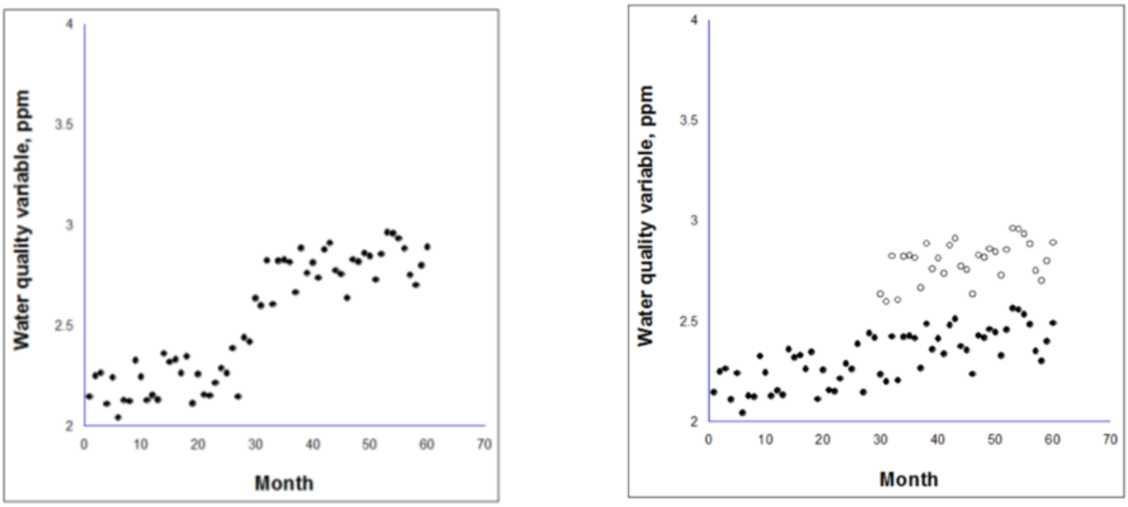


Figure 3-1: The impact on water quality trend due to a change in method (left) and results showing what would have happened if the method had not changed (right). The figures are reproduced from Davies-Colley and McBride (2016). The open circles on the right are the values associated with the new laboratory method, and the filled circles are the original protocol.

In the case study summarised in Figure 3-1, the absence of paired results derived from new and old sampling methods makes bias detection difficult. The presence of bias can lead to erroneous conclusions about environmental trends.

Several factors may influence bias when changing measurement methods, not least the methods themselves. The Australian New Zealand Guidelines (2018) for water quality note that interference and a laboratory's use of a method can contribute to bias. Examples of bias between laboratories have been documented (Davies-Colley et al. 2019; Kilroy and Daly 2020), as well as interference effects (Rus et al. 2013; Coats et al. 2016; Davies-Colley and McBride 2016)

3.1.2 Limit of detection

Achieving lower detection limits is a frequently mentioned reason for changing analytical methods. Changing an analytical detection limit can also result in step changes, which then can influence the results of state and trend assessment (Smith and McCann 2000; Chanat et al. 2016). This is illustrated in Figure 3-2. The left-hand diagram illustrates a situation where the detection limit changes from 15 units to 1 unit at 30 months, and the right-hand diagram shows actual values without censoring.

There is a distinct difference in the trend line slope between the left and right diagrams, illustrating that censoring can impact the estimated slope.

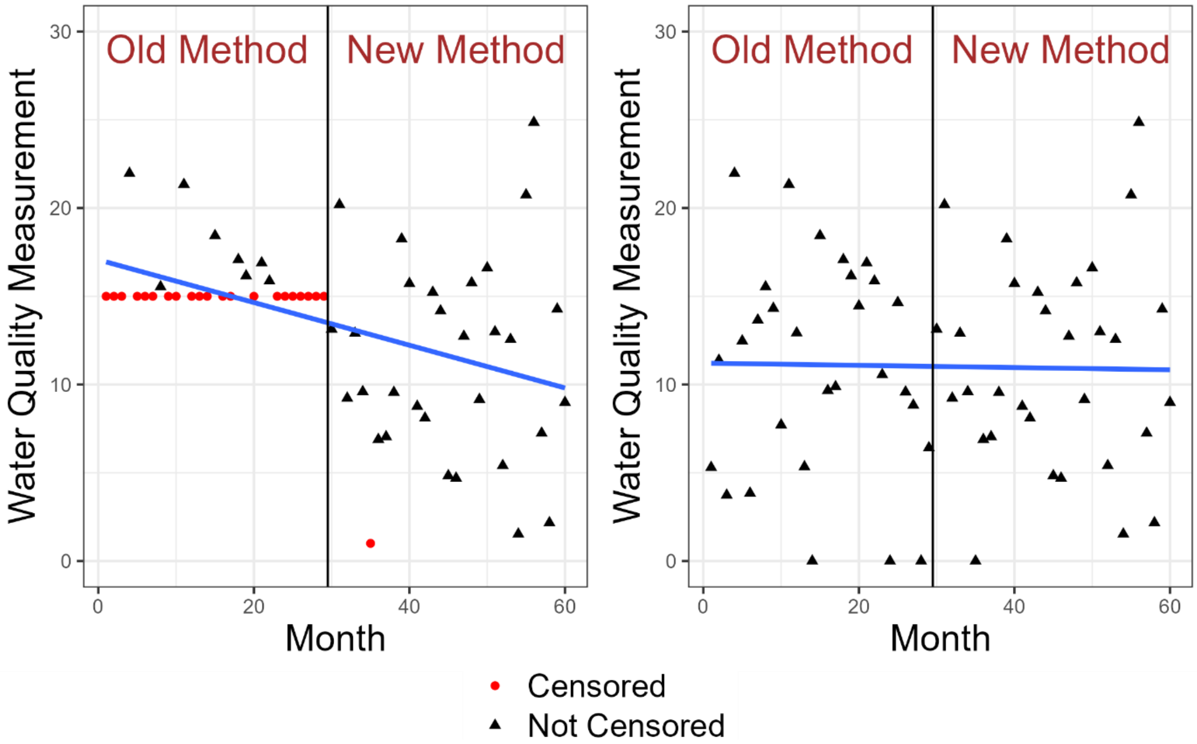


Figure 3-2: An illustrative example of the impact of change in detection limit on trend assessment; the blue line is the evaluated trend. The left diagram has two detection limits (15 - old method, 1 - new method), and the right-hand diagram is the uncensored data. There is a slight trend in the uncensored data, but the change in the detection limit artificially induces a stronger trend.

In situations like this, when the limit of detection changes, the most straightforward approach is to censor the data up to the highest reporting limit (Helsel 2011). This approach is recommended in the current guidance on trend assessment (Snelder et al. 2021). However, it also means that the benefits of lower detection limits may not be achieved while including the older data in the trend analysis.

3.1.3 Precision and reproducibility

Repeated measurements of the same physical water sample under the same conditions using the same method will yield slightly different results. The amount of variation is called precision. When there is little variation in repeated measurements, precision is high; alternatively, precision is low when there is high variation. Precision plays a part in evaluating the limits of detection of a method.

Many water quality studies do not estimate precision through repeated measurements. However, precision is one factor that can influence how well the two methods agree (Bland and Altman 1999). If the precision of one method is low, the results may differ each time the measurement is repeated. Therefore, it is very unlikely that you will get the same result if measurements were made of the same water sample using a second method, even if that method had very high precision.

We found little discussion of reproducibility and precision in the literature review. In theory, better precision would be a reason to adopt a new method, though better precision does not necessarily equate to better accuracy (Rus et al. 2013). There may be benefits in adopting methods with better precision for detecting changes in the state of the environment. However, in practice, the differences in precision between methods may make little difference in trend detection. There are multiple sources of variation in environmental data, and variation due to the precision of methods being just one. Still, it may make a difference in detecting trends at low concentrations and/or very stable environments, such as those in some groundwaters.

Snelder et al. (2021) noted that another aspect of precision relates to numerical precision and how numbers are rounded. Methods with higher precision may be reported with higher numerical precision than methods with lower precision. This needs to be considered in non-parametric trend analysis as it can influence the number of tied values. The reader is referred to *the guidance for the analysis of temporal trends in environmental data* (Snelder et al. 2021) to understand the concept of tied values in trend analysis. Rounding high precision values to align with lower precision values is the most acceptable course of action when numerical precision varies between methods.

3.2 Estimation of the level of agreement between methods

Most papers and reports identified in the literature review agreed that if methods were to change, then some form of collection of data using the new and existing methods should be carried out in parallel (Coats et al. 2016; von Bromssen et al. 2018; Davies-Colley et al. 2019). Parallel testing results in multiple pairwise observations and enables the level of agreement between methods to be estimated. When estimating the level of agreement, generally, we are comparing one method with another method and not against a gold standard as in a calibration (Francq and Govaerts 2014), so we are not trying to establish which method is the most accurate, only how well the two methods agree.

Section 5.7 of NEMS (2019b) calls for monthly parallel testing for a period of time (12 months). The best practice measures of the same NEMS (2019b) Standards call for a minimum of 12 samples if taken monthly. However, longer periods have been used. Coats et al. (2016) used two years of parallel water quality data to assess levels of agreement and to account for method changes in monitoring data of Lake Tahoe (USA). In the case of meteorology studies, one paper talks about comparisons over at least 5 years (Rhoades and Salinger 1993).

Three common approaches are used to estimate levels of agreement from parallel testing:

- correlation
- the Bland-Altman approach, and
- regression.

Other approaches, such as Youden plots (Kilroy and Daly 2020), Cumulative Sum chart CUSUM (Hunter et al. 2022) and paired t-tests, have been applied to investigate levels of agreement between methods. We restrict the discussion to the first three approaches as they are the most common; each is discussed in turn.

3.2.1 Correlation

The correlation coefficient, such as Pearson *r* or Spearman *ρ*, is commonly used to assess the level of agreement between two methods. Values of *r* or *ρ* close to one indicate better agreement than values closer to zero or minus one. However, this interpretation may be misleading (Bland and Altman 1986) for two reasons

1. *r* measures the strength or relationship along a line, not limited to the 1:1 line of numerical equivalence. It is possible to have a perfect correlation between results but completely different numeric values. Therefore, this approach does not tell us about a potential step change (bias) if we change from one method to another.
2. Correlation depends on the range of a quantity. A wide numerical range often leads to higher correlations, and water quality variables often have a high numeric range. This can disguise the existence of critical numerical differences between the methods.

The first problem can be overcome partly by using Lin's concordance correlation coefficient (Lin 1989), which estimates how close two methods are to a 1:1 line of perfect numeric agreement. A possible interpretation of Lin's concordance correlation coefficient is given in Table 3-1. However, Davies-Colley and McBride (2016) noted that the approach suffers from the same range problem as Person's *r* and is also sensitive to the sample size.

Table 3-1: Suggested verbal interpretation of Lin's Concordance Correlation Coefficient (CCC). After Davies-Colley and McBride (2016).

Lin's CCC	Strength of agreement
>0.99	Almost perfect
0.96-0.99	Substantial
0.90 – 0.95	Moderate
<0.90	Poor

3.2.2 Bland Altman

Bland and Altman introduced an alternative method to correlation in 1986 to assess the level of agreement between two methods. Their paper (Bland and Altman 1986) has been cited over 50,000 times[6] and is widely used for assessing agreement between methods, although it seems to be not commonly used in water quality studies. The method involves using plots to visually assess bias and precision and simple calculations to estimate the level of agreement.

The Bland and Altman approach involves two plots, as shown in Figure 3-3. The left graph plots method A's results against method B's (one pair of results per symbol), and the right-hand graph plots the difference between the paired results against the average value of each pair.

On the left-hand diagram of Figure 3-3 are the points for the paired results and a solid black line, the line of equality. The line of equality is where the results would lie if the two methods gave exactly the same results. The pairwise points do not all lie on the line of equality, illustrating that the two methods do not give exactly the same results.

The right-hand diagram of Figure 3-3 focuses on the differences between the paired results, which is of most interest in method comparison studies. There are three lines as well as points. The blue line is the mean difference or bias between the paired results. The two red lines are the upper and lower limits of agreement between the two methods. Each of those lines is discussed below.

[6] According to Google Scholar on 9/10/23

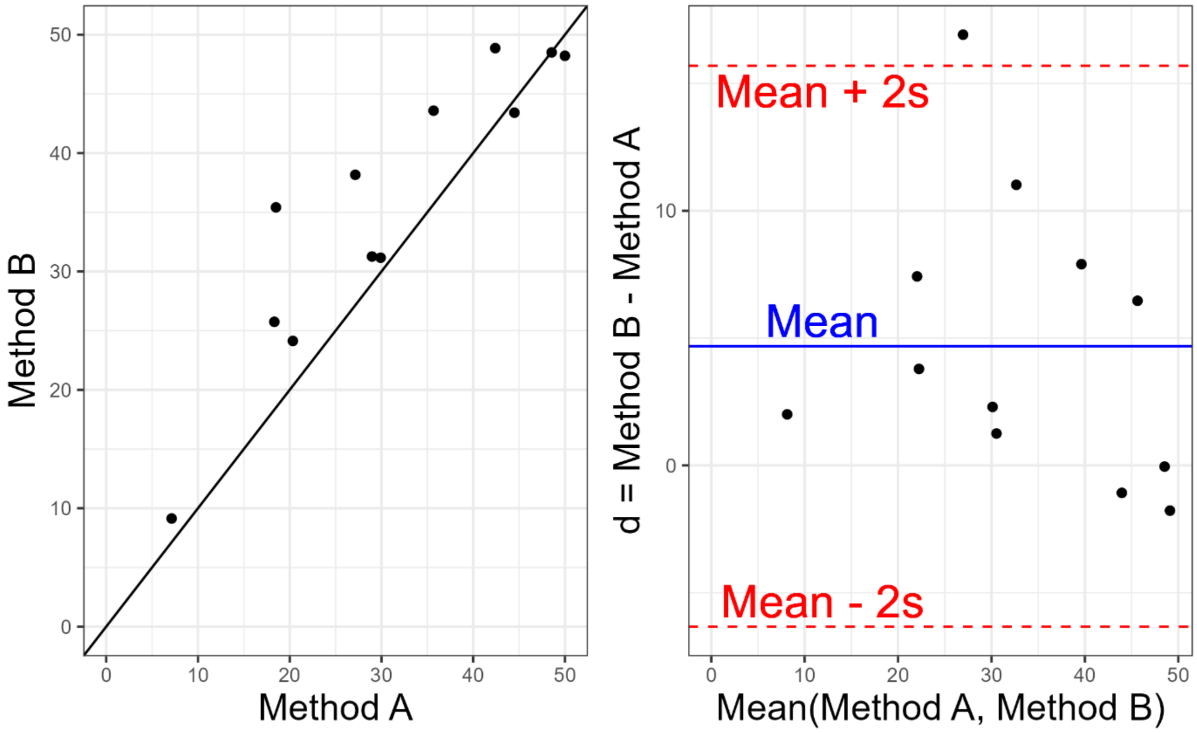


Figure 3-3: Bland and Altman plot, the left diagram shows the relationship between the two methods, and the right diagram shows the difference between methods with their average value. Note that the positive average value on the right-hand diagram implies that method B generally gives higher results than method A (this is synthetic data). The x-axis on the right-hand diagram, Mean (Method A, Method B), is the average value of each result pair.

In detail, bias, illustrated by the blue line, is calculated by taking the mean of the differences ($\bar{d}$), where the difference between paired observation, *i* is $d_i = method\ B_i - method\ A_i$ and is an estimate of the systematic difference between the two methods.

The red dotted lines are estimates of the limits of agreement – these define a range where we expect most of the individual differences between the two methods lie. It is a measure of the random differences between the two methods and is based on the standard deviation of the differences. If the differences are normally distributed, then most of the differences (approximately 95% of the observations) will be between:

$$Upper\ level\ of\ agreement = \bar{d} + 2s$$

$$Lower\ level\ of\ agreement = \bar{d} - 2s$$

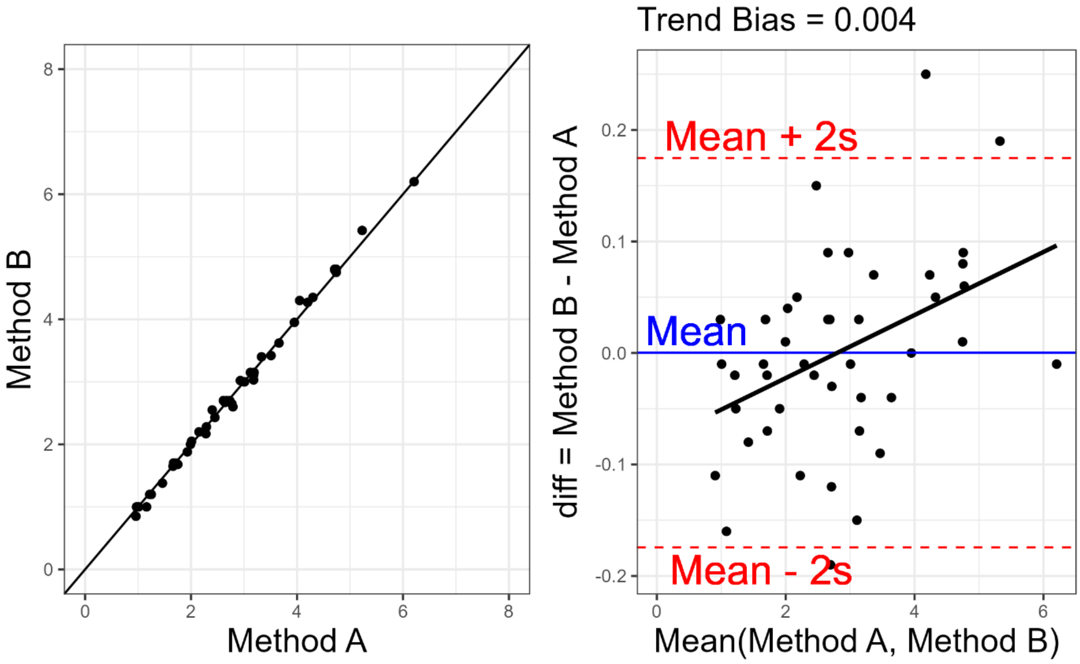


Figure 3-4: Bland and Altman plots for two methods with a trend in the bias. Data comes from Bland and Altman (1999). The p-value of the t-test of the slope in the regression line = 0.004 (black line on the right-hand diagram), which suggests sufficient evidence of a trend in the bias (Trend Bias) as a function of the measured values.

The graphical approach also allows us to visually identify any variations in bias across the observed value range. The visual assessment can also be supported by statistical analysis. For example, it is possible to add a fourth line to the right-hand Bland and Altman plot, as shown in Figure 3-4.

In Figure 3-4, the regression line in black on the right-hand figure shows a tendency for method B to give higher results than method A at higher measured values. Using linear regression, we can estimate the p-value of the slope, which helps inform us if there is any trend in the bias (trend bias). In this case, the p-value = 0.004 is evidence that the bias varies by measured levels of A and B.

Water quality variables are generally highly skewed, so a log transformation of the data is required[7]. Using the log transformation means we estimate the relative rather than absolute difference between the methods on a log scale. This is done because the range of absolute differences generally becomes larger with higher values of measured water quality variables. This makes it difficult to estimate the bias. If absolute differences are used to estimate the correction factor, occasionally, the results will be implausible, such as negative adjusted values.

To avoid this situation, the log-transformed approach uses a slightly different approach, so the difference is $d_i = method\ B_i/method\ A_i$ which becomes $\log\left(d_i\right) = \log(method\ B_i) - \log\left(method\ A_i\right)$. Once the statistics have been calculated, it is best to take the antilog and present

[7] Bland and Altman suggest "log transformation is the only transformation giving back transformed differences which are easy to interpret, and we do not recommend using any other in this context". If log transformation proves unsuitable, we suggest regression approaches should be tried.

the level of agreement as a ratio rather than a log of the ratios. A worked example and calculations can be found in Appendix B.

Using relative rather than absolute values of difference has another valuable property. The level of absolute bias tends to scale with the average value of the variable of interest. So, sites with low nutrient concentrations tend to have numerically smaller levels of bias than sites with high nutrient concentrations, making it difficult to compare bias for the same pair of methods between sites. Using relative bias overcomes both of these issues.

3.2.3 Regression

Regression methods are an obvious approach for assessing the agreement between the methods. The methods most commonly used in water quality analysis, such as Ordinary Least Squares (OLS) regression, assume errors only exist in the dependent (or Y) variable. However, errors exist in the measurements of both variables, so which one is the independent or dependent variable? Assuming that measurement from method A is the independent variable results in one regression line, whereas assuming method B is the independent variable results in a different line, as illustrated in Figure 3-5. Therefore, many of the most common methods are considered unsuitable for assessing agreement.

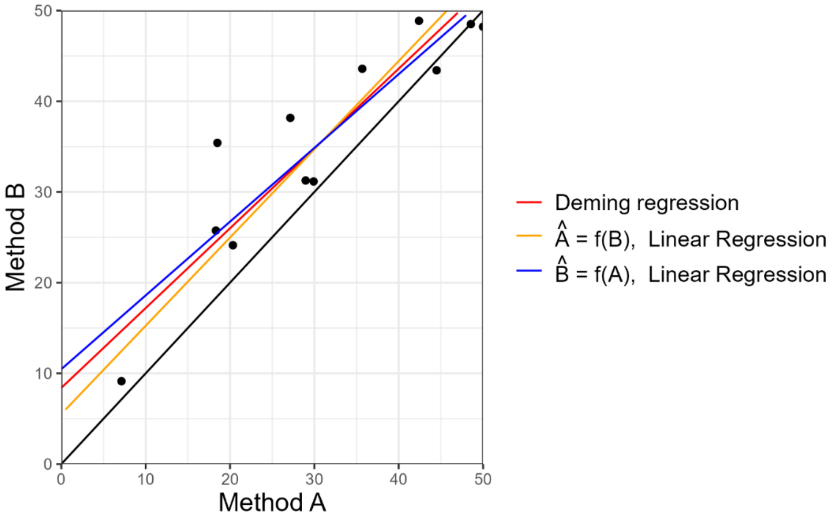


Figure 3-5: Two Ordinary Least Square (OLS) regressions and one Deming regression applied to paired results.

Several methods exist for dealing with errors in both variables, including Deming and Passing-Bablok regressions, and are applied to method comparison studies (Johnson 2008; Francq and Govaerts 2014). Davies-Colley et al. (2019) used a geometrical mean regression approach to assess the level of agreement between water quality variables measured by two agencies. According to Wicklin (2019), the US FDA (Food and Drugs Administration) encourages the use of Deming regression for method comparison studies. In addition to dealing with errors in both variables, Deming regression can also account for differences in the errors (or precision) associated with the new and old methods. However, because generally only single results are reported for old and new methods in water quality method comparison studies, precision cannot be calculated, so this is only a theoretical benefit.

It should be noted that regression analysis can suffer from issues related to range, similar to correlation analysis. A worked example of Deming regression is given in Appendix B.

3.2.4 Comparisons between methods to estimate the level of agreement

Each of the three methods discussed has its own strengths and weaknesses. Lin's Concordance Correlation Coefficient summarises the relationship between the two methods into a single number. The Bland and Altman approach and regression analysis goes further. They can both be used to assess the agreement level and correct or adjust values if necessary. None of the three methods can adequately accommodate censored data, so before any calculations using Lin’s, Deming regression or Bland Altman's approach are carried out, censored values must be removed from the dataset. However, plotting and highlighting which values are censored is still recommended.

The Bland and Altman approach provides a simple graphical summary of the difference between methods. A distinct benefit of regression approaches is that they can describe and account for confounding variables, such as bias between two laboratory methods used for estimating nitrate as a function of total suspended solids TSS (Rus et al. 2013; Davies-Colley and McBride 2016). This cannot be done within the Bland and Altman approach; however, using the extra features of a regression approach generally requires a larger number of pairwise samples than the minimum number called for by NEMS (2019b), so it is not described here.

There is no reason why Lin’s concordance, Bland and Altman, and regression analysis should not be used in parallel, and this approach is demonstrated later in the report. However, if a choice needs to be made, Bland and Altman's approach is recommended due to its simplicity and because it provides a descriptive visual representation of the level of agreement. In any case, it is recommended that the Bland and Altman approach should be conducted before using Lin’s or regression approaches.

3.3 Trend analysis in the presence of method changes (Accounting for and adjusting trends)

When it comes to analysing trends while accounting for changes in the method, two broad approaches can deal with bias (step change) and differences in detection limits:

- first a stepwise approach where differences between the methods are estimated based on paired sample results from parallel sampling, and this information is used to adjust for bias and changes in detection limits prior to trend estimation
- second approach simultaneously estimates trends whilst accounting for differences in methods; although this can be done with or without the use of paired sample results, the use of paired sample results is recommended (von Bromssen et al. 2018).

The first approach is a sequential stepwise approach, which appears most consistent with non-parametric methods to estimate monotonic trends as recommended in the *Guidance for the analysis of temporal trends in environmental data* (Snelder et al. 2021). It involves accounting for each confounding factor before trend assessment. The confounding factors associated with method changes are bias and differences in detection limits. Either Bland and Altman analysis or regression can be used to estimate for bias, and the resulting estimates can be used to adjust values (Newell and Morrison 1993; Coats et al. 2016). Detection limits provided by the laboratory need to be harmonised prior to other covariate adjustments, such as flow, before trend analyses. More details about how this is done can be found in the worked example section below (Section 6).

The second approach, which simultaneously estimates trends in the presence of a method change, has predominantly been applied to non-monotonic trends and uses methods such as General Additive Models (GAM) and General Additive Mixed Models (GAMM) as per von Bromssen et al.

(2018) and Murphy et al. (2019). Although approaches have been proposed for estimating monotonic trends in the presence of a step change without pairwise data, such as a modified Kendall test (Newell et al. 1993), we are unaware of this being used in practice. Using parallel testing is preferable for assessing trends in the presence of method changes. In the absence of parallel testing, we are left with trying to simultaneously estimate two unknowns, trend and bias, which is a difficult task (Oelsner et al. 2017). So, this approach is not recommended.

In any case, whatever approach is used, having only twelve samples (taken monthly) is a weak evidence base for making adjustments to data for state or trend analysis. Method comparison studies have often considered 30 or 40 samples as a minimum sample size for estimating bias (Passing and Bablok 1984; Linnet 1999), but ignoring the evidence from even a small number of samples is an even weaker position. Bias estimated from Bland and Altman or regression approaches should be used to adjust data before a state or trend assessment.

Sensitivity analysis may be worthwhile given the potential impact of adjusting data on the conclusions from a state or trend assessment. A sensitivity analysis systematically explores how an output (state or trend) responds to changes to an input (bias). This approach helps analysts develop an informed opinion as to whether observed changes in state or trend could be due to method changes or environmental changes.

3.3.1 Discrete data and microbiological variables

Many water quality variables, such as the level of nutrients, are measured on continuous scales. Theoretically, they can take infinite values, ignoring complications with censoring and rounding numbers. However, exceptions exist (such as microbiological enumeration data), which are measured on discrete scales. The pragmatic approach assumes that continuous and discrete data can be treated similarly, and the numerical results from one test are equivalent to those from another test, possibly with some conversion factor. However, this may not be true for some pairs of microbiological data.

Different microbial tests use different counting systems. For instance, the NEMS test method for *E. coli* and enterococci uses the MPN (Most Probable Numbers) enumeration system, while traditional systems use CFUs (Colony Forming Units). A possibly surprising result of the two different approaches is that numbers may appear on one scale (MPN) but not the other (CFU), posing a unique challenge when interpreting or comparing the results from the two methods. The differences between microbiological counting schemes mean that achieving a perfect agreement between the two methods is unlikely (McBride 2005).

Exactly how these differences between tests impact accounting for and assessing the influence of method changes has not been explored for the NEMS format of the MPN test (though not stated in NEMS, it appears to use the IDEXX Quanti-Tray 2000 format for the test) against other enumeration tests. There are theoretical reasons for the differences between CFU and MPN results (McBride 2005; Gronewold and Wolpert 2008). Gronewold and Wolpert (2008) identified the problem of combining MPN and CFU datasets to support environmental decision-making.

We have not explored the most appropriate approach for dealing with the differences arising from MPN/CFU enumeration methods. The most straightforward and pragmatic solution assumes that CFU and MPN values may be used interchangeably, possibly including a correction factor as in the case of nutrients (Prats et al. 2008; Rubini et al. 2023). Further investigation may reveal that this is not the optimal approach.

3.4 Deciding when and when not to account for method changes

Although there is discussion in the water quality literature about bias, changing limits of detection, and how these can be accounted for, there is little discussion of when to account for a method change and what the decision criteria should be. Coats et al. (2016) applied individual corrections to all sites in their study rather than trying to establish whether a correction was required or not. In contrast, Chanat et al. (2016) suggested that the decision should be made on a case-by-case basis.

The medical literature discusses acceptable differences between methods in terms of the decisions made (Westgard 1998; Bland and Altman 1999; Johnson 2008). They frame the issue of significant differences regarding practical differences in terms of outcomes rather than formal statistical differences.

There has been some discussion between the project stakeholders about how small differences between methods do not need to be accounted for and what a significant difference is. However, it is not clear how to operationalise this idea. At least in theory, minor differences between methods could, in the case of freshwater, result in changes to attribute bands and trends. So, the most straightforward approach is to always account for method changes and use sensitivity analysis to assess if uncertainty around method changes may influence an assessment against criteria such as NPF-FM s3.19, which requires specific actions if a deteriorating trend is detected.

3.5 Points raised during the workshop

An on-line meeting took place between water quality scientists, data analysts and regional council science managers. The meeting traversed a range of issues associated with using water quality data in the presence of laboratory method changes and the implications. The discussion was based on the group's collective experience and took a broader perspective than found in the brief literature review. The meeting also explored the type of advice and guidance the sector sought.

The sector was looking for broad guidance on managing issues arising from method changes, such as:

- when could two methods be deemed equivalent, what criteria should be used to decide equivalency and, more importantly, when are they not equivalent? Could we create a cutoff value and rules to inform these decisions?
- is there a need for multiple pairwise observations, and are 12 months of observations sufficient?
- if the issue is due to a change to laboratory methods, do we need to sample from multiple sites? Or would parallel testing from one site be enough?
- if we need to adjust data to account for method changes, what are the pros and cons of the different adjustment approaches?
- do we need to adjust all historical data, and could this impact NEMS Quality Codes?
- how do we deal with situations when pairwise data are absent, as in the case of historical changes in methods?

There was a consensus around:

- the importance of metadata and ensuring all changes to methods are documented, including changes which may be regarded as minor or insignificant at the time, and ensuring all adjustments made in the data analysis stage are documented
- need to work closely with laboratories to improve method comparisons (Jensen and Kjelgaard-Hansen 2006), with specific mention of comparison as low nutrient concentrations
- The use of judgment calls for dealing with method changes. At times, the most appropriate course of action is to adjust data to take into account method changes. In other situations, it may be appropriate to separately analyse for state and trend before and after a method change.

The group noted that, at times, the changes in method had not been recorded in water quality datasets, meaning the issue is often overlooked and impossible to identify in some historical datasets. There is a need for basic record-keeping and documentation regarding monitoring programme design, which should address sites, frequencies, sampling methods, storage post sampling, sample handling in the lab, as well as methods of analysis, detection limits, accuracy, repeatability to ensure stakeholders have confidence in the results.

4 Observed level of agreement

This section focuses on the level of agreement between results derived from different methods. There are many causes for differences. Although the level of agreement is specific to pairs of laboratory methods, it is also influenced by the procedures used to collect the samples and the agency that collects the samples (Davies-Colley et al. 2019). There may be differences between laboratories using the same method (Kilroy and Daly 2020). In addition, other water quality factors, such as sediment concentration (TSS), can also influence the levels of agreement between methods (Rus et al. 2013; Davies-Colley and McBride 2016).

Using data provided by councils, which came from 72 groundwater, 43 lake and 336 river monitoring sites (estuarine and coastal sites were not represented), we estimated the levels of agreement between various combinations of methods for eight water quality variables (number of sites in brackets): nitrate (244), nitrite (215), dissolved inorganic nitrogen (DIN) (43), total oxidised nitrogen (TON) (205), total Kjeldahl nitrogen (TKN) (3), turbidity (141), total nitrogen (TN) (94) and total phosphorus (TP) (78) (Figure 4-1). The number of paired samples is shown on each graph. A description of the individual methods is provided in Table A-1. Plotting on a log scale shows a general tendency for the results pairs to cluster around the 1:1 line, though there is noise or deviation from the line. Paired observations, where one or more of the values are censored, are plotted on the graph as blue dots, with censored values being represented as the numerical value of the limit of detection. For most analytes, deviation from the 1:1 line is greatest at lower values and when one or more values are censored, except turbidity and TKN.

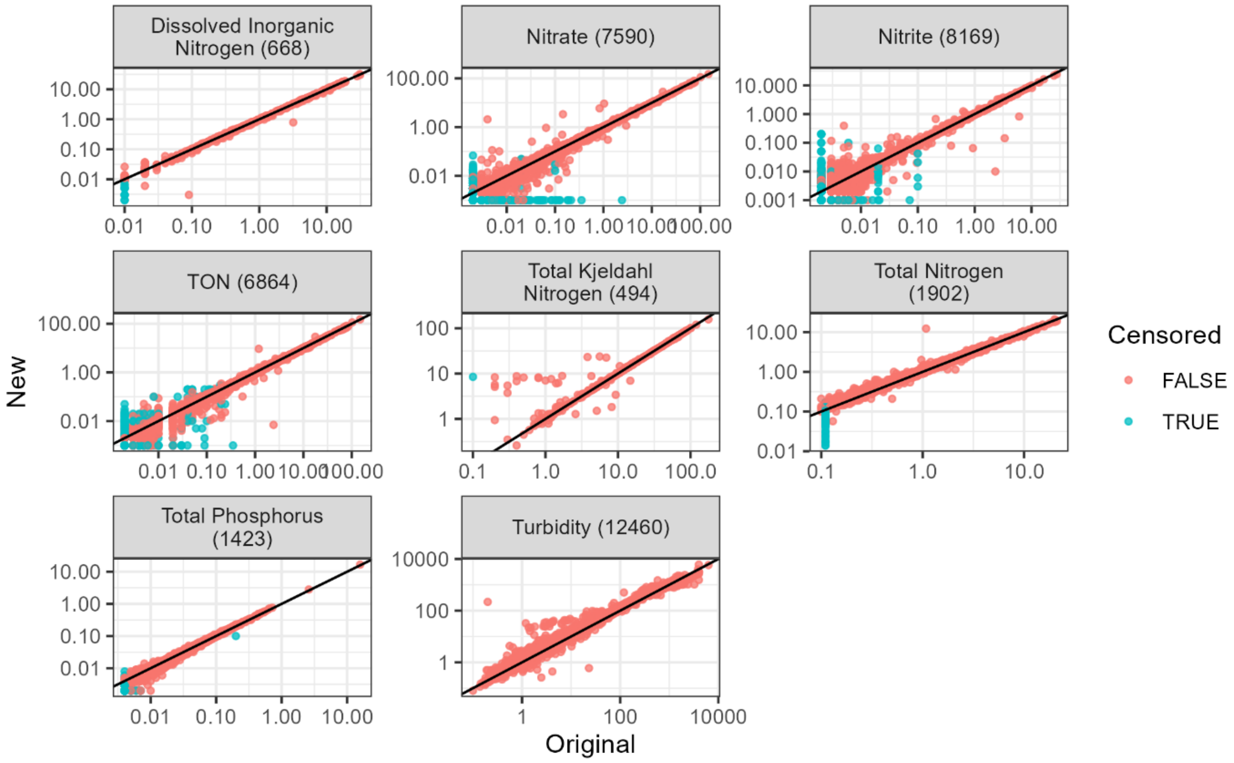


Figure 4-1: New (NEMS) and original (old) methods for measuring eight water quality variables for multiple sites. The black line is a line of equivalence, and the colours indicate if one or more of the measurements are censored. The figures are plotted on a log scale (units are mg/L for nutrients and NTU for turbidity). Most of the points plot on or around the line of equivalence. Except for turbidity, the differences become proportionally greater at low concentrations. The observation count is printed on the title, and the number of sites is given in Table 4-1.

It should be noted that these results are not fully representative of bias between method changes. There are multiple reasons as to why. Firstly, the data only includes a subset of potential method changes. Secondly, the results only came from a few regions, so they may not be representative of water quality in other regions. Thirdly, the data comes from a small selection of laboratories.

The strength of agreement from Lin's CCC was estimated for each site in the sample and shown in Figure 4-2. There were examples of very high concordance but also poor concordance (where Lin's CCC <0.90), particularly for nitrite.

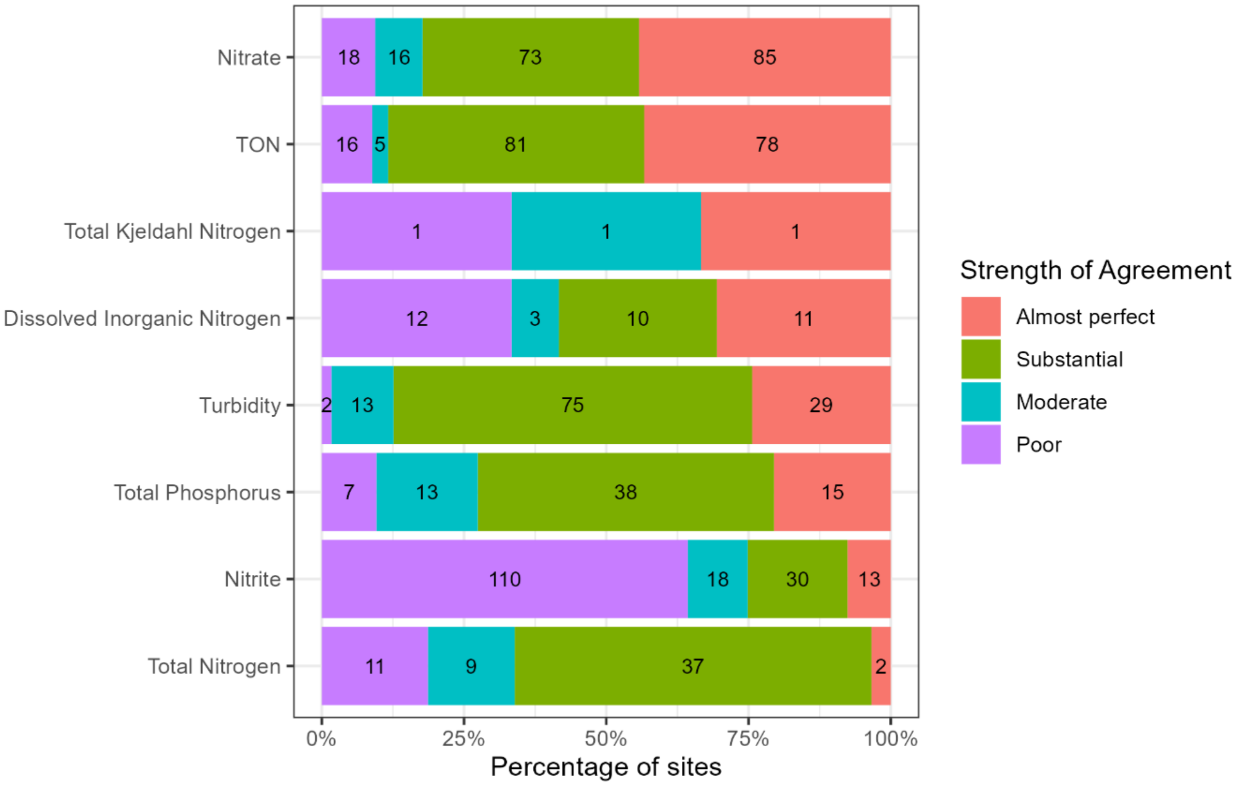


Figure 4-2: Strength of agreement based on Lin's CCC for eight analytes for multiple sites. The numbers refer to site counts and differ from those in Table 4-1 as only sites with nine or more non-censored results were included in the results, and all pairs with censored values were removed from the calculations.

Bias for each site and analyte combination was estimated using the log-transformed Bland-Altman approach (see Appendix B for a worked example), and the summary statistics of the results are presented in Table 4-1. There are examples of analytes where the new method tends to give lower results than the old method, for example, total oxidised nitrogen, where the value of the new method is 93.4% of the old method. There are examples of new methods resulting in higher values, such as turbidity, where the new method (infrared turbidity meter following ISO7027:1999) on average (mean) gives higher values than the original method (APHA 21st Edition Method 2130 B) by approximately 1.177 times.

Table 4-1: Estimated bias as a ratio of the New/Old method. Values > one imply that, on average, the new method gives higher values than the old method, and a value < one indicates that the old method tended to yield higher values. These values are unitless as they are ratios.

Water quality variables	Mean	Median	5th percentile	95th percentile	Site count (n)
Total Oxidised Nitrogen (TON)	0.934	0.966	0.741	1.025	205
Nitrate	0.941	0.964	0.793	1.028	244
Total phosphorus	0.943	0.951	0.806	1.029	78
Dissolved Inorganic Nitrogen (DIN)	0.985	0.991	0.888	1.014	43
Nitrite	1.017	1.019	0.729	1.27	215
Total nitrogen	1.051	1.044	0.949	1.183	94
Total Kjeldahl Nitrogen (TKN)*	1.104	1.076	1.007	1.222	3
Turbidity	1.177	1.169	1.067	1.329	141

*Note that a small sample size may make TKN values unsuitable for comparisons between sites.

Focusing on turbidity, Figure 4-3 shows the level of agreement across more than 100 monitoring sites, using the Bland-Altman approach for estimating bias and the level of agreement. The bias is expressed as a ratio of the new to the old value (a relative rather than absolute difference between the methods), indicated by the black dot for each site. The error bars indicate the level of agreement and provide an estimate of the random error or difference between paired observations. There is a range of values, with the bias (black dots) at all but one site being greater than one. However, the individual values, indicated by the error bars on the limits of agreement, cross over the value of one, illustrating that it is possible for individual pairs of observations for the new method to have lower values than the old method.

Figure 4-3 illustrates the systematic (bias) and random differences (upper and lower levels of agreement). In an ideal world, the random differences would be very small so we could precisely estimate the systematic differences between the two methods. The larger the random difference, the harder it is to estimate the bias accurately. Given the random differences are properties of the two methods and cannot be reduced, the only way to increase the accuracy of the bias estimate is to increase the sample size. The current NEMS recommendation is to use a minimum sample size of 12, though larger sample sizes are commonly used in method comparison studies. Some councils have taken many more pairwise samples, over 100 in some sites, though discussions with councils suggest that taking such a large number of pairwise samples is not practical in many situations.

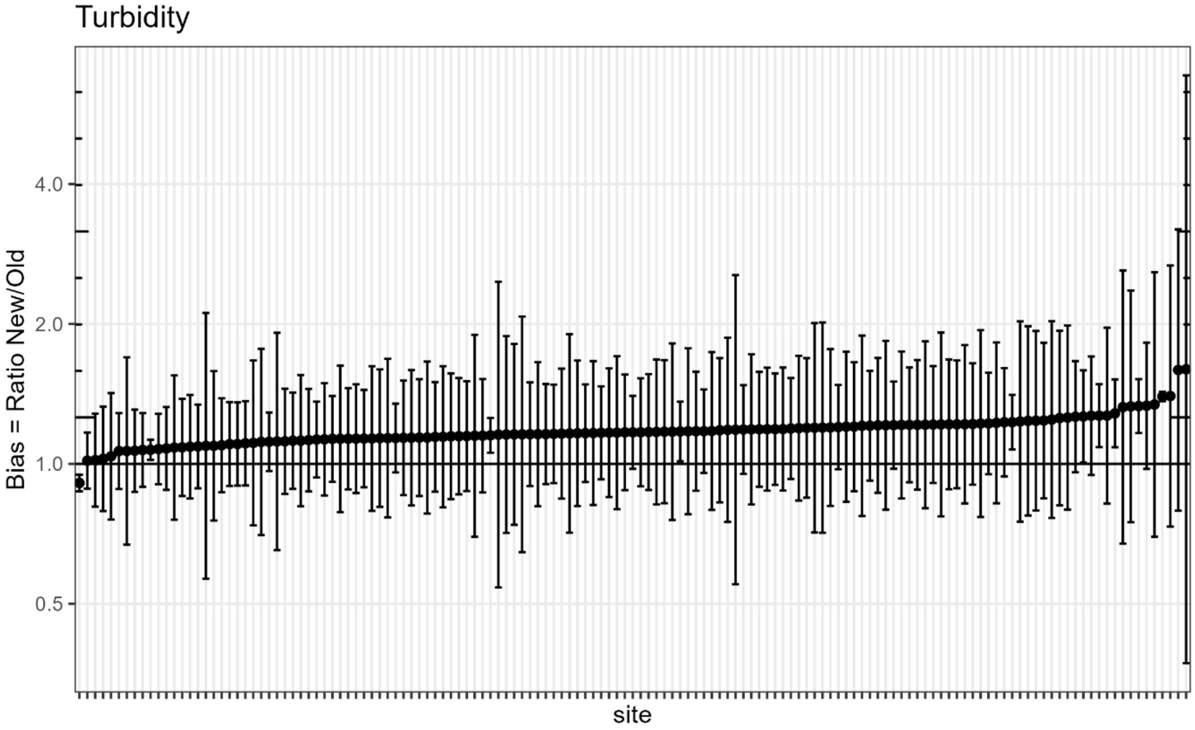


Figure 4-3: Bias and levels of agreement for turbidity. Turbidity from 141 sites. There is a tendency for the bias (indicated by dots) to be greater than one, implying new results have a tendency to have a higher value than the original methods. The bars indicate the level of agreement from a Bland and Altman analysis. Note that censored values have been removed from this analysis.

All the turbidity data are pooled in Figure 4-4, ignoring the between-site variability. Overall, Lin's Concordance Correlation Coefficient has a value of 0.991, which is regarded as almost perfect based on the descriptions provided in Table 3-1. However, the Deming Regression line on the left diagram and Bland and Altman's bias estimate on the right indicate a systematic difference between the two methods. The new method tends to result in 17% higher turbidity estimates than the original method (which is close to the mean and median bias of the individual sites). The p-value for the trend in bias (trend bias) has a value of <0.001, evidence that bias varied with the turbidity values, as illustrated by the fact that the slope in the regression line has a value other than one.

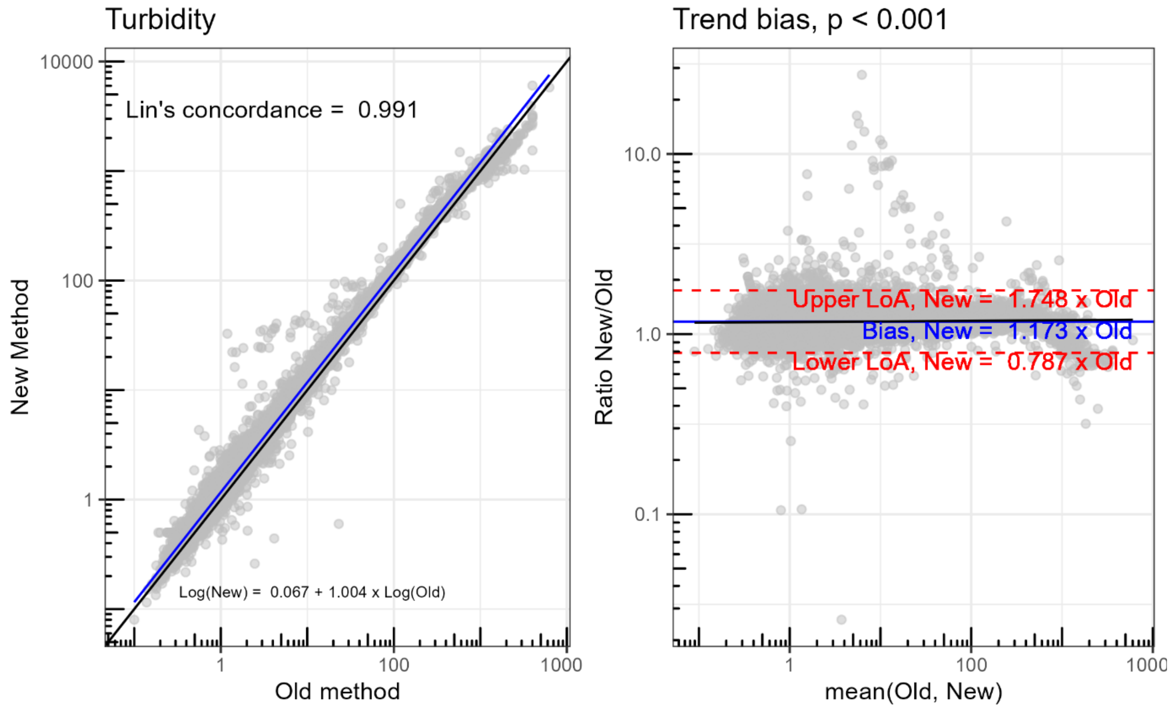


Figure 4-4: Bland and Altman's approach to presenting levels of agreement between methods for all turbidity data, augmented with regression and correlation estimates. In the left-hand diagram, the black line is the line of equivalence, and the blue line is the Deming regression line, with the equation presented on the lower half of the graph and Lin's CCC in the upper half, clearly illustrating the tendency for the new method to result in higher values than the original method. The right-hand diagram illustrates the level of agreement, the blue line (and equation) estimates the bias, and the red lines and equations show the limits of agreement. The trend bias figure tests the hypothesis that the bias is independent of the average values of the observations. In this case, bias is not independent of the observed value of turbidity at multiple sites.

Graphs with the pooled results, using the Bland and Altman approach, Lin's concordance and Deming regression, showing the bias and level of agreement for other nutrients, can be found in Appendix C.

5 Recommended approach

The general approach is laid out in Figure 5-1. In an ideal situation, the recommendation is to use site-specific parallel testing using the old and new methods, assessing the level of agreement, interpreting the results, and state and trend assessment. Along with the specifics of the approach, there is also a bigger picture of managing change, which includes collaboration between councils and laboratories, collecting information to allow a comparative analysis of changes from elsewhere, and using good data management practices.

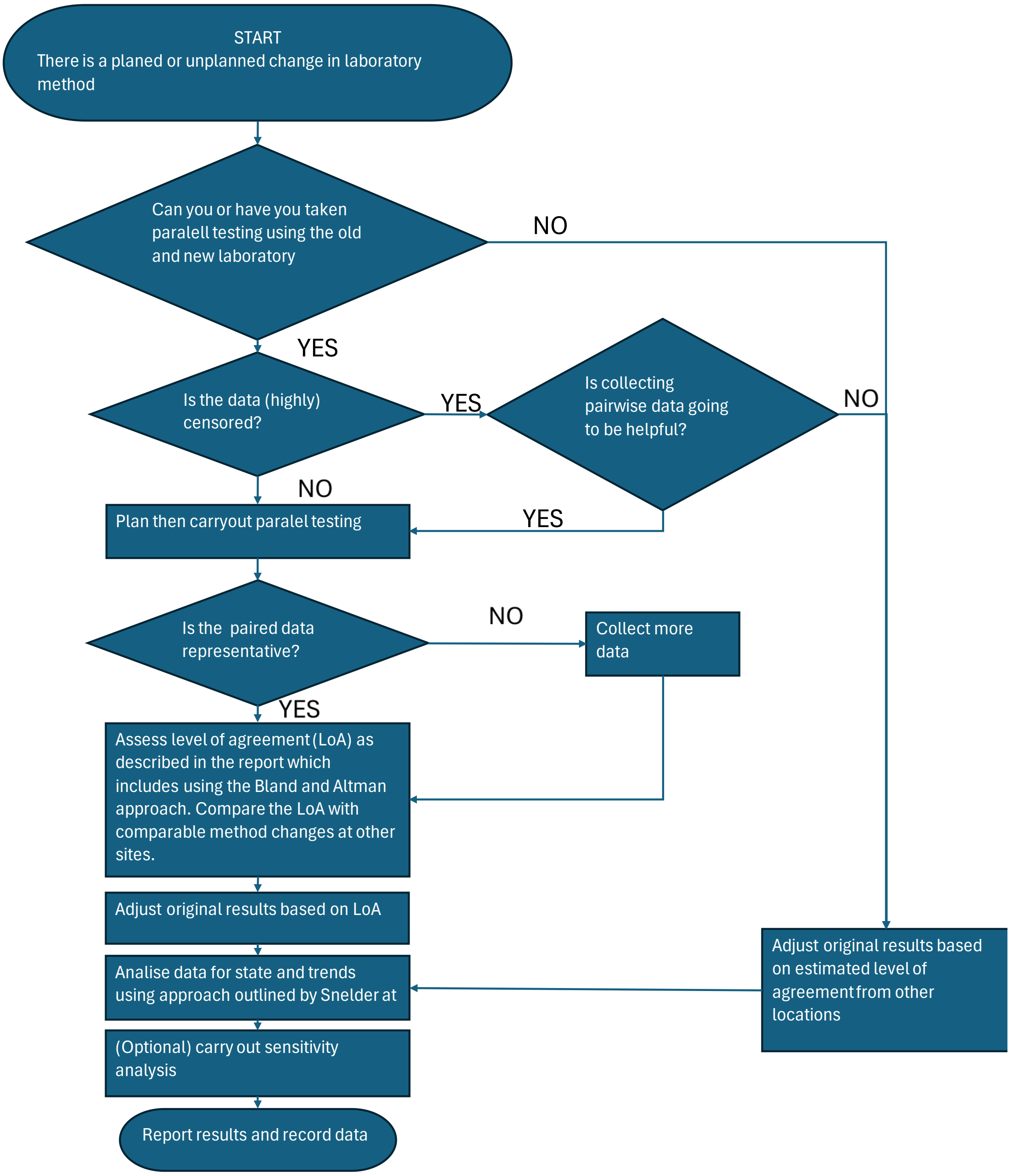

START
There is a planed or unplanned change in laboratory method
Can you or have you taken paralell testing using the old and new laboratory
NO
YES
Is the data (highly) censored?
YES
Is collecting pairwise data going to be helpful?
NO
NO
Plan then carryout paralel testing
YES
Is the paired data representative?
NO
Collect more data
YES
Assess level of agreement (LoA) as described in the report which includes using the Bland and Altman approach. Compare the LoA with comparable method changes at other sites.
Adjust original results based on LoA
Adjust original results based on estimated level of agreement from other locations
Analise data for state and trends using approach outlined by Snelder at
(Optional) carry out sensitivity analysis
Report results and record data

Figure 5-1: General process of assessing the levels of agreement (LoA) and accounting for the influence of changes in laboratory measurement methods.

However, there may be reasons to deviate from the recommended approach, possibly due to it being impossible or impracticable to carryout parallel testing. In other situations, notably when data is highly censored, it may be possible to carry out parallel testing, but the resulting paired sample data will be of limited value as we do not have techniques to assess the level of agreement when data is censored. In that case, it may be reasonable not to collect paired data. See Section 5.3 for specific guidance on what to do in the absence of paired data.

5.1 Parallel testing and paired data

In accordance with the NEMS (2019b), we strongly recommend a period of ***parallel testing,*** where laboratory analysis using old and new methods is performed in parallel for a minimum of 12 months. However, we recognise that parallel testing is not always feasible. The outputs of this parallel testing exercise will be paired data for use in estimating the level of agreement between two methods, benchmarking against other sites and the basis for making adjustments.

Parallel testing should take place at every site where method changes occur to ensure the resulting paired data is site-specific. This parallel testing adds to the body of knowledge about method changes and highlights site-specific differences that, if necessary, can be further investigated.

Every effort should be made to ensure that each pair of laboratory measurements using the old and new methods uses the same water taken from the same location at the same time. Ideally, parallel testing for method comparison studies involves collecting and splitting a water quality sample for duplicate laboratory analyses using the old and new methods.

In addition to how the individual water samples are collected and split, drawing valid conclusions also requires that the set of pairs of results are drawn from a "representative sample" of water quality. However, the NEMS recommends convenience sampling[8], taking a minimum of 12 monthly samples, which may or may not represent water quality in the long term. Therefore, we recommend comparing parallel testing results ***against historical data*** to check that, as a minimum, the range of the pairwise sample results covers the interquartile range of the historical data. If it does not, then it suggests that the paired data may not be representative of the preceding period, and it calls into question the validity of the results in accounting for differences between methods. If that is the case, further parallel testing is recommended.

As previously mentioned, site-specific parallel testing is recommended. However, there may be situations where this is not possible. One alternative approach could be sampling from a limited number of sites that exhibit a wide range of concentrations of the water quality variable being studied rather than sampling from all sites. This may involve sampling under different conditions, including high and low flow, and taking into account other factors that may affect the behaviour of the new method, such as different types of water sources (marine, surface, and groundwater) and other contaminant interference effects such as turbidity.

5.2 Assess the level of agreement and interpretation of results

The goal of parallel testing is to assess the level of agreement between the two methods. To an extent, the calculations are a mechanical exercise. As a minimum, we recommend using the Bland and Altman approach, which provides a simple visual and numerical summary of the differences and

[8] Convenience sampling is a non-probabilistic approach; observations are selected based on their availability and may not be representative of the entire population and can introduce bias. In this case we take the samples from one year and assume they are representative of the rest of time.

similarities between the two methods. However, at times, it may be helpful to augment the Bland and Altman approach with Deming regression, particularly if there is evidence of a trend in the bias.

The resulting graphs should be critically examined and compared against pairwise samples from other sites, looking for similarities and differences between pairs of methods. We do not provide any rule as to what course of action to take if the results differ from those found at other sites: this would be a judgment call and needs to take into account contextual data. The resulting findings and any decisions made should be noted with the reasoning as to why.

5.3 State and trend assessment

Caution must be exercised when interpreting states and trends in the presence of method changes, especially when parallel sampling results are unavailable. We recommend that ***data for analysing state and trends be routinely adjusted in the presence of method changes***. Method changes are treated as a confounding variable, and this adjustment should be done prior to adjustments for other confounders, such as seasonality and flow, assuming that the *Guidance for the analysis of temporal trends in environmental data* (Snelder et al. 2021) is being followed.

Site-specific adjustments cannot be used when parallel testing has not been carried out. In the absence of parallel sampling at a site, it is recommended that estimates from elsewhere are used rather than assuming there is no difference between the old and new methods.

Given that there is uncertainty on any bias assessment (the average difference between methods), it may be appropriate to conduct ***a sensitivity analysis to work through the impact of method change on estimates of state and trend***. This is an optional, though potentially useful activity. A sensitivity analysis systematically explores how a state or trend assessment results vary with the size of the bias. It is best done with a question in mind, such as how big the bias would need to be to change my assessment of state or trend? Then, you can compare this bias with comparable method changes from elsewhere to see if method changes could play an important part in the state or trend assessment. The exact approach will depend on the specifics of the state and trend assessment and the expertise and resources available, so the worked example does not cover it.

In some cases, it may not be appropriate or necessary to make adjustments. This includes situations when the paired data are highly censored, or there is evidence that the method used produces incompatible results for site-specific reasons, or there is no requirement to assess state and trend across the period of the method change. The techniques available for evaluating levels of agreement assume that the results are measured on a continuous scale. Evaluating the agreement levels becomes difficult and impracticable if the pairwise results are highly censored. If, on the other hand, the results are not censored but appear quite different from those found elsewhere despite no obvious causative factors, then it should cause us to question the results, and we might judge that it is impossible, in this case, to adjust for method changes.

Though we recommend adjusting for method change, informed judgment about when to make such an adjustment may be required. These judgments must consider technical factors, potential issues and the practical consequences of such decisions. Twelve months of data points provide only a weak evidence base for making an adjustment. Although twelve months of data points may provide a weak evidence base for making an adjustment, ignoring evidence (even though tenuous) may lead to a less technically justifiable outcome.

One of the outstanding issues with making adjustments is that we do not yet have enough evidence to know whether making adjustments for method changes will make practical differences to many water quality state and trends estimates. The impact will become clear once analysts start to account for method changes seriously. In the meantime, we recommend that data be routinely adjusted.

5.4 Bigger picture/context

Managing method changes well requires coordinated efforts from those who plan the change, collect samples, conduct laboratory analyses, and record, analyse, and use the data.

Should a method change be required, it is recommended that:

- In the case of laboratory-instigated method changes, the laboratory communicates the change as early as possible to its council clients.
- In the case of council-instigated changes, such as changes to the required laboratory- or sample collection method, the council communicates the changes to the laboratories as soon as possible.
- Information on changes to an analyte’s limits of detection (LoD) and results from previous parallel sampling exercises for that analyte should be gathered so everyone is aware of what level of systematic (bias) and random differences (upper and lower levels of agreement) might be expected.
- All water quality monitoring process changes should be documented with metadata to assist current and future data analysis. All changes are to be recorded, including but not limited to changes in methods, laboratory service providers, and field sampling protocols.
- In addition, documentation must be kept on the results of assessing the levels of agreement and any adjustments made. High-quality documentation and metadata will help stakeholders evaluate the results of state and trend assessments.

6 Worked examples

In this section, we provide examples, focussing specifically on the main components for assessing and accounting for method changes:

- Evaluating if the pairwise dataset is representative.
- Assess the level of agreement between the old and new approaches and evaluate the results.
- Adjust data as required before assessing state and trend.

Managing laboratory method changes should be covered by the Field and Office Manual or equivalent (NEMS 2019b), which lays out Standard Operating Procedures (SOP). Before implementing any changes to laboratory methods, developing a comprehensive set of SOPs is essential. These procedures should cover all aspects of sampling, measuring and processing data to ensure that any variation between the old and new laboratory methods is controlled for as much as possible. To know what should be covered, refer to NEMS (2019b). Stakeholders need to be aware of what is changing and why. The SOP should be trialled to ensure it works in practice.

Appropriate **data management** must be in place so that both raw and adjusted results are accessible and distinguishable and the adjustment methods or processes are clearly documented. These requirements are particularly important in longer-term monitoring because personnel collecting and using these data will change over time.

This worked example assumes that parallel sampling is possible and that the results will allow the level of agreement to be assessed. Parallel sampling may not always be possible. For example, if the original analytical method suddenly becomes unavailable due to an instrument failure, there is no practicable way to collect pairwise data.

On the other hand, pairwise analysis may be possible, but the results may not be very informative – this situation may exist when most observations obtained using the original method were reported as below the limit of detection, so the bulk of results would be censored. Should parallel sampling not enable the level of agreement to be evaluated, it may be decided to dispense with parallel sampling and follow the advice in Section 5.3.

6.1 Evaluating if the pairwise dataset is representative

NEMS recommends monthly parallel testing for a minimum of 12 months, which may or may not result in a dataset representative of the long-term variation in the water quality variable of interest. We will have more confidence in any assessment of state or trend if the dataset used to inform a method change adjustment is representative of the state and trend being assessed.

Exploratory data analysis (EDA) should be carried out on each monitoring site's paired and historical data. Plotting the data is an excellent way to explore the data. Unusual values, such as outliers, should be flagged, and expert judgment should be used in deciding how they are dealt with. The key question the EDA seeks to answer is if the paired data are representative. If they are representative, we have some assurance that the result can be generalised to future observations. If they are not representative, then a decision needs to be made to either:

- collect more data to make the sample representative or

- accept that the sample may not be representative and that we have less confidence in any subsequent state and trend analysis.

Given the importance of state and trend analysis, we recommend that every effort be made to try to ensure that the pairwise dataset is representative.

Rather than proposing formal statistical tests of representativeness, we suggest a simple rule of thumb. The expectation is that the range of paired data using the original method should, as a minimum, cover the interquartile range of the historical data (ideally collected over five years or more). This approach guarantees that the pairwise sample will cover a minimum of 50% of the range in observed data, but there will be situations where this approach is inadequate, and it requires the application of expert judgment to decide if this test is reasonable for any given situation.

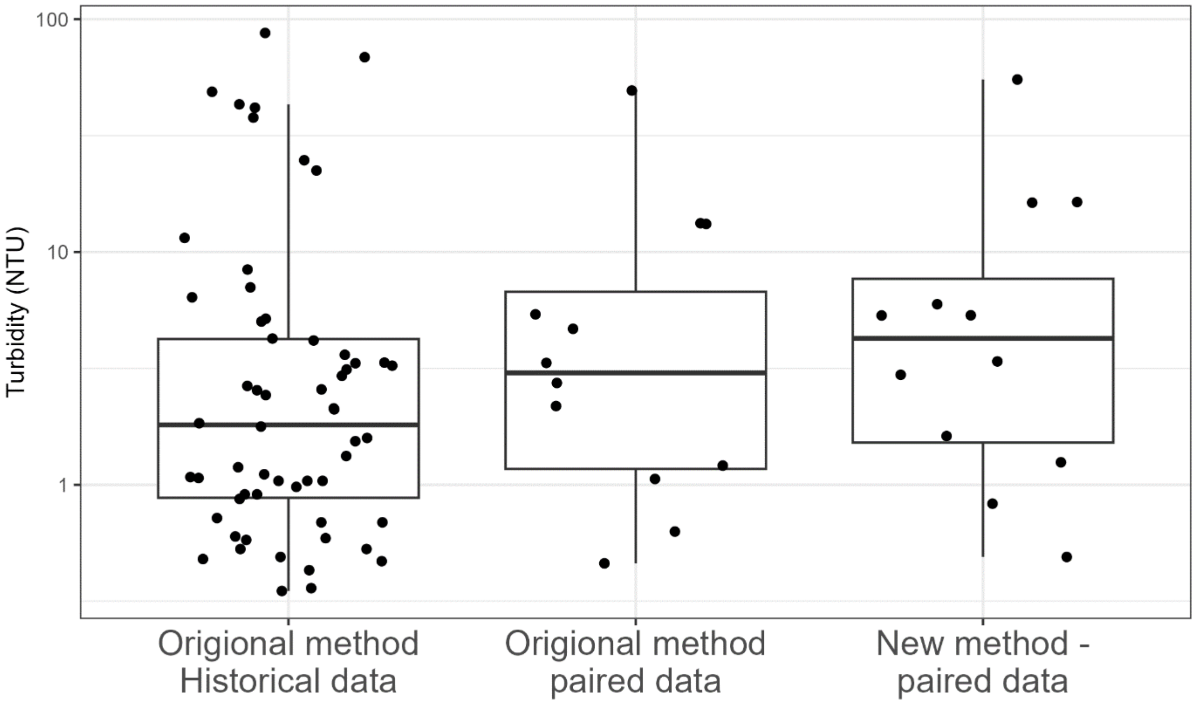


Figure 6-1: Are the paired turbidity data representative? The range of the original method paired data (middle boxplot) covers the interquartile range of the original method historical data (left-hand boxplot). The interquartile range (IQR) of the original method using historical data is represented by the box in the left boxplot, whereas the range is the distance covered between the minimum and maximum values.

Figure 6-2 shows a situation where the pairwise data are not representative of the historical data. The paired data using the original method does not cover the interquartile range of the historical data. Paired data using the original methods appears to be systematically lower than the historical data. Obviously, this could be due to a trend and values changing over time, but at the very least, it is a red flag, and further data collection and investigation should be considered.

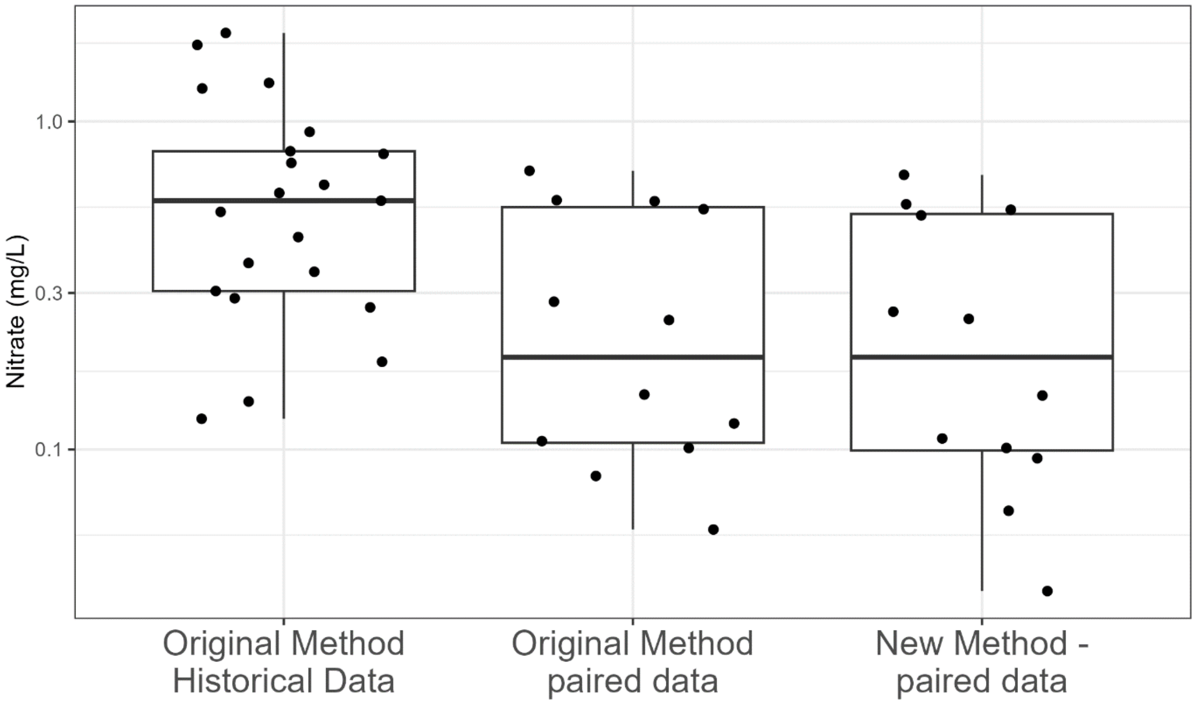


Figure 6-2: Are the paired nitrate data representative? The range of the original method paired data (middle boxplot) should cover the interquartile range of the original method historical data (left-hand boxplot). In this example, it does not; therefore, we suggest the data are not representative. The box represents the interquartile range of the data, and the points are the actual observations.

6.2 Assess the level of agreement between the old and new approach and evaluation of results

Once we are satisfied that the data are representative, such as in Figure 6-1, we follow the Bland and Altman approach to assess the similarity of methods. We recommend excluding censored values from the regression and bias value calculations.

Figure 6-3 illustrates a substantial level of concordance between the two methods (0.982), but there is a systematic difference between them (i.e., bias). On average, the Bland and Altman approach indicates that the new method values are 1.237 times greater than those of the old method, and the regression line's intercept gives a similar value of $10^{0.097}$ = 1.25. Based on the estimated values, there is little to choose between the regression and bias estimated using the Bland-Altman approach, so we recommend the Bland and Altman estimates as they are simpler to calculate. As part of the process, we should compare the estimates against results from other sites, such as those given in Table 4-1; should this comparison confirm that the estimated bias is in the range observed elsewhere and is within the 5th to 95th percentile range of values from other sites, then it gives us some confidence that the estimated level of agreement may be used to calculate state and trends.

Appendix B provides a worked example of how bias and trend bias were calculated in Figure 6-3. The calculation of Lin’s concordance and Deming regression is somewhat more complicated, and these values were estimated with *R* (R Core Team 2022) using the *mcreg* (Deming method comparison) function from the package MCR (Potapov et al. 2023) and the *epi.ccc* (Lin's concordance correlation coefficient) function from the epiR package (Stevenson and Sergeant 2023). Example code is also given in Appendix B.

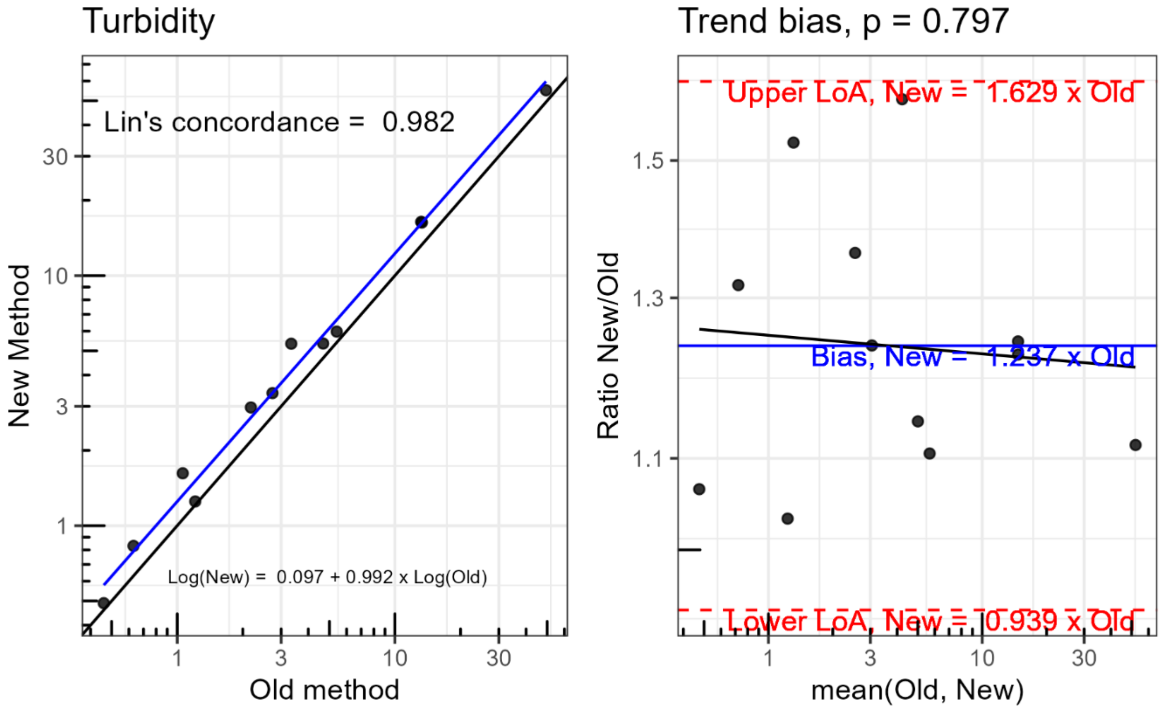


Figure 6-3: Bland and Altman Approach for assessing the level of agreement for turbidity at one site. The figure includes Lin's CCC and Deming regression on the left graph and p-value for assessing the trend in the right-hand graph; note p = 0.797, suggesting that there is no evidence of a trend in the bias.

Plotting the data may reveal that bias varies as a function of the measured value, as shown in Figure 6-4. If there are no errors in the data, it would be appropriate to use the regression line, as bias appears to be a function of analyte concentration. However, some of the results at lower nitrate concentrations (0.01 g/m^3) do appear quite different from the other observations as they don't follow the same general trend as the other observations, so they may be outliers. Inspecting the right-hand figure provides further evidence that these values may be unusual as they are outside the upper and lower levels of agreement (LoA). The analyst is left to decide whether these unusual values should be included when assessing the level of agreement and which approach to use.

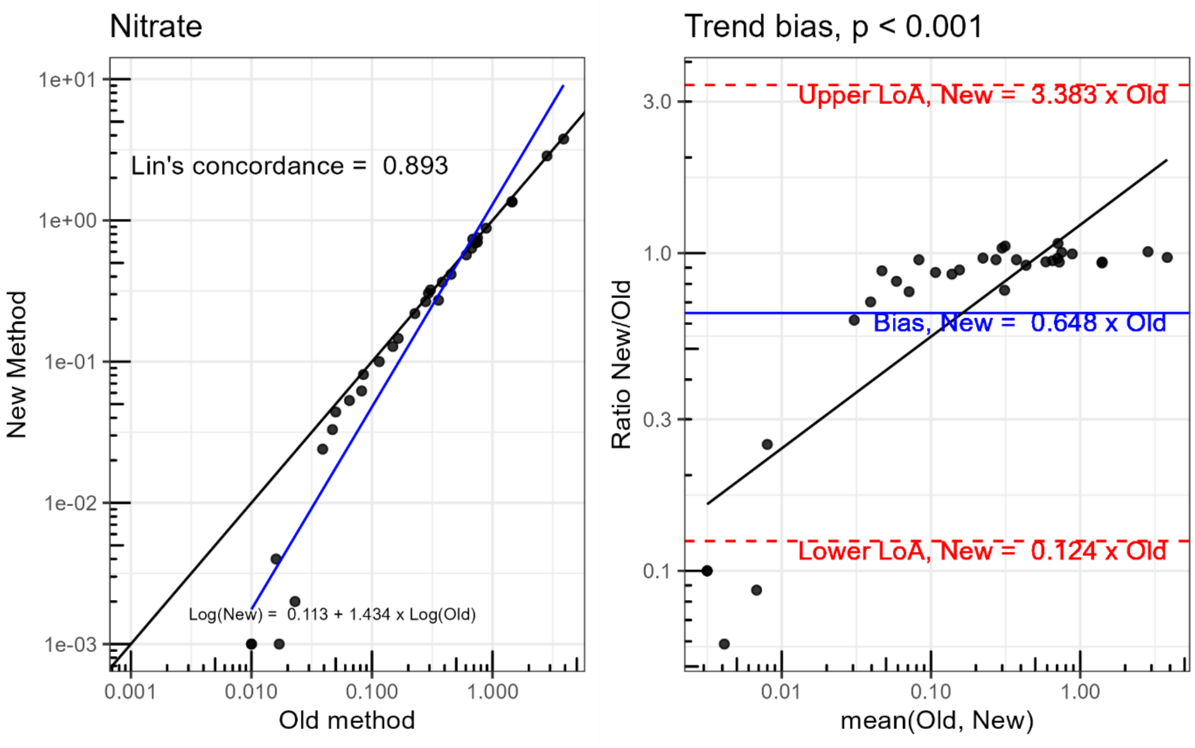


Figure 6-4: Bland Altman plot for nitrate, illustrating that bias may be a function of the concentration of the analyte. The regression line shows a deviation from the line of equivalence, and the trend bias has p <0.001. However, inspection of the data points may suggest anomalous observations (possibly outliers) at low concentrations.

At times, the regression approach can give completely misleading results. In the following case, the analyte variation has a limited range, and Lin's CCC is negative (Figure 6-5). The overall bias is estimated to be 0.998, which suggests that both methods give similar results in the range seen elsewhere on average. However, p = 0.014 for the trend bias provides some evidence that bias is not independent of analyte concentration. However, if the regression line were used to adjust data, it would imply that the two methods would give widely diverging and unrealistic results. It is an example of what is known in machine learning as overfitting.

In the type of situation illustrated in Figure 6-5, where there is a very limited range of values, it may be that the pairwise samples are not representative, and more data may be required. Alternatively, variation in water quality variables is limited in very stable environments, such as in some deep groundwaters. Regression estimates of bias in these situations are not particularly helpful, and the Bland-Altman approach is to be preferred.

As measured values of water quality variables approach a detection limit, estimates of bias become less precise. There also comes a point when it is not possible to estimate bias. Method comparison approaches, Bland and Altman and regression analysis are not designed to be used with censored data. This has implications for reliably estimating state and trend in the presence of method changes.

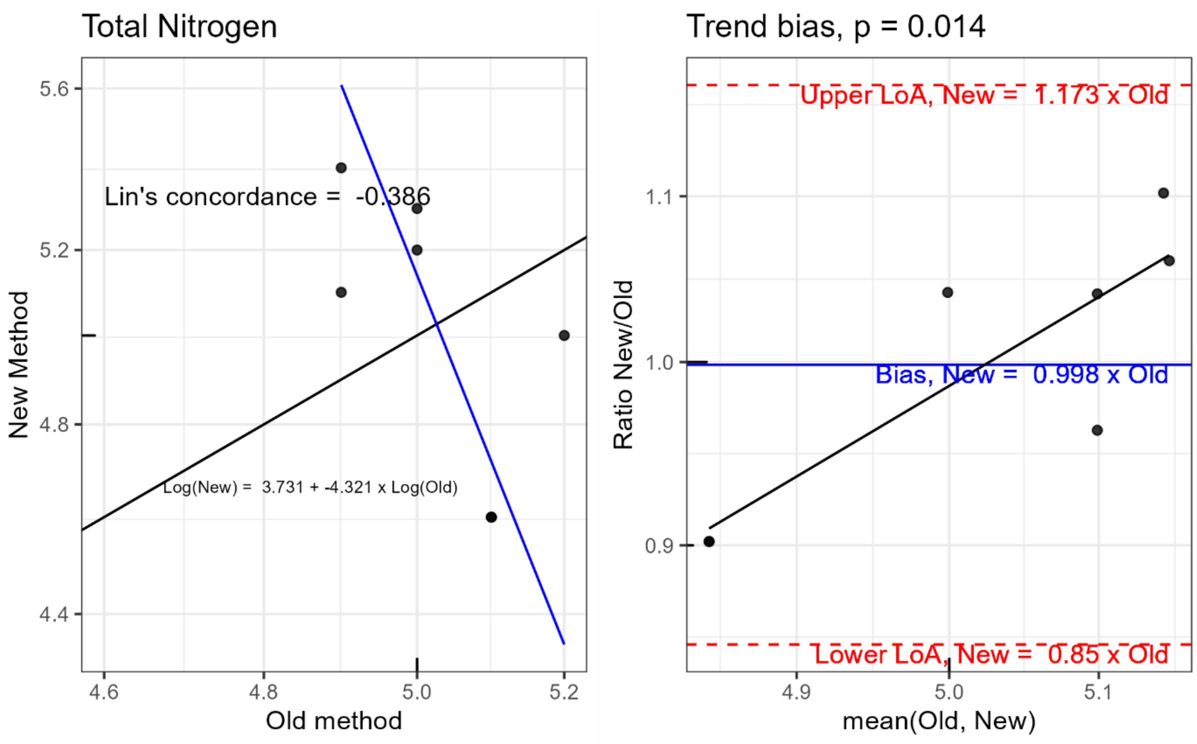


Figure 6-5: Regression analysis gives misleading results. Negative Lin's CCC (-0.386) and the blue regression line on the left-hand diagram indicate an inverse relationship between the two methods. An inverse relationship suggests that as the results of one method increase, those from the other decrease. This could be a red flag indicating that the two methods give very different results, which could result from the use of a very limited range of values and/or low measurement precision. In this example, the range of values is certainly limited.

Judgment must be exercised when deciding which level of agreement approach to use. Regression approaches have the advantage of being more flexible and can account for bias varying with measured values. However, Figure 6-4 and Figure 6-5 illustrate the problem that they can overfit, particularly when small datasets are used. The Bland and Altman approach is simpler and not as susceptible to over-fitting. Inspecting the graphical and numerical evidence will assist analysts in deciding which approaches are more appropriate for the situation.

Adjustment is impractical when many of the data are censored. High levels of censoring may make it impossible to estimate the level of agreement between methods. In this situation, all that can be done is to censor the data to the highest value in the dataset.

6.3 Make adjustments to data as required before assessing state and trend

The above analysis is not intended to describe the differences and similarities between the two methods but to provide evidence regarding data adjustment. This will ensure that method changes are accounted for when estimating environmental states and trends.

The recommended approach is to adjust the old values to align with the new ones and use the resulting estimates to assess the state and trends. This should be done within the state or trend analysis using a modified ephemeral data file rather than making a global change to the raw data, which would result in a NEMS code QC 300 (synthetic data). We do not go through the details of state and trend assessment – these processes are fully described in *Guidance for the analysis of temporal trends in environmental data* (Snelder et al. 2021).

The equations in the Bland and Altman graphs presented above provide sufficient information to make the appropriate adjustments. The blue numbers on the right-hand graph of Figure 6-5 are the output of the Bland and Altman approach, and the equation in black for the left-hand figure is the Deming approach.

We recommend using the equation for bias on the right-hand graphs, the Bland and Altman approach (see Appendix B for the mathematical details), unless you have evidence that bias is better accounted for by regression equations, as shown in the left-hand diagram.

Before estimating state and trends, the data should be censored to the highest censored value (as necessary), and corrections should be made for numerical precision. Correcting for numerical precision is essential, especially after data adjustment, as the adjustment may result in values that overstate the data's precision. Non-parametric trend assessment methods are sensitive to the ranking of values, and overstating the numerical precision can influence ranking.

State or trend can then be estimated. We recommend using site-specific data to make the adjustments; however, this is not always possible, and information from other sites must be used.

Whether adjustment makes a practical difference depends on context, what question is being asked of the data, how the data is to be used, and the magnitude of the bias. The pragmatic way to evaluate the impact of method changes is to analyse the state and trend with and without bias correction, a form of sensitivity analysis.

We suggest sensitivity analysis as an optional step in the analysis of the state and the trends to consider how sensitive these assessments are to bias. This could be as simple as estimating the state and trend of data that is and is not adjusted for bias and comparing the results. Both the adjusted and non-adjusted data should be censored to the highest censored value, and corrections should be made for numerical precision to ensure a fair comparison.

In some situations, it may be appropriate to carry out a more comprehensive sensitivity analysis, using the range of observed bias from other sites or asking questions of the data, such as how big the bias would need to be to result in a change of trend.

This type of sensitivity analysis is important as having only twelve months of site-specific paired data points is a weak evidence base for adjusting state or trend data, but ignoring the evidence is an even weaker position. Exploring how sensitive any assessment of state and trend is to bias can build confidence or highlight concerns.

7 Conclusions

In an ideal world, laboratory measurement methods would remain the same. As Oakley et al. (2003) noted, "Designing a monitoring project is like getting a tattoo: you want to get it right the first time because making major changes later can be messy and painful...". Laboratory method changes are inevitable and, to a certain extent, unavoidable due to changes in technology, equipment becoming obsolete, requirements for standardisation and the evolution of monitoring practices. However, it is important to note that changing methods has potential benefits. The guidance set out in this document is designed to help realise those benefits and reduce some of the pain arising from method changes.

The preceding guidance is best described as emerging practice. It is based on evidence from multiple sources but has yet to be extensively tested in the water quality domain. So, with time and experience, this guidance may need updating. For example, if we were to move away from the non-parametric test of trend, this guidance would need to be updated.

The guidance does not prescribe hard or absolute limits on what constitutes equivalence between methods and what does not. Instead, it provides an approach for assessing and accounting for the influence of changes in laboratory measurement methods, which can be used to help interpret long-term time-series data. A key component of the approach is the use of parallel sampling. There is a recognition that there will be instances when parallel sampling is not possible, or if it is possible, may not be informative due to high levels of censoring, so it may be omitted. Evaluating the similarities and differences between the two methods is vital to building confidence in any water quality data analysis in the presence of method changes, even if there is a decision not to adjust method changes in any subsequent analysis of state or trend.

8 Acknowledgements

We thank Ton Snelder, Sandy Haidekker, and Shirley Hayward for their feedback on the draft report. In addition, we thank Roger Young, Luke Fullard, Katie Blakemore, Emily Gray, Juliet Milne, and Sam French for their thoughts about method changes and Horizons, ECan, and Environment South councils for providing pairwise data.

9 Glossary of abbreviations and terms

Attribute

An attribute is something we can measure and monitor that tells us about the state of a river or lake. The term is used specifically in the National Policy Statement for Freshwater Management (NPS-FM).

Bias

Bias is the systematic difference in results between the two methods. In this report, this is the average ratio of the ratio of two methods rather than the average absolute differences between the methods.

Bland and Altman

The Bland-Altman method assesses agreement between two measurement methods by plotting the difference between their measurements against the average of the measurements. It provides insights into bias and variability between methods, helping to identify systematic errors and acceptable limits of agreement in a simple visual representation.

Emerging practice

There is a growing body of evidence that this practice works, but there is insufficient evidence and or agreement that this is the best practice.

Limit of Detection (LoD)

The lowest quantity (often a concentration) that can be measured within a stated confidence limit.

Line of equality

The line of equality represents the line where the difference between two measurement methods is zero, so it is where the two methods give exactly the same results.

NEMS

The National Environmental Monitoring Standards (NEMS) are a series of environmental monitoring standards prepared by the NEMS steering group on authority from the Regional Chief Executive Officers (RCEOs) and the Ministry for the Environment (MfE).

Parallel testing

Taking a sample and ideally spitting it into two and taking measurements using the old and new laboratory methods.

Precision

The closeness of agreement between repeated independent measurements of the same quantity under unchanged conditions. Depicted by the random error in the results of repetitions of the same test performed on the same sample.

Representative sample

A sample from the population which represents the characteristics of the entire population.

Sensitivity analysis

A systematic exploration of how the state or trend changes in response to estimates of bias.

Trend Bias

Trend Bias is the p-value of the slope of the estimated bias with respect to the average or measurements, and it provides evidence that the slope differs from zero. If the p-value is less than a specified value, usually 0.05, it suggests bias varies with the level of the observed values.

10 References

ANZG (2018) Laboratory Analysis. https://www.waterquality.gov.au/anz-guidelines/monitoring/laboratory-analysis

Bland, J.M., Altman, D.G. (1986) Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet*, 1(8476): 307-310. 10.1016/s0140-6736(86)90837-8

Bland, J.M., Altman, D.G. (1999) Measuring Agreement in Method Comparison Studies. *Statistical Methods in Medical Research*, 8(2): 135-160. 10.1177/096228029900800204

Bland, J.M., Altman, D.G. (2010) Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *International Journal of Nursing Studies*, 47(8): 931-936. 10.1016/j.ijnurstu.2009.10.001

Chanat, J.G., Moyer, D.L., Blomquist, J.D., Hyer, K.E., Langland, M.J. (2016) Application of a Weighted Regression Model for Reporting Nutrient and Sediment Concentrations, Fluxes, and Trends in Concentration and Flux for the Chesapeake Bay Nontidal Water-Quality Monitoring Network, Results through Water Year 2012.

Coats, R., Lewis, J., Alvarez, N., Arneson, P. (2016) Temporal and Spatial Trends in Nutrient and Sediment Loading to Lake Tahoe, California-Nevada, USA. *JAWRA Journal of the American Water Resources Association*, 52(6): 1347-1365. https://doi.org/10.1111/1752-1688.12461

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*, HAM2016-070: 34.

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(3): 437-450.

Domagalski, J.L., Morway, E., Alvarez, N.L., Hutchins, J., Rosen, M.R., Coats, R. (2021) Trends in Nitrogen, Phosphorus, and Sediment Concentrations and Loads in Streams Draining to Lake Tahoe, California, Nevada, USA. *The Science of the total environment*, 752: 141815-141815. 10.1016/j.scitotenv.2020.141815

Francq, B.G., Govaerts, B.B. (2014) Measurement Methods Comparison with Errors-in-Variables Regressions. From Horizontal to Vertical OLS Regression, Review and New Perspectives. *Chemometrics and Intelligent Laboratory Systems*, 134: 123-139. https://doi.org/10.1016/j.chemolab.2014.03.006

Graham McBride, N., Till, H.D., Andrew Ball, E., Lewis, C.D.G. (2002) Freshwater Microbiology Research Programme.

Gronewold, A.D., Borsuk, M.E. (2010) Improving Water Quality Assessments through a Hierarchical Bayesian Analysis of Variability. *Environmental science & technology*, 44(20): 7858-7864. 10.1021/es100657p

Gronewold, A.D., Wolpert, R.L. (2008) Modeling the Relationship between Most Probable Number (Mpn) and Colony-Forming Unit (Cfu) Estimates of Fecal Coliform Concentration. *Water Research*, 42(13): 3327-3334. https://doi.org/10.1016/j.watres.2008.04.011

Helsel, D.R. (2011) Three Approaches for Censored Data. *Statistics for Censored Environmental Data Using Minitab and R*. John Wiley & Sons.

Hunter, S., Fullard, L., Feck, A. (2022) Analysis of Paired Sampling Results in the Horizons Region Implications of Moving to the NEMS Methodology for Nitrogen Sampling, Horizons Regional Council report

Jensen, A.L., Kjelgaard-Hansen, M. (2006) Method Comparison in the Clinical Laboratory. *Veterinary Clinical Pathology*, 35(3): 276-286. 10.1111/j.1939-165X.2006.tb00131.x

Johnson, R. (2008) Assessment of Bias with Emphasis on Method Comparison. *Clinical biochemist reviews*, 29 Suppl 1(Suppl 1): S37-S42. https://go.exlibris.link/ZLYzrlL9

Kilroy, C., Daly, O. (2020) Inter-Laboratory Comparison of Analysis of Chlorophyll *a* from Periphyton Samples, NIWA report number 2020321CH.

Lin, L.I. (1989) A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, 45(1): 255-268. 10.2307/2532051

Lindenmayer, D.B., Woinarski, J., Legge, S., Maron, M., Garnett, S.T., Lavery, T., Dielenberg, J., Wintle, B.A. (2022) Eight Things You Should Never Do in a Monitoring Program: An Australian Perspective. *Environmental Monitoring and Assessment*, 194(10): 701. 10.1007/s10661-022-10348-6

Linnet, K. (1999) Necessary Sample Size for Method Comparison Studies Based on Regression Analysis. *Clinical Chemistry*, 45(6): 882-894. 10.1093/clinchem/45.6.882

McBride, G.B. (2005) A Proposal for Strength-of-Agreement Criteria for Lin's Concordance Correlation Coefficient. *NIWA Client Report*, HAM2005-062: 10.

Meals, D.W., Spooner, J., Dressing, S.A., Harcum, J.B. (2011) Statistical Analysis for Monotonic Trends, Tech Notes 6,: 23. https://www.epa.gov/polluted-runoff-nonpoint-source-pollution/nonpoint-source-monitoringtechnical-notes

Ministry for the Environment (2023) National Policy Statement for Freshwater Management 2020 (February 2023). In: N.Z. Government (Ed), Wellington.

NEMS (2019a) Water Quality Part 1 of 4: Sampling, Measuring, Processing and Archiving of Discrete Groundwater Quality Data.

NEMS (2019b) Water Quality Part 2 of 4: Sampling, Measuring, Processing and Archiving of Discrete River Water Quality Data.

NEMS (2019c) Water Quality Part 3 of 4: Sampling, Measuring, Processing and Archiving of Discrete Lake Water Quality Data.

NEMS (2020) Water Quality Part 4 of 4: Sampling, Measuring, Processing and Archiving of Discrete Coastal Water Quality Data.

Newell, A.D., Blick, D.J., Hjort, R.C. (1993) Testing for Trends When There Are Changes in Methods. *Water, Air, and Soil Pollution*, 67: 457-468.

Newell, A.D., Morrison, M.L. (1993) Use of Overlap Studies to Evaluate Method Changes in Water Chemistry Protocols. *Water, Air, and Soil Pollution*, 67(3-4): 433-456. 10.1007/BF00478157

Oakley, K.L., Thomas, L.P., Fancy, S.G. (2003) Guidelines for Long-Term Monitoring Protocols. *Wildlife Society bulletin*, 31(4): 1000-1003.

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., De Cicco, L.A., Mills, T.J., Farmer, W.H. (2017) Appendix 4. - Step-Trend Analysis of Changes in Laboratory Analysis and Sample Collection Methods - Water-Quality Trends in the Nation’s Rivers and Streams, 1972–2012—Data Preparation, Statistical Methods, and Trend Results. *Scientific Investigations Report*: 158. 10.3133/sir20175006

Passing, H., Bablok, W. (1984) Comparison of Several Regression Procedures for Method Comparison Studies and Determination of Sample Sizes. Application of Linear Regression Procedures for Method Comparison Studies in Clinical Chemistry, Part Ii. *Journal of Clinical Chemistry and Clinical Biochemistry*, 22(6): 431. 10.1515/cclm.1984.22.6.431

Potapov, S., Model, F., Schuetzenmeister, A., Manuilova, E., Dufey, F., Raymaekers, J. (2023) Mcr: Method Comparison Regression. R Package Version 1.3.2. https://CRAN.R-project.org/package=mcr

Prats, J., Garcia-Armisen, T., Larrea, J., Servais, P. (2008) Comparison of Culture-Based Methods to Enumerate Escherichia Coli in Tropical and Temperate Freshwaters. *Letters in applied microbiology*, 46(2): 243-248. https://doi.org/10.1111/j.1472-765X.2007.02292.x

R Core Team (2022) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rhoades, D.A., Salinger, M.J. (1993) Adjustment of Temperature and Rainfall Records for Site Changes. *International journal of climatology*, 13(8): 899-913. 10.1002/joc.3370130807

Robson, A.J., Neal, C. (1997) Regional Water Quality of the River Tweed. *Science of The Total Environment*, 194: 173-192. 10.1016/S0048-9697(96)05363-6

Rubini, S., Galletti, G., Bolognesi, E., Bonilauri, P., Tamba, M., Savini, F., Serraino, A., Giacometti, F. (2023) Comparative Evaluation of Most Probable Number and Direct Plating Methods for Enumeration of *Escherichia Coli* in *Ruditapes Philippinarum*, and Effect on Classification of Production and Relaying Areas for Live Bivalve Molluscs. *Food Control*, 154: 110005. https://doi.org/10.1016/j.foodcont.2023.110005

Rus, D.L., Patton, C.J., Mueller, D.K., Crawford, C.G. (2013) Assessing Total Nitrogen in Surface-Water Samples--Precision and Bias of Analytical and Computational Methods.

Shapiro, J., Swain, E.B. (1983) Lessons from the Silica ``Decline'' in Lake Michigan. *Science (American Association for the Advancement of Science)*, 221(4609): 457-459. 10.1126/science.221.4609.457

Smith, D.G., McCann, P.B. (2000) Water Quality Trend Detection in the Presence of Changes in Analytical Laboratory Protocols. *Proceedings of the National Water Quality Monitoring Council Conference*.

Snelder, T., Fraser, C., Larned, S., Whitehead, A. (2021) Guidance for the Analysis of Temporal Trends in Environmental Data. *NIWA Client Report*, 2021017WN: 99.

Stevenson, M., Sergeant, E. (2023) Tools for the Analysis of Epidemiological Data. R Package Version 2.0.63. https://CRAN.R-project.org/package=epiR

von Bromssen, C., Folster, J., Futter, M., McEwan, K. (2018) Statistical Models for Evaluating Suspected Artefacts in Long-Term Environmental Monitoring Data. *Environmental Monitoring and Assessment*, 190(9). 10.1007/s10661-018-6900-3

Westgard, J.O. (1998) Points of Care in Using Statistics in Method Comparison Studies. *Clinical chemistry (Baltimore, Md.)*, 44(11): 2240-2242. 10.1093/clinchem/44.11.2240

Wicklin, R. (2019) Deming Regression for Comparing Different Measurement Methods. SAS. https://blogs.sas.com/content/iml/2019/01/07/deming-regression-sas.html

Appendix A List of original and NEMS methods

Table A-1: Description of the Original and NEMS methods.

Variable	Original	NEMS
Total Nitrogen	**Total calculated** Calculation: TKN + Nitrate-N + Nitrite-N. Please note: The Default Detection Limit of 0.05 g/m^3 is only attainable when the TKN has been determined using a trace method utilising duplicate analyses. In cases where the Detection Limit for TKN is 0.10 g/m^3, the Default Detection Limit for Total Nitrogen is 0.11 g/m^3.	**Total direct** Alkaline persulphate digestion, automated Cd reduction/sulphanilamide colourimetry. APHA 4500-N C & 4500-NO3- I (modified) 23rd ed. 2017
Total Phosphorus	**Phosphorus (Total) FIA** Total phosphorus digestion, automated ascorbic acid colourimetry. Flow Injection Analyser. APHA 4500-P H 23rd ed. 2017**.**	**Phosphorus (Total) Discrete** Total phosphorus digestion, ascorbic acid colourimetry. Discrete Analyser. APHA 4500-P B & E (modified from manual analysis and also modified to include a reductant to reduce interference from any arsenic present in the sample) 23rd ed. 2017. NWASCO, Water & soil Miscellaneous Publication No. 38, 1982.
Nitrate	Ion Chromatography following USEPA 300.0 (modified)	Calculation from NNN - Nitrite (both being APHA On-line Edition Method 4500)
Nitrite	Ion Chromatography following USEPA 300.0 (modified)	Flow Injection Autoanalyser following APHA On-line Edition Method 4500-NO2 B
Soluble Inorganic Nitrogen (HRC-NEMS)	Calculation	Calculation from NNN and NH3. APHA 4500
Total Organic Nitrogen	By Calculation. TON = Nitrite N + Nitrate N	Flow Injection Autoanalyser following APHA On-line Edition Method 4500-NO3 I.

Variable	Original	NEMS
Total Kjeldahl Nitrogen	By Calculation. TKN = Total Nitrogen - NNN	Calculation from NNN and TN to APHA 4500
Turbidity	Turbidity Meter following APHA 21st Ed. Method 2130 B.	Infrared Turbidity Meter following ISO7027:1999

Appendix B Worked Example of Bland and Altman Estimate, Deming Regression and Lin’s Concordance

This section provides a worked calculation for estimating bias and Levels of Agreement (LoA) using the Bland and Altman approach, Deming regression and Lin’s Concordance. The worked example uses the data presented in Figure 6-3. The data comes from 12 turbidity samples taken monthly. The numbers highlighted in yellow relate to the numbers is Figure 6-3.

Table B-1: Worked example of the Bland and Altman approach for estimating bias.

	New	Old	Mean (New + Old)/2	Log$_{10}$New	Log$_{10}$-Old	D = log$_{10}$New-log$_{10}$Old	Ratio (Old/New) = 10^{d}	Average (log$_{10}$New,Log^{10}Old)
Sample no.	NTU	NTU	NTU			Unitless	Unitless	Unitless
1	2.97	2.18	2.575	0.473	0.338	0.134	1.362	0.406
2	1.25	1.21	1.230	0.097	0.083	0.014	1.033	0.090
3	0.83	0.63	0.730	-0.081	-0.201	0.120	1.317	-0.141
4	1.62	1.06	1.340	0.210	0.025	0.184	1.528	0.117
5	16.4	13.2	14.800	1.215	1.121	0.094	1.242	1.168
6	5.35	4.68	5.015	0.728	0.670	0.058	1.143	0.699
7	5.34	3.34	4.340	0.728	0.524	0.204	1.599	0.626
8	3.39	2.74	3.065	0.530	0.438	0.092	1.237	0.484
9	16.3	13.3	14.800	1.212	1.124	0.088	1.226	1.168
10	5.97	5.4	5.685	0.776	0.732	0.044	1.106	0.754
11	0.49	0.46	0.475	-0.310	-0.337	0.027	1.065	-0.324
12	55.1	49.4	52.250	1.741	1.694	0.047	1.115	1.717
					Σ d	1.108		
n = 12			Mean difference (bias)		(Σ d)/n	0.092		
					Standard Deviation, s	0.060		

The mean difference (log$_{10}$ New - log$_{10}$ Old) is 0.092 , so log$_{10}$ New = 0.092 + log$_{10}$ Old. The 95% limits of agreement on the mean difference (bias) ranges from - 0.027 and 0.212, and these figures are log percentages, so it is best to back-transform the data.

$$(\Sigma\ (\log_{10}\text{New} - \log_{10}\text{Old}))/n = 0.092$$

$$\log_{10}\text{New} = 0.092 + \log_{10}\text{Old}$$

$$\text{New} = 10^{0.092} \text{ x Old}$$

$$\text{New} = 1.237 \text{ x Old}$$

To test if there is a trend in the bias (Trend Bias), we regress the difference, d, against the Average ($\log_{10}$ New,$\log_{10}$ Old). The outputs of a regression are shown below. Trend Bias is the p-value of the slope, in this case, p = 0.797, so we have little evidence to suggest that the bias varies systematically with the observed turbidity level in this case.

lm(formula = d ~ Average($\log_{10}$ New,$\log_{10}$ Old), data = df)

```
Residuals:  Min        1Q         Median          3Q         Max
            -0.082169  -0.038185  0.000281        0.026295   0.112000

Coefficients:                    Estimate Std.    Error       t value    Pr(>|t|)
(Intercept)                      0.097048          0.025427   3.817      0.00339 **
Average (log10 New,log10 Old)    -0.008398        0.031801    -0.264     0.79709
```

Deming Regression - Example R code

```
# Create  example data frame, same data as in Table B-1 above
df <- data.frame(
+   New = c(2.97, 1.25, 0.83, 1.62, 16.4, 5.35, 5.34, 3.39, 16.3, 5.97, 0.4
9, 55.1),
+   Old = c(2.18, 1.21, 0.63, 1.06, 13.2, 4.68, 3.34, 2.74, 13.3, 5.4, 0.46
, 49.4)
+ )
# Deming Regression
#install.packages("mcr") # if not already installed
library(mcr)

# log transform the data
df$New <- log10(df$New)
df$Old <- log10(df$Old)

# Perform regression
deming_model <- mcreg(y = df$New, x = df$Old, method.reg = "Deming")

# Present Results
summary(deming_model)

------------------------------------------

Reference method: Method1
Test method:     Method2
Number of data points: 12

------------------------------------------

The confidence intervals are calculated with
 bootstrap  ( quantile ) method.
Confidence level: 95%
Error ratio: 1

------------------------------------------

DEMING REGRESSION FIT:

                 EST SE        LCI       UCI
Intercept 0.09665353 NA 0.05134519 0.1573641
Slope     0.99161642 NA 0.93290506 1.0556573
```

```
------------------------------------------

BOOTSTRAP SUMMARY

           global.est bootstrap.mean    bias bootstrap.se
Intercept     0.09665        0.09893 0.00227      0.02700
Slope         0.99162        0.99282 0.00120      0.03095
```

So the equation is:

Log10(New) = 0.99162 x Log10(Old) + 0.09665

Which is the same as on Figure 6-3

Lin's concordance – Example R code

```
#install.packages("epiR") # if not already installed
library(epiR)

ccc_result <- epi.ccc(df$New, df$Old)
ccc_result$rho.c

        est    lower     upper
1 0.9819408 0.948903 0.9936865
```

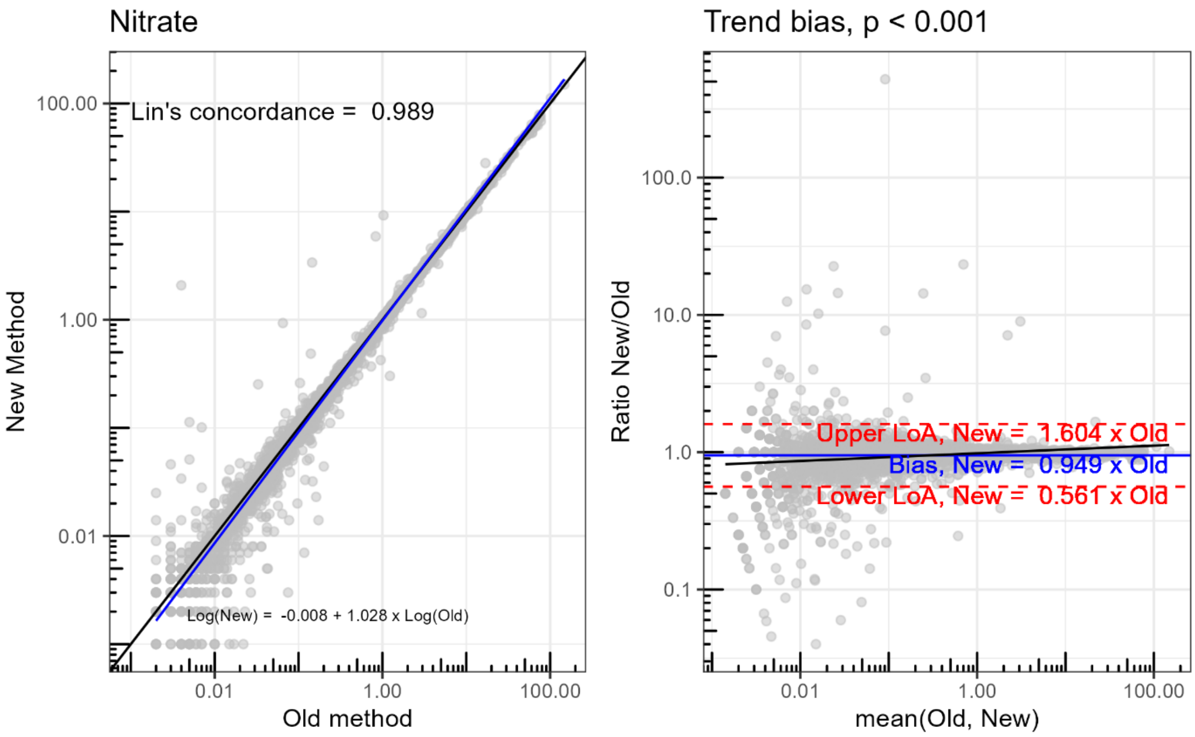
So Lin's concordance coefficient is 0.9819408, which is the same as on Figure 6-3.

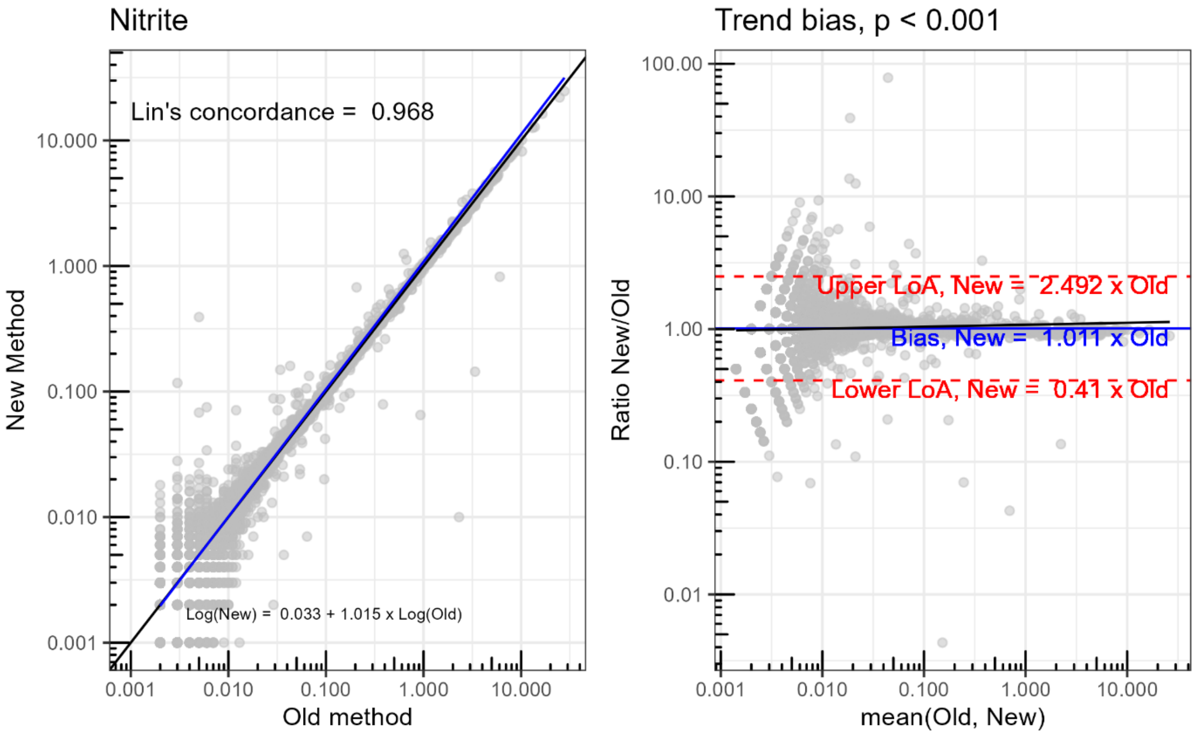
Appendix C Observer bias and levels of agreement

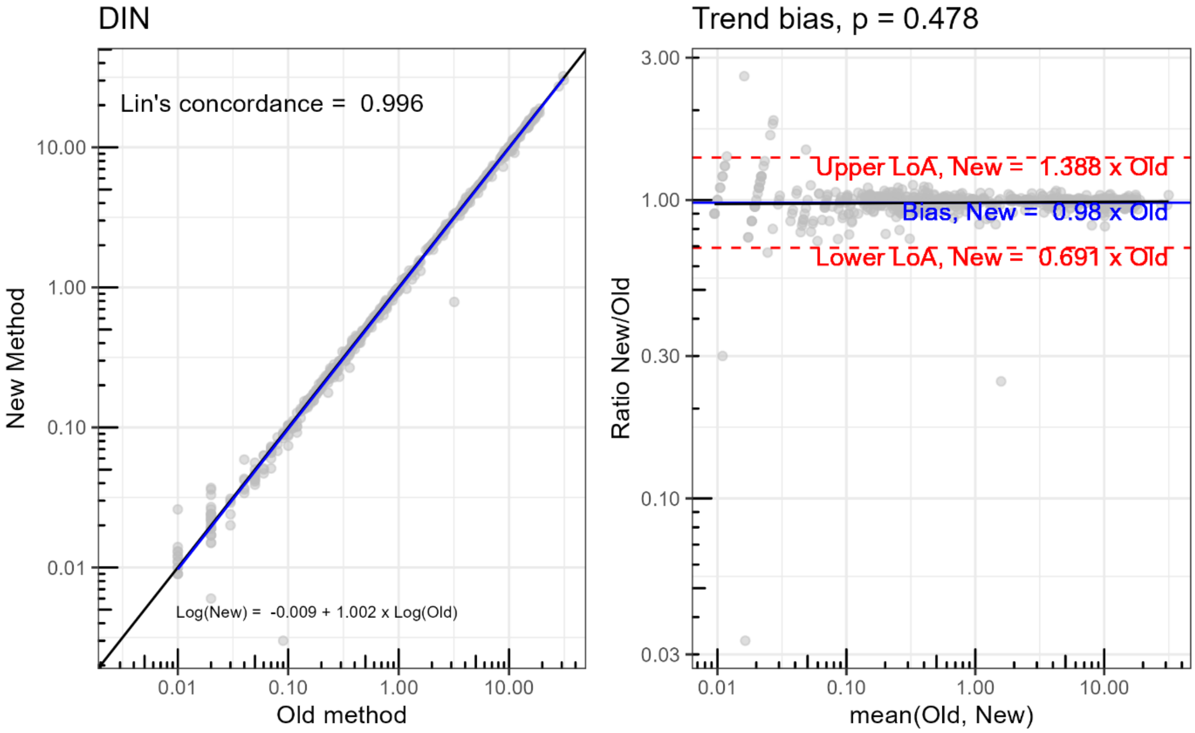
This section provides Bland and Altman graphs for combined data from all sites, rivers, lakes and groundwater, for each analyte and Bias and levels of agreement for each site in the dataset (note the sites are not named). The data are sourced from multiple water types.

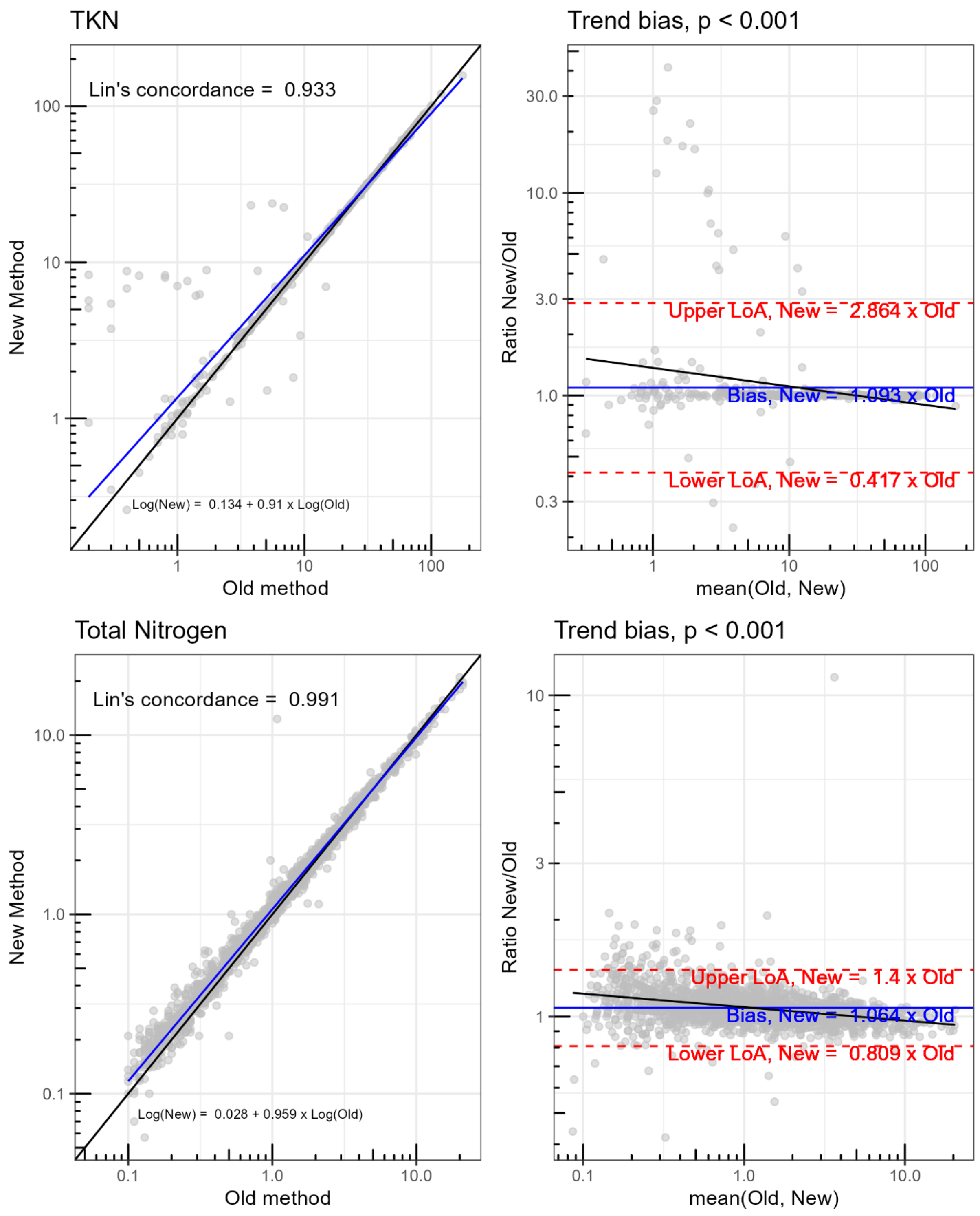
Bland and Altman Plots

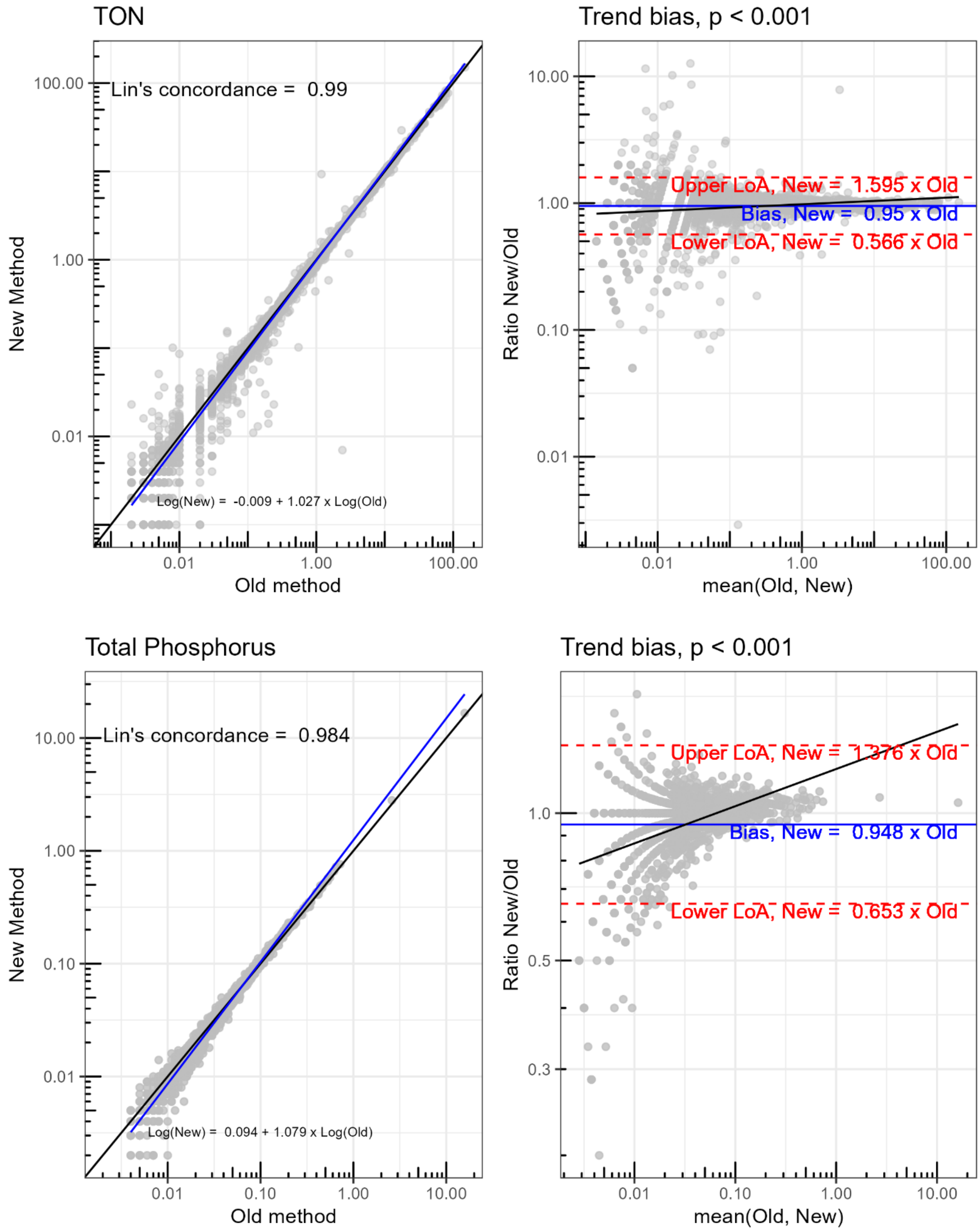
Bland and Altman plots are shown for nitrate, nitrite, DIN, TKN, total nitrogen, TON and total phosphorous. The data points come from multiple sites. The plots include Deming regression lines and estimates on Lin's concordance.




Nitrite
Lin's concordance = 0.968
Log(New) = 0.033 + 1.015 x Log(Old)
New Method
10.000
1.000
0.100
0.010
0.001
0.001
0.010
0.100
1.000
10.000
Old method
Trend bias, p < 0.001
Upper LoA, New = 2.492 x Old
Bias, New = 1.011 x Old
Lower LoA, New = 0.41 x Old
Ratio New/Old
100.00
10.00
1.00
0.10
0.01
0.001
0.010
0.100
1.000
10.000
mean(Old, New)

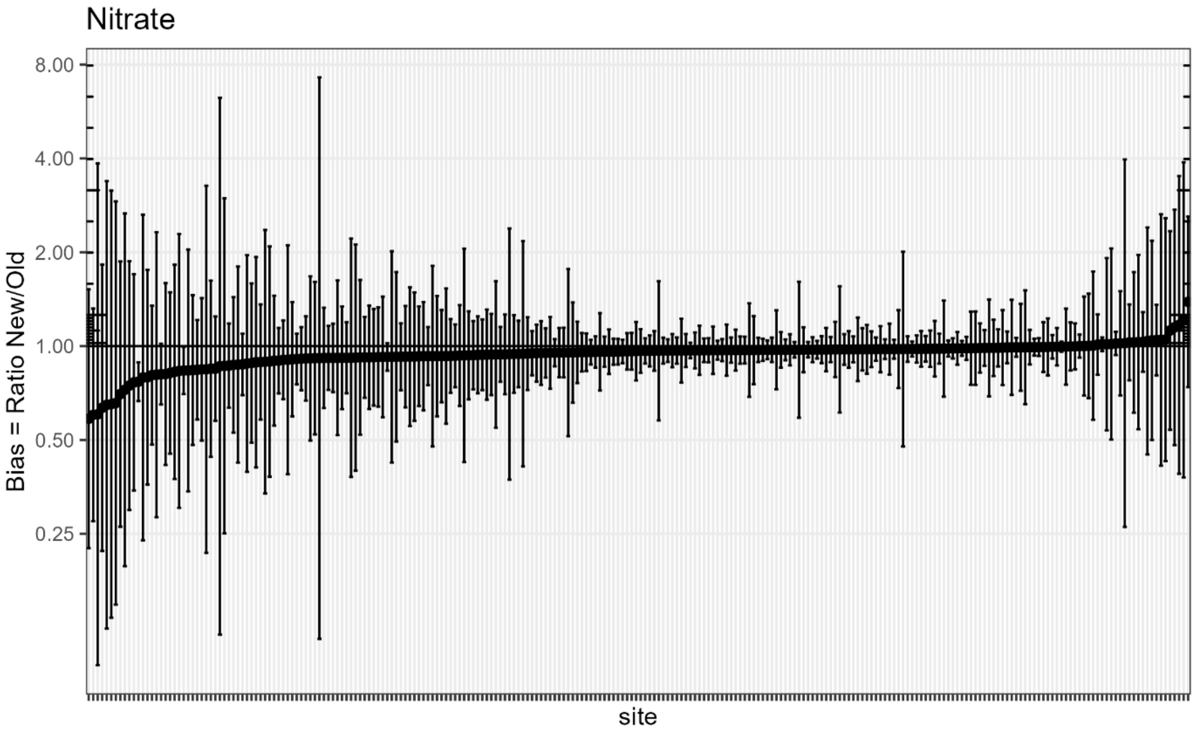

DIN
Lin's concordance = 0.996
Log(New) = -0.009 + 1.002 x Log(Old)
New Method
10.00
1.00
0.10
0.01
0.01
0.10
1.00
10.00
Old method
Trend bias, p = 0.478
Upper LoA, New = 1.388 x Old
Bias, New = 0.98 x Old
Lower LoA, New = 0.691 x Old
Ratio New/Old
3.00
1.00
0.30
0.10
0.03
0.01
0.10
1.00
10.00
mean(Old, New)

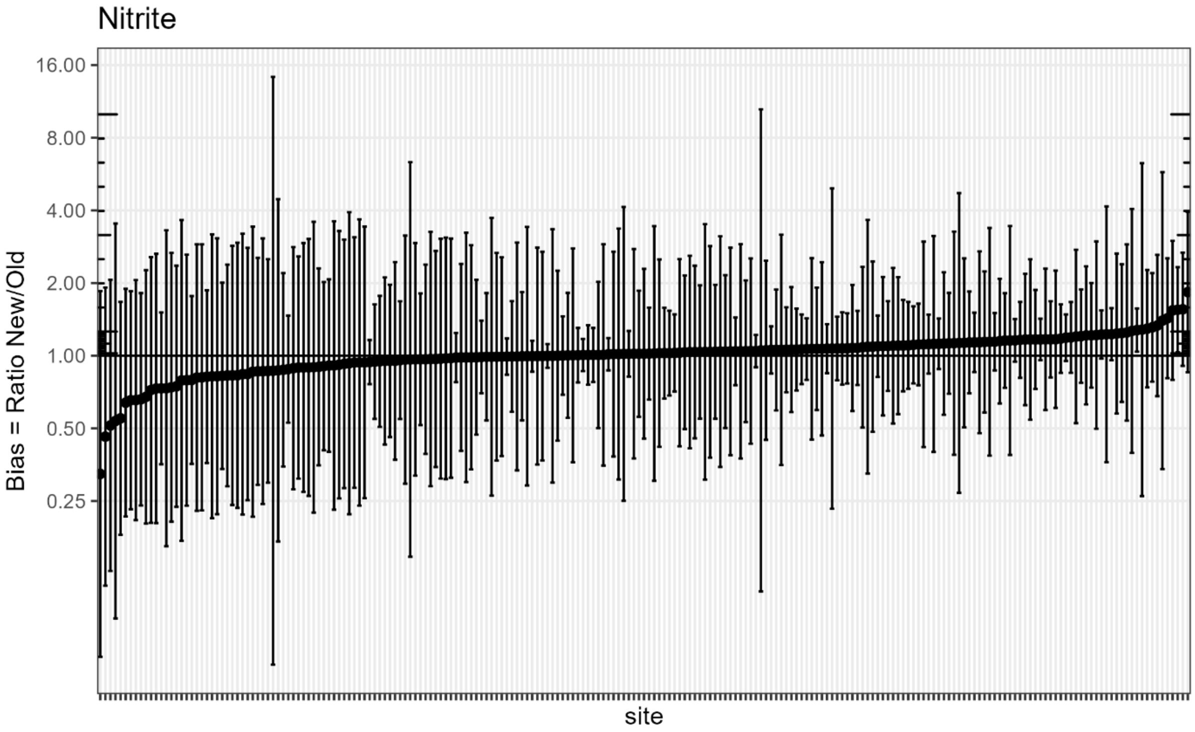

TKN
Lin's concordance = 0.933
Log(New) = 0.134 + 0.91 x Log(Old)
New Method
Old method
100
10
1
Trend bias, p < 0.001
Upper LoA, New = 2.864 x Old
Bias, New = 1.093 x Old
Lower LoA, New = 0.417 x Old
Ratio New/Old
30.0
10.0
3.0
1.0
0.3
mean(Old, New)
Total Nitrogen
Lin's concordance = 0.991
Log(New) = 0.028 + 0.959 x Log(Old)
10.0
1.0
0.1
Trend bias, p < 0.001
Upper LoA, New = 1.4 x Old
Bias, New = 1.064 x Old
Lower LoA, New = 0.809 x Old
10
3
1

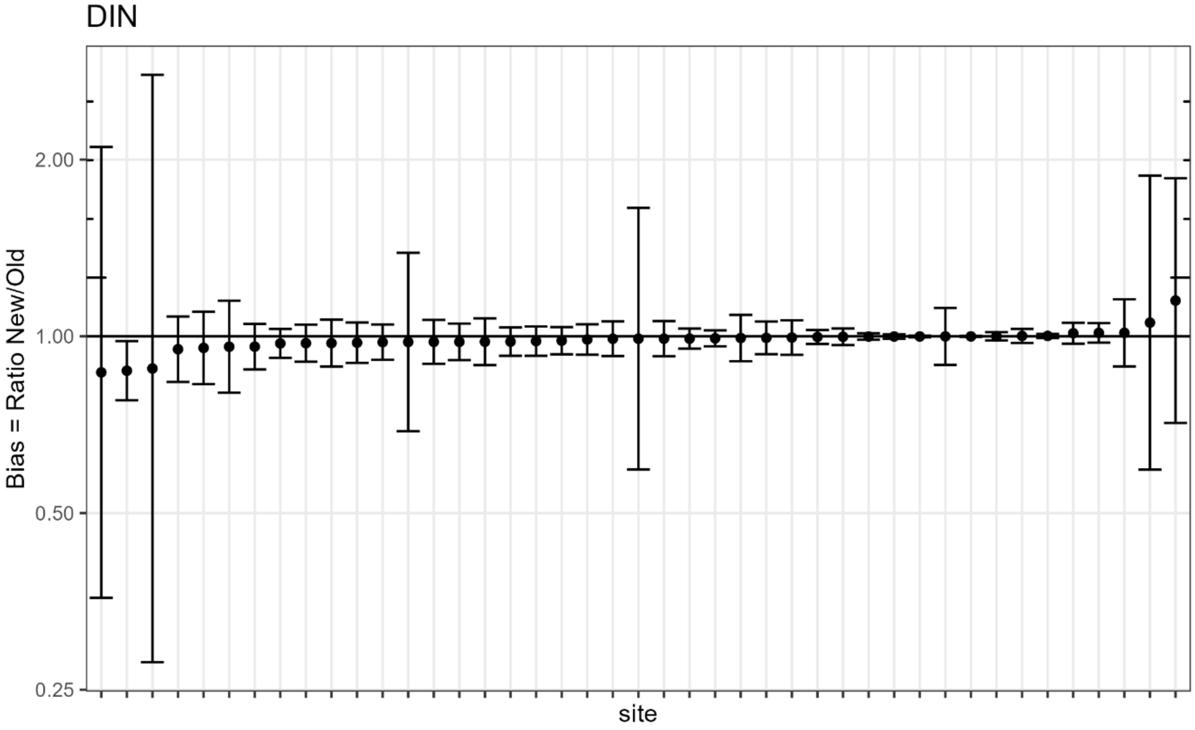

TON
Lin's concordance = 0.99
Log(New) = -0.009 + 1.027 x Log(Old)
New Method
Old method
100.00
1.00
0.01
Trend bias, p < 0.001
Upper LoA, New = 1.595 x Old
Bias, New = 0.95 x Old
Lower LoA, New = 0.566 x Old
Ratio New/Old
mean(Old, New)
10.00
1.00
0.10
0.01
Total Phosphorus
Lin's concordance = 0.984
Log(New) = 0.094 + 1.079 x Log(Old)
New Method
Old method
10.00
1.00
0.10
0.01
Trend bias, p < 0.001
Upper LoA, New = 1.376 x Old
Bias, New = 0.948 x Old
Lower LoA, New = 0.653 x Old
Ratio New/Old
mean(Old, New)
1.0
0.5
0.3
0.01
0.10
1.00
10.00

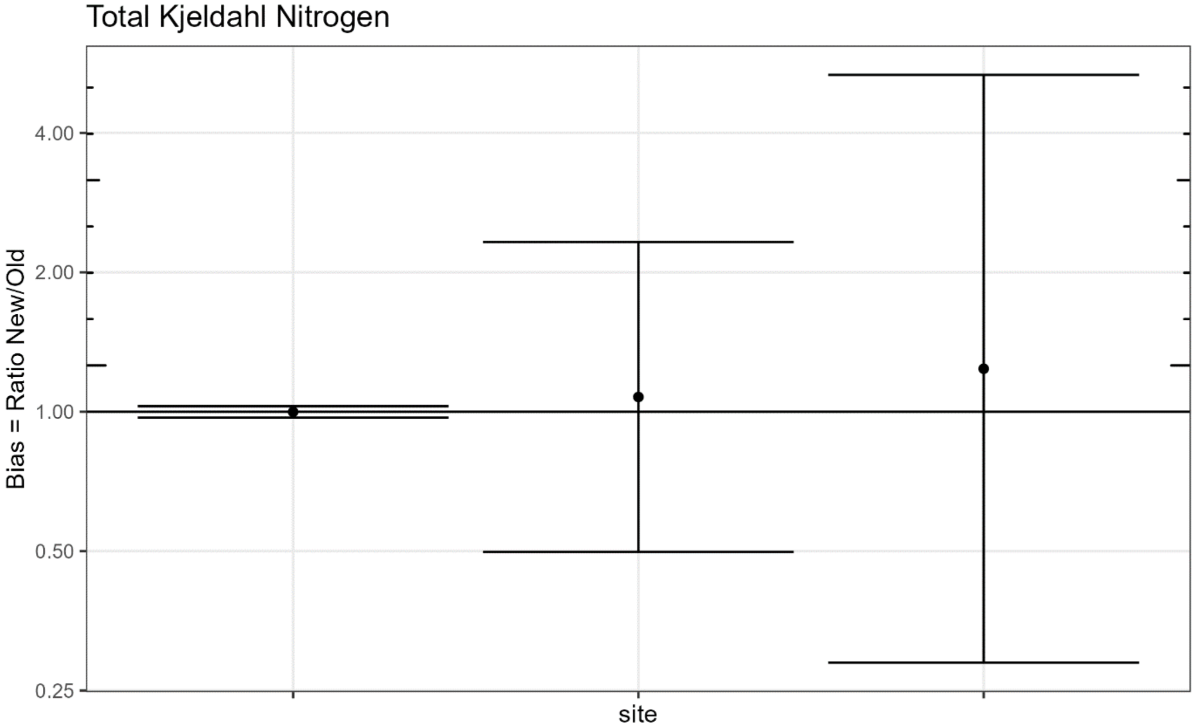
Graphs displaying the level of agreement (LoA)

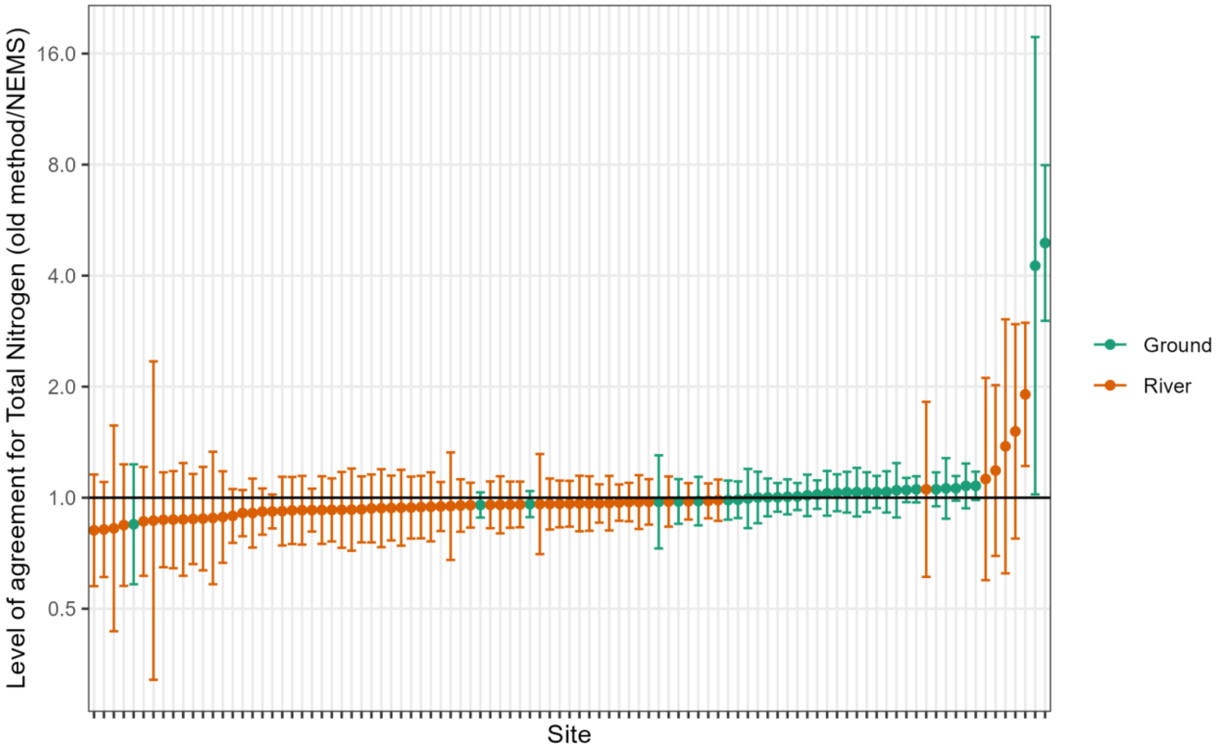
Graphs display levels of agreement for multiple sites (sites are not named) for nitrate, nitrite, DIN, TKN, total nitrogen, TON, and total phosphorus. In the case of total nitrogen, groundwater and surface water are differentiated, demonstrating that the LoA is not simply related to whether water is from surface or groundwater.

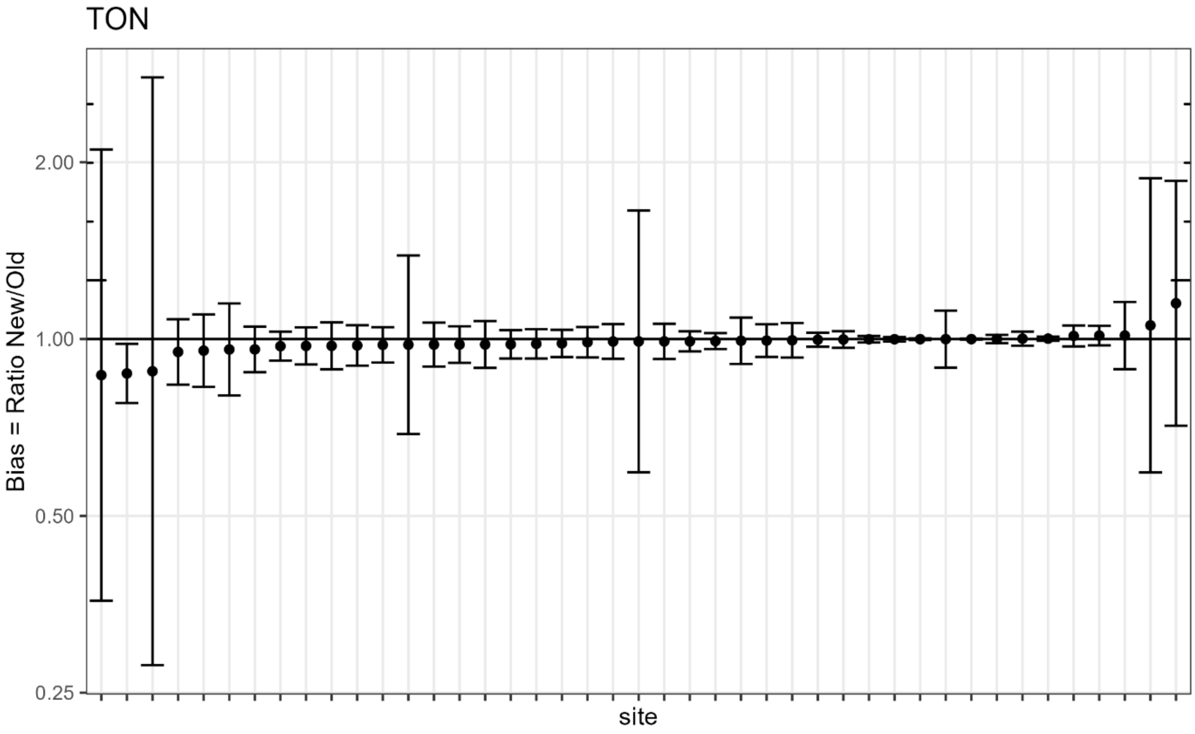


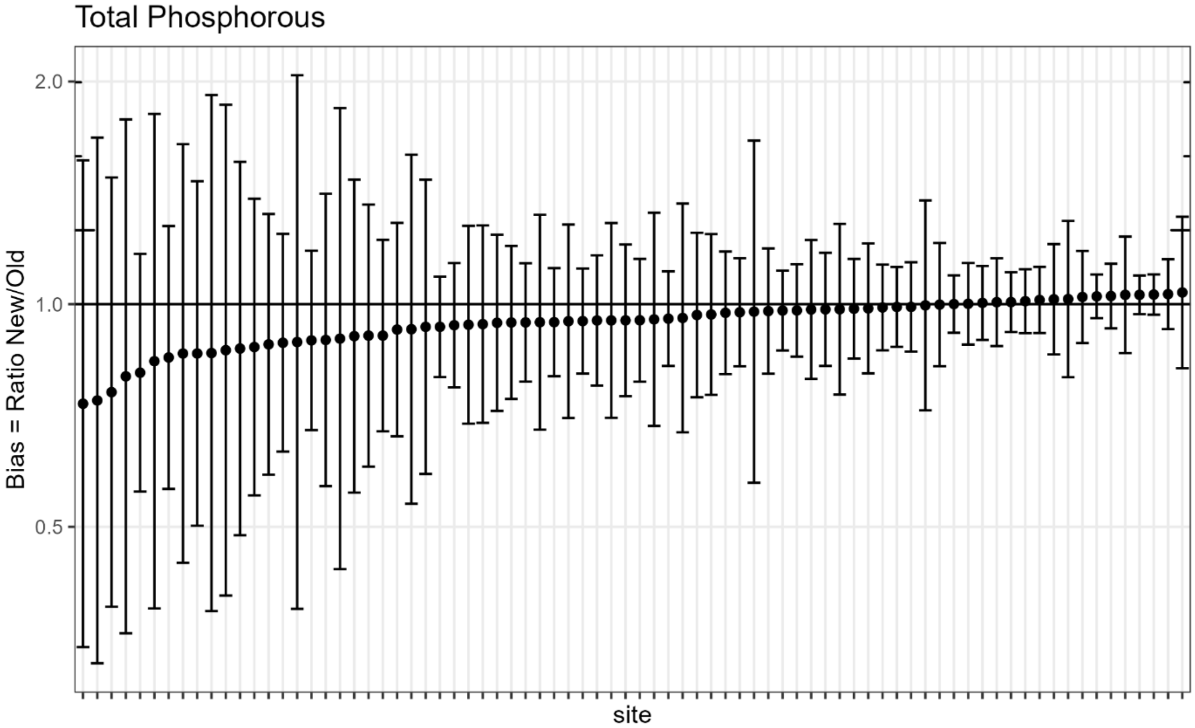



DIN
Bias = Ratio New/Old
2.00
1.00
0.50
0.25
site


Total Kjeldahl Nitrogen
Bias = Ratio New/Old
4.00
2.00
1.00
0.50
0.25
site


Level of agreement for Total Nitrogen (old method/NEMS)
16.0
8.0
4.0
2.0
1.0
0.5
Site
Ground
River


TON
Bias = Ratio New/Old
2.00
1.00
0.50
0.25
site


Total Phosphorous
Bias = Ratio New/Old
2.0
1.0
0.5
site