Accounting for changes in method in long-term nutrient data.

Recommendations based on analysis of paired SoE data from Wellington rivers.

Prepared for the Ministry for the Environment

June 2016

Authors/Contributors:
Rob Davies-Colley
Graham McBride

For any information regarding this report please contact:

Rob Davies-Colley
Principal Scientist
Water Quality
+64-7-856 1725
rob.davies-colley@niwa.co.nz

National Institute of Water & Atmospheric Research Ltd
Gate 10, Silverdale Road
Hillcrest, Hamilton 3216
PO Box 11115, Hillcrest
Hamilton 3251
New Zealand

Phone +64-7-856 7026
Fax +64-7-856 0151

NIWA Client Report No:	HAM2016-070
Report date:	June 2016
NIWA Project:	MFE16206

Contents

Executive summary **5**

1 Introduction **7**

1.1 Estimation of total nitrogen 8

1.2 Field *versus* laboratory filtration for dissolved nutrients 8

1.3 Screen vs trace level measurement of dissolved nutrients 8

2 US Geological Survey studies of total nitrogen estimation **10**

3 Methods **12**

3.1 Paired nutrient data in Wellington river SoE datasets 12

3.2 Data comparisons 12

4 Results and discussion **14**

4.1 Estimation of total nitrogen 14

4.2 Field *versus* laboratory filtration 20

4.3 Trace *versus* screen analysis of dissolved nutrients 24

5 Changing measurement protocols **28**

5.1 Why is such duplicate sampling necessary? 28

5.2 Ramifications of changed protocols and bias 29

6 Recommendations **30**

6.1 General 30

6.2 Total nitrogen 30

6.3 Dissolved nutrients 30

7 Acknowledgements **31**

8 Glossary of abbreviations and terms **32**

9 References **34**

Tables

Table 1-1: Comparison of analytical methods and detection limits (DL) for three forms of dissolved nutrients in GWRC's trials. 9

Table 3-1: Suggested verbal interpretation of Lin's Concordance Correlation Coefficient (CCC) values. 13

Table 4-1: Statistics of TN-A and TN-K and their ratio (TN-A/TN-K) in GWRC's paired dataset. 14

Table 4-2: Pearson correlation matrix for TNratio, SPM indices (TSS, turbidity, black disc visibility), and organic carbon (TOC). 16
Table 4-3: Table of correlation coefficients, field *versus* laboratory filtration. 21
Table 4-4: Statistics of field- *versus* laboratory-filtered concentrations for dissolved nutrients. 23
Table 4-5: Advantages and disadvantages of different filtration protocols for dissolved nutrient assay. 24
Table 4-6: Table of correlation coefficients, 'trace' *versus* 'screen' analysis. 25
Table 4-7: Statistics of 'screen' *versus* 'trace' results for dissolved nutrient analysis. 27

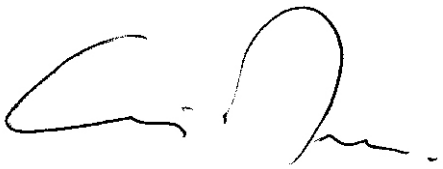
Figures

Figure 2-1: Accuracy, precision and bias in environmental measurements. 10
Figure 2-2: Under-estimation in total nitrogen estimated as TN-A as a function of TSS. 11
Figure 4-1: TN-A plotted *versus* TN-K for 55 Wellington river sites (sampled monthly for 12 months). 15
Figure 4-2: Ratio of TN-A/TN-K (TNratio) plotted against TSS. 17
Figure 4-3: Ratio of TN-A/TN-K (TNratio) plotted against visual clarity. 18
Figure 4-4: Comparison of the ratio of TN-A/TN-K (TNratio) in Wellington rivers with TN-A bias in rivers throughout the USA. 19
Figure 4-5: Field-filtered (Y) plotted *versus* lab filtered (X) concentrations for dissolved nutrient forms. 22
Figure 4-6: 'Screen' (Y) plotted against 'trace' (X) results for dissolved nutrient forms. 26
Figure 5-1: Apparent river water quality trend for five years of monthly sampling. 28
Figure 5-2: Apparent river water quality trend for five years of monthly sampling with and without a step change induced by protocol change. 29

Reviewed by

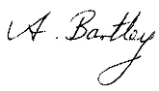
Juliet Milne

Approved for release by



Craig Depree

Formatting checked by



Executive summary

Some regional councils in New Zealand have been considering changing their river water quality monitoring protocols to those recommended in the NEMaR (National Environmental Monitoring and Reporting) project. Many of the NEMaR recommendations are likely to be ratified in the National Environmental Monitoring Standard (NEMS) for discrete water quality sampling and testing) that are currently being developed.

While there are benefits in national consistency, there are particular concerns about on-going use of historical data collected with 'old' protocols, particularly for tracking trends. Specific concerns with nutrients that are considered in this report are:

- Total nitrogen. Direct measurement by alkaline persulphate digestion (denoted TN-A), as recommended in NEMaR, *versus* estimation from total Kjeldahl nitrogen plus nitrate-nitrite-nitrogen (TN-K).
- Dissolved nutrients. (a) Laboratory filtration (as recommended in NEMaR) *versus* field filtration, and (b) Increased precision (a lower detection level – as recommended in NEMaR).

Before implementing changes, Greater Wellington Regional Council (GWRC) compared measurements of nutrients at their 55 regional river SoE sites by the NEMaR protocols *versus* their historical protocols. They made available 12 months of paired monthly monitoring data (July 2013 to June 2014) to inform decisions on nutrient protocols, nationally. Their dataset included a 12 month (July 2013 to June 2014) trial of TN methods (TN-A *versus* TN-K), and two sequential trials of dissolved nutrient protocols – a six month (July to December 2013) trial of laboratory- *versus* field-filtration, followed by a six month (January to June 2014) trial of 'trace' analysis methods (with lowered detection levels) *versus* normal laboratory protocols ('screen' methods).

Estimating total nitrogen

The direct persulphate digestion method for estimating total nitrogen (TN-A) was recommended in NEMaR reports, rather than the traditional TN-K estimate (TN-K = TKN + NNN; where NNN is nitrate-nitrite-N). This recommendation was based on the better precision of TN-A and experience in NZ's National Rivers Water Quality Network (NRWQN). However, recently the United States Geological Survey (USGS) reported that TN-A, while fairly accurate at low suspended particulate matter (SPM) concentrations, increasingly under-estimates total nitrogen as suspended particulate matter (SPM) concentrations increase above about 200 mg/l total suspended solids (TSS). If this finding applied to New Zealand rivers, which are seldom so high in SPM (about 2% of the time), TN-A and TN-K estimates of total nitrogen would be almost entirely consistent.

However, Wellington rivers show a tendency for the TN-A method to give lower estimates than the TN-K method. TN-A was similar to TN-K at low SPM concentration, but at higher concentrations (total suspended solids (TSS) > 10 mg/l; visual clarity < 0.6 m), TN-A became systematically lower than TN-K with further increases in SPM – dropping to a value of c. 75% at 100 mg/l and c. 50% at 1000 mg/l TSS. The pattern of the discrepancy in Wellington rivers is generally consistent with USGS studies, except that the downward inflection in the curve of the ratio TN-A/TN-K commences at much lower TSS (about 20-fold) for reasons that are unclear.

The finding of a pronounced underestimation by the TN-A method in turbid Wellington river waters raises issues with important ramifications for regional and national reporting if it represents the national situation. Firstly, the finding implies that all 'total nitrogen' data should, in future, be clearly identified as TN-A or TN-K. Second, it implies that 'total nitrogen' is not necessarily comparable across regions using different methods, and so may not be nationally consistent – particularly for comparatively turbid waters. Until the generality and drivers of these differences in TN-A and TN-K are better understood, we recommend councils continue with current methods. In the future it may be necessary to correct TN-A data from turbid waters to obtain unbiased and consistent estimates of total nitrogen.

NZ research is urgently needed to better understand the (inferred) TN-A bias and develop means for avoiding or correcting for it.

Dissolved nutrients

Field filtration of water samples is considered to be advantageous for slowing biochemical changes. However, NEMaR reports recommend laboratory filtration of water samples rather than in the field, mainly to avoid burdening field technicians. Laboratory filtration does incur a further laboratory charge, but field filtration costs field technician time (plus equipment and consumables).

In Wellington rivers there was good agreement between field and laboratory-filtered measurements for all three dissolved nutrient forms (nitrate-nitrite-nitrogen, NNN; ammoniacal-N and dissolved reactive phosphorus, DRP). The good overall agreement of field- *versus* laboratory filtration in Wellington rivers suggests that either protocol is suitable, and decisions on which to use can be made on pragmatic grounds. However, if filtration is done in the laboratory sample handling is even more important and prompt chilling with slush ice is recommended.

'Trace' dissolved nutrient analyses for samples from Wellington rivers (with efforts made to reduce measurement uncertainty in the laboratory) agreed closely overall with standard ('screen') analyses. Therefore improved dissolved nutrient measurement precision, with no change in chemical principle, is not anticipated to cause a step-change and confound trend analysis – although this should always be checked.

General ramifications with changed protocols

Changing field or laboratory protocols in long-term water quality datasets risks obscuring time-trends owing to a step change – as might occur for example, if TN analyses were changed from TN-K to TN-A. With any change in protocol, we recommend overlapping 'old' with 'new' methods for at least 12 monthly visits. The overlap data ensures continuity of the time-series and provides a means of converting data to the same basis for trend analyses, should an adjustment be required. These recommendations need to be incorporated into the National Environmental Monitoring Standard (NEMS) for Discrete Water Quality Sampling and Testing that is currently in preparation.

1 Introduction

The National Environmental Monitoring and Reporting (NEMaR) project made a number of recommendations to achieve national consistency in water monitoring (Davies-Colley et al. 2012 a, b; McBride et al. 2013). Many of these recommendations seem likely to be picked up and ratified in the National Environmental Monitoring Standard (NEMS) for Discrete Water Quality Sampling and Testing that is currently in preparation. In particular, NEMaR recommended:

- Direct laboratory measurement of total nitrogen (TN-A) by the alkaline persulphate digestion, rather than calculation from analyses of total Kjeldahl nitrogen (TKN) plus nitrite-nitrate nitrogen (NNN). (TN-K = TKN + NNN).
- Filtering water samples in the laboratory rather than field-filtering for dissolved inorganic nutrient analysis.
- Analysis of dissolved nutrients at sufficient precision to avoid large proportions of data being below detection limits (and therefore of low relative precision).

TN-A was recommended in NEMaR reports based on its higher precision than TN-K (and slightly lower cost) and good experience in the National Rivers Water Quality Network (NRWQN; Davies-Colley et al. 2011). This recommendation was made *before* we became aware of reports that TN-A under-estimates total nitrogen under certain water quality conditions. Laboratory filtration was recommended to avoid burdening field technicians with extra, sometimes onerous and prolonged, tasks. Additionally, the NEMaR project recommended reporting of 'best estimates' (actual concentrations) rather than censoring data (reporting "less than detection limit"), particularly for dissolved nutrient species. 'Non-detects' in censored data, rather than numbers, present difficulties for statistical analyses in state-of-environment reporting, including trend analyses.

There is a need to assess the implications of implementing these recommendations for existing regional SoE monitoring programmes, particularly as regards to analysis of time-series data for trends.

Greater Wellington Regional Council (GWRC), among several other regional authorities, have lengthy data sets containing estimates of total nitrogen from a TKN-based analytical method (TN-K), which appears to differ appreciably on occasion from direct TN measurements (TN-A). Before implementing NEMaR nutrient protocols (and following NEMaR recommendations regarding change in protocols) GWRC conducted a 12-month period of 'overlap' comparing proposed with existing nutrient methods at 55 Wellington regional river SoE sites. The purpose was to provide a comparison and (potentially) a means of numerically 'adjusting' data obtained by different protocols.

Initial work with GWRC's paired nutrient data was conducted using core funding from the NIWA Environmental Information Centre then research funding through the VMO (Values, Monitoring and Outcomes) research programme (MBIE's Landcare Research contract CO9X1003; subcontract to NIWA). The findings were presented at the NZ Water Symposium in Blenheim (November 2014) in two 'companion' oral conference presentations (Milne et al. 2014; Robinson et al. 2014).

Because of the ramifications for national as well as regional reporting, the Ministry for the Environment funded completion of the work as presented in this report. We present comparisons of nutrient measurements on Wellington rivers by NEMaR recommended protocols *versus* GWRC's historical protocols. The following comparisons are included:

(1) a 12-month (July 2013 to June 2014) comparison of two different methods for estimating total nitrogen in water samples

(2) a six-month (July to December 2013) comparison of field *versus* laboratory filtration for dissolved nutrients, and

(3) a further six-month (January to June 2014) comparison of dissolved nutrients with improved precision and lower detection limits ('trace' methods) *versus* standard protocols ('screen' methods).

The paragraphs below expand on these three areas.

1.1 Estimation of total nitrogen

NEMaR reports (e.g., Davies-Colley et al. 2012) recommend direct measurement of total nitrogen (TN) by the alkaline persulphate method (denoted TN-A) rather than estimation from total Kjeldahl nitrogen (TKN) added to nitrate-nitrite nitrogen (NNN) (TN-K = TKN + NNN). GWRC, together with a number of other regional councils, currently measure TN-K. TN-A has better precision than TN-K, but the US Geological Survey (USGS) has recently reported that it *underestimates* total nitrogen concentrations under certain water quality conditions. (That is, TN-A measurements are negatively biased; USGS 2009; Rus et al. 2012.) There is evidence in GWRC's paired datasets that TN-A is sometimes appreciably lower than TN-K. (Note, also, that there are very different detection limits for TN-A (10 mg/m^3) *versus* TN-K (110 mg/m^3) in GWRC's datasets.)

So the questions to be addressed are: (1) is the NEMaR recommendation for TN-A justifiable? And, if so, (2) can TN-K data be 'adjusted' to be comparable with TN-A data for tracking trends? Alternatively, if a change to a single, nationally consistent total nitrogen measure is *not* implemented, then the challenge for national reporting is how to adjust regional and national data collected by different protocols to the same basis.

1.2 Field *versus* laboratory filtration for dissolved nutrients

Historically several councils, including GWRC, have field filtered water samples for analysis of dissolved nutrients. Filtration removes (most) bacteria and slows biochemical degradation of samples, maintaining integrity of nutrient chemistry. However, NEMaR recommendations, based on NIWA experience in the NRWQN (Davies-Colley et al. 2011), favour laboratory filtration to avoid burdening field staff who must sample close to a nominated time-of-day on their sampling surveys and usually have to intercept couriers for overnight sample freight to the laboratory. Field filtration can be time-consuming if appreciable filter-clogging suspended particulate matter (SPM) is present (Alton Perrie, GWRC, pers. comm., September 2014); such difficulties are much more easily addressed in the laboratory. Field filtering also has the potential to introduce contamination (e.g., via wind-blown dust). A comparison of field- *versus* laboratory-filtration was conducted to check whether any step-change in dissolved nutrient data may result from changed protocols.

1.3 Screen vs trace level measurement of dissolved nutrients

NEMaR recommendations (Davies-Colley et al. 2012b) call for precision to be consistent with levels of dissolved nutrients measured, which can be quite low (approaching or below typical detection limits) in the case of 'reference' river sites and oligotrophic lakes. To avoid fairly large proportions of datasets being below detection limit for dissolved nutrients, particularly for ammoniacal-N and DRP, some councils including GWRC, are commissioning measurement with lower measurement

uncertainty (e.g., Table 1-1). However, increasing precision (and lowering the detection limit, DL) may involve subtle changes to laboratory protocols (e.g., using discrete analyser rather than flow injection). Such changes have the potential to introduce step-wise changes (discontinuity) to time-series data, which can confound time-trend analysis. Therefore a comparison of 'trace' *versus* standard ('screen') nutrient analysis (by Hill Laboratories) was conducted to check consistency of results from the two nutrient methods.

Table 1-1: Comparison of analytical methods and detection limits (DL) for three forms of dissolved nutrients in GWRC's trials. 'Screen' refers to standard analysis protocols; 'trace' refers to methods intended to improve low nutrient estimates – both at Hill Laboratories.

	'Screen' test		'Trace' test	
Dissolved nutrient	DL (mg/m^3)	Method	DL (mg/m^3)	Method
Ammoniacal-N	10	Filtered sample. Phenol/ hypochlorite colorimetry. Discrete Analyser. (NH4-N = NH4+-N + NH3-N) APHA 4500-NH_3 F (modified from manual analysis) 22nd Ed. 2012.	5	Filtered sample. Phenol/ hypochlorite colorimetry. Flow injection analyser. (NH4-N = NH4+-N + NH3-N) APHA 4500-NH_3 H, 22nd Ed. 2012.
DRP	4	Filtered sample. Molybdenum blue colorimetry. Discrete Analyser. APHA 4500-P E (modified from manual analysis) 22nd Ed. 2012.	1	Filtered sample. Molybdenum blue colorimetry. Flow injection analyser. APHA 4500-P G, 22nd Ed. 2012.
NNN	2	Total oxidised nitrogen. Automated cadmium reduction, flow injection analyser. APHA 4500-NO_3- I, 22nd Ed. 2012.	1	As per 'screen' test.

2 US Geological Survey studies of total nitrogen estimation

It is important for the current report to understand the concepts of bias and precision. Figure 2-1 (from McBride, 2005) graphically illustrates these concepts with a 'dart board' metaphor. The 'bull's eye' is analogous to the true concentration that we are trying to estimate in water quality work. Accuracy in water quality, as in other fields, incorporates both minimal bias and good precision.

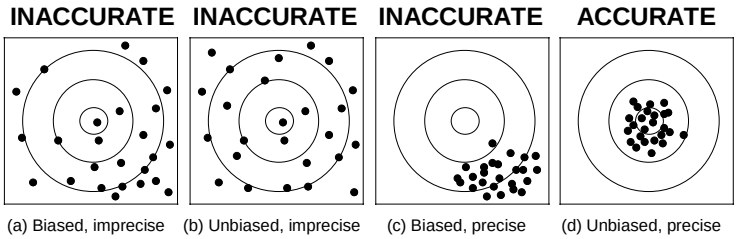


Figure 2-1: Accuracy, precision and bias in environmental measurements. These concepts are illustrated by the analogy of a dart-board in which the true concentration is represented by the bull's eye (after McBride, 2005). "Accurate" implies unbiased *and* precise.

Recently the US Geological Survey (USGS 2009) reported on accuracy of different methods of estimating total nitrogen in river water samples. A comprehensive report by Rus et al. (2012) consolidated and extended earlier USGS findings on accuracy of total nitrogen estimation, and a brief overview of the USGS findings is included here.

Three different methods of estimating total nitrogen (TN) are considered by the USGS, as follows.

(1) TN-A. Direct estimation from the alkaline-persulphate digestion of whole water samples

(2) TN-K. Calculated as the sum of total Kjeldahl nitrogen (TKN) plus nitrate- and nitrite-nitrogen (NNN) (TN-K = TNK + NNN), and

(3) TN-C. Calculated as the sum of dissolved nitrogen plus particulate nitrogen measured by a high temperature combustion method.

TN-A is most precise, but has recently been found by the USGS (2009) to under-estimate total nitrogen when suspended matter concentrations are high. TN-K is much less precise and tends to give results that are slightly high (about +3%) probably by the reduction of some nitrite- and nitrate-nitrogen (NNN) to ammoniacal-N which then becomes included in TKN (effectively resulting in this nitrogen being 'double-counted'). The lower precision of TN-K is associated with a high detection limit (typically around 100 mg/m^3) meaning that results for oligotrophic rivers and lakes are often close to detection limits. Finally, TN-C is regarded as a useful unbiased benchmark measurement of total nitrogen, but precision is again compromised by two constituent measurements. Furthermore, TN-C is more costly: by about 50% (Rus et al. 2012: Table 6; p 33). So far as we are aware TN-C has not been used routinely in NZ.

Rus et al. (2012) built on the earlier USGS work by comparing TN by different methods, using synthetic samples (with known TN) and real river samples (905 samples from 77 geographically widely spread river sites over the continental USA). TN-C was regarded as a reference measurement of the true TN concentration. They confirmed that TN-A under-estimates TN, and found that the discrepancy increased systematically with increasing TSS (Figure 2-2). At low TSS there was minimal underestimation of TN (about -4%) that was attributed to a fraction of particle-associated nitrogen

being refractory to the alkaline-persulphate digestion. At very high TSS (> 200 mg/l) the under-estimation of TN-A measurements (relative to TN-C) increased with increasing TSS, suggesting a secondary mechanism, possibly reagent exhaustion (Figure 2-2).

TN-K was found by Rus et al. (2012) to slightly over-estimate 'true' total nitrogen (as measured by TN-C), by around 3% on average, which was attributed to reduction of some oxidised nitrogen during the TKN digestion. Consistent with this interpretation, the positive bias of TN-K increased with NNN at high levels.

In the light of these findings, different approaches to estimating total nitrogen going forward were considered (Rus et al. 2012), but no single entirely satisfactory approach was identified. The 'best' approach might be including TN-C as well as on-going measurements of TN-A or TN-K, so as to provide an unbiased estimate of 'true' total nitrogen as well as continuing historical protocols. But that duplication of analyses would be very costly – incurring about an 80% increase over TN-A or TN-K alone. The next-best approach would seem to be switching from TN-A or TN-K to TN-C, but TN-C is less precise than TN-A (and more costly), and there would still be the problem for trend analysis of comparability with historical TN-A or TN-K. Rus et al. (2012) were inconclusive about how best to proceed at particular monitoring sites, but acknowledged the on-going challenge in estimating 'total' nitrogen in waters.

The present report on analysis of Wellington river TN data was guided by the Rus et al. (2012) report. We also adopt the USGS nomenclature for 'types' of total nitrogen (e.g., TN-A). We analyse GWRC's paired TN-A and TN-K data, compare findings with the USGS findings, and discuss implications for tracking time-trends over the long term. Note that, whereas the US study assessed negative bias in TN-A relative to TN-C reference measurements, this report simply compared TN-A measurements with TN-K measurement.

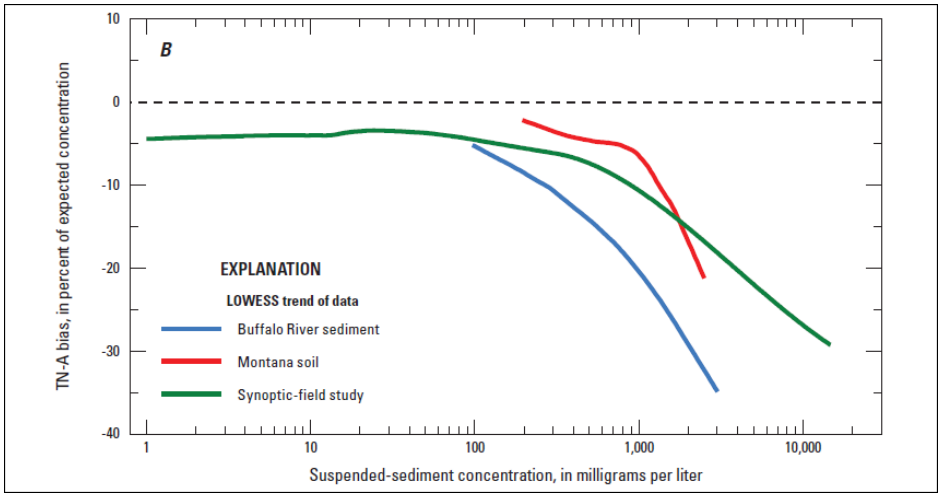


Figure 2-2: Under-estimation in total nitrogen estimated as TN-A as a function of TSS. Trend curves are from 77 widely distributed river stations in the USA (905 data points). (Figure 12B of Rus et al. 2012).

3 Methods

3.1 Paired nutrient data in Wellington river SoE datasets

GWRC provided their 2013-2014 dataset (.xlsm) for 55 regional river SoE sites at which both existing and proposed (NEMaR) nutrient protocols were applied. Therefore, 12 months of monthly data was available from each river site (maximum sample number, N = 55 x 12 = 660). This dataset incorporates three nutrient protocol comparisons, as follows

1. total nitrogen. TN-A *versus* TN-K (GWRC's historical method) for 12 months; July 2013 – June 2014;
2. dissolved nutrients. Laboratory- *versus* field-filtration (GWRC's historical method) for 6 months; July-December, 2013, and
3. dissolved nutrients. 'Trace' (more precise; lower detection limit) *versus* standard ('screen') methods of analysis for a further 6 months; January-June 2014.

Data were supplied *un*-censored – that is with 'best estimates' (actual concentrations) provided even where these were below detection limits as recommended in NEMaR (Davies-Colley et al. 2014b). 'Best estimates' of total suspended solids (TSS) and dissolved nutrients were occasionally negative – which is meaningful in a statistical sense (owing to uncertainty of measurement when concentrations are low) despite being physically meaningless.

3.2 Data comparisons

The EXCEL dataset provided by GWRC was copied into DataDesk® (Velleman 1992) for data exploration and statistical calculations. DataDesk permits datasets to be 'explored' with simple graphical displays aided by colour assignments of data groupings, and before formal statistics are computed or publication quality graphics prepared. This is particularly valuable for typically 'noisy' and highly skewed water quality data. DataDesk plots and other outputs are 'hot-linked' so that outliers or other interesting or dubious data-points can be selected or masked in order to view the change in other graphs or outputs.

Standard statistical tools available in DataDesk were used, such as correlation and linear regression. Some variables were log-transformed to reduce skewness and handle very wide data ranges (e.g., 1000-fold for total nitrogen and NNN). LOWESS (LOcally WEighted Scatterplot Smoother) was used to indicate patterns in data clouds on X-Y graphs. The root mean square error (RMSE) was used to indicate residual variance around trend lines or curves.

A spectral colour assignment applied to each of the 55 GWRC river monitoring sites in DataDesk, was helpful in identifying particular points, such as outliers, during the data exploration phase. The (coloured) DataDesk plots were replaced by publication quality (monochrome) graphs for the purposes of this report.

To measure strength of agreement of different nutrient measurement protocols, we used Pearson's product moment correlation coefficient (*r*), but also Lin's concordance correlation coefficient (CCC). Lin's coefficient measures agreement of points with the 1:1 line of perfect agreement on X-Y plots of the paired measurements, while Pearson's correlation coefficient measures closeness of points to the linear regression line. The former is considered better for assessing numerical agreement (Lin

1989). For a dataset with no net discrepancy (bias) in the long-term, *r* = CCC. Lin's correlation was calculated using the calculator at http://services.niwa.co.nz/services/statistical/concordance.

An interpretation of CCC values is given by the on-line calculator, and repeated here (Table 3-1).

Table 3-1: Suggested verbal interpretation of Lin's Concordance Correlation Coefficient (CCC) values.

CCC value	Strength-of-agreement
>0.99	Almost perfect
0.95-0.99	Substantial
0.90-0.95	Moderate
<0.90	Poor

It came to our attention during the preparation of this report that the numerical values of Lin's correlation coefficient (CCC) are affected by data range in the same way (and for the same reason) as Pearson's correlation coefficient (*r*). Both correlations may *appear* to be strong (approaching 1.0), even when visual inspection of a plot of the paired data indicates that the point-by-point agreement is actually not so strong. Indeed it may even be rather poor if the data range is very wide – say several orders of magnitude as is common for nutrient data. What would be better for comparing measurements by different protocols is a statistical measure that is unaffected by data range. For now, being unaware of any more sophisticated statistical index of agreement, we simply calculated the ratios and differences of paired measurements. We compared the mean ratio with 1.0 (or mean difference with zero) for an index of bias. An index of 'noise' (compounded measurement error) is provided by the standard deviation of the ratio and difference.

4 Results and discussion

4.1 Estimation of total nitrogen

4.1.1 Data analysis

Figure 4-1 shows TN-A plotted against TN-K for samples from the Wellington rivers. Agreement is fairly good overall, but there is more 'noise', as could be expected, at comparatively low concentrations close to the detection limit. Pearson's correlation coefficient, r = 0.972, and Lin's concordance correlation coefficient CCC = 0.970 – indicating 'substantial' agreement (based on the interpretation of CCC values given by the on-line calculator). However, *the impression of agreement is misleading* because of the wide range of data. Agreement is actually quite poor on occasion. TN-A is often lower than TN-K, and sometimes markedly so. For example, in the Awhea River at Tora Road, TN-A was only 16% of TN-K on 27 November and 17% on 30 September, 2013 (two points falling well below the 1:1 line in Figure 4-1).

There is a cluster of six outlier data points in Figure 4-1 for different rivers on different occasions, for which TN-K is very low and much lower than TN-A (i.e., the ratio of TN-A/TN-K is much greater than unity). A check of laboratory reports (Alton Perrie, GWRC, pers. comm. September 2014) revealed no transcription errors, so these anomalous data points remain unexplained. We consider these data points 'highly improbable' and have masked them for the remainder of the analysis reported here. A very low TN-A data point for the Waiohine River at Gorge (1.4 mg/m^3 on 26 November 2013) was also deemed an outlier (lower than the NNN and TKN values for that day). With these points masked, Pearson's R increased to 0.977 and Lin's CCC to 0.971 – the marginally lower value for CCC compared to R confirming a degree of bias.

Table 4-1 gives the statistics of TN-A and TN-K and their ratio (TNratio = TN-A/TN-K; excluding the unexplained outliers) for Wellington rivers. The median ratio is 0.93 (corresponding to an average under-estimation by TN-A, compared to TN-K, by about 7%).

Table 4-1: Statistics of TN-A and TN-K and their ratio (TN-A/TN-K) in GWRC's paired dataset. (N= 643/660) Missing cases include the six 'suspect' data points highlighted in Figure 4-1.

Variable	Mean	Median	StdDev	Min	Max	IntQRange
TN-A	685	348	980	30	8600	77
TN-K	750	390	1030	30	8600	83
TNratio	0.93	0.93	0.19	0.16	1.67	0.18

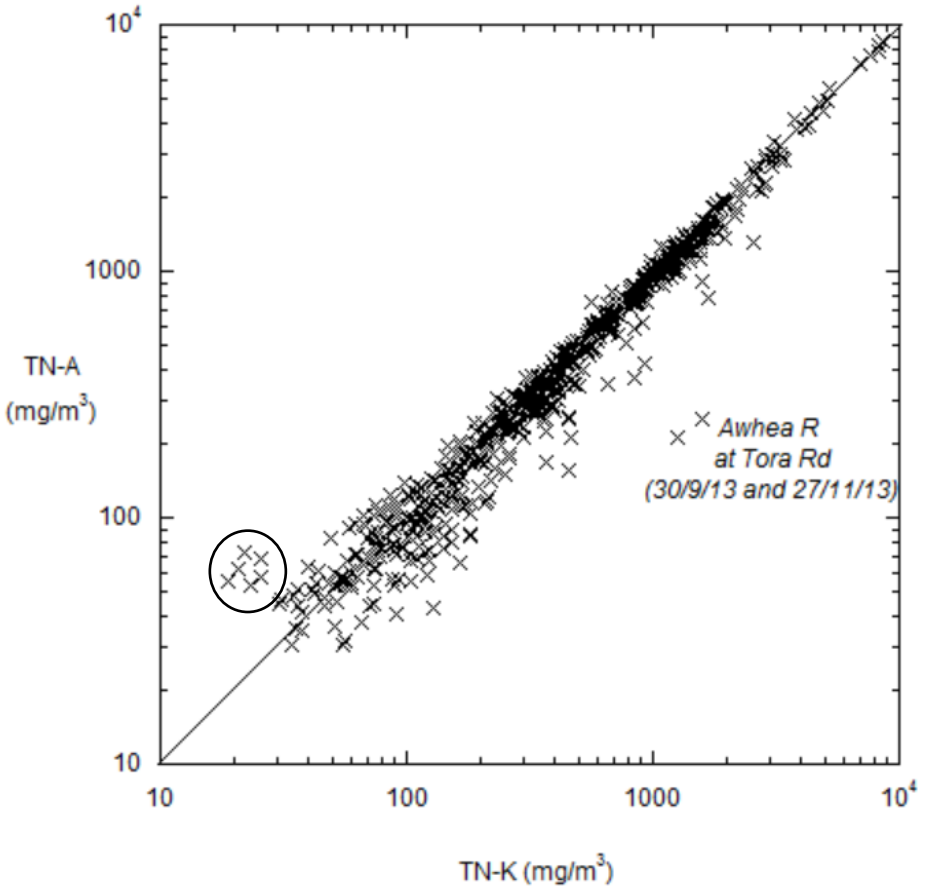


Figure 4-1: TN-A plotted versus TN-K for 55 Wellington river sites (sampled monthly for 12 months). Note the log-log grid labelled with powers of 10. A cluster of anomalous low-level points, plotting well above the line of equality, is highlighted.

The USGS (Rus et al. 2012) have reported that the under-estimation of total nitrogen by TN-A (*versus* measurements of TN-C) increases with suspended particulate matter (SPM) content of river waters, which they indexed as total suspended sediment (TSS). GWRC collect TSS data, and also measure its close optical correlates, turbidity and black disc visibility. We tested all three variables as indices of SPM and potential explanatory variables for the TN-A *versus* TN-K discrepancy. Note that 'best estimates' of TSS were used lower than the detection limit (usually 2 mg/l), which results in some negative TSS values owing to the uncertainty of measurement in very clear water. This results in the inability to display a small proportion of data – because negative values cannot be displayed on a logarithmic scale.

Plausibly the TN-A *versus* TN-K discrepancy might reflect organic content of samples (containing nitrogen), so total organic carbon (TOC – also measured by GWRC) was also investigated as a potential explanatory variable.

Table 4-2 shows the correlation matrix for TNratio, and the SPM variables and TOC – all log-transformed. There is very little difference between Pearsons correlation (on log-transformed variables) and Spearman rank correlation *in this case* because the relationships are monotonic and the logarithmic transformation does a fairly good job of linearising. So we only give the Pearson's matrix here.

Table 4-2: Pearson correlation matrix for TNratio, SPM indices (TSS, turbidity, black disc visibility), and organic carbon (TOC). (Variables were Log-transformed).

	TNratio	Log(Visibility)	Log(TSS)	Log(Turbidity)
Log(Visibility)	0.378			
Log(TSS)	-0.387	-0.884		
Log(Turbidity)	-0.406	-0.950	0.907	
Log(TOC)	-0.222	-0.665	0.582	0.635

Table 4-2 shows that TNratio is weakly (but significantly) related to indices of SPM (TSS, turbidity, and visual clarity by the black disc method) – which, themselves, are closely inter-correlated. Davies-Colley et al. (2014) reported similar strong mutual correlations of visual clarity, TSS and turbidity for the NRWQN. TNratio is less strongly related to TOC in Wellington rivers, which itself is less closely, but still moderately, related to the indices of SPM. Potentially a better index than TOC for analysing effects of SPM on total nitrogen assays would be particulate organic carbon (POC) or particulate organic nitrogen (PON) from high temperature combustion, but these data were not available.

Figure 4-2 shows the scatter plot for TNratio *versus* logTSS (r = -0.39; Table 4-2). A LOWESS (locally weighted scatterplot smoother) curve (using the default 20% span) has been added to the plot to indicate the overall curvilinear trend in data. The LOWESS smoother curve (with a mean square deviation in predicted TNratio of 0.0265 and the root mean square error – an index of residual variation – RMSE = 0.163) is slightly better than the linear regression (RMSE = 0.168) as a predictor, as expected given the apparent curvilinear trend in the data.

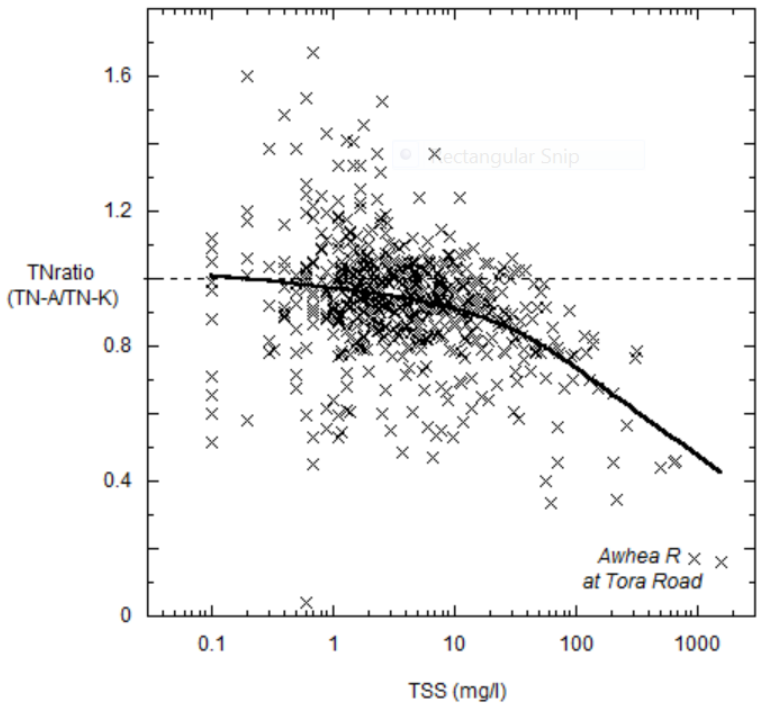


Figure 4-2: Ratio of TN-A/TN-K (TNratio) plotted against TSS.A LOWESS smoother (convex-up curve) is shown with a 20% span. TNratio is lowest for the Awhea River at Tora Road (two points on bottom right) at which TN-A was only 16% of TN-K on 27 November and 17% on 30 September, 2013 (with TSS close to 1000 mg/l on both occasions).

Turbidity and visual clarity provide alternate indices of SPM to TSS – with the advantage that neither approaches a detection limit, unlike TSS, and so are unaffected by poor precision. More importantly, visual clarity and turbidity are more nearly universally measured in regional SoE networks in NZ than TSS so are more likely to be available along with total nitrogen data.

TNratio plotted *versus* turbidity (not shown) shows a very similar pattern to that for TSS. The LOWESS smoother curve for turbidity (with RMSE = 0.166) is very similar to that for TSS.

Figure 4-3 shows TNratio plotted *versus* visual clarity, for which the residual variance is similar (RMSE = 0.168). The very low TNratio values on two occasions for the Awhea River at Tora Road (30 September and 27 November, 2013) appear to plot anomalously – suggesting that the visibility observations are too high. (This was confirmed on a plot of visibility *versus* turbidity – not shown – on which these two high turbidity data points plot as outliers.) The inferred bias in visibility probably reflects difficulties during very wet conditions – as noted on field sheets (Juliet Milne, GWRC, pers. comm. October 2014).

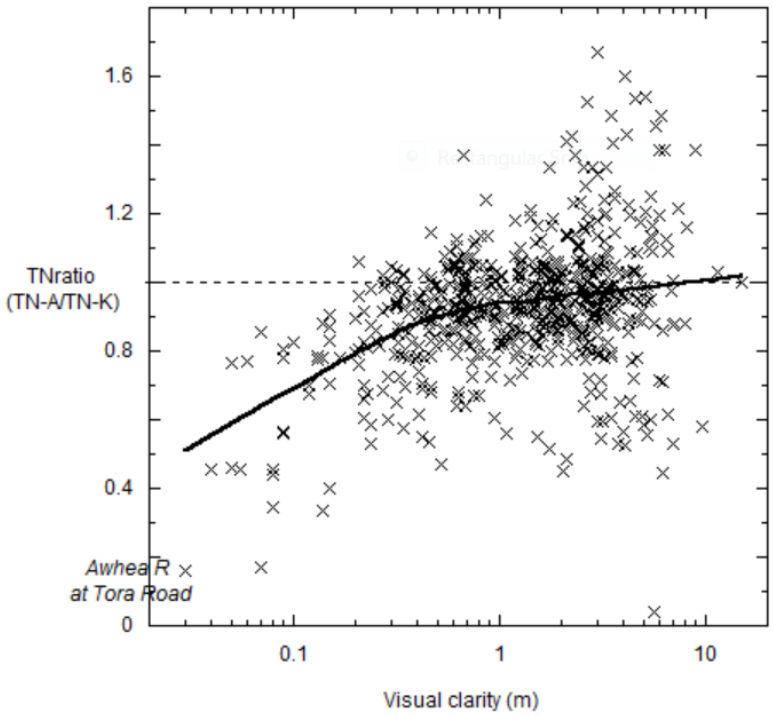


Figure 4-3: Ratio of TN-A/TN-K (TNratio) plotted against visual clarity. A LOWESS smoother (convex-up curve) is shown. The ratio is lowest for the Awhea River at Tora Road (two points on bottom left) at which TN-A was only 16% of TN-K on 27 November and 17% on 30 September, 2013, with visibility of a few cm on both occasions. (However, reported visibility on both occasions is too high to be consistent with the measured turbidity – see text.)

Note that there is considerable data scatter around equality of TN-A and TN-K (TNratio = 1.0; Figure 4-2; Figure 4-3). The scatter is greatest in clear waters (low SPM) which are also waters with low total nitrogen measured with comparatively poor precision, particularly as TN-K. That is, much of this scatter arises from uncertainty of measurement. However, some of this data scatter might reflect variation in nitrate-nitrite-nitrogen (NNN; aka oxidised nitrogen) which Rus et al. (2012) found to cause over-estimation of total nitrogen by the TN-K method (positive bias) at high concentrations owing to reduction of some oxidised nitrogen. When we tested this hypothesis (for relatively clear samples, visual clarity > 0.6 m), a dependence on NNN was detectable, but very weak (plot not shown). Accordingly, most of the data scatter around TNratio = 1.0 at low SPM can be attributed to compounding of errors in both measurements.

4.1.2 Discussion

We found that direct total nitrogen estimates (TN-A) in Wellington rivers are low compared to estimates from TN-K – to an extent depending on SPM (indexed by TSS, turbidity, or visual clarity). Based on the work of the USGS on total nitrogen (Rus et al. 2012) we interpret the discrepancy as an *under*-estimate by TN-A (i.e., a negative bias) *versus* the more-nearly unbiased (albeit less precise) estimate by TN-K.

When SPM in Wellington rivers is low (TSS < 10 mg/l; turbidity < 10 NTU; visibility > 0.6 m), the ratio TN-A/TN-K is nearly constant approaching 1.0 (i.e., TN-A ~ TN-K). However, when SPM is higher, TN-A tends to fall below TN-K. High SPM occurs in roughly 25% of the Wellington river data across all 55 river sites (167/651 visual clarity measurements < 0.6 m). This proportion is similar to rivers nationally, as indicated by data for the NRWQN sites (228/997 or 23% of visibility measurements in 2011 using the data of Davies-Colley et al. 2014), suggesting about ¼ of river TN-A measurements, nationally, will be biased low.

Total nitrogen for relatively turbid rivers, also for turbid lakes, can be expected to be appreciably under-estimated by TN-A. That is, the bias of total nitrogen by TN-A is expected to be severe for waters with high SPM.

It is informative to compare the discrepancies between TN-A and TN-K in the GWRC dataset with the TN-A bias for rivers across USA as reported by Rus et al. (2012). The LOWESS trend curve for TNratio (= TN-A/TN-K) as a function of TSS (taken from Figure 4-2) is overlain on the TN-A bias curve as a function of TSS in Figure 4-4. (The graphical comparison was achieved by expressing the *complement* of TNratio as a percentage.) The decreasing trend of TNratio in Wellington rivers (inflexion point on the curve) starts at a much lower TSS value than the TN-A bias curve for the USGS dataset.

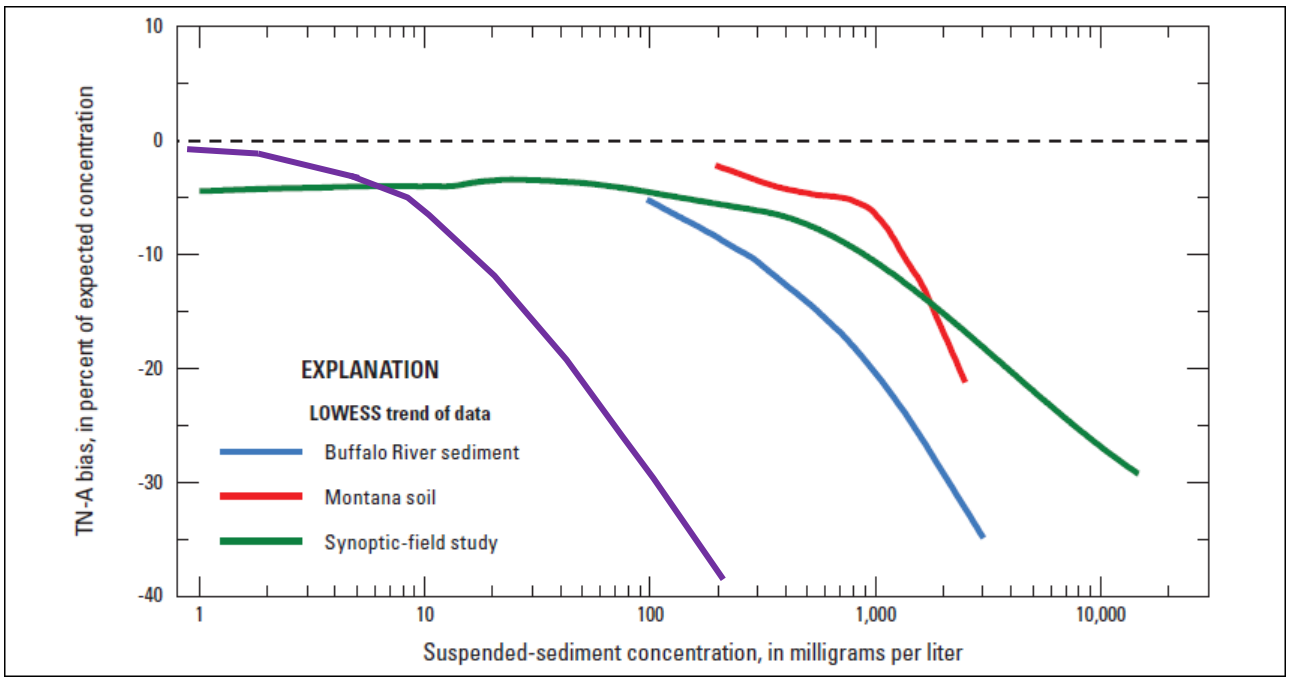


Figure 4-4: Comparison of the ratio of TN-A/TN-K (TNratio) in Wellington rivers with TN-A bias in rivers throughout the USA. The GWRC LOWESS trend for TNratio as a function of TSS (purple curve taken from Figure 4-2) is superimposed on Figure 12B of Rus et al. (2012) showing general trends of under-estimation in TN-A as a function of TSS for 77 river stations in the USA (905 samples) and in a river sediment sample and soil sample. (The TN-A bias percent scale is also used for the complement of TNratio (1.0 – TNratio) expressed as a percentage for the GWRC trend curve.).

TSS values > 200 mg/l are comparatively rare in NZ rivers, occurring mainly in high flows. For example, in the add-on of TSS to the NRWQN dataset (Davies-Colley et al. 2014) only 1.8% of measurements exceeded 200 mg/l. So if the trend for US rivers was assumed to apply to NZ rivers there would seem little reason to expect a discrepancy in changing method to TN-A as recommended

in NEMaR. Fortunately, before adopting TN-A, GWRC overlapped TN-A with TN-K measurements (following another NEMaR recommendation), and this has alerted NZ to a significant problem with total nitrogen estimation.

We do not know why the TN-A discrepancy for Wellington Region rivers is so much more 'sensitive' to SPM (about 20-fold) than rivers in the USA. This may reflect climatic and therefore soil differences, or it may reflect analytical differences in TN-A and/or TN-K measurements between the NZ and USGS datasets. Investigations to clarify the mechanism and the reasons for the greater sensitivity in NZ are highly desirable so as to improve the basis for correcting TN-A to be more nearly comparable to TN-K.

Until the discrepancy in TN-A *versus* TN-K is better understood, and ways to avoid it or reliably account for it are developed, we recommend maintaining the *status quo* on total nitrogen analysis. This means that regional reporting can continue as previously – without risk of a step change that will obscure time trends in total nitrogen. (An exception might be when total nitrogen is very low, such as in oligotrophic lakes, when TN-A is the *only* useful analytical option given the relatively high detection limit of TN-K.)

However, status quo total nitrogen measurement implies *national inconsistency* (with TN-K measured in some regions and TN-A in others) – and consequent difficulties for national reporting. Fortunately, *time trend* analyses for TN-A and TN-K seem likely to give broadly similar results (unless there is a systematic change also in SPM). *State* reporting presents more of a challenge, however. Potentially the overall trend curves (LOWESS curves) for the Wellington Region could be used (cautiously) to 'adjust' TN-A data to be comparable with TN-K. However, we do not know whether this could be applied robustly elsewhere in New Zealand. The 'adjustment' would be made as follows:

Predicted-TN-K = (Measured-TN-A)/(Predicted-TNratio)

where (Predicted-TNratio) is the ratio TN-A/TN-K predicted using the LOWESS curves presented here in terms of TSS or visual clarity.

An important ramification of the findings of this report, is that all total nitrogen data should be identified in regional and other records as TN-A or TN-K (or perhaps some other designation if and when other methods are adopted). Data columns should not be headed merely "TN" – which is ambiguous and might propagate serious confusion. Furthermore, if TN-A is used to estimate total nitrogen in comparatively turbid water bodies, an index of SPM (e.g., TSS or visual clarity) ideally needs to be measured too to enable the bias in TN-A to be corrected.

4.2 Field *versus* laboratory filtration

4.2.1 Data analysis

Figure 4-5 shows field-filtered *versus* lab-filtered concentrations of three forms of dissolved nutrients (nitrate-nitrite-nitrogen, NNN; ammoniacal-N, Am-N; and dissolved reactive phosphorus, DRP) for GWRC river sites. A logarithmic scale is used on each of the three graphs to encompass the very high variability, particularly for nitrate-nitrite-nitrogen (NNN), much of it being seasonal with winter peaks and summer troughs. (Note that some uncensored data points for Am-N are *negative*, and so cannot be displayed on a logarithmic scale.)

Table 4-3 presents statistics of agreement between laboratory- and field-filtered data (Pearson's *R* and Lins concordance correlation coefficient (CCC) both for all data and for relatively high datapoints > detection limit). There is close agreement and high concordance of data obtained by the two filtration protocols, but, again, the large data range makes the agreement look better than it really is.

Table 4-3: Table of correlation coefficients, field *versus* laboratory filtration.

Variable	N	Pearson's *R*	Lin's CCC	Interpretation of CCC*
Am-N	71	0.942	0.925	Moderate
Am-N (> DL)	56	0.968	0.961	Substantial
DRP	266	0.924	0.913	Moderate
DRP (> DL)	207	0.971	0.967	Substantial
NNN	266	0.992	0.992	Almost perfect
NNN (> DL)	257	0.997	0.996	Almost perfect

(*) A suggested interpretation of Lins CCC is as given on the web calculator at:

http://services.niwa.co.nz/services/statistical/concordance.

The agreement between filtration protocols for NNN is excellent (despite a small degree of increasing 'noise' at low concentrations – near the detection limit (DL) of 2 mg/m^3), and despite two fairly low outliers for which field-filtered NNN was appreciably lower than lab-filtered. But the agreement is less close for DRP (DL = 4 mg/m^3) and comparatively poor for ammoniacal nitrogen (Am-N; DL = 10 mg/m^3) – both of which tend to show increasing data scatter at concentrations in the vicinity of the detection limit.

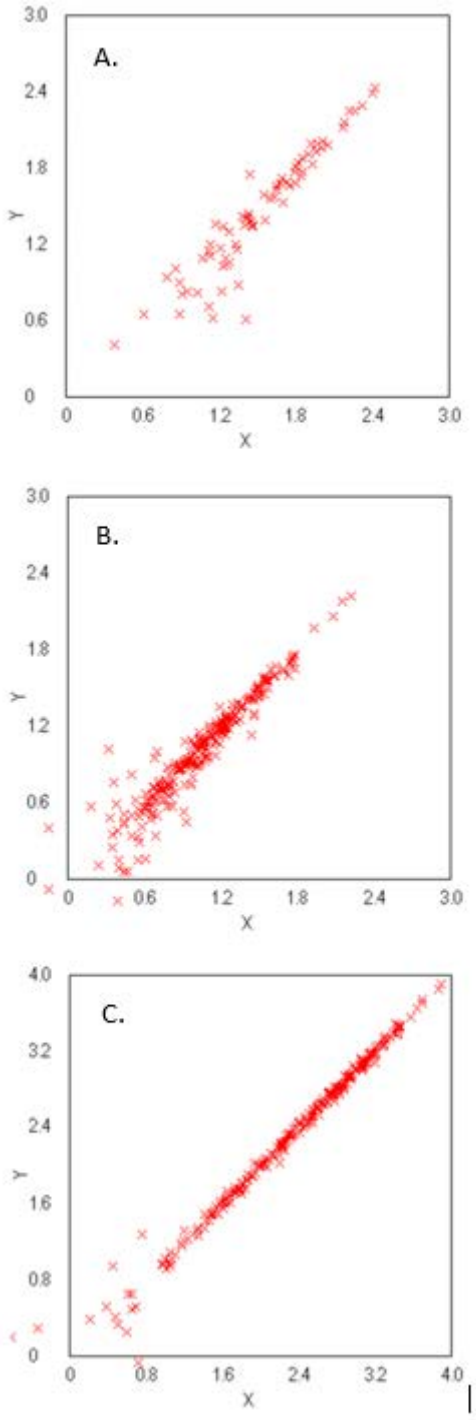


Figure 4-5: Field-filtered (Y) plotted *versus* lab filtered (X) concentrations for dissolved nutrient forms. Data were obtained in a 6-month trial from July to December 2013 (inclusive). Plots, taken from the on-line Lin's Concordance Calculator output, are on log-log scales, for A. Ammoniacal nitrogen (Am-N), B. Dissolved reactive phosphorus (DRP) and C. Nitrate-nitrite-nitrogen (NNN). All data are as mg/m^3, i.e., Log_{10}Am-N = 3.0 implies that Am-N = 10^3 = 100 mg/m^3.

Field-filtered and lab-filtered data for the three forms of dissolved nutrients were compared using both the differences (e.g., Am-N(field) – Am-N(lab)) and ratios (e.g., Am-N(field)/Am-N(lab)). Differences might be more relevant if the same degree of absolute bias occurred – for example due to contamination incidents. Ratios might be more relevant if the same degree of *relative* bias occurred in field or laboratory-filtered data. Table 4-4 shows statistics for both differences and ratios of paired dissolved nutrient data. Overall the agreement is quite good for concentrations above the detection limit. However, field-filtered values tend to be a little higher than lab-filtered values for all three dissolved nutrients. Ammoniacal-N tends most strongly to be higher on field-filtered samples (by an average of +8%) compared to DRP (+5%) and least for NNN (+3%).

Table 4-4: Statistics of field- *versus* laboratory-filtered concentrations for dissolved nutrients. A. differences and B. ratios of the dissolved nutrient species: ammoniacal nitrogen (Am-N), dissolved reactive phosphorus (DRP), and nitrate-nitrite-nitrogen (NNN). Data were obtained in a 6-month trial from July to December 2013 (inclusive). Only data above the detection limits are considered.

A. Differences (as mg/m^3)

Variable	Count	Mean	Median	StdDev	Min	Max	IntQRange
Am-N	56	4.0	3.4	8.5	-27.4	20.9	8.6
DRP	207	1.2	0.6	3.1	-6.3	16.5	2.4
NNN	257	17.3	23	84	-420	510	30

B. Ratios

Variable	Count	Mean	Median	StdDev	Min	Max	IntQRange
Am-N	56	1.12	1.08	0.24	0.50	1.72	0.26
DRP	207	1.09	1.05	0.19	0.51	2.01	0.18
NNN	257	1.04	1.03	0.13	0.31	1.54	0.12

4.2.2 Discussion

Dissolved nutrients in Wellington rivers, measured following laboratory filtration, agree closely with measurements following field filtration. There seems no compelling reason to favour either field- or laboratory filtration, so contrary to NEMaR recommendations (Davies-Colley et al. 2012b), there is no particular need to maintain national consistency on filtration protocols.

However, there is a slight tendency for higher average estimates of dissolved nutrients with field filtration than with laboratory filtration, and this pattern is consistent across all three dissolved nutrient species: NNN (+3%), DRP (+5%) and Am-N (+8%). A possible explanation is minor microbial uptake of nutrients before laboratory processing. An important rationale for field processing is to arrest biochemical changes in dissolved nutrients, and it is plausible that some net change (nutrient uptake) occurred in samples from Wellington rivers.

The choice to field- or laboratory-filter can be made on a pragmatic basis. Table 4-5 gives advantages and disadvantages of field and laboratory filtration for dissolved nutrients. If laboratory filtration is chosen, it is very important to minimise changes in unfiltered samples during freight to the laboratory by appropriate handling and storage – particularly chilling with slush ice as recommended

in NEMaR (Davies-Colley et al. 2012b). Slush ice (a mixture of ice chips and water at 0°C) is preferred to use of ice packs which have less efficient sample bottle contact resulting in slower cooling.

Table 4-5: Advantages and disadvantages of different filtration protocols for dissolved nutrient assay.

Filtration protocol	Pros (Advantages)	Cons (Disadvantages)
Field filtration	Removes micro-biota promptly and arrests biochemical action (and consequent nutrient changes).	Considerable time is sometimes required to obtain a useful volume of filtrate, particularly from very turbid samples – burdening field teams who may be on a strict schedule (e.g., to meet a courier deadline for overnight sample delivery).
	Avoids further analytical cost. (But there IS still a cost in terms of filtering materials as well as staff time).	Field vehicles (and conditions) are (often) not ideal for filtrations.
	Field filtration is standard protocol for ground waters – so might be preferable for samples from rivers undergoing exchange with ground waters.	There is the theoretical potential for gross contamination. (This was not recognisable in the GWRC data).
Laboratory filtration	In the laboratory, many samples can be filtered and otherwise processed in parallel – which is 'efficient' for very turbid or filter-clogging samples (that are slow to filter).	Relies crucially on (prompt) dark-storage and rapid chilling (but avoiding freezing) to slow changes until arrival at the laboratory.
	More convenient for many reasons – including parallel filtrations of many samples.	The laboratory will impose a filtration charge.
	Laboratory conditions are clean and otherwise near-ideal, so contamination is less likely.	Changes in groundwater-affected samples may occur during freight to the laboratory, even with effective and rapid chilling.

4.3 Trace *versus* screen analysis of dissolved nutrients

4.3.1 Data analysis

Wellington river water samples collected from January to June 2014 (inclusive) were used to compare 'trace' (more precise, lower detection level) *versus* 'screen' (routine) methods for dissolved nutrients (Hill Laboratories Ltd, Hamilton) – on laboratory filtered sub-samples.

Figure 4-6 shows 'screen' concentrations plotted *versus* 'trace' concentrations for the three forms of dissolved nutrients at GWRC's river sites. Again a logarithmic scale is used on each of the three graphs to encompass the very high variability, particularly for nitrate-nitrite-nitrogen (NNN).

Table 4-6 presents statistics of agreement between 'trace' *versus* 'screen' methods (Pearson's *R* and Lins concordance correlation coefficient (CCC) for relatively high data points (> 'screen' detection limits).

The agreement overall for NNN is excellent despite a small degree of increasing 'noise' at low concentrations ('screen' detection limit = 2 mg/m^3. But the agreement is less close for DRP ('screen' DL = 4 mg/m^3) and comparatively poor for Am-N ('screen' DL = 10 mg/m^3) – both of which tend to show increasing data scatter at levels close to their detection limits.

Table 4-6: Table of correlation coefficients, 'trace' *versus* 'screen' analysis.

Variable	N	Pearson's *R*	Lin's CCC	Interpretation of CCC*
Am-N	326	0.807	0.784	Poor
Am-N (> DL)	91	0.973	0.966	Substantial
DRP	326	0.981	0.979	Substantial
DRP (> DL)	232	0.985	0.984	Substantial
NNN	326	0.998	0.998	Almost perfect
NNN (> DL)	316	0.999	0.998	Almost perfect

(*) A suggested interpretation of Lins CCC is as given on the web calculator at:

http://services.niwa.co.nz/services/statistical/concordance.

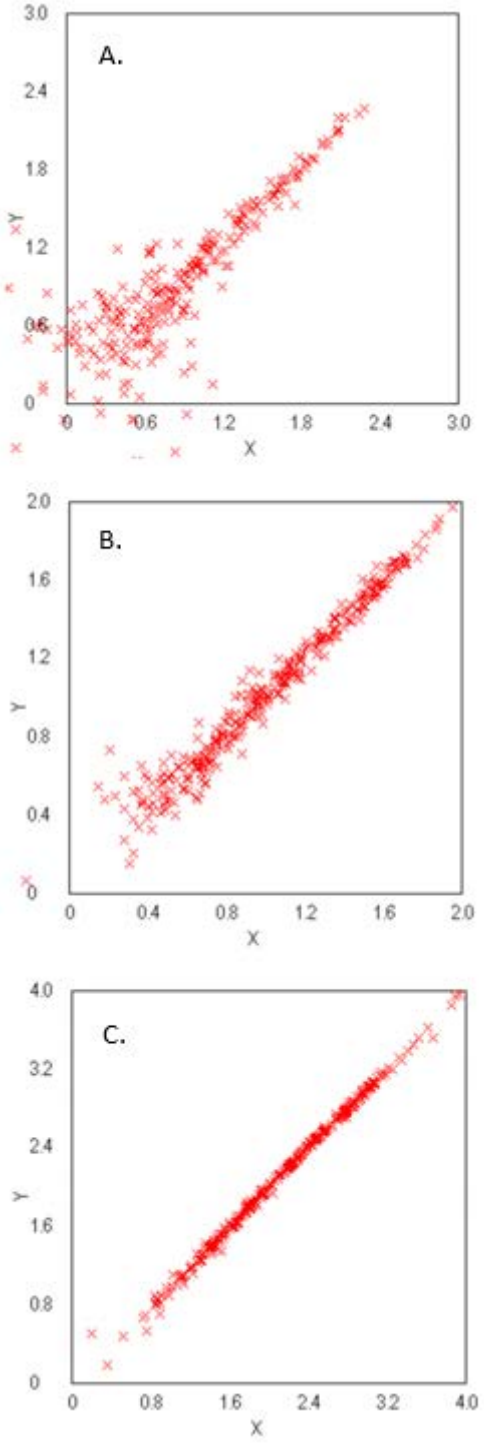


Figure 4-6: 'Screen' (Y) plotted against 'trace' (X) results for dissolved nutrient forms. Data were obtained in a 6-month trial from January to June 2014 (inclusive). Plots, taken from the on-line Lin's Concordance Calculator output, are on log-log scales, for A. Ammoniacal nitrogen (Am-N), B. Dissolved reactive phosphorus (DRP) and C. Nitrate-nitrite-nitrogen (NNN). All data are as mg/m^3, i.e., Log_{10}Am-N = 3.0 implies that Am-N = 10^3 = 1000 mg/m^3.

Table 4-7 gives statistics for both differences and ratios of paired dissolved nutrient data (concentrations greater than detection limit). Overall it can be seen that the agreement is quite good. However, 'screen' ammoniacal-N appears to be low (by an average of -10%) compared to 'trace' values. 'Screen' and 'trace' values are almost equal for DRP (-1%), and slightly higher (+3%) for NNN (medians of ratios).

Table 4-7: Statistics of 'screen' *versus* 'trace' results for dissolved nutrient analysis. A. differences and B. ratios of the dissolved nutrient species: ammoniacal nitrogen (Am-N), dissolved reactive phosphorus (DRP), and nitrate-nitrite-nitrogen (NNN). Data were obtained in a 6-month trial from January to June 2014 (inclusive). Only data above the detection limits are considered.

A. Differences (mg/m^3)

Variable	Count	Mean	Median	StdDev	Min	Max	IntQRange
Am-N	91	-2.3	-2.7	6.5	-33	23	57
DRP	232	-0.1	-0.2	2.2	-9.6	8.4	2.2
NNN	316	13	2.5	100	-609	1438	9.1

B. Ratios

Variable	Count	Mean	Median	StdDev	Min	Max	IntQRange
Am-N	91	0.93	0.90	0.19	0.62	1.70	0.24
DRP	232	0.99	0.99	0.12	0.61	1.48	0.14
NNN	316	1.05	1.03	0.10	0.84	1.75	0.087

4.3.2 Discussion

Overall there was good agreement of dissolved nutrient concentrations estimated by 'trace' *versus* screen' methods – suggesting little, if any, discrepancy attributable to the small changes in laboratory protocols required to achieve higher precision. In particular, for nitrate-nitrite-nitrogen (NNN) the results from 'screen' and 'trace' protocols agreed very well. Likewise, for dissolved reactive phosphorus (DRP) the protocols agreed closely overall albeit with more 'noise'. We can be less confident about agreement of 'screen' *versus* 'trace' protocols for ammoniacal-N because of the comparatively high average uncertainty of measurement with many values below the detection limit. Furthermore, 'screen' values (above detection level) for ammoniacal-N averaged about 10% lower than 'trace' results, for reasons that are unknown.

There is a need to improve precision, and further lower detection limits, for both DRP and Am-N so that lesser proportions of datasets are compromised by high measurement uncertainty, particularly for oligotrophic rivers and lakes.

5 Changing measurement protocols

When considering a change to a new laboratory or field method for water quality variables, one should first examine the degree of agreement between the old and the new. To do that, both the old and new protocols should be applied simultaneously on the same sampling visit or to the same bulk samples, i.e., each method is supplied an aliquot from the same (well-mixed) bulk sample. This should be done a number of times, to encompass a range of conditions (Davies-Colley et al. 2012b). Experience to date indicates that for a monthly river sampling programme at least one year of duplicate analyses ought to be made – as GWRC did for total nitrogen.

5.1 Why is such duplicate sampling necessary?

Results from SoE river water quality sampling programmes are used for two principal aims:

- to estimate the water quality 'state'
- to assess *change in* water quality (trend) over time.

Both activities can be confounded by changes in laboratory or field procedures. This may particularly affect time trend assessment. For example, Figure 5-1 displays an apparent trend through a five-year trend assessment period with monthly sampling.

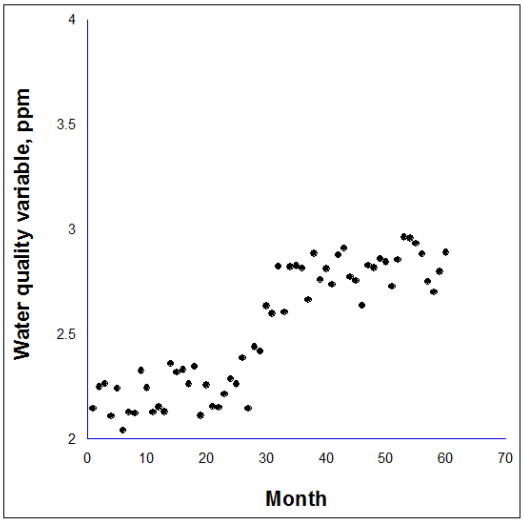


Figure 5-1: Apparent river water quality trend for five years of monthly sampling.

On Figure 5-1 there is obvious evidence of a trend over time. One could easily contemplate fitting a line through these data to infer a trend, in which case there would have been a trend on the order of 1 ppm over the five years. But in fact these data include a step change of 0.4 ppm at month 30.

Figure 5-2 shows the *same* data with and without the step change due to method bias. When the step change is taken into account the estimated trend magnitude drops by about 40%, a substantial change.

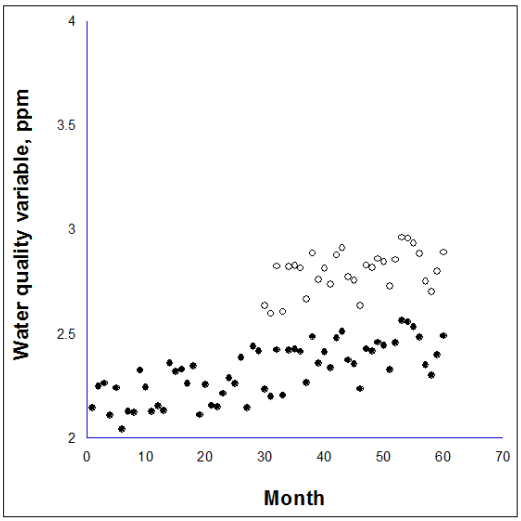


Figure 5-2: Apparent river water quality trend for five years of monthly sampling with and without a step change induced by protocol change. Data are plotted without a step change of 0.4 ppm at month 30 (solid circles) and with that change (open circles).

5.2 Ramifications of changed protocols and bias

Two points arise from the considerations above:

- The existence of the step change is not particularly obvious on Figure 5-2, even though it is substantial.
- If the step change was the result of a change in field protocols or laboratory method (a new method and/or a new laboratory), the information returned from the monitoring programme is confounded.

The following lessons can be deduced:

- If a laboratory method or field protocol has changed during a period of record, the data analyst must be made aware of this: clearly the effects of such changes may be difficult to detect yet are important in practice.
- Any bias introduced by a new method can only be accounted for if there is a parallel period of sampling-and-analysis before the new method is used on its own.

6 Recommendations

6.1 General

1. If any change in protocols or field staff or laboratory service provider is contemplated, at least a 12-month 'overlap' of the 'old' and 'new' protocols should be conducted[1]. This will provide a means of 'adjusting' data should this be found to be necessary (i.e., if a bias is uncovered).

2. The findings of this report should be taken up in the NEMS standard for Discrete Water Quality Sampling and Testing (currently under development).

6.2 Total nitrogen

1. Always specify the total nitrogen analysis method used; we suggest by denoting it TN-A or TN-K.

2. Keep the total nitrogen method regionally consistent so that real environmental change is not confounded by a spurious step change. (However, TN-A might be the *only* useful analytical option for oligotrophic waters given the much higher detection limit for TN-K.)

3. Research is needed urgently on the mechanism of TN-A bias by SPM (and reasons for the high 'sensitivity' of Wellington rivers – and NZ rivers generally?) and how best to account for this bias or whether it can be avoided.

6.3 Dissolved nutrients

1. Laboratory filtration and field filtration are almost equivalent in terms of dissolved nutrient results, so the choice of protocol can be made on pragmatic grounds.

2. Water samples (particularly if unfiltered) must be stored dark and chilled promptly after collection to avoid changes in dissolved nutrients. Prompt chilling (and shading from sunlight) is most practically achieved by placing water sample bottles into chilly bins with slush ice (an ice-water mix at 0^{o}C).

3. Continuing efforts are needed to increase precision (reduce measurement uncertainty and decrease detection limits) in analytical methods, particularly for ammoniacal-N and DRP.

[1] In practice, the extra cost incurred in such 'overlapping' is a strong disincentive to make protocol changes, unless the new protocol is substantially cheaper or more precise or less biased. This should promote stability of long-term datasets.

7 Acknowledgements

Juliet Milne (then of GWRC) first raised issues around total nitrogen in changing to NEMaR protocols, and Alton Perrie (GWRC) forwarded paired ('old' and 'new' protocol) Wellington river SoE data for analysis. Both are thanked for comments on early versions of the report and for continuing interest in the findings and the national, as well as regional, ramifications. We thank Jochen Schmidt (Chief Scientist EI, NIWA) for arranging interim (NIWA core) funding for this study. Analysis was continued using research funding through the VMO ("Values, Monitoring and Outcomes") research programme (MBIE contract CO9X1003; NIWA subcontract from LCR). The Ministry for the Environment is thanked for funding completion of the work with the preparation of this report. Lauren Long (MfE contract manager) is thanked for helpful feedback on the report. Peter Robinson (Chemist at Hill Laboratories), Mike Crump (NIWA laboratory manager) and Neale Hudson (NIWA) made helpful comments on an early version of the report. Juliet Milne (now of NIWA), and Craig Depree and Clive Howard-Williams (both of NIWA) are thanked for reviewing the current version.

8 Glossary of abbreviations and terms

Accuracy	Closeness of measurements to the true value (implies **precise** *and* unbiased).
Am-N	Ammoniacal-nitrogen (= ammonia plus ammonium ion).
Bias	Systematic deviation of replicate measurements from the true value. Also known as systematic error.
Censoring (of data)	Reporting an analytical result as "less than" (a **detection limit**) rather than reporting the analyst's best estimate. While the concept of censoring may have some merit for one-off analyses of strange substances that should not be present in the environment (dioxin, plutonium), it is a severe impediment to state-of-environment (SoE) reporting based on ongoing, indefinite monitoring (i.e., repeat sampling) of analytes that are *known* to be present, such as dissolved reactive phosphorus (DRP) in waters[2].
Detection limit (DL)	Concentration or value below which presence of an analyte cannot be 'detected' (guaranteed to be present) by the analyst.
DRP	Dissolved reactive phosphorus (usually the main form of dissolved phosphorus).
NNN	Nitrite-nitrate-nitrogen (NNN); (aka oxidised nitrogen: **NO_x-N**).
NO_x-N	Oxidised nitrogen (NO_x-N); (aka nitrite-nitrate-nitrogen: **NNN**).
Precision	Closeness of replicate measurements (long-term).
RMSE	Root mean square error. Index of the variance around a trend curve.
SPM	Suspended particulate matter in waters. A gross index of concentration of SPM is given by **TSS** or **SSC.**
SSC	Suspended sediment concentration. Measured as the dry weight of residue collected on a filter (usually a glass fibre filter) divided by total water sample volume. That is, the whole water sample is filtered rather than sub-sampling as for TSS. Ideally TSS = SSC, but in practice they can differ if the volumetric subsampling for TSS is not fully representative of the water sample.
TN	Total nitrogen (generic term). *Should be replaced in future by a clear identification of the method of nitrogen assay*, e.g., whether **TN-A** or **TN-K.**
TN-A	Total nitrogen estimated by direct analysis by the alkaline persulphate method. Shown by the USGS to be *biased low*, to an extent dependent on **SPM**.
TN-C	Total nitrogen estimated by summing the particulate nitrogen (PN) measured by an unbiased method such as high temperature oxidation, to the dissolved nitrogen (DN) (TN-C = PN + DN). TN-C is regarded by the USGS as essentially unbiased.

[2] Laboratory analysts may not want to report a <DL result when they're not confident in the actual result measured. Data analysts are often interested in a sequence of results, and masking all the lower results by <DL can hide an emerging pattern.

TN-K	Total nitrogen estimated by adding total Kjeldahl nitrogen (**TKN**) to nitrate-nitrite-nitrogen (**NNN**) (TN-K = TKN + NNN). TN-K is regarded by the USGS as minimally unbiased.
TSS	Total suspended solids. Measured as the dry weight of residue collected on a filter (usually a glass fibre filter) per unit volume filtered. Ideally TSS = SSC, but in practice they can differ if the volumetric subsampling for TSS is not fully representative of the water sample.
UoM	Uncertainty of Measurement. Confidence interval on measurements in a batch based on standard deviation of replicates (usually three) in the batch. The long-term statistics of UoM establish measurement **precision**.

9 References

Davies-Colley, R.J., Ballantine, D.J., Elliott, A.H., Swales, A., Hughes, A.O., Gall, M.P. (2014) Light attenuation – a more effective basis for the management of fine suspended sediment than mass concentration? *Water science and technology*, 69(9): 1867-1874. doi: 10.2166/wst.2014.096

Davies-Colley, R., Verburg, V., Hughes, A., Storey, R. (2012a) Variables for regional monitoring underpinning national reporting. Summary of recommendations. National Environmental Monitoring and Reporting (NEMaR) Variables, Step 1. *NIWA Client Report* (HAM2012-006) for the Ministry for the Environment. May 2012, August 2012: 67.

Davies-Colley, R., Verburg, V., Hughes, A., Storey, R. (2012b) Freshwater monitoring protocols and quality assurance (QA). National Environmental Monitoring and Reporting (NEMaR) Variables, Step 2. *NIWA Client Report* (HAM2012-092) for the Ministry for the Environment. August 2012: 96.

Helsel, D.R. (2005) Non-detects and data analysis, Wiley-Interscience, Hoboken, New Jersey: 250.

McBride, G.B. (2005) *Using Statistical Methods for Water Quality Management: Issues, Problems and Solutions,* John Wiley & Sons, New York: 313.

McBride, G.B., Davies-Colley, R., Bargh, M. (2013) Towards a nationally consistent and dependable surface freshwater monitoring programme for New Zealand. Description and costs including quality assurance. National Environmental Monitoring and Reporting (NEMaR) Phase 3. *NIWA Client Report* (HAM13-502) for the Ministry for the Environment. May 2013: 39.

Milne, J.R., Perrie, A., Robinson, P., Davies-Colley, R.J. (2014) National consistency in nutrient test methods: assessing the implications of changing methods on long-term water quality datasets – Part 1. *Water Symposium, NZ Freshwater Sciences and NZ Hydrological Society joint annual conference,* 24-28 November 2014, Blenheim.

Robinson, P., Milne, J.R., Perrie, A., Davies-Colley, R.J. (2014) National consistency in nutrient test methods: assessing the implications of changing methods on long-term water quality datasets – Part 2. *Water Symposium, NZ Freshwater Sciences and NZ Hydrological Society joint annual conference*, 24-28 November 2014, Blenheim.

Rus, D.L., Patton, C.J., Mueller, D.K. and Crawford, C.G. (2012) Assessing total nitrogen in surface-water samples - precision and bias of analytical and computational methods. *US Geological Survey Scientific Investigations Report*, 2012-5281: 38.

U.S. Geological Survey (2009) Potential bias in alkaline persulfate analysis of total nitrogen in whole water samples and recommendations for quantifying bias in whole water samples: U.S. Geological Survey Office of Water Quality Technical Memorandum 2009.03, accessed May 2014, at http://water.usgs.gov/admin/memo/QW/qw09.03.html.

Velleman, P.F. (1992) *Data Desk 4, Statistics guide,* Volume 2. Ithaca, New York, Data Description Inc.